



# Statistical perspectives on dependencies between genomic markers

Kumulative Habilitationsschrift

zur

Erlangung des akademischen Grades

doctor rerum naturalium habitata (Dr. rer. nat. habil.)

der Agrar- und Umweltwissenschaftlichen Fakultät

der Universität Rostock

vorgelegt von

Dörte Wittenburg

geboren am 16. Oktober 1978 in Rostock

aus Rostock

Rostock, 10. Februar 2021

[https://doi.org/10.18453/rosdok\\_id00003502](https://doi.org/10.18453/rosdok_id00003502)

**Gutachter:**

Prof. Dr. Klaus Wimmers (Universität Rostock, AUF, Tierzucht und Haustiergenetik;  
FBN Dummerstorf)

Prof. Dr. Hans-Peter Piepho (Universität Hohenheim, Fg. Biostatistik)

Prof. Dr. Jörn Bennewitz (Universität Hohenheim, Fg. Tiergenetik und Züchtung)

**Tag der Verteidigung und Probevorlesung:** 12. November 2021

# Contents

<b>1</b>	<b>General introduction</b>	<b>9</b>
1.1	Genomic evaluations . . . . .	9
1.2	Non-additive genetic effects . . . . .	9
1.3	Association between markers . . . . .	10
1.4	Breeding populations . . . . .	11
1.5	Statistical approaches for estimating genetic effects . . . . .	12
1.6	Objectives and outline of thesis . . . . .	13
1.7	Literature cited . . . . .	14
<b>2</b>	<b>Publications on non-additive genetic effects</b>	<b>17</b>
2.1	Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers . . . . .	18
2.2	An approximate Bayesian significance test for genomic evaluations . . . . .	33
2.3	Milk metabolites and their genetic variability . . . . .	47
2.4	Genomic additive and dominance variance of milk performance traits . . . . .	60
<b>3</b>	<b>Publications on the dependence between markers</b>	<b>77</b>
3.1	Covariance between genotypic effects and its use for genomic inference in half-sib families	78
3.2	Design of experiments for fine-mapping quantitative trait loci in livestock populations	91
3.3	Grouping of genomic markers in populations with family structure . . . . .	105
3.4	Estimation of recombination rate and maternal linkage disequilibrium in half-sibs . . .	117
3.5	Male recombination map of the autosomal genome in German Holstein . . . . .	130
<b>4</b>	<b>General discussion</b>	<b>143</b>
4.1	Model parameterisation for non-additive genetic effects . . . . .	143
4.2	Covariance between genotype codes . . . . .	147
4.3	Inferences from recombination rates . . . . .	149
4.4	Accounting for multicollinearity . . . . .	152
4.5	Concluding remark . . . . .	153
4.6	Literature cited . . . . .	154
<b>5</b>	<b>Summary</b>	<b>159</b>





# Acknowledgement

My utmost respect goes to Prof. Dr. Norbert Reinsch (FBN Dummerstorf) and Prof. Dr. Volkmar Liebscher (University of Greifswald) who have always contributed lively to discussions on any of these publications. Many ideas have been generated during such meetings and they would have not reached maturity without the invaluable feedback from N. Reinsch and V. Liebscher.

I particularly thank Dr. Friedrich Teuscher for his critical view and concrete advice which have ever been helpful since my doctoral thesis.

Furthermore, I am thankful to all my present and past colleagues at the FBN Dummerstorf for providing help whenever needed.

– Grateful to my family.



# Abbreviations

AUC Area under the curve

cM centiMorgan

DNA Deoxyribonucleic acid

EM Expectation maximisation

GBLUP Genomic best linear unbiased prediction

GWAS Genome-wide association study

HWE Hardy-Weinberg equilibrium

IPF Integrative lasso with penalty factors

LD Linkage disequilibrium

LE Linkage equilibrium

NOIA Natural and orthogonal interactions

SNP Single nucleotide polymorphism



# Chapter 1

## General introduction

### 1.1 Genomic evaluations

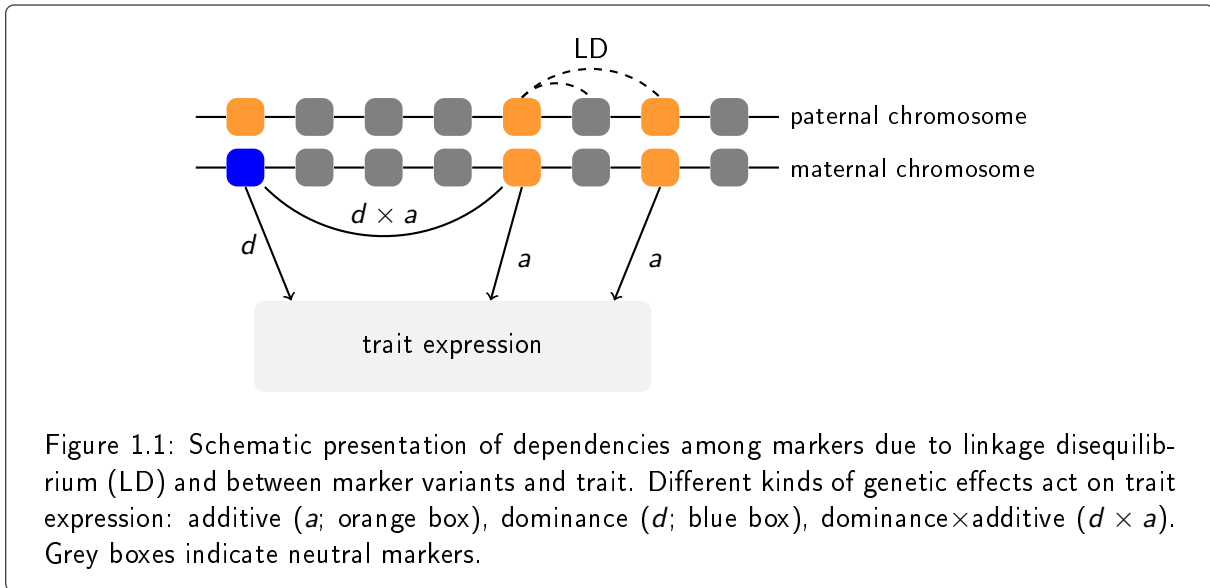
High-throughput biotechnological methods allow collecting extensive genetic information, for instance, from blood or tissue samples. From this material, the state (genotypes) of hundreds of thousands of molecular markers can be determined for an individual. Single nucleotide polymorphisms (SNPs) are the most common markers; each SNP bears two variants that can be observed in a population (e.g. alleles A and B). Hence, in diploid organisms like humans or cattle, three different marker genotypes can appear, AA, AB and BB, which are typically coded as 0, 1 and 2, respectively.

In animal breeding, SNP data are valuable for studying the association between genetic and phenotypic variation. In genome-wide association studies (GWAS), SNPs that are closely linked to causative DNA segments are identified and corresponding effect sizes are estimated. Knowing the trait-associated markers on the genome, it is possible to predict the average trait expression, e.g. of milk yield, or production lifetime of cattle in early life. Moreover, animal breeders can develop mating strategies to further improve the performance in later generations. Accurate estimates of genetic values are fundamental to meet the breeding objectives. Circumstances affecting the accuracy are highlighted in the following sections.

### 1.2 Non-additive genetic effects

A GWAS in its simplest form is based on a genome-wide regression approach: a phenotype  $y$  is regressed onto the number of reference alleles at a SNP. SNPs may be investigated one at a time or jointly. The less frequent SNP variant is often chosen as the reference allele in the investigated population but it may also be the reference allele as defined by the corresponding genome assembly. This way, a purely additive impact of the reference allele is assumed on the trait.

In addition to an additive effect of a SNP variant, non-additive effects resulting from interaction of SNP alleles at one locus (dominance) or between loci (epistasis) may be relevant for trait expression (see Figure 1.1). Considering only two-loci interactions, four types of epistatic effects can be distinguished: additive $\times$ additive, additive $\times$ dominance, dominance $\times$ additive, dominance $\times$ dominance (Cockerham, 1954). Additive effects are directly transmitted from parent to progeny. In contrast, the combination of paternally and maternally inherited alleles causes the appearance of dominance or epistatic effects on the progeny phenotype, making non-additive effects particularly important in



animal production.

To mention just a few examples, Fuerst and Sölkner (1994) investigated sources of genetic variation for different milk performance traits and breeds and found that more than 50 % of the variation of protein percentage was explained by additive genetic variation and only a small contribution of at least 5 % was caused by dominance. In an  $F_2$  crossing experiment with mice, Peripato et al. (2004) identified 8 loci which were involved in 6 interactions and another two loci with additive or dominant effect on litter size. The simultaneous consideration of additive and non-additive effects also has advantages on the accuracy of trait prediction in future generations. Representing an extreme case, Lee et al. (2008) have shown that the precision of predicting coat colour increases from 54 % to 81 % when dominance is considered in addition to additive effects in a mouse population.

Modelling and interpretation of marker effects require careful distinction between allele effects on the genotype level – they reflect the difference between genotypes – and average effects on the population level. The average effect of a marker allele contributes to the breeding value of an animal, thus being determinant of breeding decisions. In a detailed inspection, the average effect contains parts of non-additive effects which thus directly influence the additive genetic variation (Cheverud and Routman, 1995). The extent of contribution depends on the allele frequency in the underlying population. Especially in populations with a low effective population size (e.g.  $N_e = 64$ , similar to dairy cattle), selection and drift lead to a change of allele frequencies, thereby epistatic effects increasingly cause additive genetic variation (Cheverud and Routman, 1996). Jointly modelling additive, dominant and epistatic effects requires a proper coding of SNP genotypes (i.e. statistical parameterisation) in order to avoid confounding of the different genetic variance components (Álvarez-Castro and Carlborg, 2007; Cockerham, 1954; Vitezica et al., 2017).

### 1.3 Association between markers

To draw reliable conclusions about the relationship between trait expression and genetics, it is important to take into account the associations among markers on the genome. One aspect is the

order of SNPs on the chromosomes: the sorting (i.e. genome assembly) is regularly updated based on inferences from improved technologies and growing availability of DNA-sequence data. Another aspect is the distance between SNPs. It is not only important to know how many base pairs lie in between the SNPs (physical measure) but also how likely it is that the variants at two SNP loci are inherited together from parent to progeny (genetic measure). This genetic measure is expected to vary over regions on the genome. For instance, the chance of a cross-over, i.e. an exchange of information happens between two parental chromosome sets during meiosis, is expected to be higher on the telomere end of the chromosome than on the centromere (e.g. Nachman, 2002). The expected number of cross-overs between two loci defines the genetic distance given in Morgan units (or centiMorgan; cM): one cross-over is expected on a distance of one Morgan length. An odd number of cross-overs constitutes an observed recombination event. A genetic map function is used to describe the relationship between probability of recombination (i.e. recombination rate) and genetic distance between markers.

Measuring the strength of association between markers in terms of linkage disequilibrium (LD) enables us to report which SNPs are “linked”. Linkage does not necessarily refer to the proximity on the genome but linked SNPs have a higher probability of being inherited jointly to an offspring than any unlinked markers. Hence, a recombination event is unlikely to appear among closely linked markers. In a linkage analysis, genetic information from parents and offspring are analysed, and the probability of a recombination event is estimated. Information about parent-child trios are found less in animal breeding than in human studies. Non-random mating influences the population structure and the data collection strategy in livestock. For instance, paternal half-sib families are a typical family structure in dairy cattle: half-siblings have a common sire but individual dams. The sire and its ancestors are often genotyped but not necessarily the dams. Knowing the combinations of alleles (i.e. haplotypes) of parent animals enables measuring the strength of dependence between markers in the progeny generation. The recombination rate between pairs of SNPs is a central parameter for this field of research.

## 1.4 Breeding populations

In a purebred population, the prediction of average progeny performance is a benchmark parameter of parent animals whose suitability as breeding candidates is evaluated; this is the concept of “genomic selection”. An animal’s breeding value is derived as cumulative effect of markers depending on its marker genotypes. The higher the breeding value is, the more likely the animal is selected for breeding. Moreover, the distribution (mean value and variation) of breeding values in the progeny generation can be predicted from the genetic information of putative parents and additive marker effects (Bonk et al., 2016). Additional knowledge of non-additive marker effects can be taken into account in this matter.

In a crossbred population, non-additive effects have a long history. In such a breeding design, a heterosis effect, which occurs when hybrids have a higher performance than the average of the purebred parent lines, is explicitly exploited. Heterosis has a large economic importance, for instance, in pig breeding. In many crossbred populations, production and reproduction traits have a significant heterosis effect leading to lower foetal mortality, better piglet growth, higher litter size and other

advantages (Rothschild and Ruvinsky, 2011). Theoretically, only dominance effects and, if only interactions between two loci are considered, additive $\times$ additive epistasis contribute to the expression of heterosis (Melchinger et al., 2007b). A distinct influence of this phenomenon was demonstrated in a crossbreeding experiment with *Arabidopsis Thaliana* (Melchinger et al., 2007). In practice, the purebred parent animals are mainly selected for high purebred performance and the success of selection depends on the genetic correlation between purebred and crossbred performance (Duenk et al., 2019).

Furthermore, the accuracy of estimated genetic values in future generations depends on the persistence of estimated marker effects in the respective population. Linkage phases are dissolved by the degree of kinship decreasing over generations and a declining linkage disequilibrium between trait-associated DNA variant and markers, leading to a loss of prediction accuracy (Habier et al., 2010).

Especially in small populations consisting of few families, linkage is an important parameter. Because not all haplotypes that are theoretically possible are represented in the population, just the effects of chromosome segments instead of concrete marker effects are detectable with GWAS. Only in larger populations with many families, the effects of individual markers can be better distinguished and associations between markers and trait may be detected.

## 1.5 Statistical approaches for estimating genetic effects

On the one hand, extensive genetic information enables the identification of causal DNA segments but on the other hand, it leads to high model dimensions challenging any statistical approach. In a typical situation of genomic evaluation, there are many more predictor variables ( $p$  SNPs) than observations ( $n$  animals) causing linear dependencies among predictors. Due to the proximity of SNPs, linkage and LD between markers add a biologically justified source of dependence. The presence of linear dependencies among several predictors is called multicollinearity. Ignoring dependencies leads to increased standard error of parameter estimates and erroneous inferences and prediction (Dormann et al., 2013). Estimation of genomic breeding values works well for  $p > n$  (Gianola, 2013) but may lead to incorrect genomic inference of individual marker effects. In such over-parameterised models, the reliability and repeatability of statistical analyses are strongly influenced by the assumptions on a statistical model made by the experimenter.

Selection and shrinkage approaches are frequently used in genomic evaluations (de los Campos et al., 2013). Among those, the range of Bayesian methods is colorful (e.g. Erbe et al., 2012; Meuwissen et al., 2001) because they allow for high flexibility in incorporating prior knowledge. Depending on biological or other meaningful information available, prior distributions can be specified for any of the unknown parameters (Tempelman, 2015). The impact of prior assumptions, however, decreases as sample size increases.

Moreover, Frequentist (e.g. Koivula et al., 2012; Legarra et al., 2014) and also optimisation approaches (e.g. Feng et al., 2011; Waldmann et al., 2013) have proved useful. High dimensionality and high multicollinearity encouraged the development of grouped penalised regression approaches which require pre-specified groups of predictors (Yuan and Lin, 2006). This method has been further developed to allow for sparsity between and within groups (Simon et al., 2013). Hence, not only groups of highly dependent predictors but also single sites within such a group are identifiable. Dependencies between SNPs are typically measured by the population parameter LD, for instance expressed as



$r^2$  (Hill and Robertson, 1968), but this term is likely biased in a non-random mating population. Family stratification leads to different levels of LD among families. Thus, no matter what the actual grouping technique is (e.g. hierarchical clustering, k-means clustering), it needs to employ measures of association suited to the family structure in a population.

## 1.6 Objectives and outline of thesis

The first objective of this thesis was to explore statistical approaches for disentangling sources of genetic variation into additive and non-additive genetic parts. A differentiation is key to understand their importance on trait expression and to exploit the heritable potential suited to specific breeding designs. **Chapter 2** is dedicated to statistical modelling of different kinds of genetic effects in a series of four publications. Two approaches have been followed. First, genetic effects captured by single markers were studied in a Bayesian framework (Papers 2.1 and 2.2). The appropriateness of the statistical approach has been verified in a simulation study. Second, cumulative marker effects (i.e. genetic values) have been investigated for milk and milk-component traits in a sample of Holstein cows using a Frequentist GBLUP approach (Papers 2.3 and 2.4).

The second objective was to study the dependence between SNPs due to linkage and LD because the proximity of markers plays an increasing role for precise genome-based evaluations with growing number of markers available. In **Chapter 3**, dependencies between markers due to linkage and LD have been derived analytically and employed in a series of five publications. The dependence between SNPs was worked out assuming that genotypic data have been collected from half-sib families (Paper 3.1). Fields of application for the theoretical dependence between SNPs have been identified. One option is the design of future experiments to fine-map quantitative trait loci where selection candidates may be known but the number of progeny to be genotyped is wanted (Paper 3.2). As another option, grouping of markers with respect to population stratification into half- or full-sib families was compared to a population-LD approach (Paper 3.3). As recombination rate was a central parameter in these derivations, advanced methods for estimating recombination rate between SNPs were developed in Paper 3.4. Methods were also verified with empirical data at a large scale in Paper 3.5 leading to estimates of genetic distances with high confidence. This analysis additionally enabled the identification of SNPs that were putatively misplaced in the current bovine genome assembly.

In **Chapter 4**, a general discussion of results complements this habilitation thesis. The interconnecting issues between Chapters 2 and 3 are highlighted and points for further research are outlined.

## 1.7 Literature cited

- Álvarez-Castro, J. M., & Carlborg, Ö. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*, *176*(2), 1151–1167. <https://doi.org/10.1534/genetics.106.067348>
- Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., & Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. *Genet. Sel. Evol.*, *48*(1), 36. <https://doi.org/10.1186/s12711-016-0214-0>
- Cheverud, J. M., & Routman, E. J. (1995). Epistasis and its contribution to genetic variance components. *Genetics*, *139*(3), 1455–1461.
- Cheverud, J. M., & Routman, E. J. (1996). Epistasis as a source of increased additive genetic variance at population bottlenecks. *Evolution*, *50*, 1042–1051.
- Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, *39*(6), 859–882.
- 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*(2), 327–345. <https://doi.org/10.1534/genetics.112.143313>
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 027–046.
- Duenk, P., Calus, M. P., Wientjes, Y. C., Breen, V. P., Henshall, J. M., Hawken, R., & Bijma, P. (2019). Estimating the purebred-crossbred genetic correlation of body weight in broiler chickens with pedigree or genomic relationships. *Genet. Sel. Evol.*, *51*(1), 6.
- Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., Mason, B., & Goddard, M. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.*, *95*(7), 4114–4129.
- Feng, Z. Z., Yang, X., Subedi, S., & McNicholas, P. D. (2011). The lasso and sparse least squares regression methods for snp selection in predicting quantitative traits. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, *9*(2), 629–636. <https://doi.org/10.1109/TCBB.2011.139>
- Fuerst, C., & Sölkner, J. (1994). Additive and nonadditive genetic variances for milk yield, fertility, and lifetime performance traits of dairy cattle. *J. Dairy Sci.*, *77*(4), 1114–1125.
- Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, *194*(3), 573–596.
- 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.*, *42*, 5. <https://doi.org/10.1186/1297-9686-42-5>
- Hill, W., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, *38*(6), 226–231.
- Koivula, M., Strandén, I., Su, G., & Mäntysaari, E. A. (2012). Different methods to calculate genomic predictions – comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *J. Dairy Sci.*, *95*(7), 4065–4073.

- Lee, S. H., van der Werf, J. H. J., Hayes, B. J., Goddard, M. E., & Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome snp data. *PLoS Genet.*, *4*(10), e1000231. <https://doi.org/10.1371/journal.pgen.1000231>
- Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single step, a general approach for genomic selection. *Livest. Sci.*, *166*, 54–65.
- Melchinger, A. E., Piepho, H.-P., Utz, H. F., Muminovic, J., Wegenast, T., Törjék, O., Altmann, T., & Kusterer, B. (2007). Genetic basis of heterosis for growth-related traits in arabidopsis investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics*, *177*(3), 1827–1837. <https://doi.org/10.1534/genetics.107.080564>
- Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.
- Nachman, M. W. (2002). Variation in recombination rate across the genome: Evidence and implications. *Curr. Opin. Genet. Dev.*, *12*(6), 657–663.
- Peripato, A. C., De Brito, R. A., Matioli, S. R., Pletscher, L. S., Vaughn, T. T., & Cheverud, J. M. (2004). Epistasis affecting litter size in mice. *J. Evol. Biol.*, *17*(3), 593–602. <https://doi.org/10.1111/j.1420-9101.2004.00702.x>
- Rothschild, M. F., & Ruvinsky, A. (Eds.). (2011). *The genetics of the pigs* (2nd). CABI.
- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *J. Comp. Graph. Stat.*, *22*(2), 231–245.
- Tempelman, R. J. (2015). Statistical and computational challenges in whole genome prediction and genome-wide association analyses for plant and animal breeding. *J. Agric. Biol. Environ. Stat.*, *20*(4), 442–466.
- Vitezica, Z. G., Legarra, A., Toro, M. A., & Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics*, *206*(3), 1297–1307. <https://doi.org/10.1534/genetics.116.199406>
- Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.*, *4*, 270. <https://doi.org/10.3389/fgene.2013.00270>
- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.*, *68*(1), 49–67. <https://doi.org/10.1111/j.1467-9868.2005.00532.x>



## Chapter 2

# Publications on non-additive genetic effects

- 2.1 Wittenburg, D., Melzer, N., & Reinsch, N. (2011). Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. *BMC Genet.*, *12*, 74. <https://doi.org/10.1186/1471-2156-12-74>
- 2.2 Wittenburg, D., & Liebscher, V. (2018). An approximate Bayesian significance test for genomic evaluations. *Biom. J.*, *60*, 1096–1109. <https://doi.org/10.1002/bimj.201700219>
- 2.3 Wittenburg, D., Melzer, N., Willmitzer, L., Lisec, J., Kesting, U., Reinsch, N., & Repsilber, D. (2013). Milk metabolites and their genetic variability. *J. Dairy Sci.*, *96*(4), 2557–2569. <https://doi.org/10.3168/jds.2012-5635>
- 2.4 Wittenburg, D., Melzer, N., & Reinsch, N. (2015). Genomic additive and dominance variance of milk performance traits. *J. Anim. Breed. Genet.*, *132*, 3–8. <https://doi.org/10.1111/jbg.12103>



## METHODOLOGY ARTICLE

## Open Access

# Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers

Dörte Wittenburg\*, Nina Melzer and Norbert Reinsch

## Abstract

**Background:** Molecular marker information is a common source to draw inferences about the relationship between genetic and phenotypic variation. Genetic effects are often modelled as additively acting marker allele effects. The true mode of biological action can, of course, be different from this plain assumption. One possibility to better understand the genetic architecture of complex traits is to include intra-locus (dominance) and inter-locus (epistasis) interaction of alleles as well as the additive genetic effects when fitting a model to a trait. Several Bayesian MCMC approaches exist for the genome-wide estimation of genetic effects with high accuracy of genetic value prediction. Including pairwise interaction for thousands of loci would probably go beyond the scope of such a sampling algorithm because then millions of effects are to be estimated simultaneously leading to months of computation time. Alternative solving strategies are required when epistasis is studied.

**Methods:** We extended a fast Bayesian method (fBayesB), which was previously proposed for a purely additive model, to include non-additive effects. The fBayesB approach was used to estimate genetic effects on the basis of simulated datasets. Different scenarios were simulated to study the loss of accuracy of prediction, if epistatic effects were not simulated but modelled and vice versa.

**Results:** If 23 QTL were simulated to cause additive and dominance effects, both fBayesB and a conventional MCMC sampler BayesB yielded similar results in terms of accuracy of genetic value prediction and bias of variance component estimation based on a model including additive and dominance effects. Applying fBayesB to data with epistasis, accuracy could be improved by 5% when all pairwise interactions were modelled as well. The accuracy decreased more than 20% if genetic variation was spread over 230 QTL. In this scenario, accuracy based on modelling only additive and dominance effects was generally superior to that of the complex model including epistatic effects.

**Conclusions:** This simulation study showed that the fBayesB approach is convenient for genetic value prediction. Jointly estimating additive and non-additive effects (especially dominance) has reasonable impact on the accuracy of prediction and the proportion of genetic variation assigned to the additive genetic source.

## 1 Background

Molecular marker information is commonly used to draw inferences about the relationship between genetic and phenotypic variation in various species, e.g. humans [1], dairy cattle [2] or mice [3]. Assuming linkage disequilibrium (LD) between quantitative trait loci (QTL) and markers, genetic effects can be estimated and

explained as QTL effects captured by the neighbouring markers. If breeding values are the focal point, genetic effects are typically modelled as additively acting marker allele effects (e.g. [4,5]). The mode of biological action can, of course, be different from the assumption of pure additivity. One possibility to better understand the genetic architecture of complex traits is to include intra-locus (dominance) and inter-locus (epistasis) interaction of alleles when fitting a model to a trait. The importance of non-additive effects for genetic variation has recently been investigated. Knowledge about non-

\* Correspondence: [wittenburg@fbn-dummerstorf.de](mailto:wittenburg@fbn-dummerstorf.de)  
Research Unit Genetics and Biometry, Leibniz Institute for Farm Animal Biology (FBN), Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany

additive effects is essential to benefit, for example, from heterosis effects [6], especially for cross-breeding schemes (poultry, plants etc.). In general, it can be expected that the prediction of the genetic value, in particular its additive part, is improved if non-additive effects are additionally modelled. For instance, Lee *et al.* [7] reported that the accuracy of prediction increased considerably when dominance effects were included compared to a purely additive genetic model when the phenotypes coat colour (+17% accuracy) or the percentage of CD8<sup>+</sup> cells (+2% accuracy) were studied in mice. Added epistasis did not, however, contribute to the accuracy in this case. In an example with recombinant inbred lines of soybean [8], the accuracy of prediction was more than doubled under the epistatic model. Even though non-additive effects may occur on the level of gene action, most of the genetic variation might be assigned to additive effects when genes are at an extreme frequency [9]. The extent to which, for example, epistasis is involved in regulating complex traits is hardly known, but knowledge about it can be used to infer biological mechanisms and to reconstruct biological pathways [10]. In one of the first studies concerning non-additive influence on growth differences in chickens, Carlborg *et al.* [11] estimated that 10% of genetic variation in early growth (trait Gr18) was due to dominance and even 70% due to epistasis. This example showed the importance of interacting loci, though one may suppose an overestimation of the epistatic effects, a phenomenon already known as the Beavis effect [12] for single loci. Since this experiment was based on a cross of extremely different lines, further investigations are required to find evidence for interacting genes in purebreds.

Different approaches are available to model additive and non-additive genetic effects. Under the aspect of QTL detection, a genome scan can be carried out to uncover genetic effects using, for example, a variance component method [13,14]. If additive and non-additive effects are to be modelled simultaneously over the whole genome, we have to be aware of “ $p$  bigger than  $n$ ” problems, meaning there are more parameters than there are observations. To cope with the all-in-one situation, Xu presented a Bayesian approach [15], which parallels the idea of BayesA [4], and an empirical Bayes method [16] both enabling the genome-wide estimation of additive and non-additive marker effects. The Bayesian methods commonly used for the estimation of additive effects apply Markov chain Monte Carlo (MCMC) simulations which require a lot of computing time, but they convince in terms of accuracy in predicting genetic values. In particular, the BayesB approach [4] is superior to other methods, for instance ridge regression and partial least squares [17-19]. The MCMC sampling

methods may collapse under high marker density if further non-additive effects are included. As an alternative, an approximate Bayesian approach is available which applies the analytically derived posterior density for a marker effect rather than samples thereof [20]. This approach (called fBayesB) was shown to be slightly less accurate, because in an iterative procedure only a single marker effect is studied at a time while the vector of phenotypes is corrected for all other previously estimated effects. The fBayesB strategy is much faster than the conventional Bayesian methods using MCMC. This solving approach offers the possibility to additionally account for genome-wide interacting effects and to estimate them with reasonable computational effort.

The objective of this study is to explore the impact of non-additive effects on the prediction of genetic values in a livestock population. An improved estimation of additive effects and a better prediction of genetic values is intended, when additive and non-additive effects are jointly involved in fitting a model to a trait. Since methods that aim to estimate non-additive effects in arbitrary populations are just emerging, it is especially important to validate such approaches with simulations. Therefore, with this study, we pursue methodological aspects, thereby assembling facts that help to interpret results obtained with practical data in future work. We consider additive, dominance and pairwise epistatic effects captured by biallelic markers spread over the whole genome. The details of statistical modelling are presented in the first part of the paper. We extend the fast Bayesian method (fBayesB), which was developed under pure additivity [20], to include non-additive effects. fBayesB is used to estimate the genetic effects on the basis of simulated datasets which resemble a dairy cattle population. Different scenarios are simulated to study the loss of accuracy of prediction if epistatic effects are not simulated but modelled and vice versa. In the second part, we summarise the results of analysing the simulated data. The amount of genetic variation assigned to each kind of genetic effect after genome-wide estimation of marker effects is determined. To briefly show how the approach behaves in practice, we also apply fBayesB to a real data example. In the third part, we outline some constraints of estimating non-additive effects via the fBayesB approach and discuss other solving strategies.

## 2 Methods

### 2.1 Statistical model

For the statistical analysis of genetic effects in a Bayesian framework, a hierarchical model is constructed similar to that of Meuwissen *et al.* [20]. Bold symbols are used for vectors and matrices. At first, only main genetic effects (i.e. additive and dominance effects) are included. In total  $m$  loci are studied on the genome. The vector of



phenotypes  $\mathbf{y} = (y_1, \dots, y_n)'$  is modelled as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{a} + \mathbf{D}\mathbf{d} + \mathbf{e}.$$

This model is set up in the way of an  $F_\infty$  model [21]. Let  $\mu$  be a population mean and  $\mathbf{1}$  a vector of ones. The  $\mathbf{X}$  and  $\mathbf{D}$  are design matrices for allele substitution effects  $\mathbf{a} = (a_1, \dots, a_m)'$  and dominance effects  $\mathbf{d} = (d_1, \dots, d_m)'$ , respectively. The entries of the design matrices are random variables which are realised depending on the observed marker genotypes (denoted as 11, 12, 22). For a homozygous genotype at locus  $j \in \{1, \dots, m\}$  of animal  $i \in \{1, \dots, n\}$ ,  $X_{i,j} = \pm 1$  and  $D_{i,j} = 0$ ; the positive effect is assigned to the more frequent allele. For a heterozygous genotype,  $X_{i,j} = 0$  and  $D_{i,j} = 1$ .

This work relies on two assumptions. Firstly, linkage equilibrium (LE) between the different markers is assumed. Then genotypic effects at different loci are independently distributed and the estimation strategy does not depend on the order of markers. Secondly, in order to avoid the estimation of covariance components at intra-locus investigations, the additive genetic value and the dominance genetic value are assumed uncorrelated at each locus, i.e.  $\text{Cov}(X_{i,j}a_j, D_{i,j}d_j) = 0 \forall i,j$ . This assumption can be fulfilled by re-parametrising coefficients coding for the marker genotypes in advance. We apply the method of Álvarez-Castro & Carlborg [22] to obtain an orthogonal decomposition of genetic values. This method involves the genotype frequencies  $p_{11,j}$ ,  $p_{12,j}$ ,  $p_{22,j}$  at each locus  $j$  and does not necessarily depend on Hardy-Weinberg equilibrium (HWE). The method is related to Cockerham's model [23] given HWE. In an  $F_\infty$  model, the genotypic effects  $\mathbf{G}_j = (G_{11,j}, G_{12,j}, G_{22,j})'$  can be written as

$$\mathbf{G}_j = \mathbf{S} \begin{pmatrix} \mu \\ a_j \\ d_j \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}. \quad (1)$$

The second and third column of  $\mathbf{S}$  represent the possible realisations in  $\mathbf{X}$  and  $\mathbf{D}$ , respectively. The genotypic values can also be obtained in terms of an additive effect  $g_{a,j}$  and dominance effect  $g_{d,j}$  on the orthogonal scale by

$$\mathbf{G}_j = \mathbf{S}_{A,j} \begin{pmatrix} \mu^* \\ g_{a,j} \\ g_{d,j} \end{pmatrix} \quad \text{with} \quad (2)$$

$$\mathbf{S}_{A,j} = \begin{pmatrix} 1 & -p_{11,j} - 2p_{22,j} & -\frac{2p_{12,j}p_{22,j}}{\nu} \\ 1 & 1 - p_{11,j} - 2p_{22,j} & \frac{4p_{11,j}p_{22,j}}{\nu} \\ 1 & 2 - p_{11,j} - 2p_{22,j} & -\frac{2p_{12,j}p_{11,j}}{\nu} \end{pmatrix},$$

where  $\nu = p_{11,j} + p_{22,j} - (p_{11,j} - p_{22,j})^2$ . Since the representations (1) and (2) are equivalent [22], the  $F_\infty$  model can be translated into

$$\mathbf{y} = \mathbf{1}\mu^* + \mathbf{X}_a\mathbf{g}_a + \mathbf{X}_d\mathbf{g}_d + \mathbf{e}, \quad (M1)$$

where the design matrices  $\mathbf{X}_a$  and  $\mathbf{X}_d$  contain the corresponding entries of  $\mathbf{S}_{A,j}$  ( $j = 1, \dots, m$ ) and relate to the additive and dominance effects on the orthogonal scale, respectively.

To obtain numerical stability in later calculations, coefficients of the main genetic effects are additionally standardised. Let  $p_j$  denote the (estimated) allele frequency at locus  $j$ . One possibility is to divide the columns in  $\mathbf{X}_a$  and  $\mathbf{X}_d$  by the standard deviation of the random variable coding the marker genotype for the additive or dominance effects, respectively,

$$\begin{aligned} X_{a,j} &\mapsto \frac{X_{a,j}}{\sqrt{2p_j(1-p_j)}} \quad \text{and} \\ X_{d,j} &\mapsto \frac{X_{d,j}}{2p_j(1-p_j)}. \end{aligned} \quad (3)$$

Now the hierarchical structure of M1 can be characterised by the following prior distributions

$$\begin{aligned} e_i &\sim N(0, \sigma_e^2), \quad i = 1, \dots, n, \\ g_{s,j} &\sim L^*(\gamma_s, \lambda_s), \quad s \in \{a, d\}, j = 1, \dots, m. \end{aligned}$$

$L^*(\gamma_s, \lambda_s)$  denotes a mixture of a Laplace distribution with zero expectation and the point mass at zero. The mixing probability is  $\gamma_s$ , then  $\text{Pr}(g_{s,j} = 0) = 1 - \gamma_s$ . Furthermore,  $\text{Var}(g_{s,j}) = \gamma_s \frac{2}{\lambda_s^2}$ , where  $\lambda_s$  denotes a measure of uncertainty about the effects of the genetic variation source  $s$ . The hyper-parameters  $\gamma_s$  and  $\lambda_s$  are specified for each source, either additive ( $s = a$ ) or dominance ( $s = d$ ).

In a second step, the pairwise epistatic effects are modelled. The genetic effect caused by an interaction between locus  $j$  and  $k$  is denoted as  $g_{s,j,k}$  with  $s \in \{aa, ad, da, dd\}$ . The effect is considered additive  $\times$  additive ( $aa$ ), only if the individual  $i$  is homozygous at the loci  $j$  and  $k$ . It is considered additive  $\times$  dominance ( $ad$ ), when it appears at a homozygous locus  $j$  and a heterozygous locus  $k$  ( $j < k$ ) and dominance  $\times$  additive ( $da$ ) for the reverse case. The dominance  $\times$  dominance effect ( $dd$ ) appears between heterozygous loci. Using the already orthogonalised columns in  $\mathbf{X}_a$  and  $\mathbf{X}_d$ , M1 can be extended to include epistatic effects in a way similar to Kao & Zeng [21]. Let  $s \in \{aa, ad, da, dd\}$ ,

$$\mathbf{y} = \mathbf{1}\mu^* + \mathbf{X}_a\mathbf{g}_a + \mathbf{X}_d\mathbf{g}_d + \sum_s \mathbf{X}_s\mathbf{g}_s + \mathbf{e}. \quad (M2)$$

As an example,  $X_{aa,j,k} = X_{a,j} \cdot X_{a,k}$ , where the symbol  $\cdot$  denotes the element-wise multiplication of column  $j$  and

$k$  of  $X_a$ . Furthermore,  $X_{ad,j,k} = X_{a,j} \cdot X_{d,k}$  is calculated to obtain the coefficients for the effect  $g_{ad,j,k}$ . This way, in total four times  $\frac{m(m-1)}{2}$  epistatic effects are modelled.

The prior remains the same as for M1, but we assume that the probability of having non-zero epistatic effects is smaller than the  $\gamma_s$  for main effects.

## 2.2 Parameter estimation

The essence of the fBayesB approach is the iterative conditional expectation (ICE) algorithm, which is described in detail by Meuwissen *et al.* [20]. We only describe the steps which were adapted under the influence of non-additive genetics. Initially, to get rid of the population mean  $\mu^*$ , we shift the observed phenotypic values by the estimated mean value, thus  $\mathbf{y} \mapsto \mathbf{y} - 1\bar{y}$ . The vector of genetic effects  $\mathbf{g}_s$  has the length  $m_s$ , where  $m_s = m$  for  $s \in \{a, d\}$  and  $m_s = \frac{m(m-1)}{2}$  for  $s \in \{aa, ad, da, dd\}$ . In case of epistasis, the elements are stored in a vector according to a vectorised upper triangular matrix, where only elements above the diagonal are taken, i.e.

$$\mathbf{g}_s = (\mathbf{g}_{s,1,2}, \mathbf{g}_{s,1,3}, \dots, \mathbf{g}_{s,1,m}, \mathbf{g}_{s,2,3}, \mathbf{g}_{s,2,4}, \dots, \mathbf{g}_{s,2,m}, \dots, \mathbf{g}_{s,m-1,m})'$$

We carry out  $k = 1, 2, \dots, k_{\max}$  iterations and process the genetic effects in the order  $s = a, d$  based on M1 or  $s = a, d, aa, ad, da, dd$  based on M2. The genetic effect with index  $j = 1, \dots, m_s$  is estimated as the posterior expectation

$$\hat{\mathbf{g}}_{s,j}^{(k)} = E(\mathbf{g}_{s,j} | \mathbf{y} = \mathbf{y}_{-j}^{(k)}),$$

where  $\mathbf{y}_{-j}^{(k)}$  denotes the vector of observed phenotypes corrected for all estimated effects except the  $j$ -th effect in iteration round  $k$ .

Set  $Y_j = (\mathbf{X}'_{s,j} \mathbf{X}_{s,j})^{-1} \mathbf{X}'_{s,j} \mathbf{y}_{-j}^{(k)}$  and  $\sigma_j^2 = (\mathbf{X}'_{s,j} \mathbf{X}_{s,j})^{-1} \sigma_e^2$ . For convenience we denote  $Y_j^\pm = Y_j \pm \lambda_s \sigma_j^2$ .

Now the conditional expectation was determined analytically in Meuwissen *et al.* [20] as

$$E(\mathbf{g}_{s,j} | \mathbf{y} = \mathbf{y}_{-j}^{(k)}) = \frac{T_1 \Theta_U(0; Y_j^-, \sigma_j^2) + T_2 \Theta_L(0; Y_j^+, \sigma_j^2)}{T_1 + T_2 + T_3}$$

With

$$\begin{aligned} T_1 &= \exp(-\lambda_s Y_j) (1 - \Phi(0; Y_j^-, \sigma_j^2)), \\ T_2 &= \exp(\lambda_s Y_j) \Phi(0; Y_j^+, \sigma_j^2), \\ T_3 &= \frac{2(1 - \gamma_s)}{\gamma_s \lambda_s} \exp\left(-\frac{1}{2} \lambda_s^2 \sigma_j^2\right) \phi(Y_j; 0, \sigma_j^2). \end{aligned}$$

The  $\Theta_U(0; \mu, \sigma^2)$  and  $\Theta_L(0; \mu, \sigma^2)$  are the expected value of an upper and lower truncated normal distribution  $N(\mu, \sigma^2)$ , respectively, with truncation point zero. The  $\Phi(x; \mu, \sigma^2)$  denotes the normal distribution function evaluated at some point  $x$  and  $\phi(x; \mu, \sigma^2)$  is the normal density function.

We introduce a slight modification to fBayesB as we update the estimated residual variance components in each iteration  $k$  by the residual sum of squares

$$\hat{\sigma}_e^{2(k)} = \frac{1}{n-1} \left\| \mathbf{y} - \sum_s \mathbf{X}_s \hat{\mathbf{g}}_s^{(k)} \right\|^2.$$

Then  $\sigma_e^2$  is substituted by  $\hat{\sigma}_e^{2(k)}$  in the calculation of the conditional expectation. The steps above are carried out for all indices  $j$  within each source  $s$  of genetic variation. We continue until the vector of estimates  $\hat{\mathbf{g}}^{(k)} = (\hat{\mathbf{g}}_a^{(k)'}, \hat{\mathbf{g}}_d^{(k)'}, \hat{\mathbf{g}}_{aa}^{(k)'}, \hat{\mathbf{g}}_{ad}^{(k)'}, \hat{\mathbf{g}}_{da}^{(k)'}, \hat{\mathbf{g}}_{dd}^{(k)'})'$  fulfils the convergence criterion

$$\frac{\|\hat{\mathbf{g}}^{(k)} - \hat{\mathbf{g}}^{(k-1)}\|^2}{\|\hat{\mathbf{g}}^{(k)}\|^2} \leq L,$$

otherwise the iterations stop at  $k = k_{\max}$ . The (direct) genetic value  $DGV_i$  of individual  $i$  is obtained as the genome-wide sum over all genotypic values and over the different sources, i.e.

$$DGV_i = \sum_s \mathbf{X}_s \hat{\mathbf{g}}_s.$$

Eventually, as a consequence of the standardisation, the genetic variance components are estimated as  $\hat{\sigma}_s^2 = \sum_{j=1}^{m_s} \hat{\mathbf{g}}_{s,j}^2$  for each genetic variation source  $s$ . Note that this formula yields an approximation under LD because the covariance components  $\text{Cov}(X_{s,i,j} \hat{\mathbf{g}}_{s,j}, X_{s,i,j'} \hat{\mathbf{g}}_{s,j'})$  of potentially linked loci  $j$  and  $j'$  are absent. Under the given re-parametrisation, the covariance  $\text{Cov}(X_{s,i,j}, X_{s,i,j'})$  between genotype coefficients is not necessarily positive and the signs of the corresponding effects are not known. Therefore, it cannot be stated whether over- or underestimation of genetic variance components is expected. We briefly examine the impact of missing linkage information in our simulations.

The suitability of the statistical models M1 and M2 are compared among the different simulated scenarios in terms of accuracy, which is the empirical correlation between predicted and simulated  $DGV$  in a validation set. We implemented this fBayesB approach in Fortran F90.

When studying only the main genetic effects via M1, the results of fBayesB are compared with BayesB [4]. A

Fortran implementation of BayesB of Berry & Stranden is available on <http://www.genomicselection.net> (obtained Sep 4, 2009). This version was extended to include dominance effects using a concatenated matrix ( $X_a X_d$ ). In principle, it would be possible to additionally consider epistatic effects in BayesB, but this tool would probably require a few months to finish an adequate number of MCMC sampling rounds for a single simulated dataset.

### 2.3 Simulation study

#### Data generation

The simulated population is built up in such a way that it reflects a realistic dairy cattle population. We applied a mutation-drift model and simulated a population with effective population size of 100 animals and 52 273 single nucleotide polymorphisms (SNPs) on a 30 Morgan genome (in style of the Illumina Chip BovineSNP50 and based on Btau4.0 [24]). Details of the genome set-up can be found in Melzer *et al.* (Melzer, Wittenburg, Repsilber: Simulating a more realistic genotype-phenotype map for development of methods to predict phenotypes based on genome-wide marker data - the livestock perspective, submitted). Starting with homozygous loci, a mutation rate of  $2.5 \cdot 10^{-3}$  per generation was chosen for each SNP locus and 400 generations of random mating involving recombination events on the genome were carried out. About 10% of the loci were fixed due to drift. The LD was measured as  $r^2$  [25] and the average LD of adjacent SNPs was observed as  $r^2 = 0.12$ . The average SNP heterozygosity was 0.33. The training generations 401 and 402 each consisted of 50 half-sib families with 20 offspring. These individuals were genotyped and phenotyped ( $n = 2\,000$ ). The test generations 403 and 404 were built up the same way but without phenotyping the animals. Two main scenarios were set up which differed in the number of QTL. Either 23 or 230 SNP loci were randomly chosen from loci with minor allele frequency (MAF)  $> 0.02$  in generation 400 to be the QTL. Main genetic effects (i.e. additive and dominance effects) were assigned to all QTL. Motivated by the findings of Hayes and Goddard [26], allele substitution effects were drawn from a gamma distribution with shape parameter  $\alpha = 0.42$  and scale parameter  $\beta = 2.619$  (23-QTL scenario) or  $\beta = 8.282$  (230-QTL scenario) similar to Meuwissen *et al.* [4]. The sign of an allele substitution effect was drawn at random with equal chance. The degree of dominance was drawn from a normal distribution with mean  $m = 0.193$  and variance  $\tau^2 = 0.312^2$  [27]. The dominance effect was determined as the product of the absolute allele substitution effect and the degree of dominance. Epistatic effects were included optionally. This means, the genotypic information was used twice: either genotypic values were

calculated with main effects only (simulation without epistasis) or genotypic values included main and epistatic effect (simulation with epistasis). For each source of epistasis, six (57) pairs of SNPs were randomly chosen out of the 23 (230) loci to cause interactions. Epistatic effects were drawn from normal distributions with arbitrary parameters chosen such that epistasis explained approximately 25% of the total genetic variance. Different parameters were used for each source of epistatic variation; the parameters are listed in Table 1. To obtain residual error terms, which should be comparable between simulations with and without epistasis, the residual variance component was determined depending on the chosen broad-sense heritability of  $H^2 \in \{0.5, 0.3, 0.1\}$ . As an example,  $H^2 = 0.5$  results in a narrow-sense heritability of  $h^2 = 0.474$  ( $h^2 = 0.307$ ) without (with) simulated epistasis in the 23-QTL scenario. The 23-QTL scenario was repeated 100 times for every  $H^2$  and the 230-QTL scenario was repeatedly simulated only for  $H^2 = 0.5$ .

#### Scale of genetic effects

For convenience, the phenotypes were simulated on the basis of an  $F_\infty$  model, but the genetic effects were estimated on the orthogonal scale. We employed the equivalence between the representations of genotypic values in (1) and (2) to obtain the translation between scales [22]. With no epistasis simulated, the allele substitution effect  $a_j$  and dominance effect  $d_j$  were translated into the effects  $g_{a,j}$  and  $g_{d,j}$  on the orthogonal scale by

$$\begin{pmatrix} \mu^* \\ g_{a,j} \\ g_{d,j} \end{pmatrix} = S_{A,j}^{-1} S \begin{pmatrix} \mu \\ a_j \\ d_j \end{pmatrix}.$$

If epistasis was simulated, the genetic effects on the orthogonal scale were determined for all locus combinations  $j$  and  $k$  and the main genetic effects were achieved as the marginal effects. On the  $F_\infty$  scale, we denote the vector of effects  $\alpha_{j,k} = (\mu, a_p, d_p, a_k, aa_{j,k}, da_{j,k}, d_k, ad_{j,k}, dd_{j,k})'$  and on the orthogonal scale  $\alpha_{j,k}^* = (\mu^*, g_{a,j}, g_{d,j}, g_{a,k}, g_{aa,j,k}, g_{da,j,k}, g_{d,k}, g_{ad,j,k}, g_{dd,j,k})'$ . The translation for a single locus combination was

$$\alpha_{j,k}^* = (S_{A,k}^{-1} \otimes S_{A,j}^{-1})(S \otimes S)\alpha_{j,k},$$

**Table 1 Mean (m) and variance ( $\tau^2$ ) of normal distributions to simulate epistatic genetic effects**

	23-QTL scenario	230-QTL scenario
additive $\times$ additive	$m = 0.2, \tau^2 = 0.3$	$m = 0.02, \tau^2 = 0.03$
additive $\times$ dominance	$m = 0.2, \tau^2 = 0.3$	$m = 0.02, \tau^2 = 0.03$
dominance $\times$ additive	$m = 0.2, \tau^2 = 0.2$	$m = 0.02, \tau^2 = 0.02$
dominance $\times$ dominance	$m = 0.2, \tau^2 = 0.1$	$m = 0.02, \tau^2 = 0.01$

which directly led to the epistatic effects on the orthogonal scale. Due to the standardisation step in (3), the derived epistatic effect had to be multiplied by the corresponding scaling term. As an example for  $ad$ ,  $g_{ad,j,k} \mapsto g_{ad,j,k} \sqrt{2p_j(1-p_j)2p_k(1-p_k)}$ . The derivation of main genetic effects was more difficult. In order to avoid double counting, we considered the main effects separately and collected the contribution of interactions over the genome while the main effects were set to zero (this vector is denoted as  $\alpha_{j=0,k=0}$ ). The components of interest were obtained from

$$\begin{pmatrix} g_{a,j} \\ g_{d,j} \end{pmatrix} = \left[ S_{A_j}^{-1} S \begin{pmatrix} \mu \\ a_j \\ d_j \end{pmatrix} \right]_{2,3} \\ + \left[ \sum_{k=1, j < k}^m (S_{A,k}^{-1} \otimes S_{A,j}^{-1}) (S \otimes S) \alpha_{j=0, k=0} \right]_{2,3} \\ + \left[ \sum_{k=1, j > k}^m (S_{A,j}^{-1} \otimes S_{A,k}^{-1}) (S \otimes S) \alpha_{k=0, j=0} \right]_{4,7} .$$

Note that the order of loci (either  $j < k$  or  $k < j$ ) is necessary to assign the contribution of epistasis correctly to the different sources of genetic variation. Again, each main genetic effect was multiplied by the relevant standard deviation term.

#### Hyper-parameters and other settings

The parameter  $\lambda_s$ , which reflects the prior uncertainty about a genetic effect, was determined indirectly through the choice of the total prior variance. For  $s \in \{a, d, aa, ad, da, dd\}$ , we assume that

$$1 = \sum_{j=1}^{m_s} \text{Var}(g_{s,j}) = m_s \gamma_s \frac{2}{\lambda_s^2} \quad (4) \\ \Rightarrow \lambda_s = \sqrt{2m_s \gamma_s} .$$

In this study, we involved prior knowledge about the proportion of non-zero effects of the genetic variation source  $s$  in the simulated dataset and chose  $\gamma_s$  accordingly. In the 23-QTL scenario we set  $\gamma_a = \gamma_d = 0.005$  and  $\gamma_s = 10^{-6}$  for  $s \in \{aa, ad, da, dd\}$ . In the 230-QTL scenario we applied  $\gamma_a = \gamma_d = 0.05$  and  $\gamma_s = 10^{-6}$  for the epistatic effects. We will return to the issue of parameter choice in the Section Discussion.

Furthermore, to limit the number of iterations, we chose  $k_{\max} = 1\,000$  and for the convergence criterion we used  $L = 10^{-8}$  for M1. Owing to the computational effort we set  $k_{\max} = 200$  and  $L = 10^{-6}$  for M2. Results are reported only for those repetitions where convergence was achieved.

In BayesB the main genetic effects were estimated simultaneously over the whole genome. A hyper-

parameter  $\pi$  was required to give the proportion of non-zero genetic effects in total; we set  $\pi = 0.005$  ( $\pi = 0.05$ ) in the 23-QTL (230-QTL) scenario. Furthermore, we carried out 50 000 MCMC iterations (40% were neglected as burn-in) and within each iteration 1 000 rounds of the Metropolis-Hastings algorithm were employed.

#### Outline of data analysis

To begin with, we used every 10-th marker ( $m = 5\,227$ ), which included the true positions of the simulated QTL, in the statistical analysis. With this reduced genotype dataset, we evaluated differences in parameter estimation between fBayesB and BayesB based on the model with additive and dominance effects. A main issue was to study the impact of including or not including pairwise epistatic effects on the accuracy of genetic value prediction with fBayesB. The influence of a varying proportion of genetic variation on the accuracy of prediction was obtained by analysing the data produced with different broad-sense heritabilities. Further, we studied the consequences of spreading the genetic variation over a multitude of loci with almost equal amounts of variation in each source of genetic variation. In a next step, we used all SNP information ( $m = 52\,273$ ) without pre-selection of loci for the estimation of genetic effects and explored the applicability of fBayesB for a large genotype dataset. Finally, to study practical suitability, we estimated genetic effects in a sample of a heterogeneous stock of mice. Genotype and phenotype data are publicly available at [http://gscan.well.ox.ac.uk/\[28\]](http://gscan.well.ox.ac.uk/[28]).

### 3 Results

On average 567 loci per dataset had  $\text{MAF} \leq 0.01$ . These loci were omitted, but loci deviating from HWE (on average one locus per dataset) were not excluded from the analysis. The average LD between adjacent SNPs was  $r^2 = 0.07$  in the reduced genotype dataset with 5 227 SNPs.

The differences between fBayesB and BayesB on the basis of M1 are compared. Table 2 shows the average estimated variance components and the average correlation between predicted and simulated genetic values in the 23-QTL scenario. The accuracy between the methods differed only slightly,  $\rho = 0.98$  when no epistasis was simulated and  $\rho = 0.78$  with simulated epistasis. Both in simulations with and without epistasis, the estimated variance components were similarly biased with BayesB and fBayesB, i.e., the relative bias of the estimate  $\hat{\sigma}_a^2$  was -2% (-7%) and the relative bias of  $\hat{\sigma}_d^2$  was -13% (-26 to -27%) without (with) simulated epistasis. Though fBayesB only required a small fraction of computing time compared to BayesB (one second versus about six hours on a 2.93 GHz multi-user system), there was



**Table 2 Average estimated variance components (standard deviation in brackets) and average accuracy  $\rho$  of genetic value prediction\***

		Simulation without epistasis							
Method	Model	$\sigma_a^2$	$\sigma_d^2$	$\sigma_{aa}^2$	$\sigma_{ad}^2$	$\sigma_{da}^2$	$\sigma_{dd}^2$	$\sigma_e^2$	$\rho$
BayesB	M1	0.743 (0.578)	0.035 (0.039)	-	-	-	-	0.775 (0.605)	0.980
fBayesB	M1	0.742 (0.579)	0.035 (0.039)	-	-	-	-	0.752 (0.587)	0.978
fBayesB	M2	0.748 (0.583)	0.039 (0.041)	0.008 (0.013)	0.007 (0.016)	0.007 (0.014)	0.008 (0.017)	0.638 (0.484)	0.959
<i>Simulated components</i>		0.757	0.040	-	-	-	-	0.798	-
		Simulation with epistasis							
Method	Model	$\sigma_a^2$	$\sigma_d^2$	$\sigma_{aa}^2$	$\sigma_{ad}^2$	$\sigma_{da}^2$	$\sigma_{dd}^2$	$\sigma_e^2$	$\rho$
BayesB	M1	1.313 (0.681)	0.158 (0.131)	-	-	-	-	2.721 (0.874)	0.785
fBayesB	M1	1.310 (0.687)	0.161 (0.132)	-	-	-	-	2.619 (0.845)	0.781
fBayesB	M2	1.338 (0.688)	0.193 (0.142)	0.299 (0.215)	0.138 (0.111)	0.065 (0.071)	0.057 (0.070)	1.811 (0.598)	0.833
<i>Simulated components</i>		1.409	0.217	0.346	0.133	0.089	0.020	2.214	-

\*23-QTL scenario with 5 227 markers and  $H^2 = 0.5$ . Variance components for each source of genetic variation:  $\sigma_a^2$  additive genetic,  $\sigma_d^2$  dominance,  $\sigma_{aa}^2$  additive  $\times$  additive,  $\sigma_{ad}^2$  additive  $\times$  dominance,  $\sigma_{da}^2$  dominance  $\times$  additive,  $\sigma_{dd}^2$  dominance  $\times$  dominance; residual variance  $\sigma_e^2$ . M1 includes additive and dominance effects, M2 includes additive, dominance and pairwise epistatic effects.

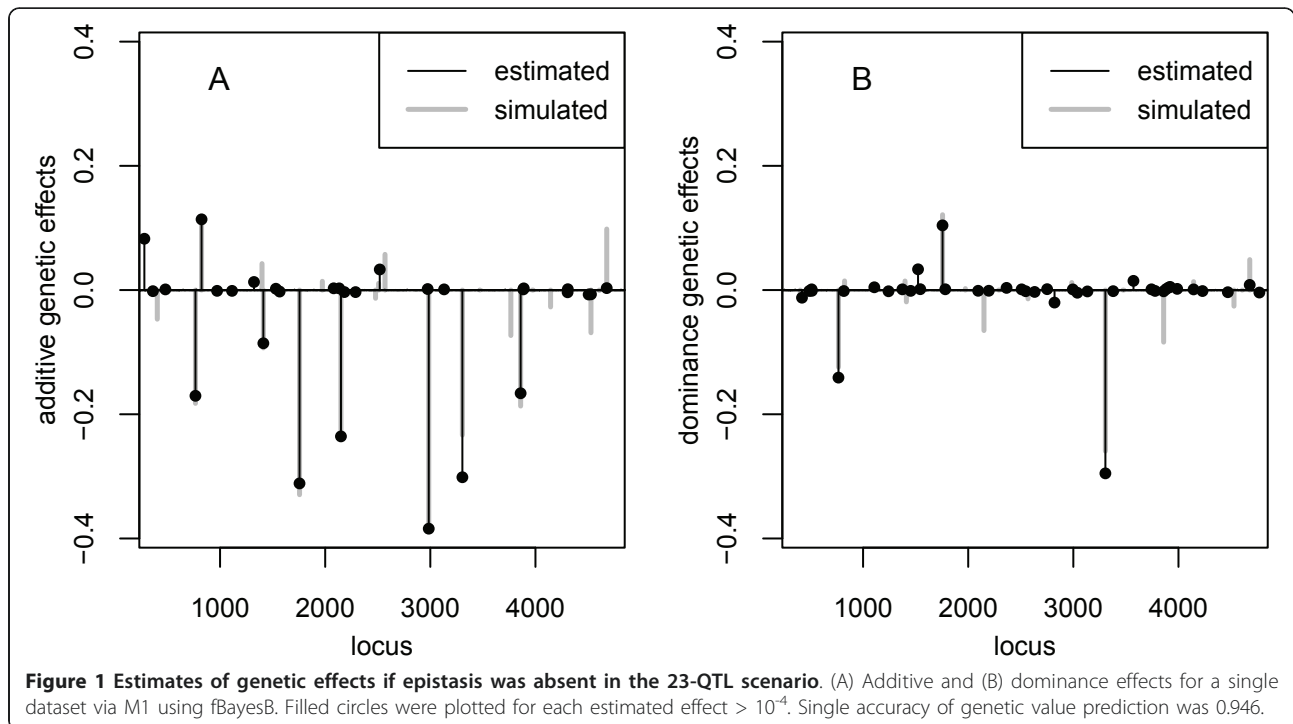
neither a lack of accuracy nor differences in bias of variance component estimation.

The additive and dominance effects were estimated equally well with both BayesB and fBayesB. As an example, Figure 1 displays results for the analysis of a single dataset via fBayesB. It shows that the size and location of large to intermediate additive effects were estimated precisely and the pivotal dominance effects were identified closely. In general, there were nearly no differences in the size of estimates of rather large effects and their position between BayesB and fBayesB. It was observed that via BayesB a lot of tiny (but with an effect size  $> 10^{-4}$ ) genetic effects were estimated over the whole genome, whereas fBayesB concentrated on the large loci.

M1 and M2 results are compared to study the impact of including or not including pairwise epistatic effects on the accuracy of predicting the genetic values in the test generations. As an example, Additional file 1 shows the estimated main genetic and epistatic effects for a single dataset when main and epistatic effects were simulated and modelled jointly. The size and position of large main or large epistatic effects were estimated quite well (visual inspection). Small effects, especially concerning dominance, were neither estimated with correct size nor at the simulated position. When epistasis was simulated in the 23-QTL scenario, we obtained an accuracy of 0.781 with M1, which was 5% less than the accuracy based on the correct model M2 for this application, see Table 2. Furthermore, the genetic variance components were underestimated to a larger extent with M1 than

with M2. The relative bias of  $\hat{\sigma}_a^2$  and  $\hat{\sigma}_d^2$  was -7% and -26%, respectively, based on M1 and -5% and -11%, respectively, based on the correct model M2. In the reverse case, when epistasis was modelled and not simulated, the accuracy was 0.959. Hence the loss of accuracy of genetic value prediction was only 2% when the incorrect model M2 was applied. The relative bias of  $\hat{\sigma}_a^2$  and  $\hat{\sigma}_d^2$  was -1% and -3%, respectively, based on the incorrect model M2 compared to -2% and -13%, respectively, based on the correct model M1. Thus, even with additionally modelled (nuisance) genetic effects in M2, the bias of variance component estimates did not increase for additive and dominance effects. In conclusion, and as expected, we obtained the best estimates of genetic variance components and the highest possible accuracy in the validation set, when M1 was applied in simulations without epistasis and M2 was used under simulated epistasis, i.e., prediction was done with the true model. The loss of accuracy was, however, low when the incorrect model was applied. The relative proportion of genetic variation that could be assigned to the variation of additive effects was estimated best if the correct model was applied. As an example, in the 23-QTL scenario with simulated epistasis, the true ratio of additive to total genetic variance was 0.613. The estimated ratio was 0.626 based on M2 but 0.884 based on M1, see Table 3.

The results obtained so far are based on  $H^2 = 0.5$ . The influence of a varying proportion of genetic variation in



terms of the broad-sense heritability  $H^2$  on the accuracy of prediction was investigated. Table 4 displays the decreasing accuracy with decreasing  $H^2$ . Simulations with  $H^2 = 0.3$  and  $H^2 = 0.5$  yielded similar accuracies with M1; accuracy differed about 3 - 5%. With M2 the differences in accuracy were 6 - 12%. With  $H^2 = 0.1$  the differences decreased further about 11 - 38%. If the proportion of the genetic variation was 0.1, fBayesB had numerical problems with M2 under the given choice of hyper-parameters; the algorithm converged to a final solution only in 40% of the repetitions (90% for  $H^2 = 0.3$ , 99.5% for  $H^2 = 0.5$ ). In repetitions that did not converge ( $H^2 = 0.1$ : 3.5%,  $H^2 = 0.3$ : 0.5%,  $H^2 = 0.5$ : 0%) a

fluctuating convergence criterion was observed. In all other cases, the algorithm collapsed for no obvious reason.

In order to prove that we benefit from additionally modelling non-additive genetic effects if those were simulated, we compared the accuracy of genetic value prediction based on M1 with accuracy obtained from a conventional model including only additive genetic effects, called M0. Except for constellations with  $H^2 = 0.1$ , accuracy of M1 was 1-2% (3-4%) higher in simulations without (with) epistasis than accuracy of M0, see Table 4. If we looked at the 10% animals with best

**Table 3** Average ratio of additive genetic variance to total genetic variance\*

Model	Simulation without epistasis	
	23-QTL scenario	230-QTL scenario
M1	0.953	0.810
M2	0.918	0.581
Simulated ratio	0.948	0.945
Model	Simulation with epistasis	
	23-QTL scenario	230-QTL scenario
M1	0.884	0.773
M2	0.626	0.401
Simulated ratio	0.613	0.648

\*fBayesB was used in both QTL scenarios with 5 227 markers and  $H^2 = 0.5$ . M1 includes additive and dominance effects, M2 includes additive, dominance and pairwise epistatic effects.

**Table 4** Average accuracy of genetic value prediction depending on broad-sense heritability  $H^2$ \*

Model	Simulation without epistasis			
	$H^2 = 0.5$	$H^2 = 0.3$	$H^2 = 0.1$	$H^2 = 0.5$ best 10%
M0	0.958	0.940	0.859	0.786
M1	0.978	0.953	0.844	0.774
M2	0.959	0.897	0.640	0.748
Model	Simulation with epistasis			
	$H^2 = 0.5$	$H^2 = 0.3$	$H^2 = 0.1$	$H^2 = 0.5$ best 10%
M0	0.741	0.707	0.581	0.618
M1	0.781	0.736	0.582	0.621
M2	0.833	0.718	0.339	0.598

\*fBayesB was used in the 23-QTL scenario with 5 227 markers. M0 includes only additive genetic effects, M1 includes additive and dominance effects, M2 includes additive, dominance and pairwise epistatic effects. In case of "best 10%" the accuracy of additive genetic value prediction was determined based on 10% animals with best predicted additive genetic value.

predicted additive genetic value (i.e. the breeding value) in simulations with epistasis and  $H^2 = 0.5$ , the accuracy of additive genetic value prediction was 0.618 with M0, 0.621 with M1 and 0.598 with the correct model M2. If we look at the 10% best animals when epistasis was not simulated, the accuracy of additive genetic value prediction was 0.786 based on M0, 0.774 with the correct model M1 and 0.748 with M2. Thus, model choice had an impact on predicting the total genetic values, but if only the extreme breeding values were of interest, e.g. for selection purposes, prediction with a conventional model (M0) was more precise than with the corresponding true model.

In a further step, we studied the consequence when the genetic variation was spread over a multitude of loci and compare results obtained with BayesB and fBayesB. Furthermore, the 230-QTL scenario is confronted with the outcomes of fBayesB in the 23-QTL case. When epistasis was not simulated in the 230-QTL scenario, highest accuracy of genetic value prediction was obtained with M1, see Table 5. Though BayesB had a higher relative bias of  $\hat{\sigma}_a^2$  (-11% vs. -1%) but crucially smaller bias of  $\hat{\sigma}_d^2$  (30% vs. 284%) compared to fBayesB, accuracy was 10% higher. Apparently, dominance has an important impact on genetic value prediction and BayesB could better cope with a larger amount of QTL. fBayesB was able to identify large to intermediate effects, see e.g. Figure 2, but small effects could not be precisely uncovered. BayesB

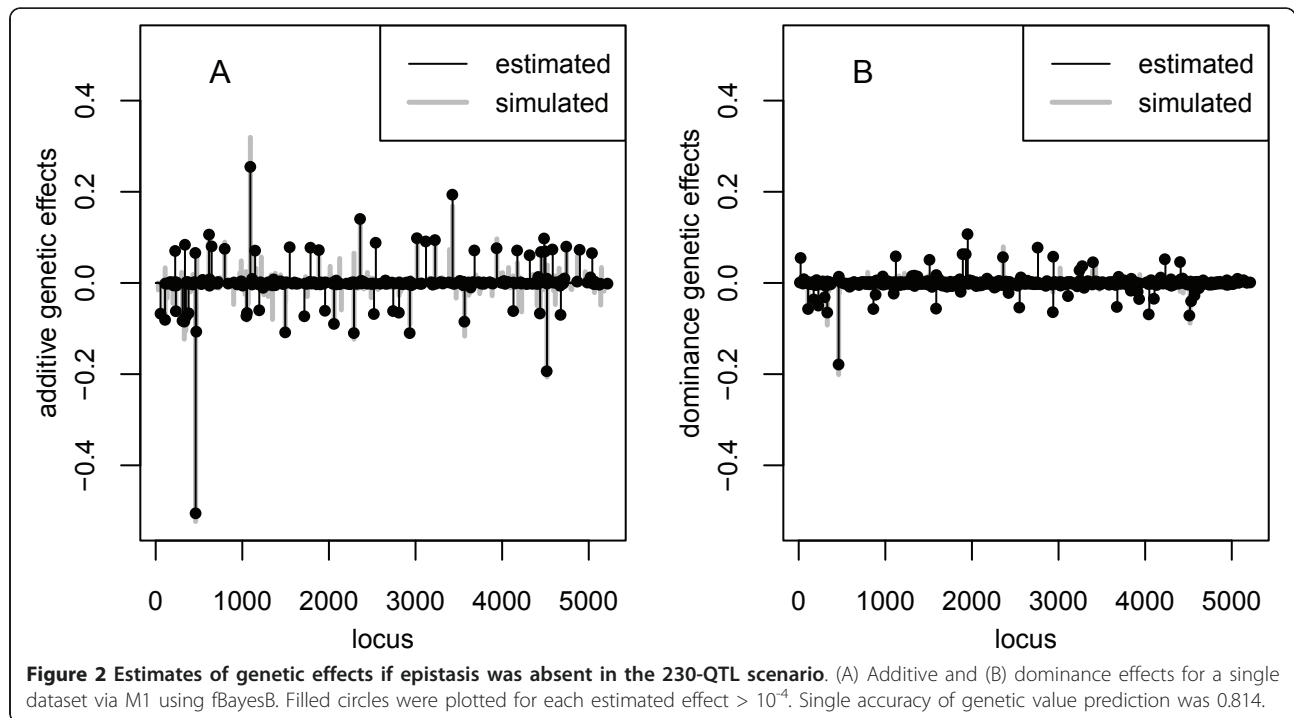
was also superior to fBayesB in terms of accuracy and bias of variance component estimation based on M1 in simulations with epistasis, see Table 5. In any case, fBayesB extremely overestimated variance components of non-additive effects. With application of M1 and fBayesB, the proportion of additive genetic variation to the total genetic variance was underestimated (overestimated) about 13% without (with) simulated epistasis (Table 3). On the basis of M2 this proportion was underestimated by about 25 to 36%.

The more QTL were simulated, the less accuracy was observed. If a 10-fold of QTL was responsible for genetic variation, the accuracy of prediction decreased about 22-24% based on M1 and 35-49% based on M2. Since the distances between QTL were smaller than in the 23-QTL scenario, we could expect that LD between loci contributed to the bias of the estimated variance components. For that reason we calculated the empirical variances obtained from the predicted effect-specific genetic values in the validation set, where the epistatic contribution was collected in one component. Table 6 shows that, if only few QTL were given, the missing LD information could be ignored, no matter if epistasis was regarded or not. In contrast, the empirical variance components clearly deviated from those estimated under LE in the 230-QTL scenario, especially if epistasis was modelled. Consequently, the reported variance components in Tables 2, 3 and 5 can only be interpreted as approximations.

**Table 5 Average estimated variance components (standard deviation in brackets) and average accuracy  $\rho$  of genetic value prediction\***

		Simulation without epistasis							
Method	Model	$\sigma_a^2$	$\sigma_d^2$	$\sigma_{aa}^2$	$\sigma_{ad}^2$	$\sigma_{da}^2$	$\sigma_{dd}^2$	$\sigma_e^2$	$\rho$
BayesB	M1	0.631 (0.204)	0.056 (0.035)	-	-	-	-	0.652 (0.180)	0.860
fBayesB	M1	0.699 (0.207)	0.165 (0.065)	-	-	-	-	0.413 (0.132)	0.760
fBayesB	M2	0.732 (0.214)	0.304 (0.112)	0.036 (0.028)	0.065 (0.036)	0.068 (0.042)	0.074 (0.046)	0.170 (0.066)	0.608
<i>Simulated components</i>		<i>0.709</i>	<i>0.043</i>	-	-	-	-	<i>0.754</i>	-
		Simulation with epistasis							
Method	Model	$\sigma_a^2$	$\sigma_d^2$	$\sigma_{aa}^2$	$\sigma_{ad}^2$	$\sigma_{da}^2$	$\sigma_{dd}^2$	$\sigma_e^2$	$\rho$
BayesB	M1	0.949 (0.250)	0.215 (0.067)	-	-	-	-	1.968 (0.266)	0.585
fBayesB	M1	0.920 (0.197)	0.267 (0.080)	-	-	-	-	1.567 (0.282)	0.543
fBayesB	M2	1.277 (0.230)	0.910 (0.269)	0.171 (0.086)	0.275 (0.106)	0.296 (0.127)	0.305 (0.126)	0.493 (0.257)	0.340
<i>Simulated components</i>		<i>1.284</i>	<i>0.192</i>	<i>0.308</i>	<i>0.103</i>	<i>0.071</i>	<i>0.014</i>	<i>1.952</i>	-

\*230-QTL scenario with 5 227 markers and  $H^2 = 0.5$ . Variance components for each source of genetic variation:  $\sigma_a^2$  additive genetic,  $\sigma_d^2$  dominance,  $\sigma_{aa}^2$  additive  $\times$  additive,  $\sigma_{ad}^2$  additive  $\times$  dominance,  $\sigma_{da}^2$  dominance  $\times$  additive,  $\sigma_{dd}^2$  dominance  $\times$  dominance, residual variance  $\sigma_e^2$ . M1 includes additive and dominance effects, M2 includes additive, dominance and pairwise epistatic effects.



Next we used the genome-wide SNP information in the statistical analysis ( $m = 52\,273$ ). An average of 5 685 loci per dataset were omitted because  $MAF \leq 0.01$ . An average of nine loci deviated from HWE, but these loci were retained. We set  $\gamma_a = \gamma_d = 0.005$  for both QTL scenarios. If only main genetic effects were simulated and modelled in the 23-QTL scenario with  $H^2 = 0.5$ , the additive genetic variance component was obtained as

$\hat{\sigma}_a^2 = 0.796$  ( $se = 0.588$ ), whereas the dominance variance component was extremely overestimated as  $\hat{\sigma}_d^2 = 0.493$  ( $se = 0.354$ ). The accuracy  $\rho = 0.723$  was still reasonably high. In the 230-QTL scenario, the accuracy of prediction reduced to  $\rho = 0.513$  and  $\hat{\sigma}_a^2 = 0.730$  ( $se = 0.251$ ), but  $\hat{\sigma}_d^2 = 0.729$  ( $se = 0.308$ ) was not estimated as well as with the reduced SNP set on the basis of M1. Including the pairwise epistatic effects via M2

**Table 6** Comparison of empirical variances of predicted genetic values and genetic variance components estimated under LE\*

		Simulation without epistasis					
		23-QTL scenario			230-QTL scenario		
Model		$\sigma_a^2$	$\sigma_d^2$	$\sigma_{epi}^2$	$\sigma_a^2$	$\sigma_d^2$	$\sigma_{epi}^2$
M1	empirical	0.743	0.035	-	0.711	0.163	-
	under LE	0.742	0.035	-	0.699	0.165	-
M2	empirical	0.749	0.038	0.030	0.805	0.338	0.278
	under LE	0.748	0.039	0.030	0.732	0.304	0.243
		Simulation with epistasis					
		23-QTL scenario			230-QTL scenario		
Model		$\sigma_a^2$	$\sigma_d^2$	$\sigma_{epi}^2$	$\sigma_a^2$	$\sigma_d^2$	$\sigma_{epi}^2$
M1	empirical	1.309	0.161	-	0.981	0.266	-
	under LE	1.310	0.161	-	0.920	0.267	-
M2	empirical	1.332	0.192	0.554	1.442	1.112	1.277
	under LE	1.338	0.193	0.559	1.277	0.910	1.047

\*fBayesB was used in both QTL scenarios with 5 227 markers and  $H^2 = 0.5$ . Estimates were obtained as empirical variances of effect-specific genetic values predicted in the validation set (rows "empirical") or as genome-wide sum of locus-specific genetic variances which coincides with the assumption of LE (rows "under LE"). Variance components for each source of genetic variation:  $\sigma_a^2$  additive genetic,  $\sigma_d^2$  dominance,  $\sigma_{epi}^2$  joint contribution of all epistatic effects. M1 includes additive and dominance effects, M2 includes additive, dominance and pairwise epistatic effects, LE linkage equilibrium.



exceeded practicability. On the basis of 5 227 SNP, more than 13 million effects had to be estimated for each source of epistatic variation and fBayesB required an average of six hours to converge. If 52 273 SNP markers are included, then approximately 1.3 billion effects have to be estimated for each of the four sources of epistasis. Though most markers or pairs of markers have no effect, their estimated genetic effects will be small but not exactly zero. It was not feasible to estimate about 5 billion effects via M2 under proper numerical precision owing to the restricted capacity of memory space. Furthermore, it is questionable how much computing time is required to execute several rounds of iteration. Thus, with 52 273 SNP markers, only M1 was applied to the simulated data with epistasis. This led to a reduced accuracy of  $\rho = 0.611$  ( $\rho = 0.380$ ) in the 23-QTL scenario (230-QTL scenario).

Finally, in the real data example, we regarded  $m = 9$  441 SNPs which passed the standard quality checks on HWE and MAF. Rarely missing genotypes for these SNPs were imputed via Beagle 3.2 [29]. We studied an immunological phenotype, i.e. percentage of CD8<sup>+</sup> cells, and standardised the vector of observations ( $n = 1$  521) to avoid numerical problems. A set of covariates was considered similar to Valdar *et al.* [30]: gender, age, family, litter, cage density, experimenter, month and year of experiment. Phenotypes were corrected for the least-squares estimates of these factors in each iteration of the fBayesB algorithm [20]. We set  $\gamma_a = \gamma_d = 0.001$  and  $\gamma_s = 10^{-6}$  for the epistatic effects. Narrow-sense heritability was similarly estimated among the models (M0:  $h^2 = 0.294$ , M1:  $h^2 = 0.295$ , M2:  $h^2 = 0.317$ ), which shows robustness of fBayesB in terms of additive genetic variation, see Table 7. Broad-sense heritability increased with growing model complexity (M1:  $H^2 = 0.347$ , M2:  $H^2 = 0.448$ ). Figures depicting estimated effect sizes are given in Additional file 2. The largest effects were observed in the MHC region on chromosome 17, which was also reported by Valdar *et al.* [28]. In total, 88% (65%) of the genetic variation was observed around the MHC with M1 (M2). Though additive genetic effect sizes were nearly the same with all models, an additional

dominance effect appeared with M2 on chromosome 17. Furthermore, a large epistatic effect occurred between chromosomes 1 and 8. Thus, adding epistatic effects to a statistical model may not necessarily improve genetic value prediction, as investigated by Lee *et al.* [7] (see Section Background), but it helps to specify sources of genetic variation and to identify loci that contribute to variation only through interactions.

## 4 Discussion

### 4.1 Hyper-parameters and convergence

When we investigated the influence of a varying proportion of genetic to phenotypic variance on genetic value prediction in the 23-QTL scenario, it was observed that fBayesB did not fulfil the convergence criterion in all situations. In the extreme case with M2 and  $H^2 = 0.1$ , only 40% of all repetitions converged to a proper final solution and it happened that fBayesB simply aborted. (Usually, the algorithm converged after 13-16 iterations with M1 and after 26-28 steps with M2, but up to a 5-fold of iteration steps were necessary in the 230-QTL scenario.) In order to avoid termination, one could tune the “free” hyper-parameter  $\lambda_s$ , which is responsible for the variation of a genetic effect a priori. For convenience, we assumed that the total prior variance was equal to one for each source of genetic variation  $s \in \{a, d, aa, ad, da, dd\}$ , see Equation (4). This prior guess depends on the hyper-parameter  $\gamma_s$  which was equal among  $s \in \{a, d\}$  and  $s \in \{aa, ad, da, dd\}$ . Thus, it seems necessary to specifically adjust  $\lambda_s$  and/or  $\gamma_s$  to each source of genetic variation.

### 4.2 Proportion of non-zero effects

A preliminary study could show that the choice of the hyper-parameter  $\gamma_s$  strongly influenced the accuracy of genetic value prediction and the ability of the fBayesB algorithm to converge (Melzer, Wittenburg, Repsilber: Simulating a more realistic genotype-phenotype map for development of methods to predict phenotypes based on genome-wide marker data - the livestock perspective, submitted). Since the aim of this paper was to investigate the suitability of fBayesB to cope with non-additive effects in general, we simply involved prior knowledge about the proportion of non-zero genetic effects when the hyper-parameter  $\gamma_s$  had to be specified for each genetic variation source  $s \in \{a, d, aa, ad, da, dd\}$ . There are several possibilities which allow for a flexible setting of this hyper-parameter. In an intuitive manner, one could determine  $\gamma_s$  via cross-validation as it was done in Melzer *et al.* Another encouraging approach was presented by Shepherd *et al.* [31], therein called emBayesB. It is a BayesB-like estimation of SNP effects without the time consuming Metropolis-Hastings algorithm, but with an EM algorithm for the estimation of  $\gamma_s$ . It employs a binary variable indicating

**Table 7 Estimated variance components for the real data example\***

Model	$\sigma_a^2$	$\sigma_d^2$	$\sigma_{aa}^2$	$\sigma_{ad}^2$	$\sigma_{da}^2$	$\sigma_{dd}^2$	$\sigma_e^2$
M0	0.169	-	-	-	-	-	0.405
M1	0.171	0.030	-	-	-	-	0.378
M2	0.174	0.046	0.000	0.000	0.026	0.000	0.303

\*Variance components for each source of genetic variation:  $\sigma_a^2$  additive genetic,  $\sigma_d^2$  dominance,  $\sigma_{aa}^2$  additive  $\times$  additive,  $\sigma_{ad}^2$  additive  $\times$  dominance,  $\sigma_{da}^2$  dominance  $\times$  additive,  $\sigma_{dd}^2$  dominance  $\times$  dominance; residual variance  $\sigma_e^2$ . M0 includes only additive genetic effects, M1 includes additive and dominance effects; M2 includes additive, dominance and pairwise epistatic effects.

whether a marker is in LD with a QTL (i.e., it is a non-zero effect). So far, this approach was verified for additive genetic effects via simulations, but it is certainly applicable to the non-additive case as well. Not only the specification of the proportion of non-zero effects in the prior setting, however, is important. This study additionally showed that the more loci were responsible for genetic variation, the worse the genetic parameters were estimated, even though we accounted for this proportion in  $\gamma_s$ . With higher proportions of small to intermediate genetic effects, the bias of estimation seriously accumulates via fBayesB. One way out is to reduce noise by eliminating zero effects. This objective is discussed in the next section.

#### 4.3 Reduction of model dimensionality

SNP density continues to increase; soon whole-genome sequences will be used for statistical analysis [32]. The ability to uncover genetic effects with Bayesian MCMC methods worsens with increasing LD due to redundancy between markers [33]. Thus, in order to deal with the huge amounts of data, it becomes important to select relevant information. Selection is conceivable in two general ways.

In order to keep as many parameters as required in the statistical model, one could apply a filtering procedure. The significance of putative non-zero effects might be determined, for example, via a stochastic variable selection approach (SVS). In the field of genomic selection, which is based only on additive effects, an SVS implementation of Meuwissen and Goddard [34] was applied by Calus *et al.* [35] to simulated data as well as by Verbyla *et al.* [36] to dairy cattle data. In case of additional non-additive effects, SVS was developed and already successfully applied to obesity data in a mouse backcross population [37]. In that work, an upper bound of model dimensionality had to be fixed and indicator variables were involved specifying which main and epistatic effect had to be included in the model. The Bayes factor then gave evidence of putative QTL.

Dimensionality can also be reduced non-parametrically. As an example, a subset of SNPs may be selected via filtering based on entropy information and wrapping using a naive Bayesian classifier [38]. Alternatively, an informative set of SNPs can be identified on the basis of LD between loci, called tagSNP [39]. This strategy would probably also reduce the bias in variance component estimation due to LD, because only one marker represents a certain chromosome segment. A haplotyping strategy based on LD information was applied to SNP data in Australian beef cattle [40] but with limited success. The authors reported that about 30 000 SNP markers (and a large number of phenotypic records) are required for accurate breeding value prediction. Thus, we have to work with some contradiction: more markers

for higher accuracy but less markers (or only the best markers) to reduce estimation errors. The best solution is probably obtained, when the models used are better able to distinguish between markers with and without effect. Meuwissen [41] presented other options to reduce a set of SNPs based on LD between loci or relatedness between individuals.

#### 4.4 Non-additive effects

This study has shown that the inclusion of dominance effects in genetic value prediction improved accuracy compared to purely additive models (Table 4). We found that the incorporation of dominance effects was less challenging than the inclusion of epistasis, and we have made a robust step towards advancing insight into the genetic architecture. Regardless of whether dominance or epistatic effects are considered, adequate data are required to estimate non-additive effects. This is also true for periodic re-estimation of genetic effects. In contrast to genomic selection, where additive effects may be obtained from average yields of progeny of genotyped parents, genotyped individuals need to have an own phenotype (e.g. cows).

In general, and also confirmed in our investigations, parametric methods have difficulties to identify and to estimate epistatic effects. One reason is that the orthogonal decomposition of genetic effects only lead to proper results under idealised conditions (LE, absence of mutation and selection etc.) which are violated in practice [42]. As reviewed and discussed by Calus [5], non-parametric methods (e.g. [43]) have the potential to outperform parametric approaches if non-additive effects are included. With an application to broiler data [44], it was shown that kernel methods had a better predictive ability than parametric methods when genome-wide markers were used. For thousands of SNPs and millions of interactions, fBayesB is still computationally feasible but it shows an inherent bias of variance component estimation. Alternatively, machine learning techniques may discover hidden patterns of gene interaction without assuming their structure [45].

Once gene interactions are discovered, they may be used for mate allocation in livestock breeding, where individuals are mated to achieve favourable non-additive gene combinations to further increase genetic gain [46]. Apart from breeding applications, improved statistical modelling [41] and our cognitive interest in the formation of complex phenotypes will benefit from knowledge about the distribution of non-additive effects over the genome and their size.

#### 4.5 Number of simulated QTL

An increase in the number of QTL was accompanied by a reduction in the quality of fBayesB for genetic value

prediction. fBayesB was able to identify only the biggest QTL effects in the simulated scenarios, in which (nearly) the same amount of genetic variation was spread over 23 or 230 QTL. Thus, effect size in the 230-QTL scenario was roughly one-tenth of that in the 23-QTL case. This complicated the identification of genetic effects in general and, in particular, of non-additive effects, which contributed very little to the genetic variance when compared with additive effects. Many tiny effects were estimated with BayesB, even if genetic variation was caused by few QTL with large effects. In both QTL scenarios, accuracy of genetic value prediction was at a high level with BayesB. It may be more realistic to assume that most livestock traits are influenced by many loci and therefore best results can be expected with BayesB.

## 5 Conclusion

This simulation study showed that the fast Bayesian method (fBayesB) is convenient for genetic value prediction. It requires only a fraction of computing time compared to a conventional MCMC approach BayesB and also enables estimating pairwise interactions.

The number of simulated QTL, the proportion of genetic to phenotypic variance as well as the quantity of SNP in statistical analyses influenced accuracy of genetic value prediction and bias of variance component estimation. Both methods obtained similar results when few QTL with additive and dominance effects were simulated; the maximum accuracy was 98%. As expected, best results were obtained on the basis of the true model corresponding to the simulated scenario, but the loss of accuracy due to using the incorrect model was limited to 2-5%. If many QTL were responsible for genetic variation, accuracy decreased about 22-49% with fBayesB compared to the few QTL scenario, depending on the model. Accuracy based on modelling only additive and dominance effects was generally superior to the complex model, no matter if epistasis was simulated or not, and an additional gain of 4-10% accuracy was observed with BayesB. To sum up, existing approaches for genome-wide estimation of additive genetic effects can easily and robustly be extended by dominance effects to improve accuracy of genetic value prediction and to get further insight into the genetic architecture. In this simulation study, the inclusion of dominance was more important than involving all pairwise interactions, which did not improve prediction in general.

## Additional material

**Additional file 1: The figure shows estimates of genetic effects and location if epistasis was present in the 23-QTL scenario: (a) additive, (b) dominance, (c) additive × additive and (d) additive × dominance**

effects for a single dataset with M2 using fBayesB. Filled circles were plotted for each estimated effect  $>10^{-4}$ . Location of (e) additive × additive and (f) additive × dominance epistatic effects. Single accuracy of genetic value prediction was 0.851.

**Additional file 2: The fBayesB approach was applied to public data on a heterogeneous stock of mice.** Genetic effects were estimated based on the different models including only additive effects (M0), additive and dominance effects (M1), additive, dominance and pairwise epistatic effects (M2).

## Acknowledgements

This study is part of the FUGATO project "Bovine Integrative Bioinformatics for Genomic Selection (BovIBI)" with financial support of the German Federal Ministry of Education and Research (BMBF).

## Authors' contributions

DW implemented the statistical methods, carried out the analysis and wrote the manuscript. NM simulated the datasets and contributed to the data analysis. NR raised the initial question, advised on the research and suggested improvements to the manuscript. All authors have read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

Received: 27 April 2011 Accepted: 25 August 2011

Published: 25 August 2011

## References

- Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, Medland S, Hill WG, Hottenga JJ, Willemsen G, Boomsma DI, Liu YZ, Deng HW, Montgomery GW, Martin NG: **Genome partitioning of genetic variation for height from 11,214 sibling pairs.** *American Journal of Human Genetics* 2007, **81**(5):1104-1110.
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: Genomic selection in dairy cattle: progress and challenges.** *Journal of Dairy Science* 2009, **92**(2):433-443.
- Legarra A, Robert-Granié C, Manfredi E, Elsen JM: **Performance of genomic selection in mice.** *Genetics* 2008, **180**:611-618.
- Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**(4):1819-1829.
- Calus MPL: **Genomic breeding value prediction: methods and procedures.** *animal* 2010, **4**(02):157-164.
- Melchinger AE, Utz HF, Piepho HP, Zeng ZB, Schön CC: **The Role of Epistasis in the Manifestation of Heterosis: A Systems-Oriented Approach.** *Genetics* 2007, **177**(3):1815-1825[http://www.genetics.org/content/177/3/1815.abstract].
- Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM: **Predicting unobserved phenotypes for complex traits from whole-genome SNP data.** *PLoS Genetics* 2008, **4**(10):e1000231.
- Hu Z, Li Y, Song X, Han Y, Cai X, Xu S, Li W: **Genomic value prediction for quantitative traits under the epistatic model.** *BMC Genetics* 2011, **12**:15 [http://www.biomedcentral.com/1471-2156/12/15].
- Hill WG, Goddard ME, Visscher PM: **Data and theory point to mainly additive genetic variance for complex traits.** *PLoS Genetics* 2008, **4**(2): e1000008.
- Carlborg Ö, Haley CS: **Epistasis: too often neglected in complex trait studies?** *Nature Reviews Genetics* 2004, **5**(8):618-625.
- Carlborg Ö, Kerje S, Schütz K, Jacobsson L, Jensen P, Andersson L: **A global search reveals epistatic interaction between QTL for early growth in the chicken.** *Genome Research* 2003, **13**(3):413-421.
- Beavis W: *QTL analyses: power, precision, and accuracy* CRC Press; 1998.
- Rönnegård L, Besnier F, Carlborg Ö: **An Improved Method for Quantitative Trait Loci Detection and Identification of Within-Line Segregation in F<sub>2</sub> Intercross Designs.** *Genetics* 2008, **178**:2315-2326.
- Zimmer D, Mayer M, Reinsch N: **Complex Genetic Effects in Quantitative Trait Locus Identification: A Computationally Tractable Random Model for Use in F<sub>2</sub> Populations.** *Genetics* 2011, **187**:261.

15. Xu S: **Estimating polygenic effects using markers of the entire genome.** *Genetics* 2003, **163**(2):789-801.
16. Xu S: **An empirical Bayes method for estimating epistatic effects of quantitative trait loci.** *Biometrics* 2007, **63**(2):513-521.
17. Fernando RL, Habier D, Stricker C, Dekkers JCM, Totir LR: **Genomic Selection.** *Acta Agriculturae Scandinavica, Section A - Animal Sciences* 2007, **57**:192-195.
18. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177**(4):2389-2397.
19. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH: **Reducing dimensionality for prediction of genome-wide breeding values.** *Genetics Selection Evolution* 2009, **41**:29.
20. Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA: **A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value.** *Genetics Selection Evolution* 2009, **41**:2[http://www.gsejournal.org/content/41/1/2].
21. Kao CH, Zeng ZB: **Modeling epistasis of quantitative trait loci using Cockerham's model.** *Genetics* 2002, **160**(3):1243-1261.
22. Álvarez-Castro JM, Carlborg Ö: **A unified model for functional and statistical epistasis and its application in quantitative trait Loci analysis.** *Genetics* 2007, **176**(2):1151-1167.
23. Cockerham CC: **An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present.** *Genetics* 1954, **39**(6):859-882.
24. The Bovine Genome Sequencing and Analysis Consortium: **The genome sequence of taurine cattle: a window to ruminant biology and evolution.** *Science* 2009, **324**(5926):522-528.
25. Hill W, Robertson A: **Linkage disequilibrium in finite populations.** *TAG Theoretical and Applied Genetics* 1968, **38**(6):226-231.
26. Hayes B, Goddard ME: **The distribution of the effects of genes affecting quantitative traits in livestock.** *Genetics Selection Evolution* 2001, **33**(3):209-229.
27. Bennewitz J, Meuwissen THE: **The distribution of QTL additive and dominance effects in porcine F2 crosses.** *Journal of Animal Breeding and Genetics* 2010, **127**(3):171-179.
28. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JNP, Mott R, Flint J: **Genome-wide genetic association of complex traits in heterogeneous stock mice.** *Nature Genetics* 2006, **38**(8):879-887.
29. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *American Journal of Human Genetics* 2007, **81**(5):1084-1097.
30. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JNP, Mott R, Flint J: **Genetic and environmental effects on complex traits in mice.** *Genetics* 2006, **174**(2):959-984.
31. Shepherd RK, Meuwissen TH, Woolliams JA: **Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers.** *BMC Bioinformatics* 2010, **11**:529.
32. Meuwissen T, Goddard M: **Accurate prediction of genetic values for complex traits by whole-genome resequencing.** *Genetics* 2010, **185**(2):623-631.
33. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R: **Additive genetic variability and the Bayesian alphabet.** *Genetics* 2009, **183**:347-363.
34. Meuwissen THE, Goddard ME: **Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data.** *Genetics Selection Evolution* 2004, **36**(3):261-279.
35. Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF: **Accuracy of genomic selection using different methods to define haplotypes.** *Genetics* 2008, **178**:553-561.
36. Verbyla KL, Hayes BJ, Bowman PJ, Goddard ME: **Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle.** *Genet Res* 2009, **91**(5):307-311.
37. Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D: **Bayesian model selection for genome-wide epistatic quantitative trait loci analysis.** *Genetics* 2005, **170**(3):1333-1344.
38. Long N, Gianola D, Rosa GJM, Weigel KA, Avendaño S: **Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers.** *Journal of Animal Breeding and Genetics* 2007, **124**(6):377-389.
39. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.** *American Journal of Human Genetics* 2004, **74**:106-120.
40. Hayes BJ, Chamberlain AJ, McPartlan H, Macleod I, Sethuraman L, Goddard ME: **Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle.** *Genetical Research* 2007, **89**(4):215-220.
41. Meuwissen T: **Use of whole genome sequence data for QTL mapping and genomic selection.** *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production* Leipzig, Germany: Gesellschaft für Tierzuchtwissenschaften e.V; 2010, [Abstract ID 0018, ISBN 978-3-00-031608-1].
42. Gianola D, de los Campos G: **Inferring genetic values for quantitative traits non-parametrically.** *Genetics Research* 2008, **90**(6):525-540.
43. Gianola D, Fernando RL, Stella A: **Genomic-assisted prediction of genetic value with semiparametric procedures.** *Genetics* 2006, **173**(3):1761-1776.
44. González-Reco O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendaño S: **Nonparametric Methods for Incorporating Genomic Information Into Genetic Evaluations: An Application to Mortality in Broilers.** *Genetics* 2008, **178**:2305-2313.
45. Gianola D, de los Campos G, González-Reco O, Long N, Okut H, Rosa GJM, Weigel KA, Wu XL: **Statistical Learning Methods For Genome-based Analysis Of Quantitative Traits.** *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production* Leipzig, Germany: Gesellschaft für Tierzuchtwissenschaften e.V; 2010, [Abstract ID 0014, ISBN 978-3-00-031608-1].
46. Goddard ME, Hayes BJ: **Genomic selection.** *Journal of Animal Breeding and Genetics* 2007, **124**:323-330.

doi:10.1186/1471-2156-12-74

**Cite this article as:** Wittenburg et al.: Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. *BMC Genetics* 2011 **12**:74.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit





# An approximate Bayesian significance test for genomic evaluations

Dörte Wittenburg<sup>1</sup>  | Volkmar Liebscher<sup>2</sup>

<sup>1</sup>Institute of Genetics and Biometry, Leibniz Institute for Farm Animal Biology, Wilhelm-Stahl-Allee 2, D-18196 Dummerstorf, Germany

<sup>2</sup>Department of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Str. 47, D-17489 Greifswald, Germany

## Correspondence

Dörte Wittenburg, Institute of Genetics and Biometry, Leibniz Institute for Farm Animal Biology, Wilhelm-Stahl-Allee 2, D-18196 Dummerstorf, Germany.  
Email: wittenburg@fhn-dummerstorf.de

## Abstract

Genomic information can be used to study the genetic architecture of some trait. Not only the size of the genetic effect captured by molecular markers and their position on the genome but also the mode of inheritance, which might be additive or dominant, and the presence of interactions are interesting parameters. When searching for interacting loci, estimating the effect size and determining the significant marker pairs increases the computational burden in terms of speed and memory allocation dramatically. This study revisits a rapid Bayesian approach (fastbayes). As a novel contribution, a measure of evidence is derived to select markers with effect significantly different from zero. It is based on the credibility of the highest posterior density interval next to zero in a marginalized manner. This methodology is applied to simulated data resembling a dairy cattle population in order to verify the sensitivity of testing for a given range of type-I error levels. A real data application complements this study. Sensitivity and specificity of fastbayes were similar to a variational Bayesian method, and a further reduction of computing time could be achieved. More than 50% of the simulated causative variants were identified. The most complex model containing different kinds of genetic effects and their pairwise interactions yielded the best outcome over a range of type-I error levels. The validation study showed that fastbayes is a dual-purpose tool for genomic inferences – it is applicable to predict future outcome of not-yet phenotyped individuals with high precision as well as to estimate and test single-marker effects. Furthermore, it allows the estimation of billions of interaction effects.

## KEYWORDS

conditional expectation, dominance, epistasis, genetic architecture, SNP

## 1 | INTRODUCTION

In animal breeding, molecular markers (e.g., single nucleotide polymorphisms; SNPs) are incorporated into statistical models to reach an improved genomic evaluation of animals. This leads to more precisely estimated breeding values of not-yet phenotyped animals and, if selection of animals is based on genomic breeding values instead of traditionally estimated breeding values, breeding costs can be drastically reduced due to the shortened generation intervals (e.g. Schaeffer, 2006). Such genomic data are also extremely valuable for elucidating the genetic architecture of some trait. Not only the size of the genetic effect and the

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2018 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

position on the genome, but also the mode of inheritance, which might be additive or dominant, and the presence of interactions are relevant for understanding the impact of a causative variant. Though there is little evidence that interactions (epistasis) contribute much to genetic variation in most populations (Hill, Goddard, & Visscher, 2008; Phillips, 2008), the revelation of epistasis may contribute to fill the gap of “missing heritability” (Zuk, Hechter, Sunyaev, & Lander, 2012).

In a typical situation of genomic evaluations, there are many more predictor variables ( $p$  SNPs) than observations ( $n$  animals); this makes testing of SNP effects a challenge. SNPs with significant impact on a trait can be identified with genome-wide association studies (GWAS) typically based on successive single-SNP investigations – one SNP is tested at a time while “walking” along the genome. Those SNPs being in high linkage disequilibrium (LD) with a causative variant will likely have a tiny  $p$ -value implying a strong association with the trait. The high LD among SNPs complicates the correct identification of a causative variant, and this kind of testing also suffers from multiplicity. Several specific methods exist for controlling the rate of false-positive detection. Ideally, the number of independent tests accounts for the LD between SNPs as was proposed by Gao, Starmer, and Martin (2008) in their simple M-method. Alternatively, a score-based test statistic accounting for the correlation between SNPs yields a ranked order of SNPs (Zuber, Duarte Silva, & Strimmer, 2012). Then, the top-ranked SNPs can be taken for further genetic investigation.

Beside different strategies for suitable  $p$ -value correction, using a statistical model that considers all SNPs jointly is a promising option. To study the genetic effect captured by SNPs, each with two alleles A and B, a whole-genome regression model is set up (e.g. de los Campos, Hickey, Pong-Wong, Daetwyler, & Calus, 2013). A trait is regressed onto the number of reference SNP alleles at all loci simultaneously. Such a model can be used for the estimation of SNP effects and/or for the selection of relevant SNPs throughout the genome. Excluding any nuisance effects the whole-genome regression model can be written as

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

with  $\mathbf{y} = (y_1, \dots, y_n)'$  being the vector of observable traits of  $n$  individuals centred by the population mean and  $\mathbf{g} = (g_1, \dots, g_p)'$  the effect at  $p$  SNPs. The  $(n \times p)$  design matrix  $X = \{X_{i,j}\}$  contains the genotype codes at locus  $j \in \{1, \dots, p\}$  for individual  $i \in \{1, \dots, n\}$ , where 1 and  $-1$  indicate homozygous genotypes AA and BB, respectively, and the heterozygote is coded as 0. Let A denote the minor allele with minor allele frequency (MAF)  $f_j \leq 1/2$ . The MAF is assumed to be known or may be estimated from SNP genotypes using, for example, the method of moments as  $f_j = \frac{1}{2n} \sum_{i=1}^n (X_{i,j} + 1)$ . The residual errors in  $\mathbf{e} = (e_1, \dots, e_n)'$  are assumed to be independent and normally distributed with zero mean and variance  $\sigma_e^2$ .

Selection-and-shrinkage approaches allow the simultaneous selection of those predictors that sufficiently map the trait and the estimation of their effect sizes. They are often implemented in a Bayesian (e.g. George & McCulloch, 1993; Habier, Fernando, Kizilkaya, & Garrick, 2011; Meuwissen, Hayes, & Goddard, 2001) or Expectation-Maximization framework (Chen & Tempelman, 2015). Such a multiple-locus model gains higher power than single-SNP GWAS; it is therefore the preferred choice (Hoggart, Whittaker, De Iorio, & Balding, 2008). When based on a Markov-chain-Monte-Carlo (MCMC) sampling scheme, Bayesian approaches, however, may become too time-consuming and memory-demanding if the model dimension is high (Meuwissen, Hayes, & Goddard, 2001; Pérez, de los Campos, Crossa, & Gianola, 2010). As an alternative to MCMC sampling, effect sizes can be approximated using a variational Bayes (vbay) approach, in which the posterior probabilities are approximated through factorisation. A vbay approach is competitive to MCMC-based methods in terms of estimating the effect size of SNPs and detecting the significant loci but it requires only a fraction of computing time (Li & Sillanpää, 2012). Applied to real and simulated data, the vbay SNP-selection approach of Logsdon, Hoffman, and Mezey (2010) could identify even SNPs with weak association to a trait. A rapid approximation of SNP-effect estimates can also be derived by a Bayesian approach in which the conditional expectation of SNP effects is calculated iteratively (Meuwissen, Solberg, Shepherd, & Woolliams, 2009). Initially developed under pure additivity, this “fastbayes” has been extended to include dominance and epistatic effects (Wittenburg, Melzer, & Reinsch, 2011). That study showed a similar precision of genetic value prediction of fastbayes and an MCMC-based version.

Model complexity and dimension are still an issue, also in times of high-performance computing. Particularly, as whole-genome sequence data are available, a causative variant shall be pinpointed to a specific base pair among millions of SNPs. Hence a fast algorithm is required to analyze the effects of all SNPs in a feasible time. The objective of the present study is to follow up the fastbayes approach and to incorporate a suitable measure of significance for testing the SNP effects. Both aspects of genomic inferences are evaluated – the ability to detect relevant loci and to predict genetic values – when different kinds of genetic effects are present (additive, dominance, epistasis). In Section 2, the components of the fastbayes approach are presented and a marginal test for genetic effects is developed. The extended approach is applied to simulated and real data, which are described in Section 3. Section 4 explains the implementation and involved software. The computing details are addressed in

Section 5. Then, in Section 6, results of genomic evaluations for different parameter settings are presented. The performance of fastbays is reviewed in Section 7, and further extensions are outlined.

## 2 | METHODS

### 2.1 | Summary of fastbays

The following investigations are based on model (1) but the entries in  $X$  contain the standardised genotype codes. They are centred and scaled such that  $X_{i,j}$  is  $2(1 - f_j)/s_j$  and  $-2f_j/s_j$  for homozygous AA and BB, respectively, and  $(1 - 2f_j)/s_j$  for heterozygous individuals with scaling term  $s_j = \sqrt{2f_j(1 - f_j)}$ .

The approximate Bayesian method “fastbays” is now further considered to include a testing procedure. According to Meuwissen et al. (2009), the idea of this iterative approach is based on a one-locus model,

$$\mathbf{y} = \mathbf{x}g + \mathbf{e}, \quad (2)$$

where  $\mathbf{x} = \mathbf{X}_j$ , the  $j$ -th column of  $X$ , and  $g = g_j$  at a single locus  $j \in \{1, \dots, p\}$ .

The likelihood function  $p(\mathbf{y}|g)$  is specified as a normal distribution with mean  $\mathbf{x}g$  and covariance matrix  $\mathbf{I}\sigma_e^2$ . The prior distribution of genetic effects is assumed to be a mixture of a Laplace distribution and a point mass at zero ( $\delta_0$ ) with  $P(g = 0) = 1 - \gamma$ . The formal prior density is

$$p(g|\gamma) = \frac{1}{2}\gamma\lambda \exp(-\lambda|g|) + (1 - \gamma)\delta_0.$$

The choice of the hyperparameters  $\gamma$  and  $\lambda$  will be discussed later.

A point estimate of the genetic effect is determined as the posterior expectation  $\hat{g} = E(g|\mathbf{y})$ , which is given in closed form (Meuwissen et al., 2009), that is

$$E(g|\mathbf{y}) = \frac{T_1\Theta_U(0; Y^-, \sigma^2) + T_2\Theta_L(0; Y^+, \sigma^2)}{T_1 + T_2 + T_3}$$

$$\text{with } T_1 = \exp(-\lambda Y) [1 - \Phi(0; Y^-, \sigma^2)],$$

$$T_2 = \exp(\lambda Y)\Phi(0; Y^+, \sigma^2),$$

$$T_3 = \frac{2(1 - \gamma)}{\gamma\lambda} \exp\left(-\frac{1}{2}\lambda^2\sigma^2\right)\phi(0; Y, \sigma^2),$$

where  $\sigma^2 = (\mathbf{x}'\mathbf{x})^{-1}\sigma_e^2$ ,  $Y = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$  and  $Y^\pm = Y \pm \lambda\sigma^2$ . The  $\Theta_U(0; \mu, \sigma^2)$  and  $\Theta_L(0; \mu, \sigma^2)$  are the expected value of an upper and lower truncated normal distribution  $N(\mu, \sigma^2)$ , respectively, with truncation point zero. The  $\Phi(x; \mu, \sigma^2)$  denotes the normal distribution function evaluated at some point  $x$ , and  $\phi(x; \mu, \sigma^2)$  is the normal density function.

In order to analyze  $p$  loci simultaneously, the model (2) is fitted iteratively to a marginalized component, similar to Meuwissen et al. (2009) and Wittenburg et al. (2011). The trait  $\mathbf{y}$  is corrected for all SNP effects except the one under investigation,

$$\mathbf{y}_j = \mathbf{y} - \sum_{i=1, i \neq j}^p \mathbf{X}_i g_i = \mathbf{X}_j g_j + \mathbf{e} \quad \text{for } j = 1, \dots, p.$$

This is done using a Gauss-Seidel-like algorithm. Set  $g_j^{(0)} = 0 \forall j$ . In iteration  $k = 1, 2, \dots$ ,

$$\begin{aligned} \mathbf{y}_j^{(k)} &= \mathbf{y} - \sum_{i=1}^{j-1} \mathbf{X}_i \hat{g}_i^{(k)} - \sum_{i=j+1}^p \mathbf{X}_i \hat{g}_i^{(k-1)} \quad \text{and} \\ \hat{g}_j^{(k)} &= E(g_j | \mathbf{y}_j^{(k)}) \end{aligned} \quad (3)$$

are calculated for all  $j = 1, \dots, p$  until relative changes are small,

$$\frac{\|\hat{g}_j^{(k)} - \hat{g}_j^{(k-1)}\|}{\|\hat{g}_j^{(k)}\|} < 10^{-4}.$$

Model (1) can be further extended to include dominance and two-locus epistatic effects; details are presented in Wittenburg et al. (2011). Then the design matrix  $X$  consists of three groups of columns, one for each kind of effect, that is  $X = (X_a, X_d, X_e)$  comprising the standardized genotype codes for additive, dominance and epistatic effects. For dominance effects, the columns in  $X_d$  are coded according to Falconer's model in a random mating population in Hardy–Weinberg equilibrium (Falconer & Mackay, 1996, p.118) and scaled afterwards. Then, at a specific locus  $j$ ,  $-2(1 - f_j)^2/s_j^2$  is used for homozygous AA,  $-2f_j^2/s_j^2$  for BB individuals and  $2f_j(1 - f_j)/s_j^2$  for heterozygous individuals. Four different types of epistatic effects can be distinguished: additive×additive, additive×dominance, dominance×additive, and dominance×dominance, leading to  $2p(1 - p)$  additional effects. A column in  $X_e$  coding for an epistatic effect contains the product of codes for the corresponding main effects. For instance, an additive×dominance effect at locus pair  $i$  and  $j$  is coded as  $-2(1 - f_i)/s_i \cdot 2f_j^2/s_j^2$  if an individual is homozygous AA at locus  $i$  and homozygous BB at locus  $j$ . Again the Gauss-Seidel-like algorithm (3) is applied to  $\mathbf{y}$  that is corrected for all other effects except the current one in order to estimate the different kinds of genetic effects.

In a real data analysis, also nongenetic factors affecting the trait have to be considered. Thus, model (1) is extended to account for potential fixed effects  $\mathbf{b} = (b_1, \dots, b_q)'$  with  $(n \times q)$  design matrix  $W$ ,

$$\mathbf{y} = W\mathbf{b} + X\mathbf{g} + \mathbf{e}.$$

Like in Meuwissen et al. (2009) and Wittenburg et al. (2011), a type of empirical Bayes method is employed to estimate the fixed effects and  $\sigma_e^2$  in iteration  $k$  of the Gauss-Seidel-like algorithm (3) as

$$\begin{aligned} \hat{\mathbf{b}}^{(k)} &= (W'W)^{-1} W'\mathbf{y}^* \quad \text{with} \quad \mathbf{y}^* = \mathbf{y} - X\hat{\mathbf{g}}^{(k)}, \\ \sigma_e^{2(k)} &= \frac{1}{n - q} \mathbf{e}^{(k)'} \mathbf{e}^{(k)} \quad \text{with} \quad \mathbf{e}^{(k)} = \mathbf{y} - W\hat{\mathbf{b}}^{(k)} - X\hat{\mathbf{g}}^{(k)}. \end{aligned}$$

Then the estimation of genetic effects continues while  $\mathbf{y}$  is corrected for the current estimates of fixed effects.

## 2.2 | Marginal test for genetic effects

As a novel contribution, the significance of SNP-effect estimates is investigated using a measure of evidence similar to De Braganca Pereira & Stern (1999). Starting with model (2), a fully Bayesian significance test is proposed to test the null hypothesis  $H_0: g = 0$ . For this purpose, the credibility  $\kappa$  of the highest posterior density (HPD) interval that is right to zero if  $\hat{g} > 0$  or left to zero if  $\hat{g} < 0$  is determined, that is

$$\kappa = P(t_1 < g < t_2 | \mathbf{y}) = \int_{(t_1, t_2)} p(g | \mathbf{y}) dg.$$

In case of  $\hat{g} > 0$ , the interval borders are  $t_1 = 0$  and

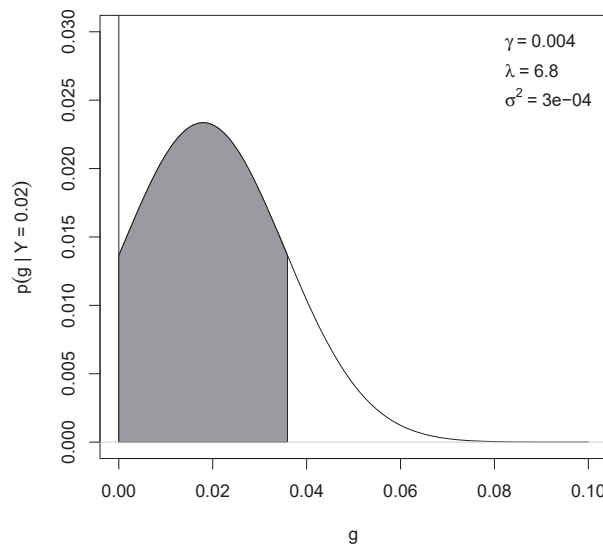
$$t_2 = \max_{g > 0} \{g : p(g | \mathbf{y}) \geq p(0 | \mathbf{y})\},$$

else if  $\hat{g} < 0$ ,  $t_2 = 0$  and

$$t_1 = \min_{g < 0} \{g : p(g | \mathbf{y}) \geq p(0 | \mathbf{y})\}.$$

Then significance of  $g$  is inferred if  $1 - \kappa \leq \alpha$ , with an appropriately chosen type-I error  $\alpha$ . Let  $\mathbb{1}$  denote the indicator function. The posterior probability is analytically derived as





**FIGURE 1** Example of a highest posterior density interval tangent to zero with arbitrary parameters mentioned within the graph

$$\begin{aligned}
 P(g < t | \mathbf{y}) &= \frac{1}{p(\mathbf{y})} \int_{(-\infty, t)} p(\mathbf{y} | g) p(g | \gamma) dg = \\
 &= \frac{1}{p(\mathbf{y})} \int_{(-\infty, t)} [f_1(g) \mathbb{1}_{\{g < 0\}} + f_2(g) \delta_0(g) + f_3(g) \mathbb{1}_{\{g > 0\}}] dg,
 \end{aligned}$$

using the functions

$$\begin{aligned}
 f_1(g) &= \frac{1}{2} \gamma \lambda \exp\left(\frac{1}{2} \lambda^2 \sigma^2 + \lambda Y\right) \phi(g; Y^+, \sigma^2), \\
 f_2(g) &= (1 - \gamma) \phi(g; Y, \sigma^2), \\
 f_3(g) &= \frac{1}{2} \gamma \lambda \exp\left(\frac{1}{2} \lambda^2 \sigma^2 - \lambda Y\right) \phi(g; Y^-, \sigma^2).
 \end{aligned}$$

Furthermore, the denominator is

$$p(\mathbf{y}) = \frac{1}{2} \gamma \lambda \exp\left(\frac{1}{2} \lambda^2 \sigma^2\right) (T_1 + T_2) + (1 - \gamma) \phi(0; Y, \sigma^2).$$

As the posterior density  $p(g | \mathbf{y})$  is bimodal due to the spike at zero, the posterior distribution function has a jump discontinuity at zero with step height  $\frac{1}{p(\mathbf{y})} f_2(0)$ ; the smaller  $Y = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{y}$ , the larger the step height. For larger values of  $Y$  the density is approximately unimodal and symmetric (Meuwissen et al., 2009). This discontinuity, however, does not affect the calculation of  $\kappa$  because the region tangent to zero is of interest here. An example presenting the HPD interval is given in Figure 1.

In the  $p$ -locus case and for the different kinds of effects, a marginal testing problem is considered for each SNP effect  $g_j$ . The corresponding credibility  $\kappa_j$  is determined conditionally on  $\mathbf{y}_j$  when convergence of the iterative approach (3) has been reached.

For comparison, the Bayes factor is calculated as the ratio of marginal likelihoods of the full model over the reduced model, that is  $g_j = 0$ , as

$$\begin{aligned}
 B_j &= \frac{\exp\left[-\frac{1}{2\sigma_e^2} (\mathbf{y}_j - \mathbf{X}_j \hat{g}_j)' (\mathbf{y}_j - \mathbf{X}_j \hat{g}_j)\right]}{\exp\left(-\frac{1}{2\sigma_e^2} \mathbf{y}_j' \mathbf{y}_j\right)} = \\
 &= \exp\left[-\frac{1}{2\sigma_e^2} \left(-2\mathbf{y}_j' \mathbf{X}_j \hat{g}_j + \hat{g}_j' \mathbf{X}_j' \mathbf{X}_j \hat{g}_j\right)\right].
 \end{aligned}$$

Following Jeffrey (1961), a substantial effect is reported if  $B_j > 3$ .

### 3 | DATA STRUCTURE AND EVALUATION CRITERIA

#### 3.1 | Simulation study

Genomic evaluations were implemented based on simulated data with realistic genetic structure, as described and used in Wittenburg et al. (2011). Briefly, following the density and distribution of SNPs on the Illumina BovineSNP50 chip, 52,773 SNPs were simulated on the cattle genome of 30 Morgan length. Then, 400 generations of random mating among 100 individuals were executed in which random recombination events according to the genetic distance between SNPs and random mutation of SNP alleles were considered to reach a realistic amount of linkage disequilibrium between SNPs. Four more generations were produced; in every generation 50 sires were mated to 20 dams in order to generate multiple half-sib families. The data were split into training (generations 401 and 402,  $n = 2,000$ ) and validation set (generations 403 and 404,  $n = 2,000$ ). Then,  $p = 5,227$  SNPs (every 10th SNP including the causative variants) were used for the evaluation. SNPs with  $MAF \leq 0.05$  were removed, and thus 793 SNPs were additionally excluded on average. Twenty-three SNPs were randomly preselected to be the causative variants, and additive and dominance effect were simulated for each of them. For each of the four kinds of epistatic effects, six SNP pairs were randomly drawn out of the 23 causative variants to interact. Dominance and epistasis explained about 10% and 29%, respectively, of the total genetic variance. Two different traits were achieved by adding different residual error terms to the sum of genetic effects, such that total genetic variation contributed either 30% (i.e. broad-sense heritability  $H^2 = 0.3$ ) or 50% ( $H^2 = 0.5$ ) to the variation of  $\mathbf{y}$ . The simulation was repeated 100 times. The fastbayes algorithm was executed using  $\gamma = 0.005$ ,  $\lambda = \sqrt{2p\gamma}$  for additive and dominance effects and  $\gamma = 10^{-6}$ ,  $\lambda = \sqrt{p(p-1)\gamma}$  for epistatic effects, as suggested by Wittenburg et al. (2011).

The main criterion for evaluation was the number of truly detected SNPs with significant impact on a trait. This was measured in terms of sensitivity and specificity for each kind of effect separately as well as overall by considering the joint set of SNPs that were significant in any kind of effect. A true positive result was obtained when the significant SNP was in a window of 100 kbp around the simulated causative variant. For instance, one gene per 100 kbp is expected in the mouse genome (Laurie et al., 2007). The impact of different values for the type-I error on sensitivity and specificity was investigated using  $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$ . Furthermore, the proportion of estimated genetic variance explained by significant SNPs to the simulated genetic variance ( $\sigma_{\text{sign}}^2$ ) was calculated. These characteristics determine the ability of an approach to select the loci relevant to genetic variation in a training set. Moreover, the ability to predict genetic values of future or not-yet phenotyped animals was verified. Accuracy of genetic value prediction was assessed as the correlation between estimated and simulated genetic values in a test set.

A permutation test approach was used to further study the sensitivity and specificity of fastbayes in the absence of any effects of causative variants. The association between genotypes and phenotypes was removed in the first simulated dataset by shuffling the SNP genotypes: the rows of the matrix  $X$  were randomly assigned to the animals. This was repeated 100 times, and the resampled datasets were analyzed as described above.

#### 3.2 | Real data

To show the performance of fastbayes on real data, data of a heterogeneous stock of mice (Valdar et al., 2006a) retrieved from <http://gscan.well.ox.ac.uk> on July 12, 2011 were analyzed. The dataset comprised genotypes at  $p = 8797$  SNPs ( $MAF > 0.05$ ) and phenotypes of  $n = 1521$  animals. Rarely missing genotypes for these SNPs were imputed via Beagle 3.2 (Browning & Browning, 2007). The percentage of CD8<sup>+</sup> cells, an immunological phenotype, was analyzed, and the vector of observations was standardised to avoid numerical problems. A set of covariates similar to Valdar et al. (2006b) was considered: gender, age, family, litter, cage density, experimenter, month, and year of experiment. A resampling scheme was used to specify a suitable hyperparameter  $\gamma$ . The data were equally split into training and test set, and animals were assigned at random to the sets. For a range of  $\gamma \in [0.001, 0.1]$ , genetic values (EGV) were estimated in the test set based on the effect estimates from the training set using a model with additive and dominance effects. The correlation between EGV and  $\mathbf{y}$  served as a measure of accuracy. The  $\gamma$  leading to the highest accuracy was chosen. Then, to specify the hyperparameter for epistatic effects, a range of  $\gamma \in \{10^{-7}, \dots, 10^{-3}\}$  was evaluated similarly. For instance,  $10^{-7}$  corresponds to an expectation of four interactions per kind of effect. Using a smaller  $\gamma$ -value would reflect the assumption of less than one interaction effect that is not envisaged. The final parameter values were employed in the genomic evaluation of the complete dataset.

**TABLE 1** Average computing time of a single analysis based on the fastbayes and vbay approach

Model	fastbayes	vbay
Simulated data ( $\varnothing$ 4,434 SNPs)		
M1	4 s	14 s
M2	7 s	29 s
M3*	2 min	10 min
M3	5.4 hr	–
Real data (8,797 SNPs)		
M2	10 s	26 min
M3	16.9 hr	–

M1: model with additive effects; M2: model with additive and dominance effects; M3: model with additive, dominance and all epistatic effects; M3\*: model with additive and additive $\times$ additive interaction effects using only every 10th SNP.

## 4 | SOFTWARE IMPLEMENTATION

The data preparation, analysis and summary of results were executed in R version 3.5.0 (R Core Team, 2015). The fastbayes approach was implemented in Fortran90 embedding CDFLIB routines (<http://biostatistics.mdanderson.org/SoftwareDownload/>); fastbayes is available online at github. The search for the HPD interval was implemented as a grid search over the interval  $[m_L, m_U]$  with  $m_L = 10 \min(\hat{g})$  and  $m_U = 10 \max(\hat{g})$ ; the step size was  $(m_U - m_L)/2000$ . The dummy variables in  $X_e$  were set up on the fly to reduce memory allocation. The fastbayes was compared to Logsdon's vbay method (Logsdon et al., 2010) being the strongest competitor in terms of computing time and accuracy of estimation; it is available as R package vbsr version 0.0.5 (Logsdon, Carty, Reiner, Dai, & Kooperberg, 2012). The vbay's approximate posterior probability of a parameter being non-zero was used to assess significance. If this probability was  $> 0.95$ , a significant effect was reported. The calculations were run on 2.1 GHz (SLES 12 64 bit) and 2.2 GHz (SLES 12 SP 3 64 bit) multiuser systems.

## 5 | COMPUTING DETAILS

Computing time was clearly in favor of fastbayes; it generally needed the least time, see Table 1. The differences in time between fastbayes and vbay were small when additive and dominance effects were studied in simulated data. For estimating approximately 39 million epistatic effects and computing their significance measures, fastbayes required on average 5.4 hr when  $H^2 = 0.5$ . More iterations were needed until convergence when  $H^2 = 0.3$ ; then computations took 5.9 hr on average. Due to memory restrictions in the Fortran implementation used in the R package for vbay (long vectors were not supported), the full model containing  $2p(p-1)$  pairwise interactions could not be analyzed. To obtain a tendency whether vbay is able to reveal epistatic effects, the model dimension was further decreased. Again, every 10th out of the 5,227 SNPs was selected but without paying attention to keeping the causative variants among the selected SNPs. Considering  $MAF > 0.05$ , 444 SNPs were retained on average. Additive and additive $\times$ additive interaction effects were analyzed, denoted as model M3\* in Table 1. In this downsized scenario, fastbayes needed 2 min computing time but 10 min were required by vbay for estimating 98,790 effects.

The difference in time increased in the real data analysis (Table 1). Additive and dominance effects were computed in 10 s using fastbayes and 26 min using vbay. Furthermore, 16.9 hr were required for estimating about 154 million interaction effects. Due to memory restrictions, vbay could not be used for studying epistatic effects.

Because a multiuser computing system was employed, the memory allocation could only be approximated roughly for simulated data. Peaks in memory allocation have been observed when epistatic effects were estimated using fastbayes ( $\leq 4$  GB on average) and when main genetic effects were computed using vbay ( $\leq 1$  GB on average).

## 6 | RESULTS OF GENOMIC EVALUATIONS

### 6.1 | Simulated data

Considering the measure of evidence, the fastbayes approach identified about one-third of the causative variants with additive contribution to the total genetic variation if  $H^2 = 0.5$ , see Table 2. If  $\alpha = 0.01$ , sensitivity for additive effects increased slightly

**TABLE 2** Average sensitivity for each kind of effect and overall specificity based on the fastbayes (measure of evidence MOE; Bayes factor BF) and vbay approach,  $H^2 = 0.5$ 

	Model	Sensitivity			Overall	Specificity	
		$a$	$d$	$e$		overall	$\sigma_{\text{sign}}^2$
fastbayes (MOE $\leq 0.01$ )	M1	0.315			0.315	1.000	0.536
	M2	0.334	0.071		0.369	1.000	0.601
	M3	0.373	0.107	0.246	0.504	0.999	0.833
fastbayes (MOE $\leq 0.05$ )	M1	0.337	–	–	0.337	1.000	0.545
	M2	0.352	0.090	–	0.393	1.000	0.615
	M3	0.390	0.116	0.256	0.522	0.998	0.872
fastbayes (MOE $\leq 0.10$ )	M1	0.347	–	–	0.347	1.000	0.550
	M2	0.361	0.098	–	0.406	1.000	0.622
	M3	0.394	0.121	0.256	0.525	0.998	0.884
fastbayes (MOE $\leq 0.20$ )	M1	0.360	–	–	0.360	1.000	0.555
	M2	0.369	0.107		0.417	1.000	0.628
	M3	0.397	0.123	0.256	0.527	0.998	0.891
fastbayes (BF > 3)	M1	0.398	–	–	0.398	0.999	0.560
	M2	0.407	0.149	–	0.467	0.998	0.639
	M3	0.418	0.153	0.273	0.563	0.994	0.927
vbay	M1	0.356	–	–	0.356	1.000	0.587
	M2	0.357	0.094	–	0.400	1.000	0.654

M1: model with additive ( $a$ ) effects; M2: model with additive and dominance ( $d$ ) effects; M3: model with additive, dominance and epistatic ( $e$ ) effects; contribution of the variance at the significant SNPs to the total genetic variance ( $\sigma_{\text{sign}}^2$ ). In total, 23 causative variants were simulated.

by about 6% when the genome-wide regression model was extended to include dominance effects but these were rarely identified; the sensitivity for dominance effects was 7%. The best outcome was observed if pairwise epistatic effects were also included in the model. Then, considering the joint set of significant SNPs over all kinds of effects, 50% of the causative variants were identified correctly. The higher the type-I error was, the higher the sensitivity turned out; the maximum was observed if  $\alpha = 0.20$ . The specificity, however, remained very high ( $\geq 99.76\%$ ) in general. Larger  $\alpha$ -values led to a further but very small increase of sensitivity (results not shown), and specificity was still very high (e.g.  $\geq 99.74\%$  if  $\alpha = 0.50$ ). This can be explained by the clear differentiation of zero and nonzero effects by the measure of evidence, see Additional File 1: effects close to zero coincidentally have a measure close to one. The genetic variance that could be explained by the significant effects was 89% at most. The accuracy of genetic value prediction achieved its maximum over all models at 82%.

The overall sensitivity of fastbayes further improved by about 7–12% when the Bayes factor instead of the measure of evidence ( $\alpha = 0.20$ ) was used for identifying the relevant SNPs. Both measures found the same significant effects but few additional effects, which were most often nonadditive, could be identified with the Bayes factor. The best result was achieved with the genome-wide regression model including all kinds of genetic effects. Then, 42% of the causative variants with additive effects were detected on average. The sensitivity for dominance effects increased to 15% but sensitivity for epistatic effects remained rather constant at 27%; 56% of the causative variants were found overall, see Table 2. The overall specificity was  $\geq 99\%$  being slightly less than with the measure of evidence. The genetic variance explained by the SNPs with substantial effect was 93%.

With  $H^2 = 0.3$ , the performance of fastbayes was worse than with  $H^2 = 0.5$ , see Table 3. About 30–33% less true positive SNPs could be identified overall using the measure of evidence and about 24–25% using the Bayes factor. The genetic variance explained by the significant SNPs decreased by 4%. The complex model including all kinds of genetic effects yielded again the best outcome regarding the ability to detect the relevant loci that capture most of the genetic variation but the accuracy of genetic value prediction was slightly reduced by 2% compared to a model containing additive and dominance effects; the correlation between simulated and estimated breeding values was 72%.

Furthermore, the data that were generated without any impact of the causative variants were analyzed. It turned out that sensitivity was 0.17% at most and specificity was  $\geq 99.88\%$  over all kinds of effects and all models using the measure of evidence and, for instance,  $\alpha = 0.05$ . Using the Bayes factor, the overall sensitivity was increased but it was at most 0.87% in the model considering all kinds of effects and specificity was  $\geq 99.42\%$ . The correlation between simulated and estimated breeding values was 0.13%.

**TABLE 3** Average sensitivity for each kind of effect and overall specificity based on the fastbayes (measure of evidence MOE; Bayes factor BF) and vbay approach,  $H^2 = 0.3$ 

	Model	Sensitivity			Overall	Specificity	
		<i>a</i>	<i>d</i>	<i>e</i>		overall	$\sigma_{\text{sign}}^2$
fastbayes (MOE $\leq 0.05$ )	M1	0.227	–	–	0.227	1.000	0.483
	M2	0.240	0.041	–	0.264	1.000	0.539
	M3	0.269	0.058	0.128	0.366	0.998	0.840
fastbayes (BF > 3)	M1	0.303	–	–	0.303	0.999	0.508
	M2	0.305	0.088	–	0.352	0.998	0.578
	M3	0.305	0.090	0.145	0.423	0.994	0.892
vbay	M1	0.253	–	–	0.253	1.000	0.564
	M2	0.253	0.047	–	0.280	1.000	0.611

M1: model with additive (*a*) effects; M2: model with additive and dominance (*d*) effects; M3: model with additive, dominance and epistatic (*e*) effects; contribution of the variance at the significant SNPs to the total genetic variance ( $\sigma_{\text{sign}}^2$ ). In total, 23 causative variants were simulated.

**TABLE 4** Average sensitivity for each kind of effect and overall specificity based on the fastbayes (measure of evidence MOE; Bayes factor BF) and vbay approach in the absence of genetic effects

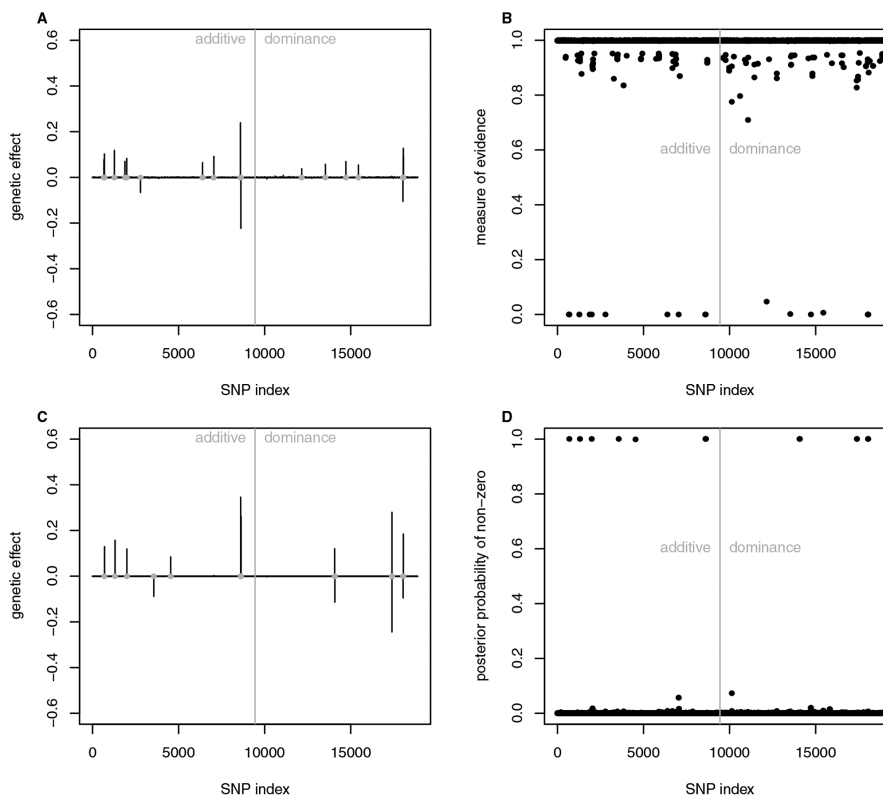
	Model	Sensitivity			Overall	Specificity	
		<i>a</i>	<i>d</i>	<i>e</i>		overall	
fastbayes (MOE $\leq 0.05$ )	M1	0	–	–	0	1.000	1.000
	M2	0	0	–	0	1.000	1.000
	M3	0	0	0.002	0.002	0.999	0.999
fastbayes (BF > 3)	M1	0.001	–	–	0.001	1.000	1.000
	M2	0.001	0.001	–	0.002	0.998	0.998
	M3	0.001	0.002	0.006	0.009	0.994	0.994
vbay	M1	0	–	–	0	1.000	1.000
	M2	0	0	–	0	1.000	1.000

M1: model with additive (*a*) effects; M2: model with additive and dominance (*d*) effects; M3: model with additive, dominance and epistatic (*e*) effects. In total, 23 causative variants were simulated.

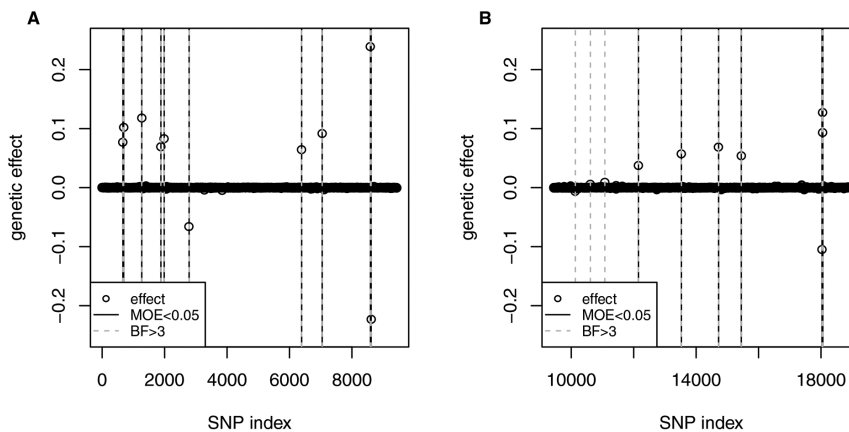
The ability to select the relevant loci and to predict the genetic values was similar using fastbayes (and the measure of evidence,  $\alpha = 0.05$ ) and vbay based on the model with additive and dominance effects, see Tables 2–4. Vbay, however, could explain up to 6–13% more of the total genetic variance considering the significant effects because intermediate effects were estimated negligibly better than using fastbayes (see also Additional File 1). Both methods could not detect the small effects correctly.

## 6.2 | Real data

The SNPs with significant effect were located near protein coding genes or in regions of quantitative trait loci (QTL) with known association to traits of the immune system (<http://www.informatics.jax.org>). Though results obtained by fastbayes ( $\alpha = 0.05$ ) and vbay have been rather similar based on simulated data, the real data analysis revealed differences regarding the dominance effects, see Figure 2 and Additional File 2. With vbay, two significant dominance SNPs were detected on chromosome 17 near the MHC region which has great impact on immunological phenotypes (Valdar et al., 2006a). The largest dominance effects (contributing 26% to the total genetic variance) were found on chromosome 15; this region was not identified by fastbayes. Fastbayes estimated a group of three SNPs near the MHC region with similar, moderate dominance effects and smaller effects on the other chromosomes. The SNPs near the MHC region also had the largest additive impact (50%) on the total genetic variance. Other SNPs with moderate additive effects were found on chromosome one and two and were in the neighborhood of known QTL (*Sle1c* and *Prdt1*, respectively). The significance measures of both methods clearly differentiate between significant and nonsignificant effects (see Figure 2B and D). In total, vbay identified eight loci with additive and six loci with dominance effect. These loci explained all of the estimated genetic variance. Fastbayes detected 10 loci with additive and seven loci with dominance effect using an optimal  $\gamma = 0.046$ , and 99% of the estimated genetic variance was explained by the significant effects.



**FIGURE 2** Results of analyzing the mouse data set ( $p = 8,797$  SNPs,  $n = 1,521$  individuals): estimated additive and dominance effects of SNPs using the fastbayes (A) and vbay (C) approach; gray dots indicate significant loci. Measure of evidence related to fastbayes (B) and posterior probability of nonzero effects related to vbay (D) reflect the significance of effects. SNP index equals SNP number for additive effects and SNP number plus  $p$  for dominance effects



**FIGURE 3** Results of analyzing the mouse dataset ( $p = 8,797$  SNPs,  $n = 1,521$  individuals): significance of additive and dominance effects of SNPs was inferred using fastbayes and (A) measure of evidence (MOE)  $\leq 0.05$  or (B) Bayes factor (BF)  $> 3$ . SNP index equals SNP number for additive effects and SNP number plus  $p$  for dominance effects

This outcome did not change using the complex model including epistasis – no significant epistatic effects were found. The optimal hyperparameter for epistatic effects was  $\gamma = 10^{-7}$ .

Using the Bayes factor for inferring significance again showed that the same significant additive effects were identified as with using the measure of evidence but additional nonadditive effects could be detected. For instance, in Figure 3B, three more dominance effects were found on chromosomes 1–3. However, each of them explained less than 0.02 % of the genetic variance and might be negligible.



## 7 | DISCUSSION

This study revisited the approximate Bayesian approach “fastbayes” that was designed for genomic evaluations. The suitability of this approach in terms of accuracy of estimating genetic values and genetic variation has been elucidated earlier, and it has been compared to an MCMC-based method in Wittenburg et al. (2011). In this study, it was shown that the extension of fastbayes to include a marginal significance test for SNP effects, which is theoretically founded, enables the detection of loci relevant to genetic variation. Though the specificity of the corresponding measure of evidence was hardly affected by changes in the type-I error level, sensitivity was better at higher levels, for instance  $\alpha = 0.20$ . Particularly regarding the nonadditive effects, sensitivity was even larger when the Bayes factor was used but then specificity was slightly decreased. However, its benefit was leveled in real data investigations. Sensitivity and specificity were also similar to a variational Bayesian method “vbay” (Logsdon et al., 2010). Fastbayes required less computing time than vbay.

For the identification of loci with significant additive or nonadditive genetic impact on a trait, MCMC-based stochastic variable selection methods are useful (e.g. Bennewitz, Edel, Fries, Meuwissen, & Wellmann, 2017; Yi et al., 2005). Depending on the number of iterations and on the model dimension, such methods are exact but may need exhausting computing time. Pairwise interaction effects lead to millions of model parameters to be estimated. To avoid this, flexible model reduction techniques exist. For instance, the reversible-jump technique can be used to avoid an oversaturated model (Balestre & de Souza, 2016). As mixing can be poor in reversible-jump MCMC algorithms (Hastie, 2005), other possibilities are sought for identifying the relevant loci. Using a single-marker regression model, which may be combined with a feature-ranking step for main and epistatic effects, is rather common in genome-wide association studies. A parallel (Schüpbach, Xenarios, Bergmann, & Kapur, 2010) and GPU-based (Kam-Thong et al., 2012; Ueki & Tamiya, 2012) implementation allow quickly scanning the genome for significant effects in ultra-high dimensions. Fastbayes differs from such a GWAS approach: while a single parameter (i.e. the effect of a single SNP or SNP pair) is considered successively, the vector of observations is marginalised (i.e. corrected for all other temporarily estimated effects) in an iterative manner. Thus, at each stage of iteration, the full genome is considered which should lead to a more precise localization of relevant loci. Considering additional kinds of genetic effects in a whole-genome regression model improves the power to map loci which are relevant to genetic variation in general. This was observed for the approximate Bayesian approaches used in this study, for MCMC-based methods (e.g. Bennewitz et al., 2017) and for other SNP-selection approaches (e.g. Sabourin, Nobel, & Valdar, 2015). However, sample size matters when nonadditive effects are studied (e.g. Van Steen, 2011).

As an option, a GWAS approach based on a single-marker regression model that considers the empirical correlation among SNPs (Zuber et al., 2012) was applied to the simulated data; it is available as R package `care` version 1.1.9. The top-ranked SNPs corresponding to a false discovery rate of 5 % were selected as significant. The rate of true positive detections was lower compared to fastbayes and vbay: sensitivity for additive effects was 26 %, for dominance 9 % and overall 32 % when  $H^2 = 0.5$ . The computation required 9 min on average.

The specification of hyperparameters may have a great impact on the outcome of Bayesian approaches (e.g. Tempelman, 2015). For the analysis of simulated data, the parameter  $\gamma$  specifying the proportion of mixing a zero and nonzero distribution of SNP effect was fixed near its true value and  $\lambda = \sqrt{2p\gamma}$  like in Wittenburg et al. (2011). For the real data analysis, a simple attempt was made to specify  $\gamma$  depending on the data. A more extensive approach is to execute cross validation like in Melzer, Wittenburg, and Reipsilber (2013). As an option toward a fully Bayesian approach, this parameter could be modeled additionally. For this purpose and similar to Scott & Berger (2006), it is assumed that  $\gamma$  is small. Then a suitable choice of the prior density, which allows a reasonable amount of variation, is  $p(\gamma) = (a + 1)(1 - \gamma)^a$  with some prior information  $a$ . The posterior expectation has been worked out in Additional File 3 and has been incorporated in the fastbayes algorithm. As the distribution of  $\gamma$  varies only little, the estimates of genetic effects and their measure of significance were not altered seriously in case of simulated data. As an example for  $H^2 = 0.5$  and based on the complex model including all kinds of genetic effects, on average one to two loci were additionally detected but the number of truly detected loci was almost unchanged. The sensitivity over all kinds of effects was 52 %. The significant loci explained about 90 % of the total genetic variance. The accuracy of genetic value prediction was 80 %.

Not surprisingly, the real data analysis of a heterogeneous stock of mice revealed results similar to Wittenburg et al. (2011). However, now the measure of evidence introduced in Section 2 rose clearly, proving the significance of the largest effects. Unlike Wittenburg et al. (2011),  $\gamma$  was specified based on a resampling approach. As a consequence,  $\gamma$  was increased for main effects and decreased for epistatic effects. Then more medium additive and dominance effects but no epistatic effects were detected in this study. The large dominance $\times$ additive epistatic effect that was reported earlier probably split into an additional additive and dominance effect.

Fastbayes is developed for, but not limited to, applications in animal breeding. It is straightforward to be applicable in plant breeding, where genome-based selection of phenotypes is an on-going issue for efficient plant production (e.g. Desta & Ortiz,

2014). This approach does not account for any population structure that might appear in a livestock population. But population stratification may cause biased allele frequency estimates used to set up  $X$ . As human genetic datasets typically consist of many unrelated individuals, this approach is perfectly suited for GWAS in human genetics. There, the elucidation of the genetic architecture of common diseases (e.g. Ripke et al., 2013) or other traits (e.g. Pickrell et al., 2016) is targeted, and the genetic merit is not focused. Furthermore, genetic data from human sibling studies can also be processed. Such data allow the precise estimation of the genetic impact of complex traits (e.g. Sariaslan et al., 2016). They are valuable to estimate the contribution of dominance to the phenotypic variation, too.

As discussed in Van Steen (2011), prior knowledge, which may be derived from biological information, would enhance the interpretation of results of genomic evaluations. Such information can be retrieved from data bases, networks or pathway analyses and can be included in a statistical model. As an option, this knowledge may be translated into weights  $w_j \in [0, \infty)$  for each locus or kind of effect  $j = 1, 2, \dots$ , where zero means exclusion and one the neutral weight. Similar to a weighted regression analysis, a vector of covariates  $\mathbf{x} = \mathbf{X}_j \sqrt{w_j}$  is considered in the one-locus model (2). This strategy of weighting has already been implemented in the fastbayes algorithm but it still needs to be validated.

## 8 | CONCLUSIONS

The extension of the approximate Bayesian approach “fastbayes” to include a marginal measure of significance allows now a quick scan of the genome to identify the SNPs that are relevant to genetic variation. The clear benefit of fastbayes is its dual-purpose use: it estimates the effect sizes for all kinds of genetic effects (additive and nonadditive) and determines the significant SNPs or SNP pairs. It requires much less computing time in higher, realistic dimensions than other approaches for whole-genome regression analyses. Due to its speed, it will also be valuable for the analysis of whole-genome sequence data.

## ACKNOWLEDGMENTS

We thank the Editor, Reproducible Research Editor and anonymous reviewer for their helpful comments. The publication of this article was funded by the Open Access Fund of the Leibniz Institute for Farm Animal Biology (FBN).

## AUTHORS' CONTRIBUTIONS

DW developed and implemented the theory, analyzed the data and wrote the manuscript. VL contributed to the theoretical investigations and suggested improvements to the manuscript. All authors have read and approved the final manuscript.

## ORCID

Dörte Wittenburg  <http://orcid.org/0000-0002-3639-2574>

## REFERENCES

- Balestre, M., & de Souza, C. L. (2016). Bayesian reversible-jump for epistasis analysis in genomic studies. *BMC Genomics*, *17*(1), 1012.
- Bennewitz, J., Edel, C., Fries, R., Meuwissen, T. H., & Wellmann, R. (2017). Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. *Genetics Selection Evolution*, *49*(1), 7.
- Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097.
- Chen, C., & Tempelman, R. (2015). An integrated approach to empirical Bayesian whole genome prediction modeling. *Journal of Agricultural, Biological, and Environmental Statistics*, *20*(4), 491–511.
- De Braganca Pereira, C. A., & Stern, J. M. (1999). Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy*, *1*(4), 99–110.
- 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*(2), 327–345.
- Desta, Z. A., & Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, *19*(9), 592–601.
- Falconer, D. S., & Mackay, T. F. C. (1996). *Quantitative genetics*, Longman, Essex, UK.
- Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, *32*(4), 361–369.
- George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of The American Statistical Association*, *88*(423), 881–889.



- Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*, 186.
- Hastie, D. (2005). *Towards automatic reversible jump Markov Chain Monte Carlo*, Ph.D. thesis, University of Bristol.
- Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, *4*(2), e1000 008.
- Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, *4*(7), e1000 130.
- Jeffrey, H. (1961). *Theory of probability*, Clarendon press, Oxford University Press.
- Kam-Thong, T., Azencott, C. A., Cayton, L., Pütz, B., Altmann, A., Karbalai, N., Sämman, P. G., Schölkopf, B., Müller-Myhsok, B., & Borgwardt, K. M. (2012). GLIDE: GPU-based linear regression for detection of epistasis. *Human Heredity*, *73*(4), 220–236.
- Laurie, C. C., Nickerson, D. A., Anderson, A. D., Weir, B. S., Livingston, R. J., Dean, M. D., Smith, K. L., Schadt, E. E., & Nachman, M. W. (2007). Linkage disequilibrium in wild mice. *PLoS Genetics*, *3*(8), e144.
- Li, Z., & Sillanpää, M. J. (2012). Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. *Genetics*, *190*(1), 231–249.
- Logsdon, B. A., Carty, C. L., Reiner, A. P., Dai, J. Y., & Kooperberg, C. (2012). A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. *Bioinformatics*, *28*(13), 1738–1744.
- Logsdon, B. A., Hoffman, G. E., & Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, *11*(1), 1.
- Melzer, N., Wittenburg, D., & Repsilber, D. (2013). Investigating a complex genotype-phenotype map for development of methods to predict genetic values based on genome-wide marker data—a simulation study for the livestock perspective. *Archiv Tierzucht*, *56*, 380–398.
- Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.
- Meuwissen, T. H. E., Solberg, T. R., Shepherd, R., & Woolliams, J. A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genetics Selection Evolution*, *41*, 2.
- Pérez, P., de los Campos, G., Crossa, J., & Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *The Plant Genome*, *3*(2), 106–116.
- Phillips, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*(11), 855–867.
- Pickrell, J. K., Berisa, T., Liu, J. Z., Séguérel, L., Tung, J. Y., & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, *48*, 709–717.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, AT. URL <http://www.R-project.org/>.
- Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., ... Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*(10), 1150–1159.
- Sabourin, J., Nobel, A. B., & Valdar, W. (2015). Fine-mapping additive and dominant SNP effects using group-LASSO and fractional resample model averaging. *Genetic Epidemiology*, *39*(2), 77–88.
- Sariaslan, A., Fazel, S., D'Onofrio, B., Långström, N., Larsson, H., Bergen, S., Kuja-Halkola, R., & Lichtenstein, P. (2016). Schizophrenia and subsequent neighborhood deprivation: Revisiting the social drift hypothesis using population, twin and molecular genetic data. *Translational Psychiatry*, *6*(5), e796.
- Schaeffer, L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, *123*, 218–223.
- Schüpbach, T., Xenarios, I., Bergmann, S., & Kapur, K. (2010). Fastepistasis: A high performance computing solution for quantitative trait epistasis. *Bioinformatics*, *26*(11), 1468–1469.
- Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, *136*, 2144–2162.
- Tempelman, R. J. (2015). Statistical and computational challenges in whole genome prediction and genome-wide association analyses for plant and animal breeding. *Journal of Agricultural, Biological, and Environmental Statistics*, *20*(4), 442–466.
- Ueki, M., & Tamiya, G. (2012). Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis. *BMC Bioinformatics*, *13*, 72.
- Valdar, W., Solberg, L. C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W. O., ... Flint, J. (2006a). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, *38*(8), 879–887.
- Valdar, W., Solberg, L. C., Gauguier, D., Cookson, W. O., Rawlins, J. N. P., Mott, R., & Flint, J. (2006b). Genetic and environmental effects on complex traits in mice. *Genetics*, *174*(2), 959–984.
- Van Steen, K. (2011). Travelling the world of gene–gene interactions. *Briefings in Bioinformatics*, p. bbr012.

- Wittenburg, D., Melzer, N., & Reinsch, N. (2011). Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. *BMC Genetics*, *12*, 74.
- Yi, N., Yandell, B. S., Churchill, G. A., Allison, D. B., Eisen, E. J., & Pomp, D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, *170*(3), 1333–1344.
- Zuber, V., Duarte Silva, A. P., & Strimmer, K. (2012). A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies. *BMC Bioinformatics*, *13*, 284.
- Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(4), 1193–1198.

## SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

**How to cite this article:** Wittenburg D, Liebscher V. An approximate Bayesian significance test for genomic evaluations. *Biometrical Journal*. 2018;60:1096–1109. <https://doi.org/10.1002/bimj.201700219>



J. Dairy Sci. 96:2557–2569

<http://dx.doi.org/10.3168/jds.2012-5635>

© American Dairy Science Association®, 2013.

## Milk metabolites and their genetic variability

D. Wittenburg,<sup>\*1</sup> N. Melzer,<sup>\*</sup> L. Willmitzer,<sup>†</sup> J. Lisec,<sup>†</sup> U. Kesting,<sup>‡</sup> N. Reinsch,<sup>\*</sup> and D. Repsilber<sup>\*1</sup>

<sup>\*</sup>Institute for Genetics and Biometry, Unit Biomathematics and Bioinformatics, Leibniz Institute for Farm Animal Biology, 18196 Dummerstorf, Germany

<sup>†</sup>Max Planck Institute for Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

<sup>‡</sup>Landeskontrollverband für Leistungs- und Qualitätsprüfung Mecklenburg-Vorpommern e.V. (LKV), 18273 Güstrow, Germany

### ABSTRACT

The composition of milk is crucial to evaluate milk performance and quality measures. Milk components partly contribute to breeding scores, and they can be assessed to judge metabolic and energy status of the cow as well as to serve as predictive markers for diseases. In addition to the milk composition measures (e.g., fat, protein, lactose) traditionally recorded during milk performance test via infrared spectroscopy, novel techniques, such as gas chromatography-mass spectrometry, allow for a further analysis of milk into its metabolic components. Gas chromatography-mass spectrometry is suitable for measuring several hundred metabolites with high throughput, and thus it is applicable to study sources of genetic and nongenetic variation of milk metabolites in dairy cows. Heritability and mode of inheritance of metabolite measurements were studied in a linear mixed model approach including expected (pedigree) and realized (genomic) relationship between animals. The genetic variability of 190 milk metabolite intensities was analyzed from 1,295 cows held on 18 farms in Mecklenburg-Western Pomerania, Germany. Besides extensive pedigree information, genotypic data comprising 37,180 single nucleotide polymorphism markers were available. Goodness of fit and significance of genetic variance components based on likelihood ratio tests were investigated with a full model, including marker- and pedigree-based genetic effects. Broad-sense heritability varied from zero to 0.699, with a median of 0.125. Significant additive genetic variance was observed for highly heritable metabolites, but dominance variance was not significantly present. As some metabolites are particularly favorable for human nutrition, for instance, future research should address the identification of locus-specific genetic effects and investigate metabolites as the molecular basis of traditional milk performance test traits.

**Key words:** metabolome, genomic relationship, single nucleotide polymorphism, heritability

### INTRODUCTION

In dairy cattle, a multitude of milk components are recorded during milk performance tests. Besides monitoring performance traits (e.g., fat or protein content), it is especially important to control udder health by means of indicator traits, such as cell count. Other milk composition traits, for instance, FA levels or acetone, are also involved in indicating management status of the cow. Most of these traits can be measured with infrared spectroscopy. Infrared spectroscopy has been extended to a more detailed analysis of milk components; for example, protein content has been further resolved into its components, such as  $\kappa$ -CN and  $\beta$ -LG (Rutten et al., 2011). The same technique is used to measure the FA composition of milk (Soyeurt et al., 2006). Although infrared spectroscopy is suitable for population-wide animal recording, novel techniques allow for a further analysis of milk (Töpel, 2004). Coupling GC-MS breaks milk down into its metabolic intermediates. Besides GC-MS, other technical processes (e.g., nuclear magnetic resonance spectroscopy; Nicholson et al., 1999) are available, which differ from GC-MS in terms of quantity, reproducibility, and sensitivity in matter determination. Klein et al. (2010) have applied GC-MS to milk samples and obtained a few metabolites for further investigation; in plants, however, GC-MS has been shown to be suitable for identification and relative quantification of several hundred metabolites in high throughput (40 min per sample; Lisec et al., 2006). Thus, GC-MS is applicable to study sources of genetic and nongenetic variation of milk metabolites in dairy cows. The milk metabolome is a snapshot of the metabolic state of a cow; thus, metabolites may primarily help to explore metabolic (production) diseases, such as ketosis, milk fever, rumen acidosis, or fatty liver syndrome (Littledike et al., 1981; Goff and Horst, 1997), and to infer the risk of a disease. For instance, 3-hydroxybutanoic acid is typically used as a biomarker for ketosis (Geishauser et al., 2000), and a related study

Received April 18, 2012.

Accepted December 13, 2012.

<sup>1</sup>Corresponding authors: wittenburg@fbn-dummerstorf.de and repsilber@fbn-dummerstorf.de

found a direct relationship between ketosis and the ratio of nuclear magnetic resonance spectroscopy-derived milk components from glycerophosphocholine to phosphocholine (Klein et al., 2012). Furthermore, some milk components may be of particular interest; for example, bovine milk oligosaccharides (**BMO**) resemble human milk oligosaccharides in structure and have a similar protective role on infants' intestines and immune system (Zivkovic and Barile, 2011) and may be a target for functional food production. Thus, milk metabolites may not only be used as biomarkers, they can also generally be seen as novel milk traits from which other production or fitness traits may be deduced and whose genetic background should be elucidated. Then, it is necessary to quantify the extent to which the observed variability of metabolic profiles is due to either genetic variation or environmental or temporal variation.

The primary interest of this study was to estimate genetic parameters and explore mode of inheritance of milk metabolites which were obtained from milk samples collected during performance testing. Toward these objectives, genetic variability of metabolite measurements was investigated in a linear mixed model approach including expected (pedigree) and realized (genomic) relationship between animals for a large sample of dairy cows. The significance of genetic variance components was tested and the ability of milk metabolites to predict genetic values for nonphenotyped animals was studied to verify their meaning for breeding purposes. The heritable effect on chemically related metabolites as well as functionally related metabolites is discussed with respect to selected groups and pathways.

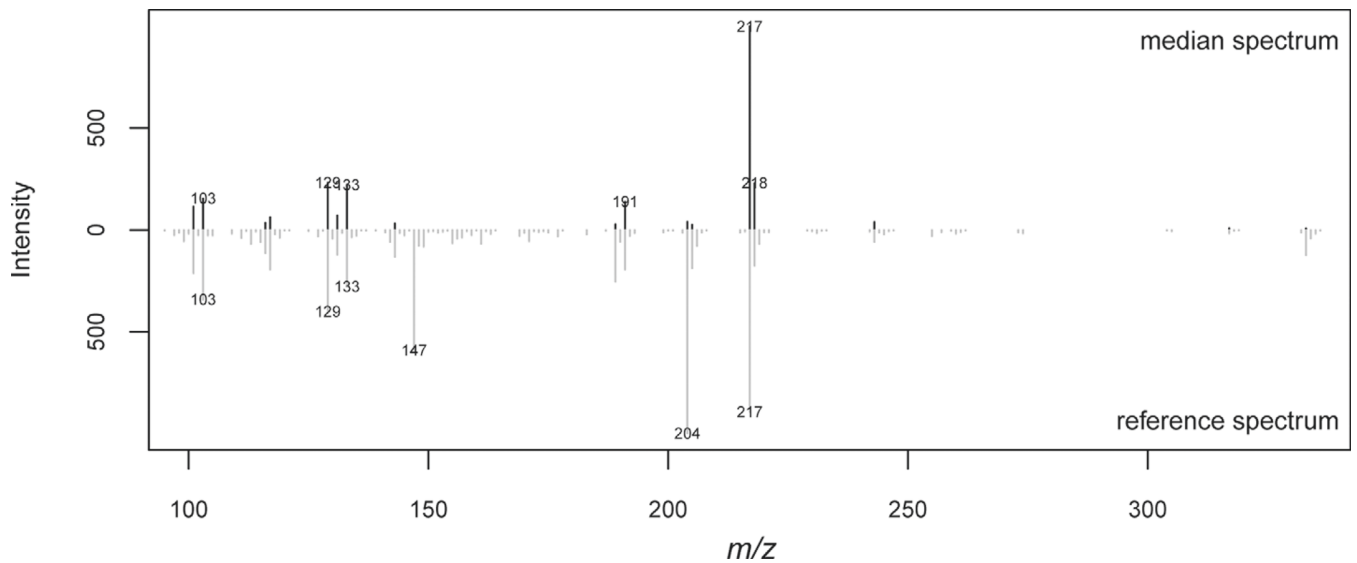
## MATERIALS AND METHODS

### *Milk Samples and Metabolite Data*

This field study was designed to gather samples under the same conditions as performance testing. Milk samples were collected from 1,344 Holstein cows held on 18 dairy farms in Mecklenburg-Western Pomerania (in the northeast of Germany) from May to November 2009. The cows were sampled between d 21 and 120 of their first lactation. Only first-lactation cows were selected to avoid variation due to parity and effects of selection due to culling of cows with low milk yield. As this project cooperated with commercial dairy farms only, a logistical challenge existed to separate samples of preselected cows that lay in the required lactation period. For this purpose, the LKV (Association for Quality Inspection, Güstrow, Germany) had access to a farm-specific animal list published and updated weekly on a website. The quality inspector drew 2 samples from the aliquot of the cow's daily milk yield—one according

to the usual procedure for milk performance tests, and the other drawn for use in the experiment. Preservatives (sodium acid) were added to the milk samples. The additional samples were transported in separate boxes to the LKV, and analyzed via infrared spectroscopy (Foss, Hillerød, Denmark). Afterward, the remaining milk was aliquoted into 2-mL tubes (Eppendorf, Germany) and frozen using liquid nitrogen until sample collection was finished (39 collection dates). Samples were deep-frozen within a few hours after being collected on a regular test day. All samples (one tube per cow) were sent to the laboratory at the Max Planck Institute for Molecular Plant Physiology (MPIMP; Potsdam, Germany) to measure the metabolite profiles. Twelve animals were later removed from the data set due to an invalid milk measurement (outside the desired lactation interval or because of decreased pH value).

Profiles for the hydrophilic fraction of metabolites were obtained from GC-MS according to Liseć et al. (2006) with minor adjustments [no ball milling, 1.1-mL total extraction volume; 100  $\mu$ L of milk sample + 1 mL of MeOH:CHCl<sub>3</sub>:H<sub>2</sub>O; 10:90 N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) derivatization]. Sample data were unbalanced in terms of farm, sample date, and half-sib families. For example, a sire had, on average, 6.6 daughters, ranging from 1 to 106. Thus, for GC-MS application, a specific randomized design based on a Latin square was developed, which was as balanced as possible with respect to the factors mentioned above (Melzer et al., 2010). Because of laboratory restrictions, the final design was slightly modified. The laboratory delivered molecule spectra, measured in 47 batches, in which molecule retention time (GC step), the mass:charge ratios, and the corresponding relative intensities of molecule fragments (MS step) were recorded for each sample. These spectra were further processed with the R package TargetSearch version 1.10 (Cuadros-Inostroza et al., 2009; R Development Core Team, 2011). The retention time of each molecule was converted into a retention index based on the retention time standards of FA methyl esters added to the sample in the GC step. Both retention index and molecule spectrum were used to annotate each molecule obtained from the MS step. Molecule spectra from narrow time windows (0.5 s), which showed highly correlated intensity values (correlation >0.95) over all samples, were combined to build a metabolite spectrum. Median values of these mass spectra were then compared with reference spectra in a database (Golm Metabolome Database, **GMD**; <http://gmd.mpimp-golm.mpg.de/search.aspx>). The assignment of a metabolite spectrum to the reference was accepted in case that the similarity score ( $\in[0; 1,000]$ ) between them was >500; otherwise, the metabolite was labeled



**Figure 1.** Metabolite spectrum of 1,6-anhydro- $\beta$ -glucose; the spectrum above the null line was obtained as median of peak intensities of all cows along the mass:charge ratio [ $m/z$ ] of molecule fragments, and the matching reference spectrum below the null line was obtained from a reference substance. The similarity score between metabolite spectrum and reference was 639.

as unknown. As an example, Figure 1 shows a median metabolite spectrum mapped to the corresponding reference. The intensity at the largest peak was taken as individual observation; metabolites with more than 20% missing values were omitted. Using these criteria, 187 identified and 3 unknown metabolites could be measured. The database mentioned above also allowed for assorting the metabolites with respect to their chemical groups. This structure was also used to give an overview of results of investigations.

#### Marker Data and Pedigree Structure

Blood samples were taken from all cows on a multiple-day tour with a veterinarian. Purification of DNA from blood was carried out with a commercially available kit (NucleoSpin Blood L, Macherey-Nagel, Düren, Germany). The concentration of DNA was controlled with NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, Wilmington, MA), ranging from 50 to 100 ng/ $\mu$ L. The laboratory at the Helmholtz Zentrum München, Germany, determined SNP genotypes using the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA). Out of 54,001 SNP on the chip, 48,713 SNP with known position were identified via BLAST analysis (Altschul et al., 1990) based on the SNP annotation Btau4.2 (The Bovine Genome Sequencing and Analysis Consortium, 2009). Then, 27 animals were skipped for having more than 10% missing SNP genotypes. Standard quality checks were applied to the SNP data (Ziegler et al., 2008);

loci that were not in Hardy-Weinberg equilibrium and loci with minor allele frequency <5%, or with more than 10% missing genotypes, were omitted. In total,  $m = 37,180$  SNP were retained. The rarely missing genotypes (0.5%) of these SNP were imputed by randomly drawing missing alleles according to observed allele frequencies. Furthermore, 10 additional animals were excluded because deviations between the realized (genomic) and expected (pedigree) relationship indicated Mendelian inconsistencies. Consistencies were assumed when 90% of the absolute differences between relationship coefficients were smaller than 0.2 (Calus et al., 2011). Thus,  $n = 1,295$  cows with proper genotypes and phenotypes remained for analyses. Cows were descendants of 192 sires, but 22 cows had unknown sires; the pedigree included 23,819 animals with up to 11 generations backward (obtained from the data center, VIT, Verden, Germany).

#### Statistical Model

The genetic variability of milk metabolites was studied taking the genomic BLUP (**GBLUP**) approach (Habier et al., 2007; VanRaden, 2008), for which a robust behavior in terms of accuracy of genetic value prediction was shown for various traits (Daetwyler et al., 2010). As metabolites were observed and analyzed on the level of genotyped cows, both additive and dominance genetic effects could be considered; this required a proper extension of the GBLUP approach to include dominance.



The raw metabolite intensities were log<sub>2</sub>-transformed to approach normality and then analyzed in a series of univariate analyses, one per metabolite. The vector  $\mathbf{y} = (y_1, \dots, y_n)'$  consisted of log<sub>2</sub> intensities for animals  $i = 1, \dots, n$ . Genetic variance captured by the markers, as well as polygenic effects, which cover residual additive genetic variation, were considered. The following linear mixed model was fitted to the data (bold indicating matrices and vectors):

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_a\mathbf{u}_a + \mathbf{Z}_d\mathbf{u}_d + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}.$$

The residuals were assumed to be independently and normally distributed,  $e_i \sim N(0, \sigma_e^2)$  for  $i = 1, \dots, n$ . The vector  $\mathbf{b}$  included systematic effects on the milk metabolites and considered fixed effects of farm  $\times$  sampling date (63 levels), metabolite measurement day (batch effect, 47 levels) and, to account for varying metabolism in different stages of lactation, linear and quadratic regression on lactation day with suitable entries in the design matrix  $\mathbf{X}$ . The design matrices  $\mathbf{Z}_a$  for the additive genetic values ( $\mathbf{u}_a$ ) and  $\mathbf{Z}_d$  for the dominance genetic values ( $\mathbf{u}_d$ ) were identity matrices of proper size because each individual was phenotyped and genotyped. The polygenic effect  $\mathbf{u}_p \sim N(\mathbf{0}, \mathbf{A}\sigma_p^2)$  with design matrix  $\mathbf{Z}_p$  involved the numerator relationship matrix  $\mathbf{A}$  obtained from pedigree information, and it was assumed to be uncorrelated with genetic effects explained by markers. The direct genetic value  $DGV_i$  of animal  $i$  was defined as the sum of effect-specific genetic values; that is,  $DGV_i = u_{a,i} + u_{d,i} + u_{p,i}$ .

Next, the covariance matrices of the effect-specific genetic values were inferred. The genetic value of type  $s \in \{a, d\}$  was defined as the sum of genetic effects depending on the individual SNP genotypes, that is,  $\mathbf{u}_s = \mathbf{M}_s\mathbf{g}_s$ . The  $(n \times m)$  matrix  $\mathbf{M}_s$  with columns  $\mathbf{M}_{s,1}, \dots, \mathbf{M}_{s,m}$  included orthogonalized genotype coefficients for the genetic effects  $\mathbf{g}_s = (g_{s,1}, \dots, g_{s,m})'$  over the whole genome. This way, the additive genetic values  $\mathbf{M}_{a,j}g_{a,j}$  and the dominance deviations  $\mathbf{M}_{d,j}g_{d,j}$  were uncorrelated at locus  $j \in \{1, \dots, m\}$ . Let 1 and 2 denote the SNP alleles at locus  $j$ , where 2 means the more frequent allele with frequency  $p_j$ . Depending on the observed marker genotypes, the entries in  $\mathbf{M}_s$  were obtained according to Zeng et al. (2005) as follows:

$$M_{a,i,j} = \begin{cases} -2p_j & \text{genotype 11} \\ 1 - 2p_j & \text{genotype 12} \\ 2(1 - p_j) & \text{genotype 22} \end{cases}$$

$$\text{and } M_{d,i,j} = \begin{cases} -2p_j^2 & \text{genotype 11} \\ 2p_j(1 - p_j) & \text{genotype 12} \\ -2(1 - p_j)^2 & \text{genotype 22} \end{cases}$$

These matrices were additionally standardized column by column:

$$\mathbf{M}_{a,j} \mapsto \frac{\mathbf{M}_{a,j}}{\sqrt{2p_j(1 - p_j)}}$$

$$\text{and } \mathbf{M}_{d,j} \mapsto \frac{\mathbf{M}_{d,j}}{2p_j(1 - p_j)}.$$

Thus, only correlations between SNP were of prime interest, meaning that SNP with rare alleles were as important as SNP with medium allele frequency for statistical investigations. If it was assumed that  $g_{s,1}, \dots, g_{s,m}$  are independent and identically distributed and each locus contributes to the genetic variation with  $\text{Var}(g_{s,j}) = \sigma_s^2$ , then the covariance matrix of the genetic values could be derived as

$$\text{Var}(\mathbf{u}_s) = \mathbf{M}_s\mathbf{M}_s'\sigma_s^2 =: \mathbf{G}_s m \sigma_s^2.$$

The genomic relationship matrix  $\mathbf{G}_a = \frac{1}{m} \mathbf{M}_a\mathbf{M}_a'$  corresponded exactly to the  $\mathbf{G}$  matrix of VanRaden (2008; version 2) and, in case of dominance,  $\mathbf{G}_d = \frac{1}{m} \mathbf{M}_d\mathbf{M}_d'$  included the correlation of dominance deviations. Thus, in the model above, it was assumed that  $\mathbf{u}_s \sim N(\mathbf{0}, \mathbf{G}_s m \sigma_s^2)$  for  $s \in \{a, d\}$ .

Let  $\sigma_A^2 = m\sigma_a^2$  denote the additive genetic variance captured by the markers and, analogously,  $\sigma_D^2 = m\sigma_d^2$  denotes the dominance variance. Narrow-sense heritability ( $h^2$ ) and broad-sense heritability ( $H^2$ ) were determined as follows:

$$h^2 = \frac{\sigma_A^2 + \sigma_p^2}{\sigma_A^2 + \sigma_p^2 + \sigma_D^2 + \sigma_e^2}$$

$$\text{and } H^2 = \frac{\sigma_A^2 + \sigma_p^2 + \sigma_D^2}{\sigma_A^2 + \sigma_p^2 + \sigma_D^2 + \sigma_e^2}.$$

Based on the full model, including marker- and pedigree-based genetic effects, goodness of fit was evaluated by metabolite-wise visual inspection of studentized residuals.

### Testing Significance of Genetic Variance Components

Assuming independent loci, and under the given parameterization of genotype coefficients, the additive genomic relationship matrix  $\mathbf{G}_a = \frac{1}{m} \mathbf{M}_a \mathbf{M}'_a$  approximates the numerator relationship matrix  $\mathbf{A}$ , where the quality of approximation depends on the number of loci  $m$  (Habier et al., 2007). Because of the finite number of loci and their dependence due to linkage, it makes sense to additionally account for genetic background in the model. To evaluate the significant presence of the polygenic variation, the null hypothesis  $H_0: \sigma_p^2 = 0$  versus the alternative hypothesis  $H_A: \sigma_p^2 > 0$  was tested via residual likelihood ratio test. This led to a nonstandard testing problem because the parameter to be tested lay at the boundary of its parameter space under  $H_0$ . The asymptotic null distribution of the test statistic is a  $\frac{1}{2} : \frac{1}{2}$  mixture of  $\chi^2$  distribution with 1 degree of freedom and point mass  $\delta_0$  at zero (Self and Liang, 1987; case 5). The significance of any additive genetic variation (i.e., both marker- and pedigree-based genetic variability) was also validated. Under the corresponding null hypothesis  $H_0: \sigma_A^2 = 0$  and  $\sigma_p^2 = 0$ , the distribution of the residual likelihood ratio test statistic asymptotically follows a  $\frac{1}{4} : \frac{1}{2} : \frac{1}{4}$  mixture of  $\delta_0$  and  $\chi^2$  distributions with 1 and 2 degrees of freedom (Self and Liang, 1987, case 9). Furthermore, the significance of the dominance variation was analyzed with  $H_0: \sigma_D^2 = 0$ . Again, the residual likelihood ratio test null distribution being  $\frac{1}{2} \delta_0 + \frac{1}{2} \chi_1^2$  was involved.

In metabolite-wise investigations, the type-I error was 5%; the  $P$ -values were adjusted to control the false discovery rate at 5% over all metabolites (Benjamini and Hochberg, 1995). Unless explicitly mentioned, all results presented below are based on false discovery rate-corrected data.

### Testing Significance of Fixed Effects

For evaluation of metabolite measurements varying over the different stages of lactation, the significance of the regression coefficients concerning lactation day was tested by an  $F$ -test with adjusted denominator

degrees of freedom (Kenward and Roger, 1997). In the software applied (see below), linear and quadratic regression on lactation day were realized with orthogonal polynomials of first and second degree. Therein, the  $F$ -test considered the null hypothesis of both regression coefficients being zero.

To verify which order of regression on lactation day was actually appropriate, nonparametric regression was applied to precorrected metabolite data and the equivalent number of parameters ( $p_e$ ) was determined (Ruppert et al., 2003). Then, a  $(p_e - 1)$ -degree polynomial was suggested to fit the data. Precorrection was obtained via linear mixed model including fixed effects of farm  $\times$  sampling date, measurement day, and a random sire effect to account for similarities among half-sibs. The method of choice was a penalized B-splines approach based on 20 knots (every fifth day).

The significance of farm  $\times$  sampling-date effect and metabolite measurement-day effect were also tested by an  $F$ -test.

### Predictive Ability

The predicted genetic value of an animal is an essential figure in breeding applications. The accuracy of predicting genetic values of nonphenotyped animals with milk metabolites observed in a training data set depends on heritability and relationship between animals. To study predictive ability of the GBLUP approach for metabolites in general, the accuracy of predicting genetic values, as well as predicting metabolite intensities, was evaluated. Predictive ability was assessed using a leave-one-out cross-validation (Hastie et al., 2009). In a successive manner, each single metabolite measurement was omitted from the training data and the cow's genetic value was predicted based on the present covariance structure. The estimates of fixed effects were also considered for predicting the metabolite intensity; after this was done for each animal, the empirical correlation ( $\rho$ ) between observed metabolite measurements and predicted genetic values served as a measure of accuracy of genetic value prediction. Accuracy of phenotype prediction was determined as empirical correlation between observed and predicted metabolite measurements. Because of the computational effort of fitting the full model, a reduced model ( $\sigma_p^2 = 0$ ) was applied to selected metabolites. Three highly heritable metabolites were chosen from the chemical groups of sugars and alcohols; then, the metabolite with highest heritability (based on the full model) and another arbitrarily chosen metabolite with medium heritability were investigated.

### Software

Quality checks on the SNP genotypes and set-up of genomic relationship matrices were implemented in Fortran 95; the imputation of rarely missing genotypes was performed using Beagle 3.2 (Browning and Browning, 2007). The estimation of systematic and genetic effects and the REML estimation of variance components were carried out with ASReml 3.0 (Gilmour et al., 2008). Model fit, likelihood ratio testing (based on the output of ASReml) and the cross validation were carried out in R (R Development Core Team, 2011). The B-spline approach was readily implemented in the R function `smooth.spline`.

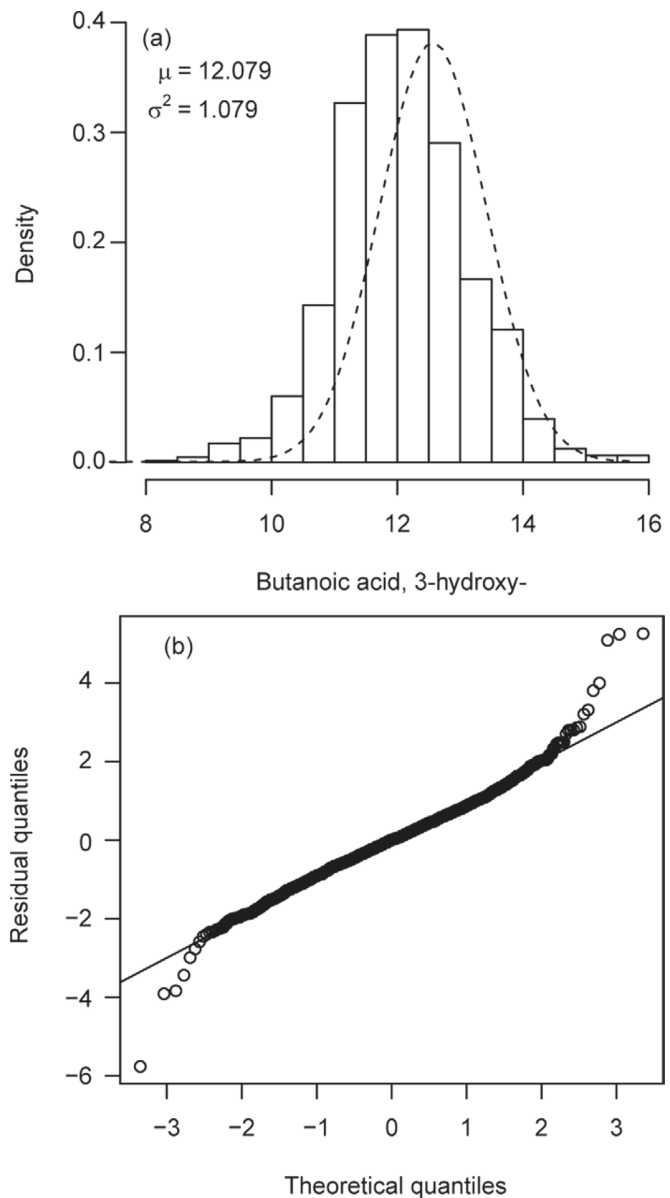
## RESULTS

### Model Fit

The raw metabolite intensities were heavily left-skewed (Supplemental Figure S1, available online at <http://www.journalofdairyscience.org/>). For most metabolites, the distribution of the  $\log_2$ -transformed metabolite measurements had a symmetric shape and approximately followed a normal distribution; for example, the histogram of 3-hydroxybutanoic acid is shown in Figure 2a. For a few metabolites, the fit at the tails was slightly insufficient due to increased frequency of extreme observations; the lack of fit was also obvious from the residual plot. The Q-Q (quantile-quantile) plot in Figure 2b shows prevailing coincidence of studentized residuals with a standard normally distributed variable. From the Q-Q plots of studentized residuals, it was concluded that the normal assumption made for the full model was generally fulfilled, such that the GB-LUP approach was suitable to explore genetic effects on metabolites.

### Genetic Variability

As the estimated genetic variance components cannot be compared among metabolites when expressed on the original scale of  $\log_2$  intensities, the proportion of genetic variation to the total phenotypic variance ( $h^2$  or  $H^2$ ), which is a universal measure, is emphasized below. Estimates of variances and heritability based on the full model of all 190 metabolites can be found in Supplemental Figure S2 (available online at <http://www.journalofdairyscience.org/>). Estimates of  $H^2$  varied from zero to 0.699, with a median of 0.125 (standard errors ranged from zero to 0.196), and  $h^2$  varied from zero to 0.569, with a median of 0.063 (standard errors ranged from zero to 0.117). To give an overview of substance groups present in the data, results were sorted according to the 51 identified chemical groups of metabolites (refer-



**Figure 2.** (a) Histogram of  $\log_2$ -transformed metabolite intensities of 3-hydroxybutanoic acid; (b) Q-Q (quantile-quantile) plot of studentized residuals. Dashed line in (a) corresponds to density function of normal distribution with mean ( $\mu$ ) and variance ( $\sigma^2$ ) estimated empirically.

ring to GMD; see Supplemental Figure S3, available online at <http://www.journalofdairyscience.org/>). A shortened list, where metabolite groups are assembled into 17 superordinate chemical groups, is presented in Table 1. This table illustrates the effect of dominance on some metabolites; for example, AA, alcohols, and lactams differed little in terms of estimated  $h^2$  but  $H^2$  clearly showed the contribution of dominance variance to alcohols ( $h^2 = 0.141$ ,  $H^2 = 0.251$ ), lactams ( $h^2 =$



## GENETICS OF MILK METABOLITES

2563

**Table 1.** Average estimated heritabilities and their interquartile range within superordinate metabolite groups<sup>1</sup>

Group	Group size	$h^2$	IQR ( $h^2$ )	$H^2$	IQR ( $H^2$ )
Alcohol	5	0.141	0.201	0.251	0.328
Aldehyde	1	0.024	—	0.081	—
Amine	3	0.096	0.178	0.295	0.032
Amino acid	18	0.154	0.173	0.171	0.187
Carboxylic acid	19	0.121	0.138	0.197	0.133
Conjugate	2	0.048	0.035	0.101	0.140
Indole	1	0.107	—	0.150	—
Lactam	3	0.136	0.252	0.291	0.182
Nucleoside	3	0.030	0.083	0.038	0.081
Nucleotide	2	0.000	0.000	0.055	0.110
Other acid	16	0.155	0.234	0.185	0.263
Polyol	5	0.157	0.184	0.222	0.292
Purine	2	0.019	0.038	0.043	0.002
Pyrimidine	4	0.079	0.072	0.139	0.170
Sugar	18	0.161	0.288	0.197	0.340
Terpenoid	1	0.000	—	0.006	—
Unspecified <sup>2</sup>	87	0.088	0.125	0.144	0.165

<sup>1</sup> $h^2$  = narrow-sense heritability,  $H^2$  = broad-sense heritability, IQR = interquartile range.

<sup>2</sup>Unspecified denotes metabolites that could not be assigned to a chemical group.

0.136,  $H^2 = 0.291$ ), and, to a smaller extent, to AA ( $h^2 = 0.154$ ,  $H^2 = 0.171$ ). The boxplot of  $H^2$  in Figure 3a shows the range of  $H^2$  estimates within each metabolite group. The interquartile range of  $H^2$  estimates varied strongly between zero and 0.340; for instance, interquartile range was 0.133 for carboxylic acids and 0.328 for alcohols (Table 1).

Significant additive genetic variation was found in 55 of 190 metabolites, meaning that  $\sigma_A^2$  or  $\sigma_p^2$  were greater than zero. False discovery rate-corrected  $P$ -values are given in Supplemental Figure S2 (available online at <http://www.journalofdairyscience.org/>). Significant results were mainly found in sugars (6), amino (10), carboxylic (6), and other acids (8). Significant dominance variation was detected only at a metabolite-wise level; for example, in 2-oxoglutaric acid or benzoic acid. In such cases, 37.6 to 100% of genetic variation was due to dominance (with a moderate level of residual variance). Variation of dominance or polygenic effects was not significantly present at a false discovery rate level. Figure 3b, which compares estimates of  $h^2$  and  $H^2$ , also represents these type-III testing results. Significant contributions of additive genetic variation to phenotypic variation were observed for highly heritable metabolites.

The average computing time was 33 min for the full model, including the  $F$ -test, but 5 min were required for the reduced model under  $H_0: \sigma_p^2 = 0$  (on a 2.93-GHz multi-user system). Therefore, checking the predictive ability concentrated on the reduced model. Accuracy of genetic value prediction was modest for 3-(4-hydroxyphenyl)-lactic acid ( $\rho = 0.28$ ), glycerol-3-phosphate ( $\rho = 0.21$ ), and ribulose-5-phosphate ( $\rho = 0.19$ ), but low for the other selected metabolites (Table

2). Accuracy of phenotype prediction varied to a lesser extent, with correlations ranging from 0.67 to 0.77. Further, the mean squared error of prediction showed that lowly heritable metabolites were as good as highly heritable metabolites for predicting total metabolite intensities.

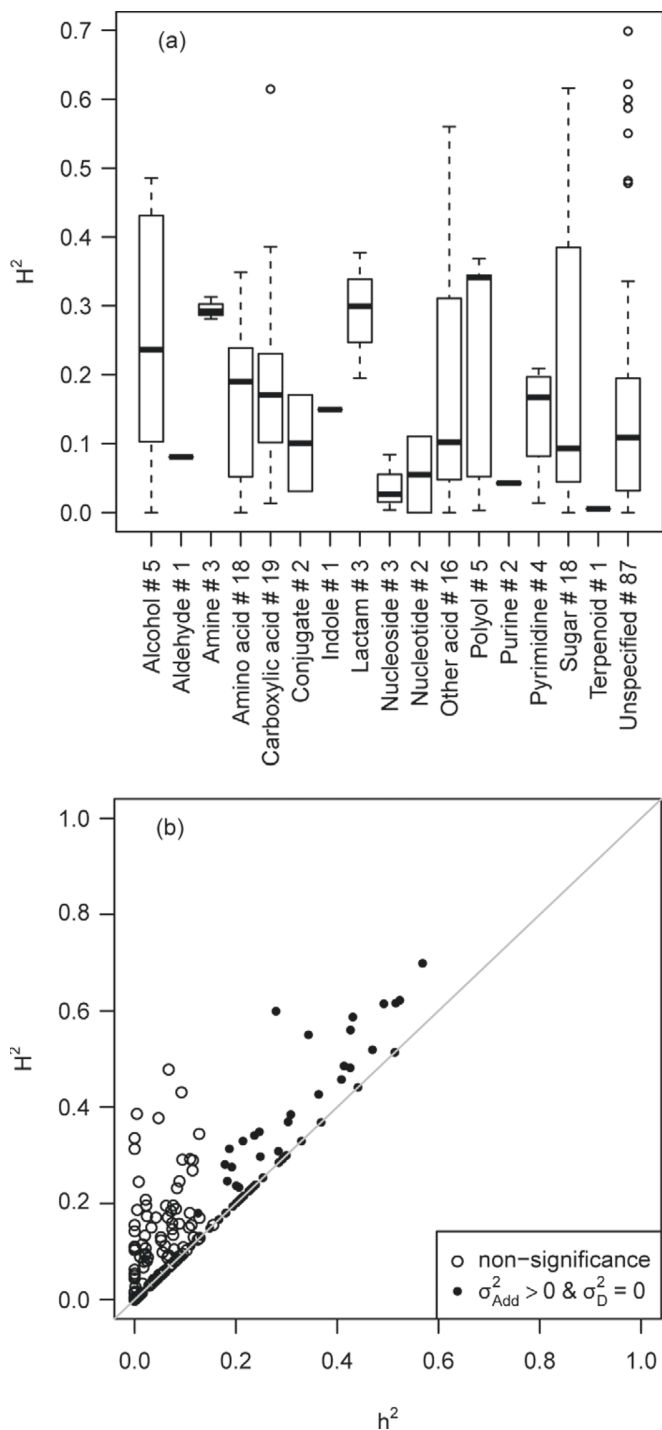
### Fixed Effects

The outcome for testing the effects of farm  $\times$  sampling date and measurement day was homogeneous—both effects were significantly present for all metabolites.

In total, 154 of 190 metabolites were significantly affected by lactation day. For 100 metabolites,  $p_e = 2$ , meaning that a linear regression was sufficient; for 27 metabolites,  $p_e = 3$ , arguing in favor of an additional quadratic term. Hence, higher order regression was only suitable for a minority of the metabolites. Mostly, the precorrected metabolite measurements appeared rather flat over the lactation period. Figure 4 gives 2 arbitrary examples, for which the quadratic regression on lactation day was appropriate despite the flat or almost linear behavior of the median of precorrected metabolites over the lactation period, which was divided into 5-d intervals.

### DISCUSSION

Milk metabolites are measurable using a GC-MS approach; they represent a new class of milk traits with unknown genetic background. Before investigating relationships of milk metabolites to classical milk traits—to performance or functional traits—it is necessary to characterize their genetic basis. The current study tackled this question, focusing on heritability and mode of



**Figure 3.** (a) Boxplot of broad-sense heritability within groups of metabolites; (b) comparison of narrow-sense heritability  $h^2$  versus broad-sense heritability  $H^2$ . Variance components: additive genetic variance  $\sigma_{\text{Add}}^2$ , which comprises marker-based ( $\sigma_A^2$ ) and polygenic ( $\sigma_p^2$ ) variance, dominance variance  $\sigma_D^2$ . “Unspecified” denotes metabolites that could not be assigned to a chemical group.

inheritance for these new molecular milk composition traits and their suitability for genetic evaluations.

### Genetic Variability

The genomic relationship between animals was used to model the covariance structure for additive and dominance genetic values. Residual additive genetic variation not captured by the SNP markers was modeled using the pedigree relationship. Heritabilities for the investigated milk metabolites in the sample of commercial dairy cows were mostly small to intermediate; half the metabolites had broad-sense heritability estimates smaller than 0.125. Almost 30% of the metabolites showed significant contribution of additive genetic variation to the phenotypic variation, but dominance variance was not significantly present. When the full model presented in this paper is compared with a typical animal model (including only pedigree-based genetic effects), then additional additive genetic variation could be explained by markers (Figure 5a). The average ratio of  $h^2$  estimated with both models was 1.689. In contrast to the animal model, the full model considered realized relationship among animals and therefore also accounted for Mendelian sampling effects. Thus, as expected, the increase in heritability also went along with a prevalently reduced standard error of  $h^2$  estimates (Figure 5b). For 97 of 138 metabolites, a lower standard error was observed; for the remaining 52 metabolites, the standard error was not estimable because the polygenic variance was estimated close to zero in the animal model.

The few examples studied for predictive ability showed rather limited power to predict genetic values; the accuracy of genetic value prediction was at most 0.28, despite  $H^2 = 0.584$  (Table 2). For comparison, the expected accuracy ( $r$ ) was approximated according to Daetwyler et al. (2010) using an effective number of chromosome segments  $M_e = 2N_eL$  with effective population size  $N_e = 100$  and chromosome length  $L = 30$  M. For the extreme examples,  $r = 0.33$  ( $H^2 = 0.584$ ) and  $r = 0.23$  ( $H^2 = 0.263$ ). These values were higher than the observed ones, which might be due to, among other reasons, an overestimated heritability or misspecified  $M_e$ . The accuracy of phenotype prediction was more convincing for the selected metabolites; the minimum correlation was 0.67. This outcome implied that more than 80% of the variation of metabolites could be explained by linear relationships using GBLUP, leaving a minor part unexplained.

A conceivable approach to increase predictive ability is combining selected metabolites with performance traits with which they are highly correlated in multivariate investigations. Therein, genetic correlations

## GENETICS OF MILK METABOLITES

2565

**Table 2.** Predictive ability of the genomic BLUP approach for selected metabolites via leave-one-out cross-validation<sup>1</sup>

Metabolite	$H^2$ (SE)	$\text{cor}(y, \hat{u}_a + \hat{u}_d)$	$\text{cor}(y, \hat{y})$	MSE
Lactic acid, 3-(4-hydroxyphenyl)-	0.584 (0.122)	0.284	0.700	0.599
Glycerol-3-phosphate	0.485 (0.129)	0.209	0.665	0.325
Ribulose-5-phosphate	0.484 (0.137)	0.185	0.709	1.036
Glucose, 2-amino-2-deoxy-	0.321 (0.147)	0.092	0.668	0.586
Glucose, 1,6-anhydro, $\beta$ -	0.263 (0.172)	0.013	0.769	0.663

<sup>1</sup>Parameters: broad-sense heritability  $H^2$  based on the reduced model (polygenic variance  $\sigma_p^2 = 0$ ) and its SE; observed ( $y$ ) and predicted ( $\hat{y}$ ) metabolite intensity, predicted additive genetic value  $\hat{u}_a$  and predicted dominance genetic value  $\hat{u}_d$ ; mean squared error (MSE) of prediction:  $\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ . Here,  $i$  denotes the sample number and  $n$  the total number of samples.

between multiple traits and phenotypic correlations are considered, because metabolites and milk traits are likely measured from the same milk sample. This issue requires a large sample size to properly estimate correlations, but will lead to an improved genetic value prediction of metabolites.

Milk metabolites, in general, might not be suitable for breeding purposes, but many metabolites showed significant genetic variability. Some of those components may be used to permanently indicate health risks, such as mastitis, but a permanent relation to diseases remains to be verified at different lactation stages. Davis et al. (2004) analyzed the relationship between mastitis and the concentration of lactic acid; as heritability of lactic acid was estimated at a low level in the sample data (i.e.,  $H^2 = 0.013$ ), this outcome suggests that lactic acid can be better seen as a temporary indicator for mastitis than as a permanent indicator. Metabolites that mostly reflect environmental or temporal variable status (as lactic acid does) may be useful to indicate variable disorders, and they might gain practical importance for monitoring and other management purposes (Melzer et al., 2013).

To improve the robustness of cows to metabolic disorders through specific breeding programs, concentration of selected metabolites with proved relation to a disease could be measured regularly during milk performance tests, for example, via infrared spectroscopy—as it is done for milk performance traits. This will also lead to a large collection of samples, thereby improving the predictive ability of metabolites. For instance, assuming  $H^2 = 0.03$ , about 6,700 individual observations are required to achieve an expected accuracy of 0.5.

Investigations incorporated SNP with minor allele frequency  $\geq 5\%$ . This implied that  $\mathbf{G}_a$  and  $\mathbf{G}_d$  were not positive definite as required for solving the mixed model equations. The ASReml options were used to regularize these covariance matrices; ASReml formed an expanded singular representation of the corresponding inverse matrices (Gilmour et al., 2008) and the outcome

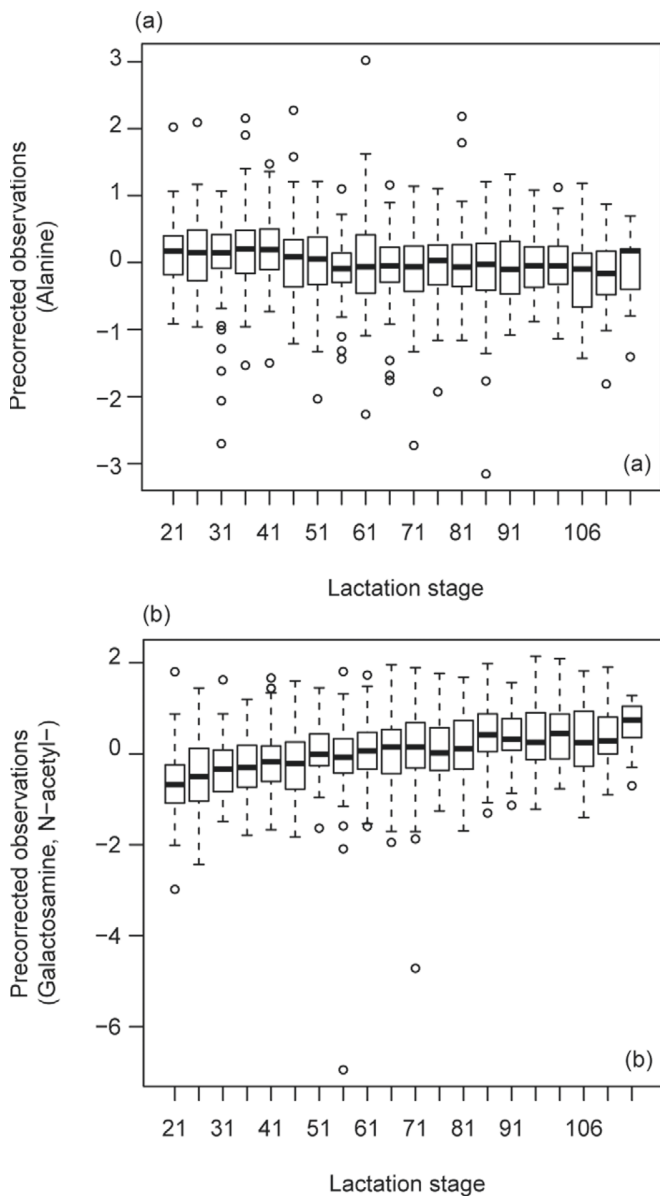
was similar to that of a regularization method, where a small value  $\varepsilon$  was added to the diagonal elements of the respective covariance matrix. As an alternative, one could select SNP with minor allele frequency  $\geq 10\%$ . In test runs, identical results were observed when adding  $\varepsilon = 0.001$  to the diagonal; other results were similar in terms of estimated heritabilities.

### Fixed Effects

Similar to a test-day model (Ptak and Schaeffer, 1993), some fixed effects on milk metabolite intensities were considered. Due to the large amount of data, effects of farm and sampling date were considered to be cross-classified. As all cows on a specific farm were fed the same food, an effect of diet was also considered by the effect of farm  $\times$  sampling date. Further, linear and quadratic regressions on lactation day were considered to account for a variable metabolite state between d 21 and 120 of lactation. Outside this lactation period a stronger deviation of metabolite intensities might be observed among animals requiring higher order regression. This study concentrated on the limited lactation period, assuming that the metabolism is mainly stable and not affected by metabolic diseases often appearing at early lactation (Goff and Horst, 1997). As milk metabolites were analyzed, the test-day model was extended to include the effect of measurement day, which accounts for batch effects mainly due to detector sensitivity when measuring the molecule intensities via MS (Steinfath et al., 2008).

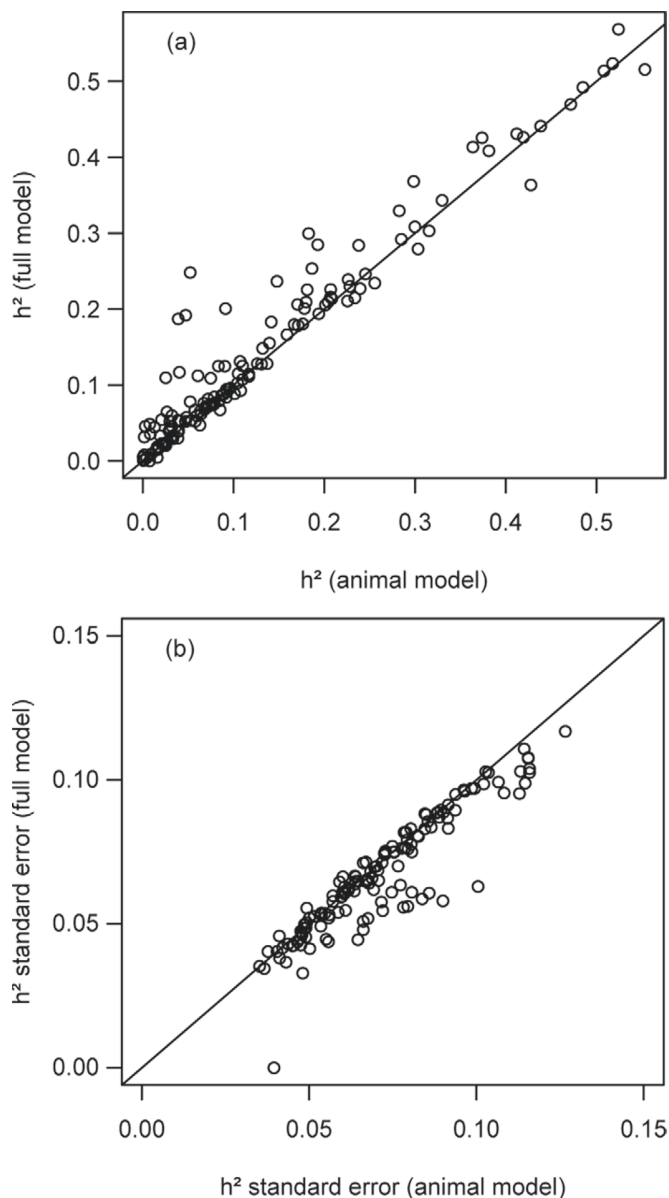
### Different Types of Metabolic Substances

According to the public database GMD, the metabolites could be assigned to 51 chemical groups. The AA represented the main group in the sample data, comprising 12 proteinogenic (e.g., alanine) and 6 nonproteinogenic (e.g., sarcosine) substances. The next frequently allocated groups were hydroxy (6) and dicarboxylic



**Figure 4.** Boxplot for precorrected metabolite measurements depending on lactation interval: (a) alanine and (b) *N*-acetylgalactosamine. Lactation period was divided into 5-d intervals. Precorrection was obtained via linear mixed model including fixed effects of farm  $\times$  sampling date and measurement day and a random sire effect.

acids (6; see Supplemental Figure S3, available online at <http://www.journalofdairyscience.org/>). In the shortened presentation of results (Table 1), where the estimates referred to superordinate metabolite groups, di- and tricarboxylic, aromatic, hydroxyl, and oxoacids were collected into carboxylic acids. Furthermore, all sugars and their derivatives, as well as modified sugars, were subsumed to sugars in general.



**Figure 5.** (a) Narrow-sense heritability based on full model (additive and dominance variance captured by markers and polygenic variance) versus heritability based on animal model (polygenic variance); (b) standard error of heritability estimates via full model versus animal model.

Compared with classical milk traits of the same animal sample (Melzer et al., 2013), such as content of protein ( $h^2 = 0.309$ ), fat ( $h^2 = 0.227$ ), or lactose ( $h^2 = 0.136$ ), heritabilities of milk metabolites were mostly below these reference values in the population studied. A few metabolites show larger estimates; the maximum value was estimated for 3-(4-hydroxyphenyl)-lactic acid at  $h^2 = 0.569$ . Amino acids are mainly associated with



protein production and breakdown, and on average  $h^2 = 0.154$  in this chemical group (Table 1). For FA components, which can be mainly found in the group of carboxylic acids, it was estimated that  $h^2 = 0.121$ . In the group of sugars,  $h^2 = 0.161$ . In each group, the estimated  $h^2$  strongly varied as it was also observed for  $H^2$  (Figure 3a).

### Pathways

To find reasons for the wide range of estimated heritabilities, functional associations between metabolites were investigated exemplarily using the KEGG database (<http://www.genome.jp/kegg>). A main metabolic pathway, glycolysis, converts glucose into pyruvates releasing energy in form of ATP. This process includes, among others, the following metabolites that were identified in the data: 2-phosphoglyceric acid ( $H^2 = 0.046$ ), 3-phosphoglyceric acid ( $H^2 = 0.308$ ), phosphoenolpyruvic acid ( $H^2 = 0.441$ ), and pyruvic acid ( $H^2 = 0.155$ ). Even though these components are closely functionally related in glycolysis, heritability strongly differed between them. Metabolites were positively correlated in this subset; correlations between residuals obtained from the full model varied from 0.064 to 0.84. Thus, the  $\log_2$ -transformed sum of (raw) metabolite intensities identified in glycolysis was investigated, and the genetic variability was determined based on the full model. For the derived variable,  $H^2 = 0.238$ , which coincided roughly with the average heritability of single metabolites.

In the pentose phosphate pathway, a few highly heritable metabolites were recognized: fructose-6-phosphate ( $H^2 = 0.686$ ), 6-phosphogluconic acid ( $H^2 = 0.426$ ), and ribulose-5-phosphate ( $H^2 = 0.519$ ). Some metabolites with small to intermediate heritability were also recognized: gluconic acid ( $H^2 = 0.297$ ), pyruvic acid, and 3-phosphoglyceric acid. Except gluconic acid, all residual metabolite intensities were positively correlated; correlations ranged from 0.045 to 0.84. The  $\log_2$ -transformed sum of positively correlated metabolite measurements resulted in  $H^2 = 0.425$ . This condition suggests that at least parts of the energy metabolism might be affected by genetics.

As some metabolites contribute to different pathways (for instance, pyruvic acid participates in glycolysis and pentose phosphate pathway), it might be an interesting point to further explore the relationship between pathways and their genetic characteristic. Eventually, metabolic flux modes (Hoffmann et al., 2006), which could be understood as realized paths of metabolite conversion, might show more elevated levels of genetic determination than the single metabolite intensities of the present study.

### Favorable Metabolites

Milk and its metabolic products, such as BMO or vitamins, are important for human nutrition. The sample data included few metabolites that contribute to vitamins: pantothenic acid ( $B_5$ ), pyridoxal and pyridoxamine ( $B_6$ ), as well as one component which is related to thiazol ( $B_1$ ), 4-methyl-5-hydroxyethyl-thiazol. These components seem to be affected more by management (e.g., feed and environment), than by genetics. Intermediate heritability was, however, found for pyridoxal ( $H^2 = 0.233$ ), and 4-methyl-5-hydroxyethyl-thiazol ( $H^2 = 0.226$ ; see Supplemental Figure S2, available online at <http://www.journalofdairyscience.org/>).

Among all metabolites, 3 main substances were found that contribute to the possible structures identified for BMO [Tao et al., 2008; i.e., *N*-acetylneuraminic acid (**Neu5Ac**), *N*-acetylgalactosamine (**GalNAc**), and *N*-acetylglucosamine (**GlcNAc**)]. As the concentration of beneficial oligosaccharides is lower in bovine than human milk, extraction of BMO from dairy sources is useful to enrich, for instance, infant formula (Zivkovic and Barile, 2011). The absolute quantity of substances found in the sample could not be determined, however, because a calibration for metabolite intensities in milk was not implemented in this study. Martín et al. (2001) reported that sialic acids, with their predominant component, Neu5Ac, amount to approximately 170 mg/kg of mature milk. For comparison, the BMO components glucose and galactose, which were not measured in this study, contribute up to 60 and 20 mg per kg, respectively (Töpel, 2004). Two of the metabolites, GalNAc and GlcNAc, had high estimates of heritability,  $H^2 = 0.587$  and  $H^2 = 0.482$ , respectively (see Supplemental Figure S2, available online at <http://www.journalofdairyscience.org/>). Food producers may desire milk of daughters of bulls with extraordinarily high breeding value with respect to favorable substances on special farms. The total oligosaccharide intensity is assumed to decrease from colostrum to later lactation milk (Tao et al., 2009). Colostrum was not available in this study, but GalNAc and GlcNAc increased slightly over the different stages of lactation (Figure 4b), and Neu5Ac was at least constant until d 100. Future studies may investigate which period of lactation BMO abundance is most suitable for extraction.

### CONCLUSIONS

In our study of the genetic background of 190 milk metabolites, heritability was found at a low to intermediate level. Even though the sample size of our experiment was small ( $n = 1,295$ ), results indicate the mode of inheritance. Most genetic variability was explained by additive genetic sources comprising pedi-

gree- and marker-based genetic variance. Significant additive genetic variation was found in 55 metabolites. Dominance variance was not significantly present. The few highly heritable metabolites studied for predictive ability showed modest to low accuracy of genetic value prediction. At the same time, accuracy of predicting total metabolite intensities was at a high level, confirming the general suitability of the GBLUP approach for studying variation of metabolite measurements. Future research should address the identification of locus-specific genetic effects on selected metabolites. This could be achieved by simultaneously estimating SNP effects over the whole genome (e.g., via BayesB; Meuwissen et al., 2001), instead of studying total genetic values, as was done in the current study. Relations of milk metabolite profiles to standard milk performance test traits should also be further investigated to discover whether it will be possible to elucidate the molecular basis of these classical traits using new, detailed milk composition traits.

#### ACKNOWLEDGMENTS

This study was part of the FUGATO plus project “Bovine Integrative Bioinformatics for Genomic Selection” with financial support of the German Federal Ministry of Education and Research (BMBF). The authors thank the participating herd owners and co-operation partners. The contribution of Å. Eckardt (Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany), who implemented GC-MS technology for milk samples and measured the metabolite profiles, is acknowledged. The SNP genotypes were measured by the working group of T. Meitinger and P. Lichtner (Helmholtz Zentrum München, Germany). S. Jakubowski, S. Hartwig, and S. Wolf (LKV Güstrow, Germany) provided milk performance data and contributed to discussions during the project. We acknowledge the support of F. Reinhardt (vit Verden), who assembled the pedigree data. Special thanks are given to colleagues at the Leibniz Institute for Farm Animal Biology (Dummerstorf, Germany): R. Grahl, who collected blood and milk samples; C. Reiko, who helped to prepare blood samples for DNA extraction; the working group of J. Vanselow, who provided technical facilities for DNA preparation; A. Rief, who built the SNP marker map; C. Kühn, who gave valuable remarks; and especially M. Nimz, M. Anders, and M. Spitschak, who gave assistance in preparation. We also thank the anonymous reviewers for their comments.

#### REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

- Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* 57:289–300.
- Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81:1084–1097.
- Calus, M. P. L., H. A. Mulder, S. McParland, E. Strandberg, E. Wall, and J. W. M. Bastiaansen. 2011. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. *Genet. Sel. Evol.* 43:34.
- Cuadros-Inostroza, A., C. Caldana, H. Redestig, M. Kusano, J. Liseč, H. Pena-Cortez, L. Willmitzer, and M. A. Hannah. 2009. Target-search—A Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data. *BMC Bioinformatics* 10:428.
- Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185:1021–1031.
- Davis, S. R., V. C. Farr, C. G. Prosser, G. D. Nicholas, S.-A. Turner, J. Lee, and A. L. Hart. 2004. Milk L-lactate concentration is increased during mastitis. *J. Dairy Res.* 71:175–181.
- Geishauser, T., K. Leslie, J. Tenhag, and A. Bashiri. 2000. Evaluation of eight cow-side ketone tests in milk for detection of subclinical ketosis in dairy cows. *J. Dairy Sci.* 83:296–299.
- Gilmour, A., B. Gogel, B. Cullis, and R. Thompson. 2008. ASReml User Guide Release 3.0. VSN International Ltd., Hemel Hempstead, UK.
- Goff, J. P., and R. Horst. 1997. Physiological changes at parturition and their relationship to metabolic disorders. *J. Dairy Sci.* 80:1260–1268.
- Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, New York, NY.
- Hoffmann, S., A. Hoppe, and H.-G. Holzhütter. 2006. Composition of metabolic flux distributions by functionally interpretable minimal flux modes (minmodes). *Genome Inform.* 17:195–207.
- Kenward, M. G., and J. H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53:983–997.
- Klein, M. S., M. F. Almstetter, G. Schlamberger, N. Nürnberger, K. Dettmer, P. J. Oefner, H. H. D. Meyer, S. Wiedemann, and W. Gronwald. 2010. Nuclear magnetic resonance and mass spectrometry-based milk metabolomics in dairy cows during early and late lactation. *J. Dairy Sci.* 93:1539–1550.
- Klein, M. S., N. Buttchereit, S. Miemczyk, A.-K. Immervoll, C. Louis, S. Wiedemann, W. Junge, G. Thaller, P. J. Oefner, and W. Gronwald. 2012. NMR metabolomic analysis of dairy cows reveals milk glycerophosphocholine to phosphocholine ratio as prognostic biomarker for risk of ketosis. *J. Proteome Res.* 11:1373–1381.
- Liseč, J., N. Schauer, J. Kopka, L. Willmitzer, and A. R. Fernie. 2006. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protoc.* 1:387–396.
- Littledike, E. T., J. W. Young, and D. C. Beitz. 1981. Common metabolic diseases of cattle: Ketosis, milk fever, grass tetany, and downer cow complex. *J. Dairy Sci.* 64:1465–1482.
- Martín, M.-J., S. Martín-Sosa, L.-A. García-Pardo, and P. Hueso. 2001. Distribution of bovine milk sialoglycoconjugates during lactation. *J. Dairy Sci.* 84:995–1000.
- Melzer, N., S. Jakubowski, S. Hartwig, U. Kesting, S. Wolf, G. Nürnberg, N. Reinsch, and D. Reipsilber. 2010. Design, infrastructure and database structure for a study on predicting of milk phenotypes from genome wide snp markers and metabolite profiles. Abstract ID 0427 in Proc. 9th World Congress on Genetics Applied to Livestock Production. German Society for Animal Science, ed. zwonul media GbR, Leipzig, Germany.
- Melzer, N., D. Wittenburg, S. Hartwig, S. Jakubowski, U. Kesting, L. Willmitzer, J. Liseč, N. Reinsch, and D. Reipsilber. 2013. Investigating associations between milk metabolite profiles and milk

## GENETICS OF MILK METABOLITES

2569

- traits of Holstein cows. *J. Dairy Sci.* 96:1521–1534. <http://dx.doi.org/10.3168/jds.2012-5743>.
- 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.
- Nicholson, J. K., J. C. Lindon, and E. Holmes. 1999. 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29:1181–1189.
- Ptak, E., and L. Schaeffer. 1993. Use of test day yields for genetic evaluation of dairy sires and cows. *Livest. Prod. Sci.* 34:23–34.
- R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Ruppert, D., M. P. Wand, and R. J. Carroll. 2003. *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.
- Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. *J. Dairy Sci.* 94:5683–5690.
- Self, S. G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* 82:605–610.
- Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. *J. Dairy Sci.* 89:3690–3695.
- Steinfath, M., D. Groth, J. Lisek, and J. Selbig. 2008. Metabolite profile analysis: From raw data to regression and classification. *Physiol. Plant.* 132:150–161.
- Tao, N., E. J. DePeters, S. Freeman, J. B. German, R. Grimm, and C. B. Lebrilla. 2008. Bovine milk glycome. *J. Dairy Sci.* 91:3768–3778.
- Tao, N., E. J. DePeters, J. B. German, R. Grimm, and C. B. Lebrilla. 2009. Variations in bovine milk oligosaccharides during early and middle lactation stages analyzed by high-performance liquid chromatography-chip/mass spectrometry. *J. Dairy Sci.* 92:2991–3001.
- The Bovine Genome Sequencing and Analysis Consortium. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* 324:522–528.
- Töpel, A. 2004. *Chemie und Physik der Milch: Naturstoff, Rohstoff, Lebensmittel*. Behr's Verlag, Hamburg, Germany.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423.
- Zeng, Z.-B., T. Wang, and W. Zou. 2005. Modeling quantitative trait loci and interpretation of models. *Genetics* 169:1711–1725.
- Ziegler, A., I. R. König, and J. R. Thompson. 2008. Biostatistical aspects of genome-wide association studies. *Biom. J.* 50:8–28.
- Zivkovic, A. M., and D. Barile. 2011. Bovine milk as a source of functional oligosaccharides for improving human health. *Adv. Nutr.* 2:284–289.





# Genomic additive and dominance variance of milk performance traits

D. Wittenburg, N. Melzer, N. Reinsch

Institute of Genetics and Biometry, Leibniz Institute for Farm Animal Biology, 18196 Dummerstorf, Germany

Corresponding author	Dörte Wittenburg
Address	Leibniz Institute for Farm Animal Biology Institute of Genetics and Biometry Wilhelm-Stahl-Allee 2 18196 Dummerstorf, Germany
Phone	+49 38208 68930
Fax	+49 38208 68902
Email	wittenburg@fhn-dummerstorf.de

**Published in:** *Journal of Animal Breeding and Genetics* (2015) 132: 3-8. <https://doi.org/10.1111/jbg.12103>

### **Abstract**

Milk performance traits are likely influenced by both additive and non-additive (e.g. dominance) genetic effects. Genetic variation can be partitioned using genomic information. The objective of this study was to estimate genetic variance components of production and milk component traits (e.g. acetone, fatty acids, etc.), which are particularly important for milk processing or which can provide information on the health status of cows. A genomic relationship approach was applied to phenotypic and genetic information of 1295 Holstein cows for estimating additive genetic and dominance variance components. Most of the 17 investigated traits were mainly affected by additive genetic effects, but protein content and casein content also showed a significant contribution of dominance. The ratio of dominance to additive variance was estimated as 0.64 for protein content and 0.56 for casein content. This ratio was highest for SCS (1.36) although dominance was not significant. Dominance effects were negligible in other moderately heritable milk traits.

**Keywords:** genetic variance, genomic relationship, SNP, Holstein cow

## Introduction

Phenotypic variation of milk traits may be caused by additive but also non-additive genetic effects. In particular, dominance effects have been a target of research for some time. As an example, depending on breed and lactation, dominance accounted for 2–24% of the phenotypic variance for fat percentage in a study by Fuerst & Sölkner (1994). Likewise, dominance accounted for 8% of the phenotypic variation observed in stature of dairy cattle (Miształ, 1997). Estimates of dominance variance, however, were sensitive to adjustments for other factors and particularly to sample size (Miształ, 1997; Miształ *et al.*, 1997). The use of genome-wide marker information has recently opened new possibilities for partitioning genetic variance. Ertl *et al.* (2013) employed a genomic relationship approach (VanRaden, 2008; Vitezica *et al.*, 2013) and examined the variation of genomic breeding values and dominance deviations of milk production and conformation traits in Fleckvieh; significant influence of dominance was found on milk production traits. Genetic variation for milk metabolites, which can be considered novel milk component traits, were analysed in a methodologically related paper (Wittenburg *et al.*, 2013). Most of their genetic variability, however, was caused by additive genetic effects. In this study, which supplements Wittenburg *et al.* (2013), a genomic relationship approach was fitted to milk performance traits in Holstein Friesian cows, and the role of dominance variance captured by single nucleotide polymorphisms (SNPs) was explored. Genetic variation of other milk component traits such as acetone content and fatty acid levels were also evaluated, as they provide additional information on milk quality and health status of cows (e.g. ketosis, Geishauser *et al.*, 2000).

## Material and Methods

The dataset contained phenotypes of  $n = 1295$  Holstein Friesian cows and their genotypes at  $m = 37180$  SNPs. The cows were sampled between 21st and 120th day of their first lactation from May to November 2009. During the regular milk performance test, one milk sample of each cow was analysed via infrared spectroscopy (FOSS, Hillerød, Denmark). The following traits were measured: milk yield, fat, protein, lactose, urea, casein, pH value, acetone, somatic cell count (SCC), from which somatic cell score (SCS) was determined as  $SCS = \log_2(SCC / 100000) + 3$ , saturated (SFA) and unsaturated fatty acids (UFA). The infrared measurements of SFA and UFA were not standardised according to a correct calibration probe because probes were not available at the time of sample collection in 2009. Thus, these raw data are not comparable to other studies but fully sufficient for studying sources of phenotypic variation. Furthermore, energy content per kg milk determined as  $energy = 0.24 \times \text{protein content} + 0.39 \times \text{fat content} + 0.17 \times \text{lactose content}$  and energy-corrected milk (ECM),  $ECM = energy \times \text{milk yield} / 3.1$ , were investigated (Kirchgessner, 1997, p.307). More details on sample collection can be found in Wittenburg *et al.* (2013).

In total, 17 traits were examined in a series of univariate analyses. The following genomic model was fitted to each milk performance trait  $\mathbf{y} = (y_1, \dots, y_n)'$ ; bold symbols were used for matrices and vectors:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_a\mathbf{u}_a + \mathbf{Z}_d\mathbf{u}_d + \mathbf{e} .$$

The residuals were assumed to be independently and normally distributed,  $e_i \sim N(0, \sigma_e^2)$  for  $i = 1, \dots, n$ . The vector  $\mathbf{b}$  considers fixed effects of farm  $\times$  sampling date (63 levels) and linear

and quadratic regression on lactation day with corresponding entries in the design matrix  $\mathbf{X}$ . The design matrices  $\mathbf{Z}_a$  and  $\mathbf{Z}_d$  relate the observed milk traits to the breeding values  $\mathbf{u}_a$  and dominance deviations  $\mathbf{u}_d$ , respectively. As each individual was phenotyped and genotyped,  $\mathbf{Z}_a$  and  $\mathbf{Z}_d$  were identity matrices of dimension  $n$ . In order to identify the correlation structure, breeding values  $\mathbf{u}_a = \mathbf{Z}_a \mathbf{g}_a$  and dominance deviations  $\mathbf{u}_d = \mathbf{Z}_d \mathbf{g}_d$  can be written in terms of average effects of gene substitution  $\mathbf{g}_a$  and dominance effects  $\mathbf{g}_d$ , respectively. Let 1 and 2 denote the SNP alleles at locus  $j \in \{1, \dots, m\}$ , where 2 means the more frequent allele with frequency  $p_j$ . The  $(n \times m)$ -coefficient matrices  $\mathbf{M}_a$  and  $\mathbf{M}_d$  contain the entries

$$M_{a,i,j} = \begin{cases} -2p_j & \text{genotype 11} \\ 1-2p_j & \text{genotype 12} \\ 2(1-p_j) & \text{genotype 22} \end{cases} \quad \text{and} \quad M_{d,i,j} = \begin{cases} -2p_j^2 & \text{genotype 11} \\ 2p_j(1-p_j) & \text{genotype 12} \\ -2(1-p_j)^2 & \text{genotype 22} \end{cases}$$

for animal  $i$  at locus  $j$  according to Falconer's model in a random mating population in Hardy-Weinberg equilibrium (Falconer & Mackay, 1996, p.118). This model was also called G2A model by Zeng *et al.* (2005). The columns of  $\mathbf{M}_a$  and  $\mathbf{M}_d$  were additionally scaled,

$$\mathbf{M}_{a,j} \mapsto \frac{\mathbf{M}_{a,j}}{\sqrt{2p_j(1-p_j)}} \quad \text{and} \quad \mathbf{M}_{d,j} \mapsto \frac{\mathbf{M}_{d,j}}{2p_j(1-p_j)} .$$

From this parametrisation, the genomic relationship matrices can be deduced. Assuming that each locus contributes equally to the additive genetic variance, i.e.  $\text{Var}(g_{a,j}) = \frac{1}{m} \sigma_A^2 \quad \forall j$ , the covariance matrix of breeding values equals  $\mathbf{G}_a \sigma_A^2$ , where the additive relationship matrix  $\mathbf{G}_a := \frac{1}{m} \mathbf{M}_a \mathbf{M}_a'$  also corresponds to the genomic relationship matrix of VanRaden (2008, version 2). The dominance relationship matrix is set up analogously and leads to the covariance

matrix  $\text{Var}(\mathbf{u}_d) = \mathbf{G}_d \sigma_D^2$  similar to, for example, Vitezica *et al.* (2013) and Da *et al.* (2014). This way of model parametrisation in terms of breeding values and dominance deviations was recently investigated by Vitezica *et al.* (2013) who found that genetic variance components were estimated correctly and were comparable to classical parametrisations (e.g. Falconer & Mackay, 1996). Dominance effects were tested for significance, i.e.  $H_0: \sigma_D^2 = 0$  versus  $H_A: \sigma_D^2 > 0$ , via likelihood ratio test, for which the test statistic was assumed to approximately follow a mixture of  $\chi^2$  distributions with 0 and 1 degree of freedom (Self & Liang, 1987).

The analyses were carried out with R (R Development Core Team, 2011) and ASReml 3.0 (Gilmour *et al.*, 2008), which was also used to approximate the standard error (SE) of estimates of variance ratios via the Delta method.

## Results

Estimates of genetic and residual variance components are given in Table 1. As these absolute figures were not comparable among traits, the ratio of dominance to additive genetic variance was calculated. The highest ratio was observed for SCS ( $\sigma_D^2 : \sigma_A^2 = 1.36 \pm 2.54$ ), followed by protein content ( $\sigma_D^2 : \sigma_A^2 = 0.64 \pm 0.37$ ), casein content ( $\sigma_D^2 : \sigma_A^2 = 0.56 \pm 0.35$ ), energy ( $\sigma_D^2 : \sigma_A^2 = 0.51 \pm 0.56$ ) and fat content ( $\sigma_D^2 : \sigma_A^2 = 0.35 \pm 0.70$ ). Smaller values were observed for pH value and UFA, but SE was more than twofold the estimate of  $\sigma_D^2 : \sigma_A^2$ . The quotient for the remaining traits approached zero. Without accounting for multiplicity, a significant contribution of dominance variance was detected only for protein content ( $P = 0.03$ ) and casein content ( $P = 0.05$ ), but not for SCS ( $P = 0.27$ ). The latter was also indicated by high SE of estimating  $\sigma_D^2$ , see Table 1. After FDR correction, the maximum  $q$ -value of the top-two

traits was 0.42. Thus, less than one false positive result is expected among protein content and casein content.

[Table 1 can be included here.]

Narrow ( $h^2$ ) and broad sense heritability ( $H^2$ ) were estimated with the genomic model. For comparison, narrow sense heritability was also calculated based on an animal model, which included the same fixed effects and polygenic effects based on the pedigree relationship of 23 819 animals. The estimates are listed in Table 2. If marker-based additive genetic effects were considered, SE of estimating  $h^2$  was slightly but consequently smaller than in an animal model for all traits. Furthermore, for protein content, casein content and urea,  $h^2$  increased by 4–5 percentage points, and it more than doubled for lactose content (Table 2). For the other traits, estimates of  $h^2$  decreased by 2–7 percentage points with the genomic model. Broad sense heritability was observed at a high level for protein content ( $H^2 = 0.60$ ), casein content ( $H^2 = 0.58$ ), pH value ( $H^2 = 0.45$ ) and energy ( $H^2 = 0.36$ ), see Table 2. Model fit was briefly investigated by visually inspecting the distribution of studentised residuals. For example, Figure 1 presents the Q-Q plot for urea and SCS. The lack of model fit for SCS (a log2 transformation) is obvious, because residuals are slightly right-skewed. In most other cases, the Q-Q plot indicated an appropriate fit of the genomic model.

[Table 2 can be included here.]

[Figure 1 can be included here.]

## Discussion

This study verified the contribution of dominance variance captured by SNPs to the phenotypic variance of milk traits. Without accounting for multiplicity, dominance effects were

significantly present for protein content and casein content. Ertl *et al.* (2013) found a significant impact of dominance on the the variation of milk, fat and protein yield and SCS in Fleckvieh using genomic data, although they used a different experimental design. The kind of genetic variation likely depends on breed. An influence of dominance on yield traits was also expected in Holstein cows. If only pedigree information was considered, and traits were recorded over several lactations, then dominance variance was 15–17% of additive genetic variance for milk, fat and protein yield and somewhat smaller for SCS (Van Tassell *et al.*, 2000). Thus, as the underlying dataset consisted of one milk sample per cow, results only indicate what kind of genetic effect was most important for trait expression in the first third of lactation. Increasing records per cow could unravel dominance which might be more important in other stages of lactation. Furthermore, the present pedigree structure gave only little information for estimating dominance as there were hardly any full-sibs. Based on genomic data, the distribution of dominance relationship coefficients was right-skewed with almost zero mean value. Only four coefficients larger than 0.2 were observed; the inter-quartile range was 0.01.

Estimates of non-additive genetic variance components can be severely biased in small samples (Chang, 1988), and this might be true for the underlying dataset. The accuracy of genetic value prediction and of variance component estimates were investigated based on simulated data of similar size and structure as the presented study (Da *et al.*, 2014;  $n = 1654$ ,  $m = 40544$ ). Although dominance variance was estimated with bias in the different scenarios with varying heritability (the relative bias ranged from -50 to 25%), prediction of total genetic values (i.e. breeding value plus dominance deviation) was still more accurate than genomic prediction based on a purely additive model. The unbiasedness of predicted total genetic values was also found in real pig data (Su *et al.*, 2012;  $n = 1484$  in training set). Hence increasing sample size



will doubtlessly improve estimation of dominance variance but genomic prediction for breeding purposes already benefits from modelling non-additive effects in general.

In addition to milk traits, which are part of the regular performance test, levels of acetone, UFA and SFA were analysed. Acetone content had intermediate genetic impact ( $H^2 = 0.18$ ) which was due to additive genetic variance. Both environmental factors (e.g. farm effects) and genetics may explain the discrepant results of Klein *et al.* (2010), who studied the association between levels of acetone and ketosis. They found that some cows appeared to have a more stable metabolism than others, especially in early lactation, and those animals were less prone to ketosis. The fatty acids differed in their genetic background. SFA had intermediate genetic impact ( $H^2 = 0.19$ ) with no contribution of dominance, whereas UFA was little determined by genetics ( $H^2 = 0.07$ ) which was also caused by some dominance ( $\sigma_D^2 : \sigma_A^2 = 0.03$ ). Matching to this, Soyeurt *et al.* (2008) observed that (mono-) UFA was consistently less heritable than SFA at every stage of lactation when a test-day model was applied. Estimates of  $h^2$  were larger than in the presented study (SFA:  $0.26 \leq h^2 \leq 0.50$  depending on days in milk), but  $h^2$  of (mono-) UFA was about 0.10 in the first third of lactation. Fatty acid levels are important measures in milk processing. In particular, SFA:UFA is a main parameter for butter production, and the higher content of UFA at the beginning of lactation has positive impact on butter quality (Bobe *et al.*, 2003; Soyeurt *et al.*, 2007). A slight increase of SFA and decrease of UFA (pre-corrected for systematic effects) were also observed in this analysis (Fig. 2). Butter made from milk produced in this lactation interval is therefore likely to be less spreadable (harder) due to the higher SFA:UFA.

[Figure 2 can be included here.]

## Conclusions

Using a genomic relationship approach, it was possible to study the contribution of additive and dominance variance to the variance of different milk traits in the first third of lactation. Dominance variance relative to additive genetic variance was highest for SCS, protein content and casein content. A significant contribution of dominance variance, however, was only observed for protein content and casein content; similar genetic variance components were estimated for both traits. Yield traits did not show any dominance variance. The milk component traits acetone content, UFA and SFA were moderately heritable, but these traits were almost unaffected by dominance. Those traits with non-vanishing dominance variance may be considered in suitable mating programmes to maximise the overall genetic merit for commercial production (Sun *et al.*, 2013). Due to the small sample size, the results of this study only indicate the presence or absence of dominance for dairy traits and need to be validated based on larger datasets.

## Acknowledgements

This study was part of the FUGATO project “Bovine Integrative Bioinformatics for Genomic Selection (BovIBI)” with financial support of the German Federal Ministry of Education and Research (BMBF). The authors thank the participating herd owners and cooperation partners: Dr. S. Jakubowski, Dr. U. Kesting and S. Wolf (LKV, Güstrow) for providing milk performance data and their contribution to discussions during the project, Dr. H. Göft (FOSS, Hillerød, Denmark) for delivering fatty acid levels, F. Reinhardt (vit, Verden) for assembling the pedigree data and R. Grahl (Leibniz Institute for Farm Animal Biology, Dummerstorf) for blood sample collection.

## References

- Bobe, G., Hammond, E. G., Freeman, A. E., Lindberg, G. L. & Beitz, D. C. (2003) Texture of butter from cows with different milk fatty acid compositions. *J. Dairy Sci.*, **86**(10), 3122–3127.
- Chang, H.-L. A. (1988) *Studies on estimation of genetic variances under non-additive gene action*. Ph.D. thesis, University of Illinois at Urbana-Champaign.
- Da, Y., Wang, C., Wang, S. & Hu, G. (2014) Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS ONE*, **9**(1), e87666.
- Ertl, J., Legarra, A., Vitezica, Z. G., Varona, L., Edel, C., Emmerling, R. & Götz, K.-U. (2013) Genomic analysis of dominance effects in milk production and conformation traits of Fleckvieh cattle. *Interbull Bull.*, **47**, 28–31.
- Falconer, D. S. & Mackay, T. F. C. (1996) *Quantitative Genetics*. Longman, Essex, England.
- Fuerst, C. & Sölkner, J. (1994) Additive and nonadditive genetic variances for milk yield, fertility, and lifetime performance traits of dairy cattle. *J. Dairy Sci.*, **77**(4), 1114 – 1125.
- Geishauser, T., Leslie, K., Tenhag, J. & Bashiri, A. (2000) Evaluation of eight cow-side ketone tests in milk for detection of subclinical ketosis in dairy cows. *J. Dairy Sci.*, **83**(2), 296–299.
- Gilmour, A., Gogel, B., Cullis, B. & Thompson, R. (2008) *ASReml User Guide Release 3.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, URL [www.vsnl.co.uk](http://www.vsnl.co.uk).
- Kirchgessner, M. (1997) *Tierernährung*. DLG Verlag.
- Klein, M. S., Almstetter, M. F., Schlamberger, G., Nürnberger, N., Dettmer, K., Oefner, P. J., Meyer, H. H. D., Wiedemann, S. & Gronwald, W. (2010) Nuclear magnetic resonance and mass spectrometry-based milk metabolomics in dairy cows during early and late lactation. *J. Dairy Sci.*, **93**(4), 1539–1550.
- Misztal, I. (1997) Estimation of variance components with large-scale dominance models. *J. Dairy Sci.*, **80**(5), 965–974.
- Misztal, I., Lawlor, T. J. & Fernando, R. L. (1997) Dominance models with method R for stature of Holsteins. *J. Dairy Sci.*, **80**(5), 975–978.

- R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL <http://www.R-project.org/>, ISBN 3-900051-07-0.
- Self, S. G. & Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, **82**, 605–610.
- Soyeurt, H., Dardenne, P., Dehareng, F., Bastin, C. & Gengler, N. (2008) Genetic parameters of saturated and monounsaturated fatty acid content and the ratio of saturated to unsaturated fatty acids in bovine milk. *J. Dairy Sci.*, **91**(9), 3611–3626.
- Soyeurt, H., Dehareng, F., Bertozzi, C. & Gengler, N. (2007) Genetic parameters of butter hardness estimated by test-day model. *Interbull Bull.*, **37**, 49–53.
- Su, G., Christensen, O. F., Ostersen, T., Henryon, M. & Lund, M. S. (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS ONE*, **7**(9), e45293.
- Sun, C., VanRaden, P., O’Connell, J., Weigel, K. & Gianola, D. (2013) Mating programs including genomic relationships and dominance effects. *J. Dairy Sci.*, **96**(12), 8014 – 8023.
- Van Tassell, C. P., Misztal, I. & Varona, L. (2000) Method R estimates of additive genetic, dominance genetic, and permanent environmental fraction of variance for yield and health traits of Holsteins. *J. Dairy Sci.*, **83**(8), 1873–1877.
- VanRaden, P. M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**(11), 4414–4423.
- Vitezica, Z. G., Varona, L. & Legarra, A. (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, **195**(4), 1223–1230.
- Wittenburg, D., Melzer, N., Willmitzer, L., Lisek, J., Kesting, U., Reinsch, N. & Repsilber, D. (2013) Milk metabolites and their genetic variability. *J. Dairy Sci.*, **96**, 2557–2569.
- Zeng, Z.-B., Wang, T. & Zou, W. (2005) Modeling quantitative trait loci and interpretation of models. *Genetics*, **169**(3), 1711–1725.

**FIGURES**

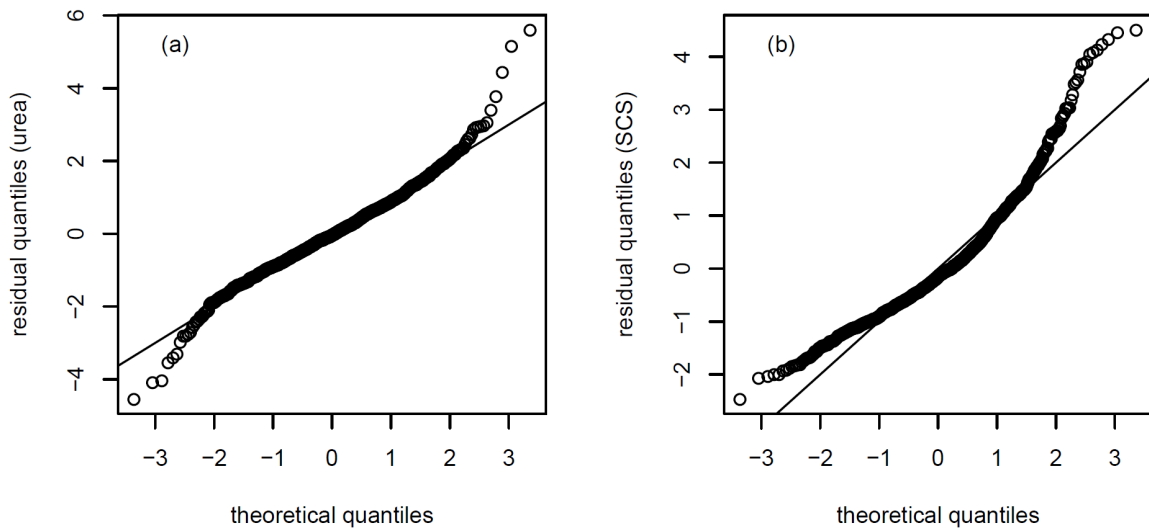


Figure 1: Q-Q plot of studentised residuals: (a) urea and (b) SCS.

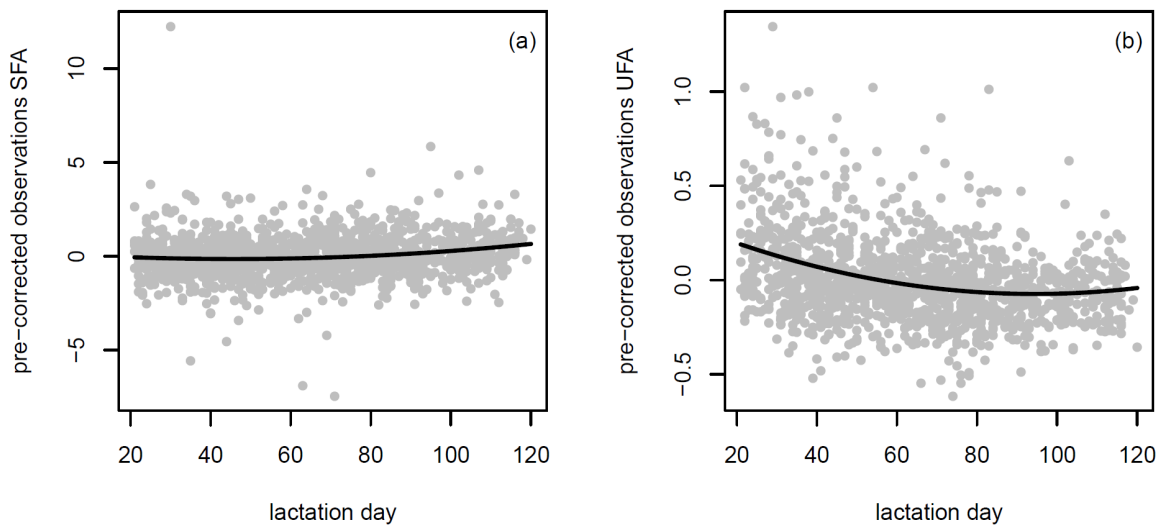


Figure 2: Correlation between lactation day and pre-corrected levels of (a) SFA and (b) UFA.

Pre-correction was obtained using a linear mixed model including fixed effects of farm×sampling date and a random sire effect. A potential quadratic regression curve was added (solid line).

## TABLES

Table 1: Estimated variance components (with standard error in brackets) and  $P$ -value for testing the presence of dominance captured by SNPs.

Trait	$\sigma_A^2$	$\sigma_D^2$	$\sigma_e^2$	$P$
milk kg	2.81 (1.22)	0.00 (0.00)	25.78 (1.48)	1.00
fat kg×10	0.51 (0.21)	0.00 (0.00)	4.82 (0.27)	1.00
fat content kg×10/kg milk	5.58 (1.66)	1.95 (3.84)	22.79 (4.06)	0.31
protein kg×10	0.22 (0.10)	0.00 (0.00)	2.19 (0.12)	1.00
protein content kg×10/kg milk	1.83 (0.35)	1.16 (0.62)	2.03 (0.63)	0.03
lactose kg×10	0.69 (0.30)	0.00 (0.00)	6.39 (0.37)	1.00
lactose content kg×10/kg milk	0.58 (0.13)	0.00 (0.00)	1.29 (0.11)	1.00
casein kg×10	0.15 (0.06)	0.00 (0.00)	1.37 (0.08)	1.00
casein content kg×10/kg milk	1.44 (0.28)	0.80 (0.48)	1.64 (0.49)	0.05
ECM kg	2.30 (1.00)	0.00 (0.00)	23.86 (1.29)	1.00
energy MJ NEL×10/kg milk	1.31 (0.33)	0.67 (0.70)	3.52 (0.73)	0.17
urea mg/(L×10)	3.39 (1.05)	0.00 (0.00)	15.59 (1.06)	1.00
pH value×10	0.12 (0.02)	0.02 (0.04)	0.17 (0.04)	0.33
SCS	0.15 (0.10)	0.20 (0.34)	2.33 (0.35)	0.27
acetone content kg×10/kg milk	0.66 (0.19)	0.00 (0.00)	2.93 (0.19)	1.00
UFA×10	0.35 (0.20)	0.01 (0.63)	4.55 (0.66)	1.00
SFA	0.28 (0.09)	0.00 (0.00)	1.22 (0.09)	1.00

$\sigma_A^2$  additive genetic variance;  $\sigma_D^2$  dominance variance;  $\sigma_e^2$  residual variance.

Table 2: Estimated heritability (with standard error in brackets).

Trait	$h_*^2$	$h^2$	$H^2$
milk kg	0.13 (0.06)	0.10 (0.04)	0.10 (0.04)
fat kg	0.12 (0.05)	0.10 (0.04)	0.10 (0.04)
fat content	0.23 (0.08)	0.18 (0.05)	0.25 (0.13)
protein kg	0.12 (0.06)	0.09 (0.04)	0.09 (0.04)
protein content	0.31 (0.10)	0.36 (0.06)	0.60 (0.13)
lactose kg	0.13 (0.06)	0.10 (0.04)	0.10 (0.04)
lactose content	0.13 (0.08)	0.31 (0.06)	0.31 (0.06)
casein kg	0.12 (0.06)	0.10 (0.04)	0.10 (0.04)
casein content	0.32 (0.10)	0.37 (0.06)	0.58 (0.13)
ECM kg	0.11 (0.05)	0.09 (0.04)	0.09 (0.04)
energy MJ NEL/kg	0.28 (0.09)	0.24 (0.06)	0.36 (0.14)
urea mg/L	0.13 (0.06)	0.18 (0.05)	0.18 (0.05)
pH value	0.44 (0.10)	0.40 (0.06)	0.45 (0.13)
SCS	0.10 (0.06)	0.05 (0.04)	0.13 (0.13)
acetone content	0.25 (0.08)	0.18 (0.05)	0.18 (0.05)
UFA	0.10 (0.06)	0.07 (0.04)	0.07 (0.13)
SFA	0.22 (0.08)	0.19 (0.06)	0.19 (0.06)

Narrow sense heritability from animal model ( $h_*^2$ ) and genomic model ( $h^2$ ); broad sense heritability from genomic model ( $H^2$ ).





## Chapter 3

# Publications on the dependence between markers

- 3.1 Wittenburg, D., Teuscher, F., Klosa, J., & Reinsch, N. (2016). Covariance between genotypic effects and its use for genomic inference in half-sib families. *G3 Genes Genom. Genet.*, *6*, 2761–2772. <https://doi.org/10.1534/g3.116.032409>
- 3.2 Wittenburg, D., Bonk, S., Doschoris, M., & Reyer, H. (2020). Design of experiments for fine-mapping quantitative trait loci in livestock populations. *BMC Genet.*, *21*, 66. <https://doi.org/10.1186/s12863-020-00871-1>
- 3.3 Wittenburg, D., Doschoris, M., & Klosa, J. (2021). Grouping of genomic markers in populations with family structure. *BMC Bioinf.*, *22*, 79. <https://doi.org/10.1186/s12859-021-04010-0>
- 3.4 Hampel, A., Teuscher, F., Gomez-Raya, L., Doschoris, M., & Wittenburg, D. (2018). Estimation of recombination rate and maternal linkage disequilibrium in half-sibs. *Front. Genet.*, *9*, 186. <https://doi.org/10.3389/fgene.2018.00186>
- 3.5 Qanbari, S., & Wittenburg, D. (2020). Male recombination map of the autosomal genome in German Holstein. *Genet. Sel. Evol.*, *52*, 73. <https://doi.org/10.1186/s12711-020-00593-z>



## Covariance Between Genotypic Effects and its Use for Genomic Inference in Half-Sib Families

Dörte Wittenburg,<sup>1</sup> Friedrich Teuscher, Jan Klosa, and Norbert Reinsch

Leibniz Institute for Farm Animal Biology, Institute of Genetics and Biometry, 18196 Dummerstorf, Germany

**ABSTRACT** In livestock, current statistical approaches utilize extensive molecular data, e.g., single nucleotide polymorphisms (SNPs), to improve the genetic evaluation of individuals. The number of model parameters increases with the number of SNPs, so the multicollinearity between covariates can affect the results obtained using whole genome regression methods. In this study, dependencies between SNPs due to linkage and linkage disequilibrium among the chromosome segments were explicitly considered in methods used to estimate the effects of SNPs. The population structure affects the extent of such dependencies, so the covariance among SNP genotypes was derived for half-sib families, which are typical in livestock populations. Conditional on the SNP haplotypes of the common parent (sire), the theoretical covariance was determined using the haplotype frequencies of the population from which the individual parent (dam) was derived. The resulting covariance matrix was included in a statistical model for a trait of interest, and this covariance matrix was then used to specify prior assumptions for SNP effects in a Bayesian framework. The approach was applied to one family in simulated scenarios (few and many quantitative trait loci) and using semireal data obtained from dairy cattle to identify genome segments that affect performance traits, as well as to investigate the impact on predictive ability. Compared with a method that does not explicitly consider any of the relationship among predictor variables, the accuracy of genetic value prediction was improved by 10–22%. The results show that the inclusion of dependence is particularly important for genomic inference based on small sample sizes.

### KEYWORDS

autoregressive  
prior  
Bayesian  
statistics  
linkage  
disequilibrium  
recombination  
rate  
SNP effect

In whole-genome regression analyses, it is often the case that the number of genomic markers,  $p$ , exceeds that of the observations,  $n$ . Moreover, linkage and linkage disequilibrium (LD) between loci adds a second source of dependency among the predictors. The number of genomic markers such as single nucleotide polymorphisms (SNPs) is still growing, e.g., ~26 million SNPs have been identified in the whole-genome sequences of cattle (Daetwyler *et al.* 2014). Genomic prediction works reasonably well when based on a huge number of explanatory variables (e.g., Gianola 2013), but the high dependency among predictors, which

is often called multicollinearity, may lead to the incorrect genomic inference of marker effects because the standard error of the estimated effects is likely to be high.

For “ $p > n$ ” problems, various methods are available that implicitly consider dependencies by selecting relevant predictors and/or shrinking the effect sizes (for a thorough review, see de los Campos *et al.* 2013). Bayesian (e.g., Meuwissen *et al.* 2001; Habier *et al.* 2011) and penalized (e.g., Gianola *et al.* 2006; Piepho 2009) methods are the most common choices for genomic prediction. However, explicitly exploiting dependencies, especially those due to the proximity of SNPs, relies on the appropriate order of loci. Using the order of SNPs, and clustering them according to their adjacency combined with the Group Lasso method can obtain better performance than other penalized approaches, where clusters that contain causal variants may be identified with more confidence (Dehman *et al.* 2015). Alternatively, haplotype-based approaches (e.g., Calus *et al.* 2008; Cuyabano *et al.* 2014) exploit the connections between SNPs, which may improve the accuracy of genetic value prediction. Associations throughout the genome can also be modeled using a first-order antedependence correlation structure (Yang and Tempelman 2012). A penalty term placed on successive differences between

Copyright © 2016 Wittenburg *et al.*

doi: 10.1534/g3.116.032409

Manuscript received March 19, 2016; accepted for publication June 22, 2016; published Early Online July 7, 2016.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental material is available online at [www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.032409/-/DC1](http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.032409/-/DC1)

<sup>1</sup>Corresponding author: Leibniz Institute for Farm Animal Biology, Institute of Genetics and Biometry, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany. E-mail [wittenburg@fbn-dummerstorf.de](mailto:wittenburg@fbn-dummerstorf.de)

the coefficients is employed to consider the natural order of effects, thereby yielding estimates with smooth transitions. In addition, sparsity in terms of nonzero effect estimates can be achieved by penalizing the  $L_1$  norm of differences in the fused lasso approach (Tibshirani *et al.* 2005). In a Bayesian smoothing framework, random-walk priors were suggested by Fahrmeir and Lang (2001), which also allow nonequal distances between predictors. In particular, for genetic applications, Gianola *et al.* (2003) described an autoregressive process with variable distances between markers in a mixed-model formulation. Incorporating the dependencies among SNPs can improve the outcomes of genomic evaluations (Yang and Tempelman 2012), but the assumed pattern of the covariance structure has been based on general assumptions in previous studies.

The objective of the present study was to extend the Bayesian approach by using an autoregressive prior for marker effects, and to determine the autocorrelation parameter explicitly according to genetic theory. The population structure influences the extent of associations. Family stratification leads to different levels of LD among families, and may result in a biased population-wide measure of LD. Thus, dependencies caused by LD were theoretically derived for a single half-sib family that is part of a typical livestock population (*e.g.*, dairy cattle). This required the haplotypes and recombination rates of the common parent (sire) as well as the LD of gametes in the population from which the individual parent (dam) was derived. The obtained covariance structure was then integrated into a statistical model for genomic evaluation. Genome segments with significant impacts on a quantitative trait were inferred, and the precision of the parameter estimates as well as the accuracy of genomic prediction were evaluated. This report ends with a discussion of partial successes, drawbacks, and further options for considering dependencies.

## MATERIALS AND METHODS

### Statistical model

To study the genetic effects captured by SNPs, each with two alleles, A and B, the following whole genome regression model is fitted to a trait  $y = (y_1, \dots, y_n)'$ ,

$$y = \mathbf{1}\mu + X\mathbf{m} + \mathbf{e}. \quad (1)$$

The design matrix  $X = \{X_{ij}\}_{i,j}$  contains the genotype codes at locus  $j \in \{1, \dots, p\}$  for individual  $i \in \{1, \dots, n\}$ , where 1 and -1 indicate homozygous genotypes AA and BB, respectively, and the heterozygote is coded as 0. The vector  $\mathbf{m} = (m_1, \dots, m_p)'$  contains the additive genotype effect at each SNP (*i.e.*, half the difference between homozygotes). Furthermore,  $\mu$  denotes the overall mean, and  $\mathbf{1}$  is a vector of  $n$  ones. The residuals are assumed to be independent and normally distributed,  $e_i \sim N(0, \sigma_e^2)$  for  $i = 1, \dots, n$ . For convenience, other effects are omitted.

### Covariance between SNP genotypes

In this section, model (1) is extended to consider the dependencies among predictor variables in  $X$  due to linkage and LD specifically for a paternal half-sib family design. The covariance between genotype codes can be derived theoretically for each pair of SNPs. The derivation is based on Bonk *et al.* (2016), who deduced the covariance between SNP genotypes coding for additive and/or dominance effects in a general mating, where the haplotypes of both parents are available. In the main result for the additive effects of SNP alleles, the covariance can be split into paternal ( $s$ ) and maternal ( $d$ ) contributions because the alleles are

inherited independently. In extension to Bonk *et al.* (2016), the maternal contribution must be generalized. It is assumed that a dam is drawn randomly from the population. Half-sibs have a common sire, so the covariance between loci  $j$  and  $k$  is determined according to the sire's diplototype  $\mathcal{S}$ ,

$$\begin{aligned} K_{j,k} &:= \text{cov}(X_{i,j}, X_{i,k} | \mathcal{S}) \\ &= \text{cov}(X_{i,j,s}, X_{i,k,s} | \mathcal{S}) + \text{cov}(X_{i,j,d}, X_{i,k,d}), \end{aligned} \quad (2)$$

where  $X_{i,j,s}$  and  $X_{i,j,d}$  take a value of  $\frac{1}{2}$  if the A allele was inherited, but  $-\frac{1}{2}$  otherwise, and  $X_{i,j} = X_{i,j,s} + X_{i,j,d}$ .

To determine the paternal contribution to the covariance in Equation (2), three different types of sire diplotypes are distinguished (the results were taken from Bonk *et al.* 2016).

1. Double homozygous sire (haplotypes AA and AA): all half-sibs inherit the same paternal haplotype AA, and the paternal covariance is zero.
2. The sire is heterozygous at one locus (haplotypes AA and AB): two haplotypes AA and AB can be observed among half-sibs, where each is equally frequent. The paternal covariance is also zero.
3. Double heterozygous sire (haplotypes AA and BB or AB and BA): all possible haplotypes appear among daughters with a probability that depends on the recombination rate  $\theta_{j,k}$ . Thus,  $\text{cov}(X_{i,j,s}, X_{i,k,s} | AA/BB) = \frac{1}{4}(1 - 2\theta_{j,k})$  or  $\text{cov}(X_{i,j,s}, X_{i,k,s} | AB/BA) = -\frac{1}{4}(1 - 2\theta_{j,k})$ . If  $j = k$ , then  $\theta_{j,k} = 0$ .

Using the allele frequency,  $p_j^A$ , and haplotype frequencies of the maternal gametes,  $p_{j,k}^{XY}$ , the second part of Equation (2) is:

$$\text{cov}(X_{i,j,d}, X_{i,k,d}) = E(X_{i,j,d}X_{i,k,d}) - E(X_{i,j,d})E(X_{i,k,d}),$$

$$E(X_{i,j,d}X_{i,k,d}) = p_{j,k}^{AA}\frac{1}{4} - p_{j,k}^{AB}\frac{1}{4} - p_{j,k}^{BA}\frac{1}{4} + p_{j,k}^{BB}\frac{1}{4},$$

$$E(X_{i,j,d}) = p_j^A - \frac{1}{2},$$

$$E(X_{i,k,d}) = p_k^A - \frac{1}{2}.$$

Combining the terms yields  $\text{cov}(X_{i,j,d}, X_{i,k,d}) = p_{j,k}^{AA}p_{j,k}^{BB} - p_{j,k}^{AB}p_{j,k}^{BA} = D_{j,k}$ , the LD of maternal gametes. If  $j = k$ , then  $\text{cov}(X_{i,j,d}, X_{i,j,d}) = \text{var}(X_{i,j,d}) = p_j^A(1 - p_j^A)$ .

In summary, the covariance between SNP genotypes among half-sibs can be split into a linkage part contributed by the sire and an LD part added by the mother. The covariance matrix  $\mathbf{K} = \{K_{j,k}\}_{j,k=1}^p$  is set up with the following elements:

$$K_{j,k} = \begin{cases} D_{j,k} + \frac{1}{4}(1 - 2\theta_{j,k}), & \text{for sire with haplotypes AA and BB} \\ D_{j,k} - \frac{1}{4}(1 - 2\theta_{j,k}), & \text{for sire with haplotypes AB and BA} \\ D_{j,k}, & \text{else.} \end{cases}$$

The linkage phase of sire, the corresponding recombination rate ( $\theta_{j,k}$ ), and the LD ( $D_{j,k}$ ) of maternal gametes are assumed to be known. If they are not available, these population parameters may be estimated, such as using the maximum likelihood approach (Gomez-Raya *et al.* 2013). Now it is necessary to determine whether genomic evaluations can be improved by knowing this covariance structure.

**Specification of prior assumptions**

To estimate the unknown parameters of model (1), a Bayesian shrinkage approach is employed. The matrix  $\mathbf{K}$  can be incorporated as a scale matrix during the specification of the prior assumptions. Thus, a hierarchical structure is defined for model (1):

$$\mathbf{y}|\boldsymbol{\mu}, \mathbf{m}, \sigma_e^2 \sim N(\mathbf{1}\boldsymbol{\mu} + \mathbf{X}\mathbf{m}, I\sigma_e^2),$$

$$\mathbf{m}|\boldsymbol{\Psi} \sim N(\mathbf{0}, \boldsymbol{\Psi}),$$

$$\boldsymbol{\mu} \propto \text{constant},$$

$$\sigma_e^2 \sim \chi^{-2}(-2, 0),$$

where  $\chi^{-2}(\nu, S)$  denotes the inverse  $\chi^2$  distribution with  $\nu$  degrees of freedom and scaling parameter  $S$ . Let  $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\mathbf{K})$  be a covariance matrix that depends on the elements of  $\mathbf{K}$ .

The SNP effects can be estimated as the mean of their posterior distribution, which can be obtained by Gibbs sampling from the conditional distribution (e.g., Sorensen and Gianola 2002)

$$\mathbf{m}|\mathbf{y}, \boldsymbol{\Psi}, \boldsymbol{\mu}, \sigma_e^2 \propto N\left(\boldsymbol{\Sigma}\mathbf{X}'(\mathbf{y} - \mathbf{1}\boldsymbol{\mu}), \boldsymbol{\Sigma}\sigma_e^2\right) \quad \text{with} \quad (3)$$

$$\boldsymbol{\Sigma} = \left(\mathbf{X}'\mathbf{X} + \boldsymbol{\Psi}^{-1}\sigma_e^2\right)^{-1}.$$

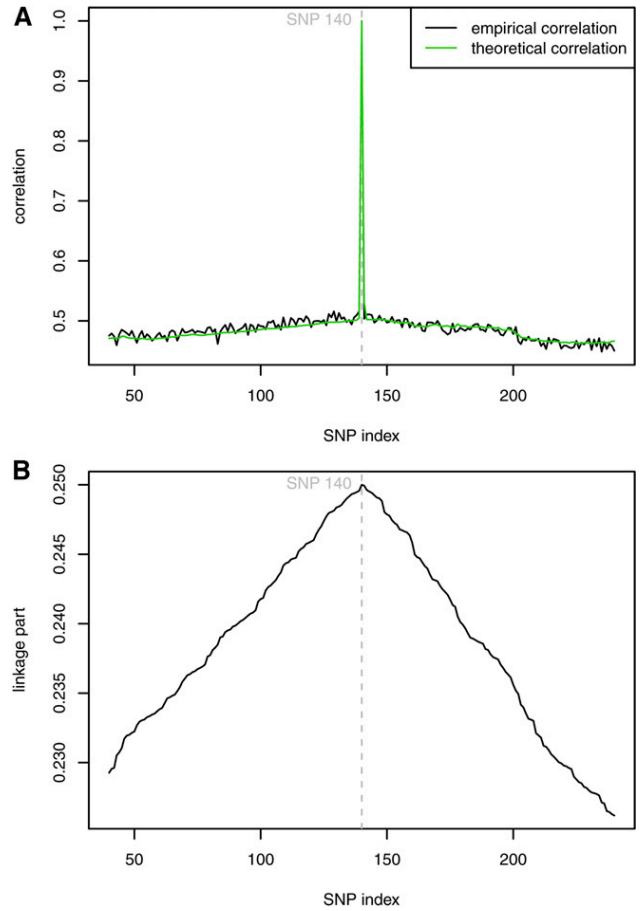
Several specifications of  $\boldsymbol{\Psi}$  are suitable, which differ in terms of the dynamics of their regularization parameters, as follows.

- (P1) Uncorrelated prior  $\boldsymbol{\Psi} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ . An inverse prior is stipulated for the shrinkage parameters  $\sigma_j^2$ , i.e.,  $p(\sigma_j^2) \propto \frac{1}{\sigma_j^2}$ ,  $j = 1, \dots, p$ . Hence, the posterior distribution is  $\sigma_j^2|\text{else} \propto \chi^{-2}(1, m_j^2)$ . This approach is similar to that proposed by Xu (2003), which represents a baseline model.
- (P2) Correlated prior  $\boldsymbol{\Psi} = \mathbf{K}\sigma^2$ . A flat prior is used for the variance component, i.e.,  $p(\sigma^2) \propto \chi^{-2}(-2, 0)$ , which is similar to that proposed by Wang *et al.* (1994). This parameter is distributed *a posteriori* as  $\sigma^2|\text{else} \sim \chi^{-2}(p - 2, \mathbf{m}'\mathbf{K}^{-1}\mathbf{m})$ .
- (P3) Correlated and adaptive prior  $\boldsymbol{\Psi} = \boldsymbol{\Gamma}\mathbf{K}\boldsymbol{\Gamma}$  with  $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \dots, \gamma_p)$ . The prior of the regularization parameters is specified as  $p(\gamma_j) \propto \chi_1^{-2}$ ,  $j = 1, \dots, p$ . Therefore, the pattern of the posterior density  $p(\gamma_j|\text{else})$  is unknown, but it is possible to employ a hybrid strategy, where a Metropolis step replaces the Gibbs step if sampling a parameter is impossible (Tierney 1994). For each  $\gamma$ , a single Metropolis iteration is used. A log-normal density is applied to the  $\gamma$ s as the proposal density  $q(\gamma_j^*|\gamma_j^{(t-1)})$  to obtain a random walk chain. By single-site updating during each iteration  $t$ , the  $\gamma$ s are individually drawn from the proposal distribution:

$$x \sim N\left(\ln\gamma_j^{(t-1)}, \varepsilon\right),$$

$$\gamma_j^* = \exp(x).$$

The tuning parameter  $\varepsilon$  is set to one. (The choice of  $\varepsilon$  is discussed below.) Let  $\boldsymbol{\tau}^*$  denote the vector of  $\tau_k = 1/\gamma_k$ ,  $k = 1, \dots, p$ , where the  $j$ th component is replaced by the current proposed value, and  $\boldsymbol{\tau}^{(t-1)}$  refers to the vector of the samples obtained from the last iteration. The ratio  $R$  is obtained as (see Appendix):



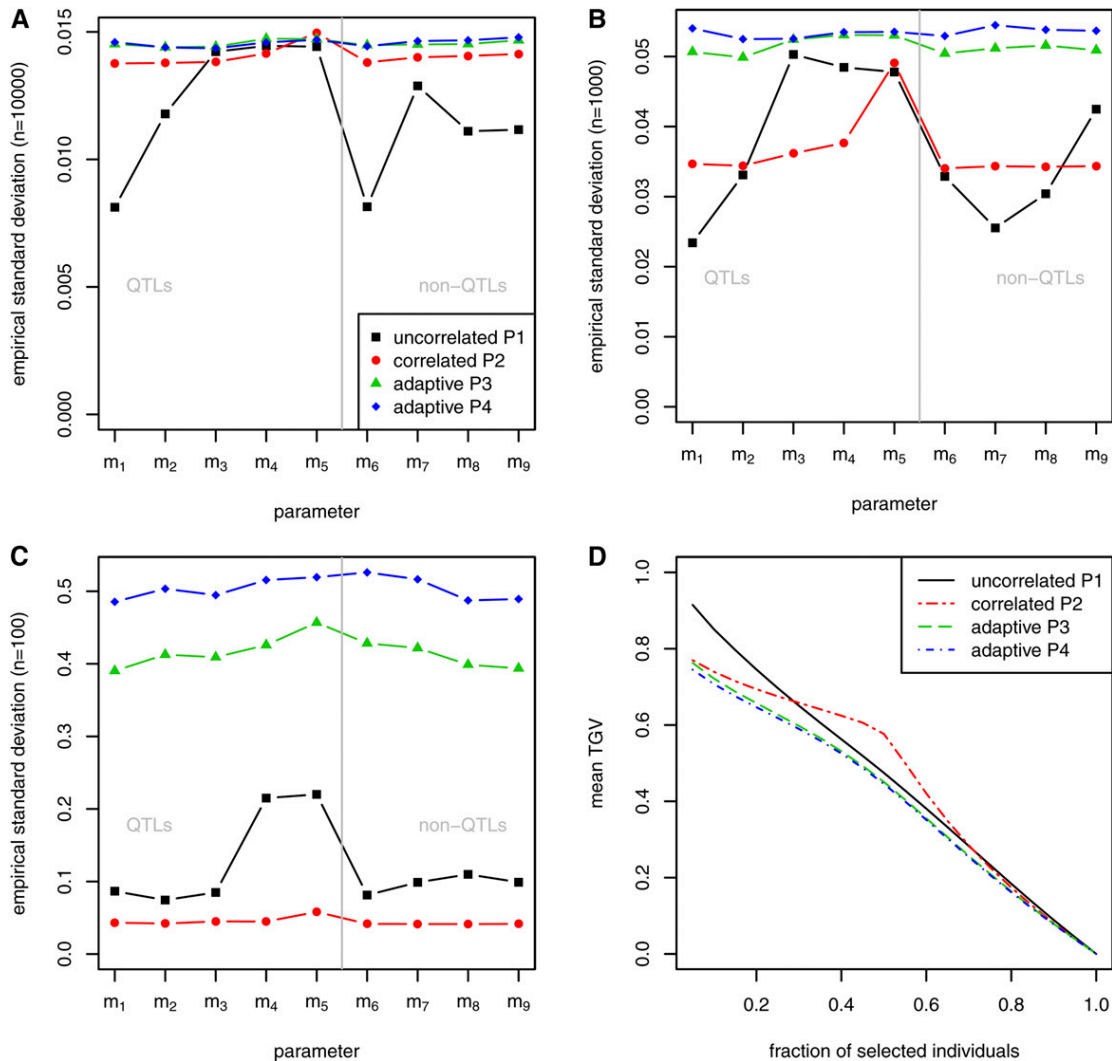
**Figure 1** (A) Theoretical vs. empirical correlation for a randomly selected SNP based on the simulated genotypes. (B) Contribution of linkage (paternal part) to the covariance between SNPs.

$$R = \frac{p\left(\gamma_j^*|\gamma_j^{(t-1)}, \mathbf{m}^{(t-1)}, \mathbf{y}\right) q\left(\gamma_j^{(t-1)}|\gamma_j^*\right)}{q\left(\gamma_j^*|\gamma_j^{(t-1)}\right) p\left(\gamma_j^{(t-1)}|\gamma_j^{(t-1)}, \mathbf{m}^{(t-1)}, \mathbf{y}\right)}$$

$$= \frac{\exp\left(-\frac{1}{2}\boldsymbol{\tau}^{*\prime}\tilde{\mathbf{K}}\boldsymbol{\tau}^* + \frac{3}{2}\ln\tau_j^* - \frac{\tau_j^*}{2}\right)}{\exp\left(-\frac{1}{2}\boldsymbol{\tau}^{(t-1)\prime}\tilde{\mathbf{K}}\boldsymbol{\tau}^{(t-1)} + \frac{3}{2}\ln\tau_j^{(t-1)} - \frac{\tau_j^{(t-1)}}{2}\right)},$$

with  $\tilde{\mathbf{K}} = \mathbf{K}^{-1}\#\mathbf{m}\mathbf{m}'$  using the Hadamard product ( $\#$ ). The acceptance ratio is then determined as  $\alpha = \min(R, 1)$ . The proposed value is accepted,  $\gamma_j^{(t)} = \gamma_j^*$ , if a random sample from a uniform distribution is lower than  $\alpha$ ; otherwise,  $\gamma_j^{(t)} = \gamma_j^{(t-1)}$ . The idea is related to the weighting of variables: with values starting at one, the proposed values of  $\gamma$  move slowly away from the initial estimate toward zero if there is evidence of a null effect, or they increase for nonzero SNP effects. Finally, after  $j = 1, \dots, p$  Metropolis steps, the vector  $\mathbf{m}^{(t)}$  is sampled from the conditional distribution (3).

- (P4) Correlated and adaptive prior  $\boldsymbol{\Psi}^{-1} = \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}'$  with  $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \dots, \gamma_p)$  and  $\mathbf{L}\mathbf{L}' = \mathbf{K}^{-1}$ . The prior of the shrinkage parameters is assumed to be  $p(\gamma_j) \propto \chi_1^2$  for  $j = 1, \dots, p$ . The posterior density is then recognized as the kernel of a scaled



**Figure 2** Simulation with five QTL. SD of estimated effects at key SNPs for different sample sizes based on one MCMC run:  $n = 10,000$  (A),  $n = 1000$  (B),  $n = 100$  (C). (D) Mean of TGV of individuals that were selected by their EGV based on 100-fold cross-validation (size of training set  $n = 100$ ).

$\chi^2$  distribution with  $\nu = 2$  and  $S = (\tilde{m}_j^2 + 1)^{-1}$ , where  $\tilde{\mathbf{m}} = \mathbf{L}'\mathbf{m}$  (see Appendix).

If covariances between SNPs are not considered, *i.e.*,  $\mathbf{K} = \mathbf{I}$ , the prior P2 is similar to a ridge-regression-type of model, while P3 and P4 are similar to, *e.g.*, BayesA-type of model (Meuwissen *et al.* 2001). In P3 and P4, two different types of modified Cholesky decompositions of the scale matrix  $\mathbf{K}$  are incorporated (Pourahmadi 2007). For P3, where the correlations between SNPs are known and stationary, the covariances between SNPs are affected by the  $\gamma$ s which are sampled during Markov chain Monte Carlo (MCMC) simulations. Shrinkage based on P4 is related to differences in the effects of adjacent markers because the  $\gamma$ s influence the entries of the lower triangular matrix  $\mathbf{L}$ .

Furthermore, the overall mean and residual variance component are distributed *a posteriori* as

$$\mu|\text{else} \sim N\left(\frac{1}{n}\mathbf{1}'(\mathbf{y} - \mathbf{X}\mathbf{m}), \frac{\sigma_e^2}{n}\right),$$

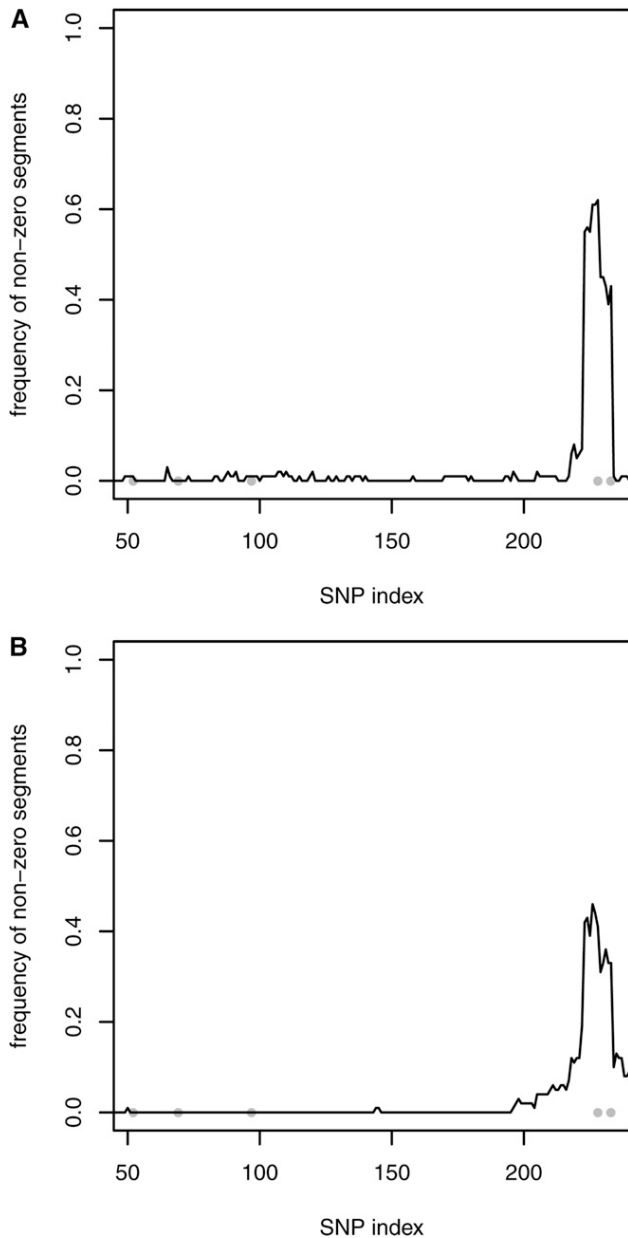
$$\sigma_e^2|\text{else} \sim \chi^{-2}(n-2, \mathbf{e}'\mathbf{e}) \text{ with residuals } \mathbf{e} = \mathbf{y} - \mathbf{1}\mu - \mathbf{X}\mathbf{m}.$$

The columns of  $\mathbf{X}$  are centered to favor mixing of the MCMC algorithm (Stranden and Christensen 2011).

### Criteria for evaluation

The accuracy of genetic value prediction was verified by cross-validation. As a measure of accuracy, the mean correlation was calculated between the estimated and simulated genetic values in test data sets,  $\rho = \text{cor}(\mathbf{X}\hat{\mathbf{m}}, \mathbf{X}\mathbf{m})$ . In addition, to compare the predictive ability of the chosen priors, the mean true genetic value (TGV) was calculated for individuals, which were selected based on their estimated genetic value (EGV) according to a given fraction of selection ( $r$ ). A range of  $r$  values between 0.05 and 1 was considered for evaluation. It is expected that the different methods would yield different rankings for the EGV. In general, the maximum mean TGV is obtained when only a few individuals are selected as parents of future offspring, and it approaches zero if more individuals are selected. Thus, the performance of a method can be determined by comparing the relationship between the mean TGV and  $r$ .

The precision of the estimates of the residual and genetic variance component was determined as the mean squared error (MSE) based on repeated simulations:



**Figure 3** Simulation with five QTL,  $n = 100$ , and 100 repetitions. Detection of nonzero segment effects using the uncorrelated prior P1 (A) and correlated prior P2 (B). Gray dots indicate the simulated QTL positions.

$$\text{MSE}(\sigma_e^2) = \frac{1}{B} \sum_{l=1}^B (\hat{\sigma}_e^{2(l)} - \sigma_e^2)^2,$$

$$\text{MSE}(\sigma_a^2) = \frac{1}{B} \sum_{l=1}^B (\hat{\sigma}_a^{2(l)} - \sigma_a^2)^2,$$

where  $\hat{\sigma}_e^{2(l)}$  and  $\hat{\sigma}_a^{2(l)}$  are the estimated residual and genetic variance in the  $l$ th training block, respectively,  $l = 1, \dots, B$ . Given the estimated marker effects, the genetic variance was estimated as  $\hat{\sigma}_a^2 = \hat{\mathbf{m}}' \mathbf{K} \hat{\mathbf{m}}$ . The true genetic variance was calculated based on simulated marker effects as  $\sigma_a^2 = \mathbf{m}' \mathbf{K} \mathbf{m}$  (Bonk *et al.* 2016). This formula considers the contribution of LD to the genetic variance (Gianola *et al.* 2013).

**Table 1** Estimated variance components based on one MCMC run and  $n = 10,000$  observations

Prior	5 QTL		50 QTL	
	$\sigma_e^2$	$\sigma_a^2$	$\sigma_e^2$	$\sigma_a^2$
Uncorrelated P1	0.503	0.488	0.486	0.510
Correlated P2	0.507	0.475	0.488	0.499
Adaptive P3	0.506	0.497	0.488	0.522
Adaptive P4	0.506	0.497	0.488	0.522
Simulated	0.500	0.487	0.500	0.489

$\sigma_e^2$ , residual variance;  $\sigma_a^2$ , additive genetic variance.

Furthermore, the precision of the estimated effects at selected key SNPs was verified by the SD of traced samples based on one MCMC run. The SNPs at the five simulated quantitative trait loci (QTL) with largest effects and, for each of the two largest QTL, two SNPs being in high or low LD with the QTL, and within a window of 20 SNPs to both sides were selected as key SNPs.

An appropriate measure of significance was required to evaluate the suitability of the suggested priors for understanding the genetic architecture, particularly relevant genomic regions. First, because the SNP effects were correlated, segment effects  $s_j^{(l)} = \sum_{k=0}^{10} m_{j+k}^{(l)}$  were calculated for a sliding window with an arbitrary width, which covered 11 SNPs, in the  $l$ th sampling round. These effects resemble haplotype effects. Second, the posterior probability of being positive was obtained from the Gibbs samples:

$$h_j = \frac{1}{N} \sum_{l=1}^N I(s_j^{(l)} > 0),$$

where  $I(\cdot)$  denotes the indicator function, and  $N$  the number of Gibbs samples after the burn-in period. If  $h_j > 0.95$  (positive effect size), or  $h_j < 0.05$  (negative effect size), the segment was declared to have a nonzero effect on  $\mathbf{y}$ . This quantity represents a measure of evidence, which is a Bayesian analog of the  $P$ -value, in a similar manner to that described by De Braganca Pereira and Stern (1999).

**Data**

Two sets of data were explored. First, simulated data were used to evaluate the performance of the Bayesian approach depending on the different prior choices. Second, real genotype data were used to study the pattern of covariance between SNPs in a real half-sib family and for a particular chromosome. A real phenotype was not analyzed because the present study focused on the impact of the covariance matrix on the outcomes of genomic evaluations. Thus, challenges related to real observations (other nuisance effects or uncertainty about genetic effects on the selected chromosome) were eliminated completely. Hence, a phenotype was simulated based on the real genotypes to investigate the feasibility of the method in a general manner.

**Simulated data:** The genomic data were simulated using a synthetic and simplified approach, but the structure obtained for the dependencies resembled a realistic setting. Further details are provided in Supplemental Material, File S1. In total, 500 marker genotypes for 10,000 progeny were simulated on a chromosome segment with a length of 12 cM, but only the loci at which the sire was heterozygous were considered in further analyses ( $p = 259$ ). The SNP alleles were recoded, so the sire haplotypes were AA/BB regardless of the allele frequencies. The coding of alleles only affected the sign of the covariance and not the estimated effect size.



■ **Table 2 Mean squared error (MSE) of the estimated variance components**

Prior	5 QTL			50 QTL		
	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$
Uncorrelated P1	0.145	0.074	0.760	0.157	0.080	0.747
Correlated P2	0.015	0.021	0.839	0.007	0.013	0.909
Adaptive P3	0.055	0.065	0.692	0.057	0.100	0.739
Adaptive P4	0.014	0.072	0.682	0.014	0.107	0.730

Correlation ( $\rho$ ) between the predicted and simulated genetic values; 100-fold cross-validation (size of training set  $n = 100$ ).  $\sigma_e^2$ , residual variance;  $\sigma_a^2$ , additive genetic variance.

To simulate phenotypic observations, the effects of either five or 50 QTL were drawn randomly from a gamma distribution with shape parameter  $\alpha = 0.420$ , and scale parameter  $\beta = 2.619$ . The sign was sampled with equal likelihood. In the five-QTL scenario, the second largest QTL was placed adjacent to the largest QTL, with four SNPs between them to complicate its detection. A residual error was added, and the resulting phenotype was scaled to have a variance of one. Finally, the simulated residual variance component  $\sigma_e^2$  was 0.500, and, considering the formula for additive genetic variance given above, the simulated additive genetic variance  $\sigma_a^2$  was 0.487 in the five-QTL scenario, and 0.489 in the 50-QTL scenario. The QTL were taken from the SNP set, and they remained in the data.

The accuracy of the predicted genetic value, and the estimated variance components was evaluated using a  $B$ -fold cross-validation. For that, the complete data set ( $n = 10,000$ ) was split into successive blocks with a training set size of  $n = 100$  (small sample size,  $B = 100$  repetitions) or  $n = 1000$  (medium sample size,  $B = 10$  repetitions).

**Semireal data:** The data set comprised a single half-sib family of Holstein-Friesian cows ( $n = 106$ ), which were initially genotyped with a 50K SNP chip, where the complete data set was described by Wittenburg *et al.* (2013). The sire was phased based on the daughter genotypes using the R package *hsphase* (Ferdosi *et al.* 2014). For convenience, only  $p = 903$  SNPs at which the sire was heterozygous were selected from BTA1. The SNP alleles were recoded corresponding to the sire haplotypes AA/BB. The paternal recombination rate and LD of maternal gametes were estimated by numerical maximization (NM) of the log-likelihood function using the R function *optim* (R Core Team 2014) (see File S1). It was found that 407 of the 903 eigenvalues of  $K$  were negative, so the bending algorithm proposed by Jorjani *et al.* (2003) was employed to obtain a positive definite approximation of the covariance matrix. A phenotype was simulated based on five QTL ( $\sigma_a^2 = 0.521$ ,  $\sigma_e^2 = 0.500$ ), as described above.

The data are provided as File S2 and File S3. The physical order of the SNPs on BTA1 followed the Btau4.2 annotation (File S4).

### MCMC computing

(Metropolis-within-) Gibbs sampling algorithms were implemented in Fortran 90 embedding LAPACK 3.5.0 ([www.netlib.org/lapack](http://www.netlib.org/lapack)) and module *random* 1.13 ([jblevins.org/mirror/amiller](http://jblevins.org/mirror/amiller)). The program ran on a 2.93 GHz multi-user system. A single chain comprising 50,000 sampling rounds was executed, and 20,000 iterations were omitted as the burn-in. The R package *coda* was used for MCMC diagnostics. In particular, the effective sample size (ESS) and Heidelberger and Welch's test, which tests the null hypothesis that the samples were drawn from a stationary distribution, were employed to determine the convergence of a Markov chain.

■ **Table 3 MSE of the estimated variance components**

Prior	5 QTL			50 QTL		
	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$
Uncorrelated P1	0.00092	0.00116	0.974	0.00133	0.00358	0.964
Correlated P2	0.00090	0.00575	0.916	0.00042	0.00277	0.943
Adaptive P3	0.00056	0.01849	0.880	0.00069	0.02758	0.889
Adaptive P4	0.00060	0.02581	0.862	0.00047	0.03698	0.874

Correlation ( $\rho$ ) between the predicted and simulated genetic values; 10-fold cross-validation (size of training set  $n = 1000$ ).  $\sigma_e^2$ , residual variance;  $\sigma_a^2$ , additive genetic variance.

### Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are given in the Supplemental Material (File S2, File S3, and File S4).

## RESULTS

### Simulated data

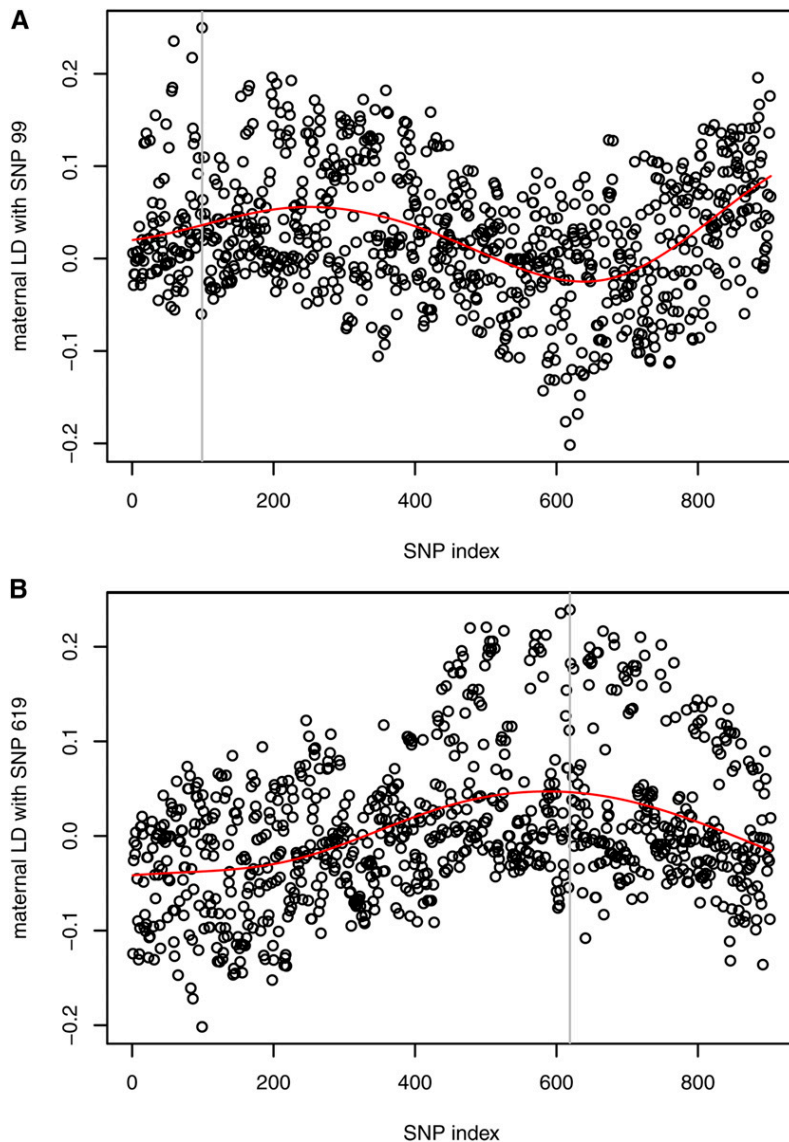
The diagonal elements of the theoretical covariance matrix  $K$  ranged from 0.473 to 0.500, with a mean value of 0.492. The off-diagonal elements varied from 0.174 to 0.331 around a mean of 0.231. The covariance decreased gradually with increasing distance between the SNPs. After conversion into correlations, the off-diagonal entries ranged from 0.352 to 0.667, with a mean value of 0.470. As shown in Figure 1A, the theoretical and observed correlations were compared for a randomly selected SNP in a window of 200 SNPs, where the theoretical correlation decreased rapidly from the diagonal to off-diagonal elements, but this agreed with the values observed based on the genotypes. Figure 1B shows the paternal contribution (*i.e.*, linkage) to the covariance between SNPs, as discussed later.

MCMC diagnostics indicated the convergence of the MCMC algorithm for all of the proposed priors. ESS at key SNP effects was rather high, *i.e.*,  $> 10,309$ , and  $> 30,000$  for a few exceptions; these values hardly differed between SNPs being the QTL or not. Heidelberger and Welch's test was generally passed. The test failed only at the key SNP with the smallest simulated effect in the five-QTL scenario when the prior P4 was selected. Furthermore, for the hybrid strategy with prior P3, the acceptance rate  $\alpha$  ranged from 0.54 to 0.79 (from 0.55 to 0.80) in the five-QTL (50-QTL) scenario.

For all of the selected priors, and with five and 50 simulated QTL, the estimated SNP effects agreed well with the simulated effects when the complete data set ( $n = 10,000$ ) was used. Most of the shrinkage effects were observed with the correlated prior P2. Many spurious effects were estimated with small effect sizes using all of the correlated prior choices. As expected, the empirical SD of key SNP effects was lowest when based on the complete data set, and it increased as the sample size decreased (*e.g.*, see Figure 2, A–C for the five-QTL scenario). In most cases, the SD of SNP effects at the QTL was lowest with the correlated prior P2, irrespective of the sample size. At non-QTL, smallest SD was obtained with P2 only for the small sample size. This was attributed to the very high shrinkage due to the prior P2.

For the five-QTL scenario and based on one MCMC run, few significant chromosome segments were falsely detected when one of the correlated priors was selected. However, for small sample sizes, repeated simulations showed that using the uncorrelated prior P1 and correlated prior P2, only one significant segment at the end of the chromosome was identified correctly (Figure 3). The relevant region was smaller with the uncorrelated prior assumption. No significant segments were found with priors P3 and P4. For the 50-QTL scenario





**Figure 4** Estimated maternal LD with two selected SNPs on BTA1 for the real genotypes. The vertical line refers to the SNP with which pairwise LD was calculated. Smoothing via B-splines visualizes the trend of the data (red curves). (A) LD pattern shows a potential error in the marker map. (B) Maximum LD was observed around the reference SNP indicating correct positioning.

and using all priors, hardly any chromosome segments were detected due to the high level of uncertainty. For a medium sample size, the identification of significant segments was generally improved, and it even reached 100% for the largest QTL with all of the priors (results not shown).

The estimated variance components differed only slightly among the four methods for the complete data set, which were estimated to range from 0.475 to 0.497 for  $\sigma_a^2$ , and from 0.503 to 0.507 for  $\sigma_e^2$  in the five-QTL scenario (Table 1). In the 50-QTL scenario, the estimates ranged from 0.499 to 0.522 for  $\sigma_a^2$ , and from 0.486 to 0.488 for  $\sigma_e^2$  (Table 1). The MSE of the estimated variance components showed that the precision was greatest with the correlated prior P2 irrespective of the number of QTL when the sample size was small, as well as with a medium sample size and 50 QTL (see Table 2 and Table 3).

The accuracy of the genetic value predictions was highest for the correlated prior P2 only with  $n = 100$ ; otherwise, the highest accuracy was obtained using the uncorrelated prior P1. The decrease in the mean TGV according to the fraction ( $r$ ) of selected candidates gave mixed results (e.g., see Figure 2D), where no method performed especially well with five QTL. The mean TGV was highest with the correlated prior P2 only when  $r$  was between about 30% and 70% in the five-QTL scenario, but

this prior was generally better with 50 simulated QTL. The adaptive priors P3 and P4 generally performed worse in terms of all the evaluation criteria.

Additional figures showing the results are provided in File S1. Moreover, as part of this material, a simple simulation study was conducted to determine the shape of the theoretical covariance according to the recombination rate between a pair of SNPs. The average theoretical covariance agreed well with the empirical covariance obtained based on the progeny genotypes.

### Semireal data

There was a good agreement between the theoretical and empirical correlations between SNP genotypes. The pattern of the correlations is shown in File S1. A gradual decrease in the entries was found with increasing distance between the SNPs. The off-diagonal values ranged from  $-0.411$  to  $1.000$ , with a mean value of  $0.219$ . The low minimum value was due to an extreme estimate of the LD, *i.e.*,  $D = -0.202$  between SNP 99 and 619, where the corresponding covariance was  $-0.193$ . If the distance between SNPs is rather large, then an extreme covariance indicates a potential error in the marker map. This hypothesis is supported by the values in the 99th row of  $D$ , which is the

■ **Table 4** MSE of the estimated variance components when the inverse covariance matrix was sparse

Prior	5 QTL			50 QTL		
	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$
Uncorrelated P1	0.145	0.074	0.760	0.157	0.080	0.747
Correlated P2	0.036	0.022	0.831	0.014	0.013	0.904
Adaptive P3	0.087	0.053	0.779	0.083	0.065	0.772
Adaptive P4	0.010	0.014	0.745	0.011	0.029	0.795

Correlation ( $\rho$ ) between the predicted and simulated genetic values; 100-fold cross-validation (size of training set  $n = 100$ ). For comparison, results with P1 were taken from Table 2.  $\sigma_e^2$ , residual variance;  $\sigma_a^2$ , additive genetic variance.

matrix of the pairwise LD values. It was expected that the maximum at around position 99 would decrease more or less smoothly on both sides, but this was not the case for this SNP, unlike SNP 619 (Figure 4).

Using MCMC computing, ESS at key SNP effects varied between 324 and 39,910, and was lowest with prior P2. However, ESS differed slightly between SNPs at the QTL and non-QTL. Heidelberg and Welch's test was passed by all of the priors, which indicated that the Markov chain converged to a stationary distribution. Furthermore, the acceptance rate  $\alpha$  ranged from 0.13 to 0.25 with prior P3.

The estimated SNP effects were far from their simulated effect sizes, which was due mainly to the small sample size causing a high amount of uncertainty. Using the correlated prior P2, which was the approach with the highest degree of shrinkage, a significant segment at the end of the chromosome was identified correctly. The variance components were estimated as  $\sigma_a^2 = 0.419$  and  $\sigma_e^2 = 0.505$ . The other prior choices yielded variance component estimates  $> 2 \times$  these values.

## DISCUSSION

Our theoretical investigations have shown that the covariance between SNPs depends on the genetic distance and the maternal LD between them. In this study, the covariance matrix derived for a single half-sib family was employed in genomic evaluations of a simulated phenotype with heritability of about 50%. The results depended greatly on the sample size, but the inclusion of the covariance matrix was clearly beneficial for small sample sizes ( $n = 100$ ) in terms of both the MSE of the additive and residual variance components as well as the accuracy of the genetic value predictions, although there were no advantages with larger sample sizes ( $n = 1000$ ). These variable outcomes may have several explanations, which are related to the nature of covariance in the selected population design and the different prior assumptions.

### Covariance matrix

Previous studies have suggested that some covariance or correlation structure for marker effects should be included in regression models to estimate their effects. Candidates selected from covariance structures have proved useful in other fields, such as time series analysis, *i.e.*, equally spaced autoregressive, Toeplitz and Gaussian decay (Gianola *et al.* 2003), and first-order antedependence (Yang and Tempelman 2012). All of these methods assume some decay of the correlations with increasing distance between the markers, thereby considering the fact that recombination becomes more frequent when markers are more distant. However, in contrast to these previous methods, our derivation makes explicit use of well-established genetic arguments, and it is based on the recombination frequencies, Haldane's underlying mapping function (Bonk *et al.* 2016), and the population-wide pairwise LD. Basically, the covariance between SNPs is determined as the sum of a paternal and maternal part. Assuming many half-sibs, the paternal part is considered for coinheritance (linkage), whereas the maternal part

■ **Table 5** MSE of the estimated variance components when the inverse covariance matrix was sparse

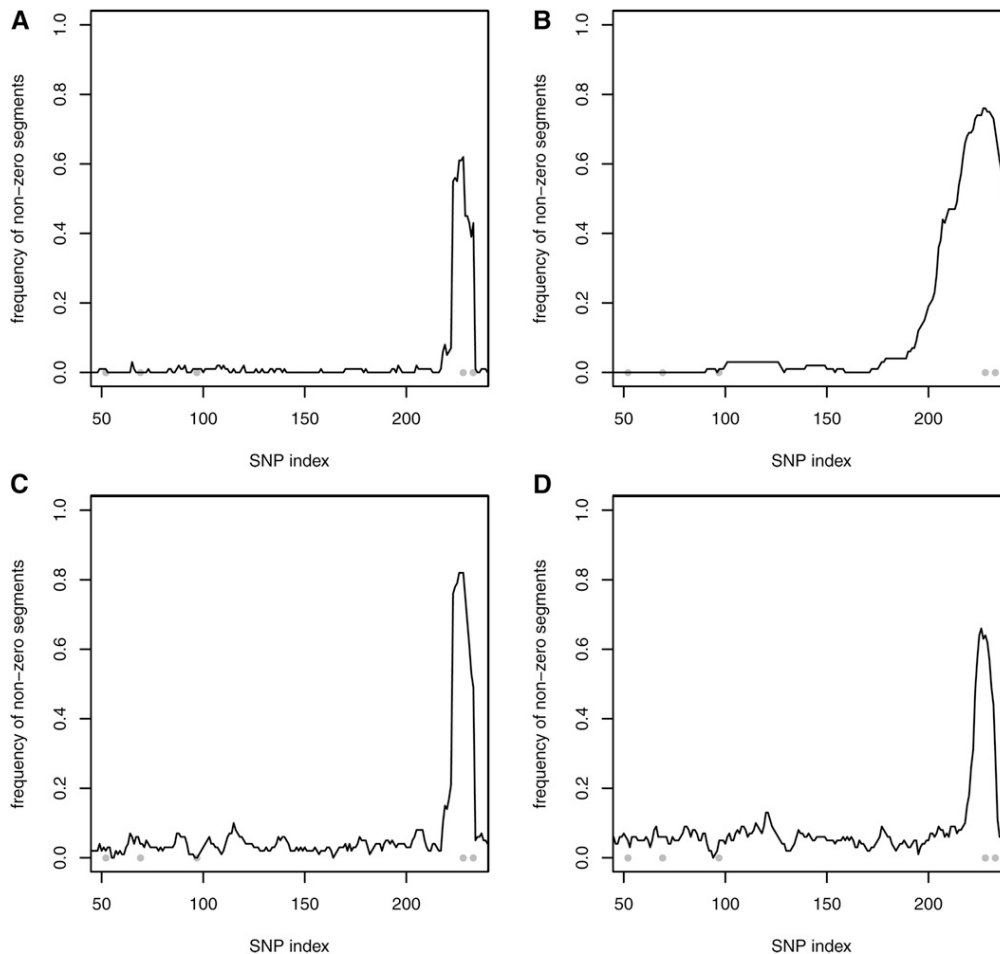
Prior	5 QTL			50 QTL		
	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$	MSE $\sigma_e^2$	MSE $\sigma_a^2$	$\rho$
Uncorrelated P1	0.00092	0.00116	0.974	0.00133	0.00358	0.964
Correlated P2	0.01499	0.01413	0.858	0.00563	0.00465	0.911
Adaptive P3	0.00084	0.00110	0.969	0.00116	0.00390	0.960
Adaptive P4	0.00063	0.00070	0.965	0.00045	0.00266	0.951

Correlation ( $\rho$ ) between the predicted and simulated genetic values; 10-fold cross-validation (size of training set  $n = 1000$ ). For comparison, results with P1 were taken from Table 3.  $\sigma_e^2$ , residual variance;  $\sigma_a^2$ , additive genetic variance.

models the population-wide LD. The sign and size of the correlations are known *a priori* (with the exception of the second adaptive method, P4) in contrast to the antedependence model proposed by Yang and Tempelman (2012). In the latter model, the sign and size of the regression for a marker effect on that of its predecessor are determined during each iteration by sampling from a normal distribution. Both adaptive versions allow for some variation in the covariances because each sampled gamma parameter affects both the diagonal and off-diagonal elements of the inverse covariance matrix. In non-QTL regions, small  $\gamma$ s are expected, whereas large  $\gamma$ -values are related to QTL regions. This was also observed in our data analyses (see File S1). Therefore, such a covariance structure may be classified as nonstationary in a similar manner to the antedependence model because the covariances depend on the distances between markers and on the marker positions, *i.e.*, their proximity to QTL.

Unlike haplotype-based approaches (*e.g.*, Calus *et al.* 2008; Cuyabano *et al.* 2014), which require that all of the individuals are phased, only the sire's haplotypes need to be known when using the proposed method. In the present study, we focused on a half-sib family structure, but other population structures are also relevant for livestock, *e.g.*, full-sib families often occur in chickens. Gametes of full-sibs are derived from a single sire and dam, and therefore the covariance between SNPs is restricted to the linkage part. (Paternal half-sibs represent a special case of a population with full-sib family structure, and the maternal gametes can be seen as random samples of a population with corresponding LD.) Then, the haplotypes of both parents should be considered when setting up the covariance matrix, as shown by Bonk *et al.* (2016). A typical livestock population comprises a mixture of families, so population-specific effects may be required instead of family-specific effects. The extension of the covariance matrix to a multiple-family study will be investigated in future research. The population covariance matrix could be set up as a weighted average of family-specific covariance terms, where the weight depends on the family size relative to the total sample size. Family sizes are often much smaller than 100, especially for full-sibs, so including the covariance between SNPs in genomic evaluations is expected to be beneficial, but this should be confirmed in a subsequent study.

All of the parameters required for the covariance matrix  $\mathbf{K}$  were available for the simulated data. However, the recombination rate and LD of the maternal gametes had to be estimated from the progeny genotypes for the real data. In this study, estimates were obtained by NM of the log-likelihood function. Alternatively,  $\theta$  and  $D$  could be estimated by the expectation-maximization (EM) algorithm (Gomez-Raya *et al.* 2013). However, this iterative approach is more time consuming than the iterations required for NM. There was good agreement between the theoretical and empirical covariances obtained from the progeny genotypes, but the covariance matrix was indefinite for the real data genotypes (and positive definite for the simulated data), and thus it



**Figure 5** Simulation with five QTL,  $n = 100$ , and 100 repetitions. Detection of nonzero segment effects using the **sparse** inverse covariance matrix and uncorrelated prior P1 (A), correlated prior P2 (B), adaptive prior P3 (C), and adaptive prior P4 (D). Gray dots indicate the simulated QTL positions.

was forced to be positive definite by bending. This was probably caused by errors in the estimation of the required parameters, which could be estimated more precisely using larger families.

When the alleles of the sire haplotypes were coded according to the observed minor allele frequency in the sample, both of the sire diplotypes AA/BB and AB/BA appeared in the double heterozygous case. The covariance matrix had then a block structure, which represented the nonrecombinant genome segments.

### Prior assumptions

In this study, each of the correlated priors incorporated a dense covariance matrix, and its inverse is also dense. According to Lang *et al.* (2002), locally adaptive random-walk priors penalize the difference between successive effects, thereby allowing local regularization to avoid oversmoothing. The inverse  $\mathbf{K}^{-1}$ , which is also called the precision matrix, could be adjusted to have a banded structure to allow better local adaptivity in genomic evaluations. As an option, sparseness can be achieved by eliminating the “noise”. In the present study, the paternal contribution (*i.e.*, linkage) to the covariance matrix was more important than the maternal part (*i.e.*, population LD), where the interquartile range of  $D$  was zero in the simulations (mean = 0) and 0.064 in the real genotypes (mean = 0.016). Thus, an approximation of

$\mathbf{K}$  retained only the paternal part  $\left\{ \frac{1}{4} (1 - 2\theta_{j,k}) \right\}_{j,k=1}^P$ . Figure 1B shows

the pattern for a randomly selected SNP. This simple structure can be inverted theoretically by exploiting, for instance,

$1 - 2\theta_{1,3} = (1 - 2\theta_{1,2})(1 - 2\theta_{2,3})$ , based on Haldane’s mapping function. The dependence throughout the genome represents an autoregressive process, so the inverse covariance matrix depends only on the predecessor and successor, thereby yielding a three-band matrix that fully considers the unequal distances between SNPs. This type of sparse structure has also been derived and implemented using a Gibbs sampling strategy for a backcrossed population by Reichelt *et al.* (2015). Using this structure, a more local impact of the regularization parameters is obtained using the adaptive priors P3 and P4, and the error propagation caused by numerical imprecisions along the chromosome appears to be reduced. The influence of this structure on the performance of the correlated prior selections is shown in Table 4 and Table 5, and in File S1. In particular, the MSE was smallest using the adaptive priors P3 and P4. Furthermore, the identification of significant chromosome segments was improved, and slightly superior to the uncorrelated prior P1 (*e.g.*, see Figure 5 for the five-QTL scenario). During the analysis of real data, this approximation has an additional positive impact because only the recombination rate needs to be estimated. Unlike classical linkage analysis or linkage/LD analysis (*e.g.*, Meuwissen *et al.* 2002), which is applied to a marker bracket instead of all markers simultaneously, this restriction to the linkage part, works well, even though the parental origin of the SNP alleles has not been identified.

A second option for approximation uses the modified Cholesky decomposition of the inverse matrix,  $\Psi^{-1} = \mathbf{L}\mathbf{\Gamma}\mathbf{L}'$ , which is suitable for P4. Wu and Pourahmadi (2003) derived a smoothing algorithm for

$L$  along its subdiagonals for longitudinal data and  $p < n$ . Future studies should investigate whether this banded structure can also be derived for  $p > n$ . The sparse structure of the inverse matrix is also a consequence of the antependence specification proposed by Yang and Tempelman (2012). Exploiting a sparse structure in the MCMC algorithm will also improve the computational speed (e.g., for  $n = 100$ , the current implementation required on average 15 min based on prior P1, and 48 min based on prior P4.)

The prior covariance matrix used in P3 is similar to that described by George and McCulloch (1993) in terms of the dynamic regularization parameters ( $\gamma_s$ ), although it is implemented as a variable selection approach. In the present study, variable selection was not executed because it was assumed that each marker contributes to the genetic variation, at least indirectly through LD. However, this strategy led to poor mixing in the MCMC algorithm compared with the adaptive prior P4 using semireal data (results not shown). The trace plots contained repetitions of effect samples with low variability, followed by samples with very high variability, which was not observed to the same extent for the simulated data with small sample sizes. The correlated prior P2 was the most robust prior choice for both the simulations and semireal data because it could detect significant chromosome segments and estimate the variance components with the smallest MSE. To facilitate realistic genome-wide evaluations, we recommend specifying chromosome-wise priors, i.e.,  $m_c | \Psi_c \sim N(\mathbf{0}, \Psi_c)$  and  $\Psi_c = \mathbf{K}_c \sigma_c^2$  for  $c = 1, \dots, n_{\text{chr}}$ . Genome-wide shrinkage, which employs a single regularization parameter  $\sigma^2$ , may be too strict if only a few QTL are present.

Bayesian approaches require the specification of prior distributions but the use of hyper-parameters was minimized in the present study. The sensitivity to the selected priors has been evaluated and discussed previously, e.g., by Knürr *et al.* (2013) and Gianola (2013). In the adaptive prior P3, the tuning variance was set to  $\varepsilon = 1.0$ , which made this approach sensitive, and it compromised the repeatability of the experiments. For example, for different choices of  $\varepsilon \in \{0.01, 0.1, 1.0\}$ , a varying acceptance rate  $\alpha$  was obtained, where  $\alpha$  was higher when  $\varepsilon$  was smaller. However, the impact on a single-effect estimate was negligible for both the simulated and semireal data, whereas the estimated variance components  $\sigma_e^2$  and  $\sigma_a^2$  differed in terms of the second decimal place (results not shown). With  $\varepsilon = 1.0$ , the range of  $\alpha$  roughly agreed with the rule of thumb given by Besag *et al.* (1995) who showed that an acceptance rate between about 30% and 70% was often satisfactory. Alternatively, the tuning variance may be adjusted during the burn-in phase to obtain an intermediate  $\alpha$  (25–50%) in a similar manner to Yang *et al.* (2015).

### Sample size

The sample size had a major effect on the outcome. For small sample sizes ( $n = 100$ ), the effect estimates were poor, and a high EGV was obtained even for those individuals with low TGV. Thus, the curves of the mean TGVs (e.g., in Figure 2D based on five simulated QTL) started below 0.900 at a selection rate  $r = 5\%$ , although the 95% quantile for the TGV was 1.106. The shape of the curve based on the correlated prior choice P2 was unusual, where it decreased very slightly until  $r = 50\%$ , and it then declined rapidly, which may be explained by the following two reasons. First, the SNP effects were greatly reduced toward zero due to the major effect of the shrinkage parameter  $\sigma^2$ . Second, the sire was heterozygous at all of the loci considered (with A alleles on one strand and B alleles on the other), so there were only few recombinant offspring, and the distribution of EGV was bimodal (see File S1). The variation around the two modes was due to variation in the maternally inherited gametes. Thus, the individuals selected with

the highest EGV comprised a mixture of high/medium/low-TGV individuals, and the curve obtained for the mean TGVs decayed slowly. The curve approached the population mean rapidly above  $r = 50\%$ . This outcome was also observed, but to a much greater extent, in simulations with 50 QTL as well as when combined with the sparse inverse covariance matrix (File S1). This pattern was not detected with a medium sample size ( $n = 1000$ ), probably because of the improved estimation of the parameters, and the occurrence of more recombinant offspring. This phenomenon is unlikely to occur if more than one chromosome is investigated, thereby causing greater variation in the paternal gametes.

### Conclusion

In this study, the dependence between SNP genotypes was derived theoretically from the genetic parameters for a half-sib family. Integrating this information into genomic evaluations improved the estimates of the variance components and the genetic value predictions when the sample size was small ( $n = 100$ ). Thus, in small populations, for which the parameter estimates are typically affected by high uncertainty when the number of predictors is large, additional information about the population structure can increase the precision, thereby following the general Bayesian principle. The proposed correlated prior choices could potentially obtain better overall performance if a locally adaptive approximation of dependence is employed.

### ACKNOWLEDGMENTS

Special thanks are given to our colleagues at the Leibniz Institute for Farm Animal Biology (FBN, Dummerstorf, Germany), S. Bonk and M. Reichelt, who contributed to discussions during the project. We also thank the anonymous reviewers for their helpful comments. The real genotype data were obtained in the Fugato-plus project “BovIBI” funded by the German Federal Ministry of Education and Research (BMBF). The publication of this article was funded by the Open Access fund of the Leibniz Association and the FBN.

Author contributions: D.W. developed the theory, implemented the statistical methods, performed the analysis, and wrote the manuscript. F.T. contributed to the theoretical investigations of the covariance between markers, and the estimation of the required parameters. J.K. developed the simulation design. N.R. raised the initial question and contributed to the discussions. All of the authors have read and approved the final manuscript. The authors declare that they have no competing interests.

### LITERATURE CITED

- Besag, J., P. Green, D. Higdon, and K. Mengersen, 1995 Bayesian computation and stochastic systems. *Stat. Sci.* 10: 3–66.
- Bonk, S., M. Reichelt, F. Teuscher, D. Segelke, and N. Reinsch, 2016 Mendelian sampling covariability of marker effects and genetic values. *Genet. Sel. Evol.* 48(1): 1–11.
- Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178(1): 553–561.
- Cuyabano, B. C. D., G. Su, and M. S. Lund, 2014 Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* 15: 1171.
- Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen *et al.*, 2014 Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 46(8): 858–865.
- De Braganca Pereira, C. A., and J. M. Stern, 1999 Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy (Basel)* 1(4): 99–110.



- 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(2): 327–345.
- Dehman, A., C. Ambroise, and P. Neuvial, 2015 Performance of a block-wise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics* 16: 148.
- Fahrmeir, L., and S. Lang, 2001 Bayesian inference for generalized additive mixed models based on Markov random field priors. *Appl. Stat.* 50(2): 201–220.
- Ferdosi, M., B. Kinghorn, J. van der Werf, S. Lee, and C. Gondro, 2014 hspase: an R package for pedigree reconstruction, detection of recombination events, phasing and imputation of half-sib family groups. *BMC Bioinformatics* 15(1): 172.
- George, E. I., and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* 88(423): 881–889.
- Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194(3): 573–596.
- Gianola, D., M. Perez-Enciso, and M. A. Toro, 2003 On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* 163: 347–365.
- Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173(3): 1761–1776.
- Gianola, D., F. Hospital, and E. Verrier, 2013 Contribution of an additive locus to genetic variance when inheritance is multi-factorial with implications on interpretation of GWAS. *Theor. Appl. Genet.* 126(6): 1457–1472.
- Gomez-Raya, L., A. M. Hulse, D. Thain, and W. M. Rauw, 2013 Haplotype phasing after joint estimation of recombination and linkage disequilibrium in breeding populations. *J. Anim. Sci. Biotechnol.* 4(1): 30.
- Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12: 186.
- Jorjani, H., L. Klei, and U. Emanuelson, 2003 A simple method for weighted bending of genetic (co)variance matrices. *J. Dairy Sci.* 86(2): 677–679.
- Knürr, T., E. Läärä, and M. J. Sillanpää, 2013 Impact of prior specifications in a shrinkage-inducing Bayesian model for quantitative trait mapping and genomic prediction. *Genet. Sel. Evol.* 45(1): 24.
- Lang, S., E.-M. Fronk, and L. Fahrmeir, 2002 Function estimation with locally adaptive dynamic models. *Comput. Stat.* 17(4): 479–500.
- Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4): 1819–1829.
- Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161(1): 373–379.
- Piepho, H.-P., 2009 Ridge regression and extensions for genomewide selection in maize. *Crop Sci.* 49: 1165–1176.
- Pourahmadi, M., 2007 Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance–correlation parameters. *Biometrika* 94(4): 1006–1013.
- R Core Team, 2014 *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: <http://www.R-project.org/>. Accessed: July 4, 2014.
- Reichelt, M., M. Mayer, F. Teuscher, and N. Reinsch, 2015 Bayesian modelling of marker effects in backcross experiments by means of their covariance matrix. Abstract Volume of the 61st Biometrical Colloquium, Biometrics and Communication: From Statistical Theory to Perception in the Public, Dortmund, Germany, p. 124.
- Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*, Springer, New York.
- Stranden, I., and O. F. Christensen, 2011 Allele coding in genomic evaluation. *Genet. Sel. Evol.* 43: 25.
- Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight, 2005 Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. B* 67(1): 91–108.
- Tierney, L., 1994 Markov chains for exploring posterior distributions. *Ann. Stat.* 22: 1701–1728.
- Wang, C., J. Rutledge, and D. Gianola, 1994 Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genet. Sel. Evol.* 26(2): 91–115.
- Wittenburg, D., N. Melzer, L. Willmitzer, J. Lisek, U. Kesting *et al.*, 2013 Milk metabolites and their genetic variability. *J. Dairy Sci.* 96: 2557–2569.
- Wu, W. B., and M. Pourahmadi, 2003 Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* 90(4): 831–844.
- Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. *Genetics* 163(2): 789–801.
- Yang, W., and R. J. Tempelman, 2012 A Bayesian antedependence model for whole genome prediction. *Genetics* 190(4): 1491–1501.
- Yang, W., C. Chen, and R. Tempelman, 2015 Improving the computational efficiency of fully Bayes inference and assessing the effect of misspecification of hyperparameters in whole-genome prediction models. *Genet. Sel. Evol.* 47(1): 13.

Communicating editor: D. J. de Koning

## APPENDIX

### Prior P3

The target distribution  $p(\gamma_j | \text{else})$  is obtained from

$$p(\gamma_j | \boldsymbol{\gamma}_{-j}, \mathbf{m}, \mathbf{y}) \propto p(\mathbf{m} | \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \propto p(\mathbf{m} | \boldsymbol{\gamma}) p(\gamma_j) \propto |\boldsymbol{\Psi}|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{m}' \boldsymbol{\Psi}^{-1} \mathbf{m}\right) \gamma_j^{-3/2} \exp\left(-\frac{1}{2\gamma_j}\right).$$

For convenience, it is  $\tau_j = \frac{1}{\gamma_j}$  and  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_p)'$ . Hence,

$$\mathbf{m}' \boldsymbol{\Psi}^{-1} \mathbf{m} = \mathbf{m}' \boldsymbol{\Gamma}^{-1} \mathbf{K}^{-1} \boldsymbol{\Gamma}^{-1} \mathbf{m} = \text{tr}\left(\mathbf{K}^{-1} \boldsymbol{\Gamma}^{-1} \mathbf{m} \mathbf{m}' \boldsymbol{\Gamma}^{-1}\right) = \sum_{j=1}^p \sum_{k=1}^p \tilde{K}_{j,k}^{-1} \frac{m_j}{\gamma_j} \frac{m_k}{\gamma_k} = \sum_{j=1}^p \sum_{k=1}^p \tilde{K}_{j,k} \tau_j \tau_k = \boldsymbol{\tau}' \tilde{\mathbf{K}} \boldsymbol{\tau}$$

with  $\tilde{\mathbf{K}} = \mathbf{K}^{-1} \# \mathbf{m} \mathbf{m}'$  employing the Hadamard product (#). Thus,

$$p(\gamma_j | \boldsymbol{\gamma}_{-j}, \mathbf{m}, \mathbf{y}) \propto \tau_j \exp\left(-\frac{1}{2} \boldsymbol{\tau}' \tilde{\mathbf{K}} \boldsymbol{\tau} + \frac{3}{2} \ln \tau_j - \frac{\tau_j}{2}\right).$$

The proposal density is lognormal, *i.e.*,

$$q(\gamma_j^* | \gamma_j^{(t-1)}) = (2\pi\varepsilon)^{-1/2} \exp\left(-\frac{1}{2\varepsilon} (\ln \gamma_j^* - \ln \gamma_j^{(t-1)})^2\right) \frac{1}{\gamma_j^*}.$$

Except the last ratio, this function is symmetric in  $\gamma_j^*$  and  $\gamma_j^{(t-1)}$ . Let  $\boldsymbol{\tau}^*$  denote the vector of  $\tau$ s, where the  $j$ th component is replaced by the current proposed value, and  $\boldsymbol{\tau}^{(t-1)}$  refers to the vector of samples from the last iteration. Finally, the ratio  $R$  is

$$R = \frac{p(\gamma_j^* | \boldsymbol{\gamma}_{-j}^{(t-1)}, \mathbf{m}^{(t-1)}, \mathbf{y})}{q(\gamma_j^* | \gamma_j^{(t-1)})} \frac{q(\gamma_j^{(t-1)} | \gamma_j^*)}{p(\gamma_j^{(t-1)} | \boldsymbol{\gamma}_{-j}^{(t-1)}, \mathbf{m}^{(t-1)}, \mathbf{y})} = \frac{\exp\left(-\frac{1}{2} \boldsymbol{\tau}^{*'} \tilde{\mathbf{K}} \boldsymbol{\tau}^* + \frac{3}{2} \ln \tau_j^* - \frac{\tau_j^*}{2}\right)}{\exp\left(-\frac{1}{2} \boldsymbol{\tau}^{(t-1)'} \tilde{\mathbf{K}} \boldsymbol{\tau}^{(t-1)} + \frac{3}{2} \ln \tau_j^{(t-1)} - \frac{\tau_j^{(t-1)}}{2}\right)}.$$

### Prior P4

The posterior density of regularization parameters is derived as

$$p(\gamma_1, \dots, \gamma_p | \text{else}) \propto p(\mathbf{m} | \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \propto |\boldsymbol{\Psi}|^{1/2} \exp\left(-\frac{1}{2} \mathbf{m}' \boldsymbol{\Psi}^{-1} \mathbf{m}\right) \prod_{j=1}^p \gamma_j^{-1/2} \exp\left(-\frac{\gamma_j}{2}\right) \propto \exp\left(-\frac{1}{2} \text{tr}(\boldsymbol{\Gamma} \mathbf{L}' \mathbf{m} \mathbf{m}' \mathbf{L})\right) \prod_{j=1}^p \exp\left(-\frac{\gamma_j}{2}\right).$$

Set  $\tilde{\mathbf{m}} = \mathbf{L}' \mathbf{m}$ , then


$$p(\gamma_1, \dots, \gamma_p | \text{else}) \propto \exp\left(-\frac{1}{2} \sum_{j=1}^p \gamma_j \tilde{m}_j^2\right) \prod_{j=1}^p \exp\left(-\frac{\gamma_j}{2}\right) = \prod_{j=1}^p \exp\left(-\frac{1}{2} \gamma_j (\tilde{m}_j^2 + 1)\right).$$

Thus, the  $\gamma$ s are conditionally independent, each distributed as scaled  $\chi^2$ .

## METHODOLOGY ARTICLE

## Open Access

# Design of experiments for fine-mapping quantitative trait loci in livestock populations

Dörte Wittenburg<sup>1\*</sup> , Sarah Bonk<sup>2</sup>, Michael Doschoris<sup>1</sup> and Henry Reyer<sup>3</sup>

## Abstract

**Background:** Single nucleotide polymorphisms (SNPs) which capture a significant impact on a trait can be identified with genome-wide association studies. High linkage disequilibrium (LD) among SNPs makes it difficult to identify causative variants correctly. Thus, often target regions instead of single SNPs are reported. Sample size has not only a crucial impact on the precision of parameter estimates, it also ensures that a desired level of statistical power can be reached. We study the design of experiments for fine-mapping of signals of a quantitative trait locus in such a target region.

**Methods:** A multi-locus model allows to identify causative variants simultaneously, to state their positions more precisely and to account for existing dependencies. Based on the commonly applied SNP-BLUP approach, we determine the z-score statistic for locally testing non-zero SNP effects and investigate its distribution under the alternative hypothesis. This quantity employs the theoretical instead of observed dependence between SNPs; it can be set up as a function of paternal and maternal LD for any given population structure.

**Results:** We simulated multiple paternal half-sib families and considered a target region of 1 Mbp. A bimodal distribution of estimated sample size was observed, particularly if more than two causative variants were assumed. The median of estimates constituted the final proposal of optimal sample size; it was consistently less than sample size estimated from single-SNP investigation which was used as a baseline approach. The second mode pointed to inflated sample sizes and could be explained by blocks of varying linkage phases leading to negative correlations between SNPs. Optimal sample size increased almost linearly with number of signals to be identified but depended much stronger on the assumption on heritability. For instance, three times as many samples were required if heritability was 0.1 compared to 0.3. An R package is provided that comprises all required tools.

**Conclusions:** Our approach incorporates information about the population structure into the design of experiments. Compared to a conventional method, this leads to a reduced estimate of sample size enabling the resource-saving design of future experiments for fine-mapping of candidate variants.

**Keywords:** Single nucleotide polymorphism, Statistical power, Target region, SNP-BLUP, Linkage disequilibrium

\*Correspondence: [wittenburg@fbn-dummerstorf.de](mailto:wittenburg@fbn-dummerstorf.de)

<sup>1</sup>Leibniz Institute for Farm Animal Biology, Institute of Genetics and Biometry, 18196 Dummerstorf, Germany

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Genomewide association studies (GWAS) help exploring the relationship between genetic and phenotypic variation. Genetic variation is often expressed in terms of genomic markers such as single nucleotide polymorphisms (SNPs). Identified variants may or may not be part of known genes. In a candidate-gene approach, variants are then assigned to the closest known gene and their functional importance can be studied further (e.g., [1]). The functional meaning of a variant may be differently interpreted if, due to statistical uncertainty, it was identified a few kbp upstream or downstream of its position. In general, it could be a complicated task to detect single loci as reported by Sahana et al. [2] in a study on udder health in dairy cattle. Instead of identifying important SNPs for clinical mastitis, only target regions were found. For instance, a window of about 1 Mbp length was detected on BTA6. A statistical reason for this complication lies in the high multicollinearity among predictor variables due to linkage and linkage disequilibrium (LD) between SNPs (e.g., [3]). Region-based aggregation tests in biologically relevant regions (e.g., genes; [4]) or fine-mapping approaches in independent partitions of the genome [5] have been suggested as powerful options. To eventually unravel which of the variants in a target region might be truly related to the trait, a follow-up experiment is recommended. The experimental design should account for the dependence between SNPs to ensure sufficient statistical power. This will be reflected in the sample size required. Statistical tools for the design of experiments (e.g., QUANTO; [6]) could not provide this until now. However, the denser the SNP chip is, the higher will be the correlation between SNPs. For instance, the target region on BTA6 of Sahana's paper covers 17 SNPs using a 50k SNP panel, 192 SNPs based on a 700k SNP panel and 21 796 SNPs in case of DNA sequence [2, 7].

In theory, it can be determined what sample size is needed for discovering a new variant in a single-locus model at a given power, e.g., 80%. Such investigations are based on ANOVA (one way classification; [8]) and can also account for a hypothetical degree of LD between causative variant and SNP [9, 10]. Proposals for an optimum experimental design have been made for mapping of a quantitative trait locus (QTL) in different population structures (e.g., F2, backcross or daughter design; [11]). However, it is not clear what sample size is required to distinguish multiple independent signals of a QTL using dense marker data.

Moreover, the power of association analysis depends not only on sample size and population parameters (e.g., heritability) but also on the underlying statistical model. Among myriad options for whole genome regression models, SNP-BLUP is an obvious choice for estimating genetic effects captured by all SNPs simultaneously. Also

because of its direct relationship to GBLUP (e.g., [12]), it is widely used in livestock (e.g., [13, 14]) and beyond (e.g., [15, 16]). Being enormously relevant in practice, it has been upgraded to comprise information of individuals with and without genotypic data in the framework of single-step methods [17, 18]. Though directly or indirectly estimated SNP effects are tested for being significantly different from zero [18], reports on statistical power of the underlying study design are lacking.

This paper addresses the question how to design a follow-up experiment based on a SNP-BLUP approach knowing that the predictor variables are so highly correlated. Our objective is the theoretical inference of optimal sample size to fine-map a QTL signal or to find evidence for multiple independent signals in a specified chunk of DNA. Eventually, it should be possible to detect variants at their actual position with high power. This paper concentrates on the case study of paternal half-sib families which is a typical family structure in livestock (e.g., dairy cattle). But the methodology developed enables sample size calculation for any population structure (e.g., full siblings, half siblings, mixture of both, unrelated individuals). Given the number of families, SNPs, signals of QTL and heritability, the optimal sample size is then presented as overall number of progeny. We validated our approach using simulated data. Furthermore, publicly available bovine HD SNP chip data helped verifying that the simulated linkage blocks resemble the genome structure in dairy cattle. A discussion of our achievements complements this study.

## Methods

The design of experiment requires a statistical model that combines phenotype with genotype data. Here, we assume a multiple-SNP approach that considers information of as many SNPs as desired simultaneously. For comparing the outcome with a conventionally used approach, a single-SNP model is specified.

### Multi-SNP model

For a joint association analysis of  $p$  SNPs with additive effects, a regression model is fitted to a phenotype  $y = (y_1, \dots, y_n)'$  of  $n$  individuals,

$$y = X\beta + e.$$

The  $n \times p$  design matrix  $X$  contains the genotype codes:  $X_{j,k} \in \{1, 0, -1\}$  for  $j = 1, \dots, n$  and  $k = 1, \dots, p$ . The columns of  $X$  and the vector  $y$  are centered within family and scaled afterwards to ensure  $\frac{1}{n}X'_{.,k}X_{.,k} = 1 \forall k$  and  $\frac{1}{n}y'y = 1$ . This way the model becomes independent of allele frequency. The residual error term is  $e \sim N(0, I_n\sigma_e^2)$ . Then the coefficient vector  $\beta$  is estimated using a ridge



regression approach as

$$\widehat{\beta} = (X'X + \lambda I_p)^{-1} X'y.$$

This step requires a penalty term  $\lambda$  which is practically obtained via cross-validation or REML approach.

Next, we investigate a multiple testing problem. For each SNP  $k$ ,  $k = 1, \dots, p$ , it is tested

$$H_0 : \beta_k = 0 \text{ vs. } H_A : \beta_k \neq 0 \quad (1)$$

with a suitable test statistic which is defined as the estimator of the  $k$ -th regression coefficient over its standard deviation, i.e.,

$$T_k = \frac{\widehat{\beta}_k}{SD(\widehat{\beta}_k)}. \quad (2)$$

The calculation of power  $\pi$  requires the distribution of  $T_k$  under  $H_A$ , then

$$\pi(\mu_k) = \Pr(T_k \geq q_{1-\alpha/2}) + \Pr(T_k < q_{\alpha/2}),$$

where  $q_{1-\alpha/2}$  and  $q_{\alpha/2}$  denote the upper and lower threshold, respectively, of the distribution of  $T_k$  under  $H_0$  with respect to a type-I error  $\alpha$ . Due to the ridge approach, requirements for fulfilling a  $t$  distribution do not hold ([19], p. 57). Hence the distribution of  $T_k$  is approximated as normal with mean  $\mu_k$  and variance 1. The distribution mean  $\mu_k$  is obtained from the expectation and variance of the estimator  $\widehat{\beta}_k$ . The moments are

$$\begin{aligned} E(\widehat{\beta}) &= (X'X + \lambda I_p)^{-1} X'X\beta, \\ V(\widehat{\beta}) &= (X'X + \lambda I_p)^{-1} X'X(X'X + \lambda I_p)^{-1} \sigma_e^2. \end{aligned}$$

The central point of our investigation is to substitute the correlation matrix  $\frac{1}{n}X'X$  to be observed in the progeny generation by the theoretical correlation matrix  $R$ . For any SNP pair  $k, l \in \{1, \dots, p\}$ ,  $\frac{1}{n}X'_{:,k}X_{:,l}$  is a plausible approximation to its expectation  $E(X_{j,k}X_{j,l}) = \text{cor}(X_{j,k}, X_{j,l})$  because of centered and scaled genotype codes. The derivation of  $R$  is shown in the [Appendix](#); it requires a genetic map and genetic information of parents.

Then the mean of the test statistic becomes

$$\mu_k = \frac{\{E(\widehat{\beta})\}_k}{\sqrt{\{V(\widehat{\beta})\}_{k,k}}} = \frac{\sqrt{n}}{\sigma_e} \frac{\{(R + \frac{\lambda}{n}I_p)^{-1}R\beta\}_k}{\sqrt{\{(R + \frac{\lambda}{n}I_p)^{-1}R(R + \frac{\lambda}{n}I_p)^{-1}\}_{k,k}}}. \quad (3)$$

Under  $H_0$ ,  $\mu_k = 0$ . In order to calculate the optimal sample size, the experimenter has to specify a set of parameters: number of SNPs ( $p$ ) in the investigated window of DNA, number of QTL signals to be detected ( $\kappa$ ), proportion of variance explained by the QTL signals in that window ( $h^2$ ) and number of families (e.g.,  $N$  sires). The input parameters for statistical power calculation are inferred from this experimental set-up:

1.  $R$  requires haplotypes of  $N$  sires (plus genetic map and maternal LD in general).

2. We assume that all variants corresponding to the QTL signals contribute equally to the genetic variance. Hence the relative effect size is determined at  $\kappa$  QTL signals as

$$\frac{\beta_l}{\sigma_e} = \sqrt{\frac{h^2}{\kappa(1-h^2)}} \quad \text{for } l \text{ in the set of QTL signals.} \quad (4)$$

The remaining  $\beta$ 's are 0.

3. The shrinkage parameter is derived corresponding to Hoerl et al. [20],

$$\begin{aligned} \lambda &= p \frac{\sigma_e^2}{\beta'\beta} \\ &= p \frac{1-h^2}{h^2}. \end{aligned}$$

This is a rough approximation assuming linkage equilibrium between variants corresponding to the QTL signals.

We circumvent doing any assumption about the unknown positions of QTL signals by taking a random sample of  $\kappa$  positions. Then the optimal sample size is calculated over a range of  $n$ 's (e.g.,  $1 - 5000$ ) employing the method of bisection. The minimum  $n$  that exceeds the given power is selected as "optimal" and denoted as  $n_{\text{opt}}$ . Here, we considered a power level of 80% which is arbitrary but often used for statistical analysis (e.g., [21]). In order to get a reliable estimate of optimal sample size, sampling is repeated 100 times, and the median of  $n_{\text{opt}}$  is suggested as final  $n_{\text{opt}}^*$ . The overall type-I error was  $\alpha = 0.01$ .

### Single-SNP model

For comparison, we consider a single SNP  $k \in \{1, \dots, p\}$  in a sliding window over the target region. Using the parameter definitions as above, the linear model in its simplest form is

$$y = X_k\beta_k + e.$$

Then the regression coefficient is estimated via ordinary least squares as

$$\widehat{\beta}_k = (X_k'X_k)^{-1} X_k'y.$$

The null hypothesis testing problem (1) and the corresponding test statistic (2) also apply in the single-SNP analysis. The test statistic is  $t$ -distributed with  $n - 1$  degrees of freedom and non-centrality parameter  $\delta_k$  ([19], pp. 110),

$$\delta_k = \frac{\beta_k}{\sigma_e} \sqrt{n}.$$

This approach neglects any impact of the other SNPs in the target region on  $y$ . Thus, a reduced pointwise error level ( $\alpha_k$ ) has to be employed to keep the overall type-I

error at  $\alpha$ . Knowing the effective number of independent tests ( $M_{\text{eff}}$ ), a suitable type-I-error correction is

$$\alpha_k = \alpha / M_{\text{eff}}.$$

In accordance with the simple  $\mathcal{M}$  method of Gao et al. [22], we suggest using  $R$  instead of  $\frac{1}{n}X'X$  for calculating  $M_{\text{eff}}$ . More precisely, the number of eigenvalues of  $R$  that contribute at least 99.5% to the sum of all eigenvalues yields  $M_{\text{eff}}$ .

### Data and validation study

The software R version 3.6.1 [23] was employed in this study. Unless otherwise stated, we implemented own R scripts.

Population genetic data were simulated using the R package AlphaSimR version 0.11.0 [24]. In total, 300 SNPs were uniformly spread in a chunk of DNA of 1 cM length corresponding approximately to 1 Mbp. The SNP density roughly resembled HD data. Five traits were simulated simultaneously, one for each number of QTL signals affecting the trait,  $\kappa = 1, \dots, 5$ . The founder population comprised 2 000 individuals (gender ratio 1:1) and constituted the parent generation. Other population parameters were kept at default settings (e.g., effective population size 100, mutation rate  $2.5 \cdot 10^{-8}$ ). Using this information, the coalescent simulation program MaCS [25] was internally called: it generated parental haplotypes with realistic amount of LD. As the data simulation yielded no consistent pattern of SNP dependence, the simulation of the parent population was repeated 100 times. The maternal LD in terms of  $r^2$  between adjacent SNPs was on average 0.45 and reflected high multicollinearity. In each repetition,  $N = 10$  sires were selected with best phenotypes with respect to  $\kappa = 1$  and mated with 1 000 dams. In order to resemble dairy cattle, one progeny per cross was simulated yielding 100 half-siblings per family. At each SNP, the major allele was coded as reference. Then, haplotypes of all selected sires, or a subset thereof if  $N = 1$  or  $N = 5$ , and maternal haplotypes of 1 000 progeny were used to set up the  $R$  matrix. Few loci with no variation were disregarded. Optimal sample size was estimated based on  $R$ . Separately for each  $\kappa$  but using the same parent generation,  $N$  males were selected based on their phenotype as sires of half-siblings in the progeny generation. The number of dams was determined according to optimal sample size required and assuming balanced family sizes. The simulation of the progeny generation was also repeated 100 times to estimate and test SNP effects for validation purposes; this yielded  $100 \times 100$  data sets in total. Heritability was  $h^2 \in \{0.1, 0.2, 0.3\}$  which was partitioned into  $\kappa$  QTL effects of equal size. QTL positions were drawn at random out of the segregating sites. For each  $h^2$ , data sets were simulated independently.

Additionally, to explore a direct relationship between positions of QTL signals and  $n_{\text{opt}}$ , we selected arbitrarily a single repetition of simulation with  $h^2 = 0.1$  and  $N = 10$ . For this particular data set, we determined  $n_{\text{opt}}$  for each SNP position (i.e., assuming one QTL signal) and for all possible SNP pairs (i.e., assuming two QTL signals).

The R package `asreml` version 3.0 [26] was used for association analysis. Other suitable R packages, such as `rrBLUP` [27] or `ridge` [28], had difficulties to converge or produced almost zero variance components due to the high multicollinearity of predictor variables. The multi-SNP model was applied to all simulated scenarios as described in [Multi-SNP model](#) section. Unlike in [Single-SNP model](#) section, the single-SNP model considered an additional factor  $u \sim N(0, A\sigma_a^2)$  that accounts for background genetic effects due to the relationship between individuals. This was modeled similarly, e.g., in `EMMAX` [29] but we used the numerator relationship matrix  $A$  for computational convenience. The pointwise testing of SNP effects was followed by  $p$ -value correction according to Benjamini & Hochberg [30].  $P$ -values from the multi-SNP model were not altered. The outcome was used to assess sensitivity and specificity of the multi-SNP and single-SNP model. For this, a window of 0.01 cM to both sides of a QTL signal (covering 2-3 SNPs) was specified in order to accept a significant SNP as a true positive result. Then, the true-positive rate (TPR) reflected sensitivity. Specificity was obtained as  $1 -$  the false-positive rate (FPR), and ROC curves were produced from TPR and FPR.

To evaluate how realistically the simulation of genetic data worked, empirical HD SNP chip data from the Dryad repository have been used [31]. These data included 1 151 dairy cows with no pedigree specification. We selected an arbitrary window on BTA7 comprising 300 SNPs on 1.16 Mbp and phased haplotypes of all animals using `AlphaPhase` [32]. We selected randomly 10 animals and marked them as sires in order to set up a matrix  $R$ . Because of the high SNP density, genetic distances were approximated linearly, i.e., 1 Mbp  $\approx$  1 cM. Maternal LD was roughly approximated from haplotype frequencies of all animals. Furthermore, this  $R$  matrix was used for the inspection of optimal sample size assuming  $\kappa = 1, \dots, 5$  QTL signals and  $h^2 = 0.1$  and following the workflow of [Multi-SNP model](#) section.

### Results

The optimal sample size suggested by the single-SNP model required the effective number of independent tests which was on average  $M_{\text{eff}} = 53$  if  $h^2 = 0.1$  and rather constant for  $R$  set up from  $N = 1, 5$  or 10 sires ( $h^2 = 0.2$ :  $M_{\text{eff}} = 54$ ;  $h^2 = 0.3$ :  $M_{\text{eff}} = 56$ ). Hence results are reported for  $M_{\text{eff}}$  based on  $N = 10$ . Table 1 presents the

**Table 1** Median of optimal sample size for detecting different number of QTL signals from 100 repetitions of simulations

QTL	$h^2 = 0.1$				$h^2 = 0.2$				$h^2 = 0.3$			
	$N = 1$	$N = 5$	$N = 10$	Single	$N = 1$	$N = 5$	$N = 10$	Single	$N = 1$	$N = 5$	$N = 10$	Single
1	128	126	127	195	57	58	57	91	34	33	34	56
2	275	269	273	382	125	126	122	175	73	70	73	106
3	421	426	436	569	214	201	205	259	126	120	120	155
4	613	540	584	756	291	288	281	342	177	170	170	204
5	763	713	685	943	385	349	344	426	228	208	207	253

Results are based on the multi-SNP approach ( $N = 1, 5, 10$  families) or single-SNP approach. In each repetition, sample size was repeatedly determined for randomly drawn QTL positions and the median was calculated

median of  $n_{\text{opt}}^*$  from 100 repetitions of simulation. The median increased almost linearly with number of QTL signals but reduced with increasing heritability, and it was rather unaffected by the number of families. As an example, 127 individuals were required to fine-map a single QTL signal based on the multi-SNP model if  $h^2 = 0.1$ . Almost twice as much were required to distinguish two signals if  $h^2 = 0.1$  or only 34 individuals were required to detect a single signal correctly when  $h^2 = 0.3$  instead of  $h^2 = 0.1$ . Optimal sample size suggested by the multi-SNP model was 17% to 39% less than estimated from the single-SNP model. Figure S.1 (Additional file 1) visualizes the dependence of optimal sample size estimated from the single-SNP model on heritability. It also shows that a much larger sample was required if QTL heritability was less than 0.2.

In case of  $h^2 = 0.1$ , the distribution of  $n_{\text{opt}}$  is represented in Fig. 1; a separate panel is shown for each number of QTL signals to be detected. Based on  $100 \times 100$  estimates of  $n_{\text{opt}}$ , we derived a bimodal distribution of optimal sample size in the multi-SNP model. The median of  $n_{\text{opt}}$  was consistently less than sample size estimated from single-SNP investigations. With increasing heritability, the first mode approached the median of  $n_{\text{opt}}$  but was still less than optimal sample size based on the single-SNP model, see Figures S.2 ( $h^2 = 0.2$ ) and S.3 ( $h^2 = 0.3$ ) (Additional file 1). The second mode appeared due to strong negative correlations between SNPs. Particularly this outcome was observed when all possible pairs of SNPs were evaluated for detecting two QTL signals in a single repetition of simulation. Figure 2a shows the correlation matrix for a single data set. Those entries of  $R$  have been selected that belonged to 10% of the highest estimates of sample size, i.e.,  $n_{\text{opt}} \geq 864$  ( $h^2 = 0.1$ ). Correspondingly, Fig. 2b indicates that, with few exceptions, negative correlations caused this outcome. The separation of SNP dependence into maternal and paternal contribution revealed further insight, and most often negative maternal LD was the driving term (Fig. S.4, Additional file 1). The distance between two QTL signals hardly influenced  $n_{\text{opt}}$  (Fig S.5, Additional file 1); any possible association was overlaid by

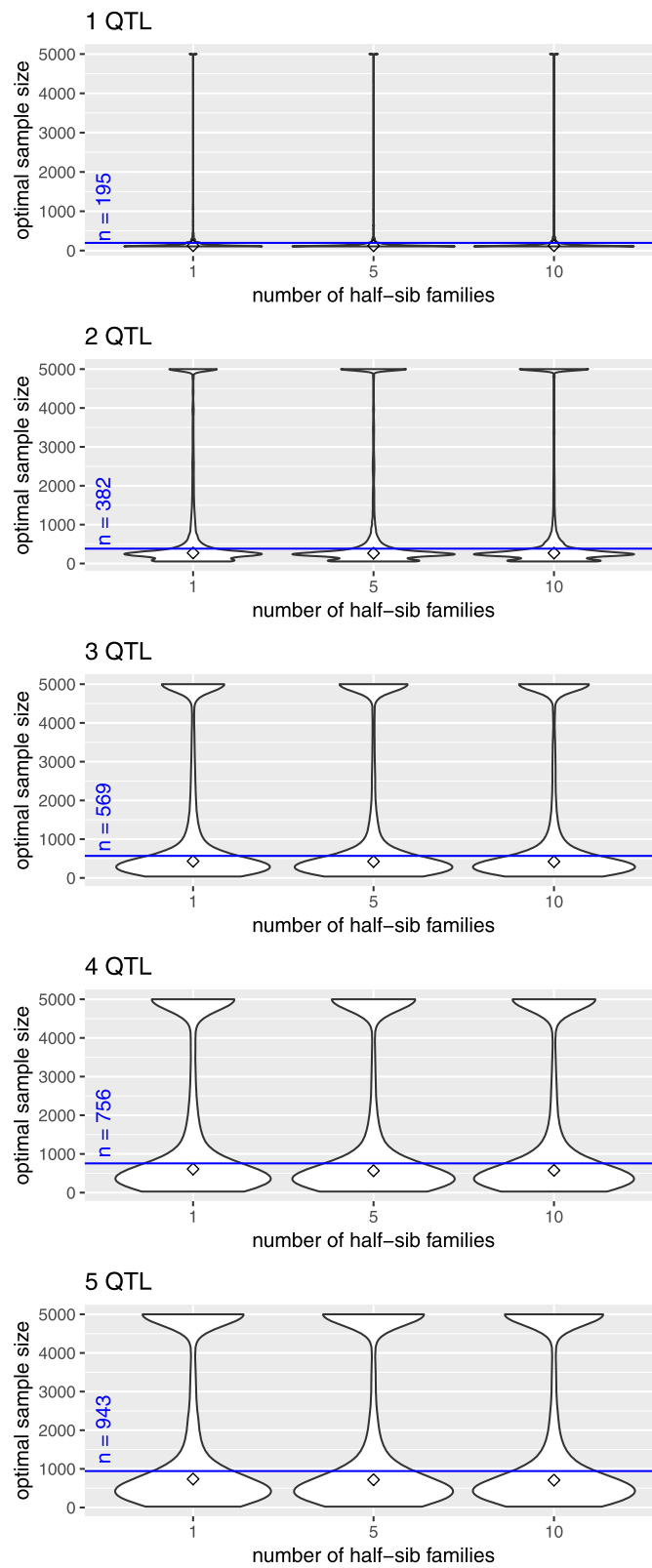
the strong impact of correlation between loci. An additional inspection of the relationship between position of a single QTL signal and  $n_{\text{opt}}$  was not conclusive. Neither extreme maternal allele frequency nor missing sire heterozygosity led to obviously increased  $n_{\text{opt}}$  for detecting one QTL signal (Fig. S.6, Additional file 1).

The association analysis of data sets of optimal sample size was validated in terms of sensitivity and specificity of testing SNP effects. The shape of ROC curves was similar for all investigated simulation scenarios. As an example, if  $N = 10$  and  $\kappa = 2$ , the median of  $n_{\text{opt}}^*$  was 273, and the outcome is displayed in Fig. 3. The analysis showed superiority of the multi-SNP model over the single-SNP model. In general, it was observed that the smaller  $n_{\text{opt}}^*$  was estimated, the larger both TPR and FPR turned out for the single-SNP model. The multi-SNP model performed rather robust against changes in sample size. However, the flat appearance of the ROC curve complicates fine-mapping of QTL signals based on the suggested multi-SNP approach. For instance, a TPR of 80% is accompanied with a FPR larger than 20%.

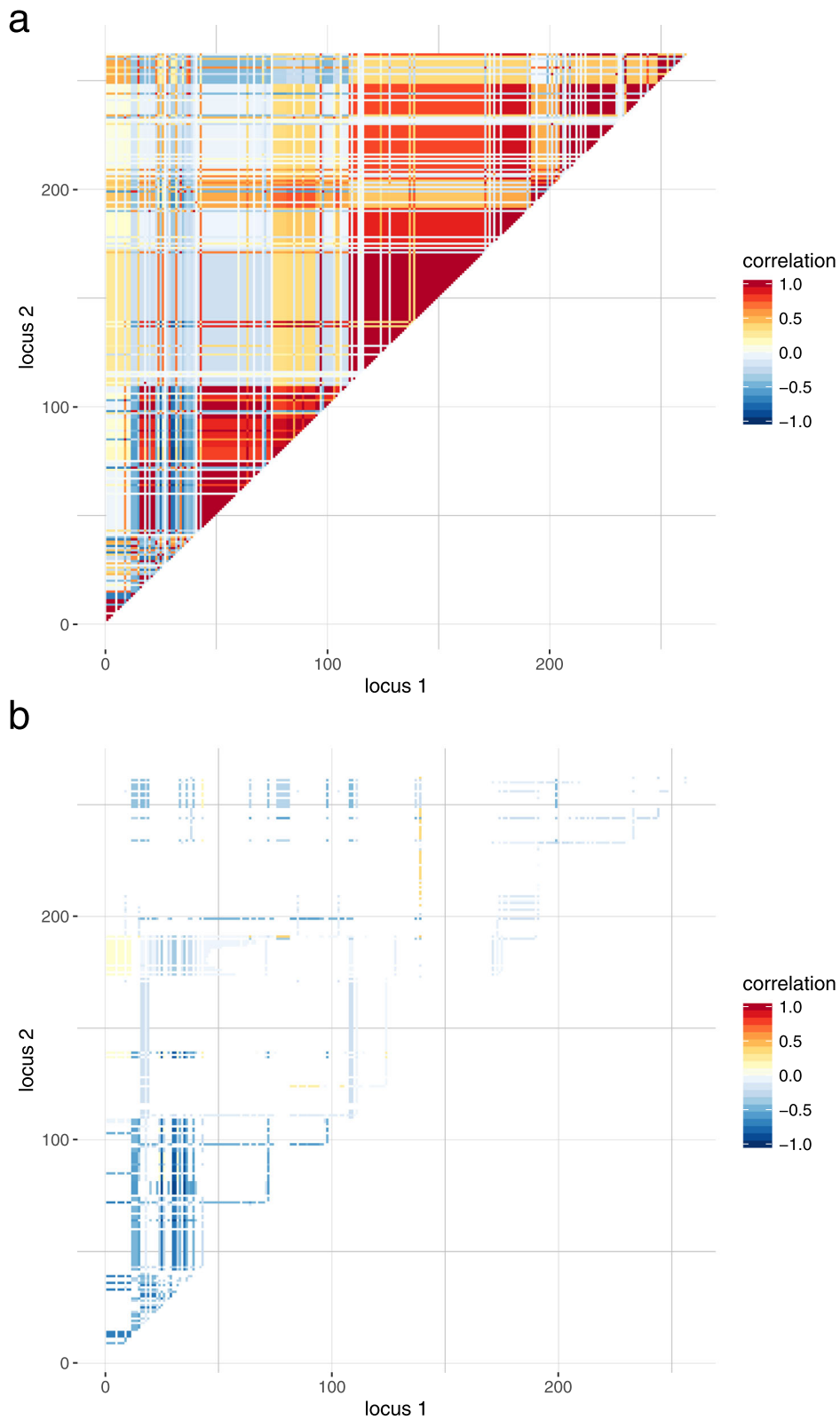
Blocks of varying linkage phases, as shown in Fig. 2, might be an artifact of data simulation. Based on empirical bovine HD SNP chip data, a possible  $R$  matrix was set up, see Fig. 4. The blocking structure was less pronounced. Using this  $R$  for estimating  $n_{\text{opt}}^*$  led to results being similar to the simulation study for one and two QTL signals but larger samples were required to detect more QTL signals:  $n_{\text{opt}}^*$  was 123 (1 signal), 288 (2 signals), 516 (3 signals), 800 (4 signals) and 1342 (5 signals) if  $h^2 = 0.1$ . The number of repetitions of randomly drawing the positions of QTL signals did not substantially affect the final  $n_{\text{opt}}^*$ . For instance, the median deviated less than 4% if  $n_{\text{opt}}^*$  was calculated 1000 instead of 100 times.

## Discussion

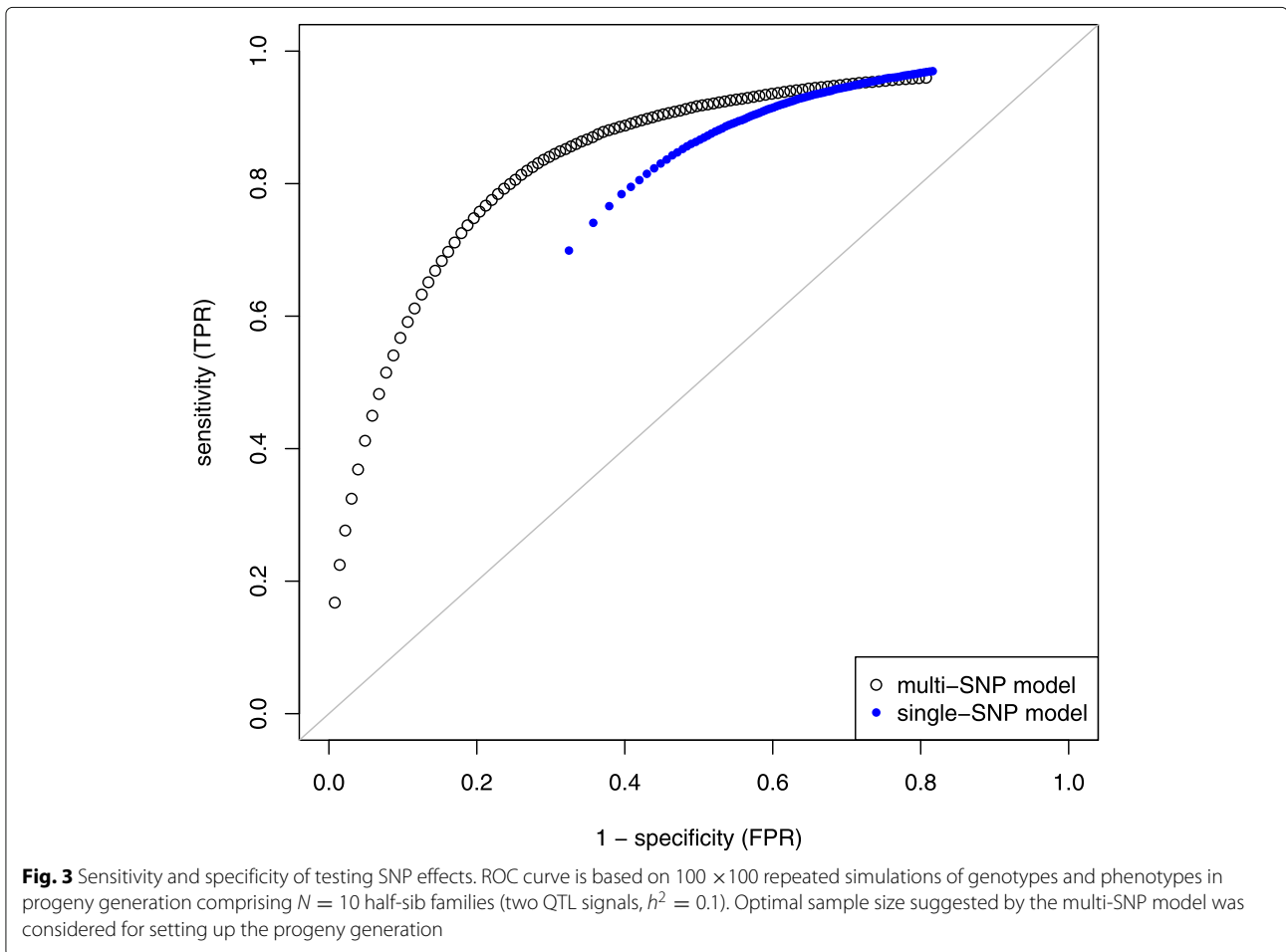
Our investigation contributes to the design of powerful experiments for fine-mapping of causative variant(s) in a genomic target region. We incorporated the expected dependence among SNPs in this region and estimated



**Fig. 1** Distribution of optimal sample size. Violinplot of  $n_{opt}$  vs. number of half-sib families for different numbers of QTL signals in a multi-SNP model. The parent generation was simulated 100 times and 100 random draws of positions of QTL signals were analyzed in each run,  $h^2 = 0.1$ . The diamond indicates the median of  $n_{opt}$  and the blue line marks the results based on a single-SNP model



**Fig. 2** Dependence between SNPs in a single simulated data set with  $N = 10$  sires. **a** Correlation matrix  $R$ , **b** entries selected from  $R$  which belong to 10% highest sample size ( $n_{opt} \geq 864$ ). All possible SNP pairs were evaluated to detect two QTL signals ( $h^2 = 0.1$ )



optimal sample size based on a SNP-BLUP approach. The outcome was compared to a single-SNP model. Negative correlations between SNPs, which were mainly due to negative maternal LD, caused essentially inflated sample size estimates. In case of positive correlations, the majority of sample size estimates was less than sample size estimated from the single-SNP approach. The less the heritability, the higher the deviation between models was.

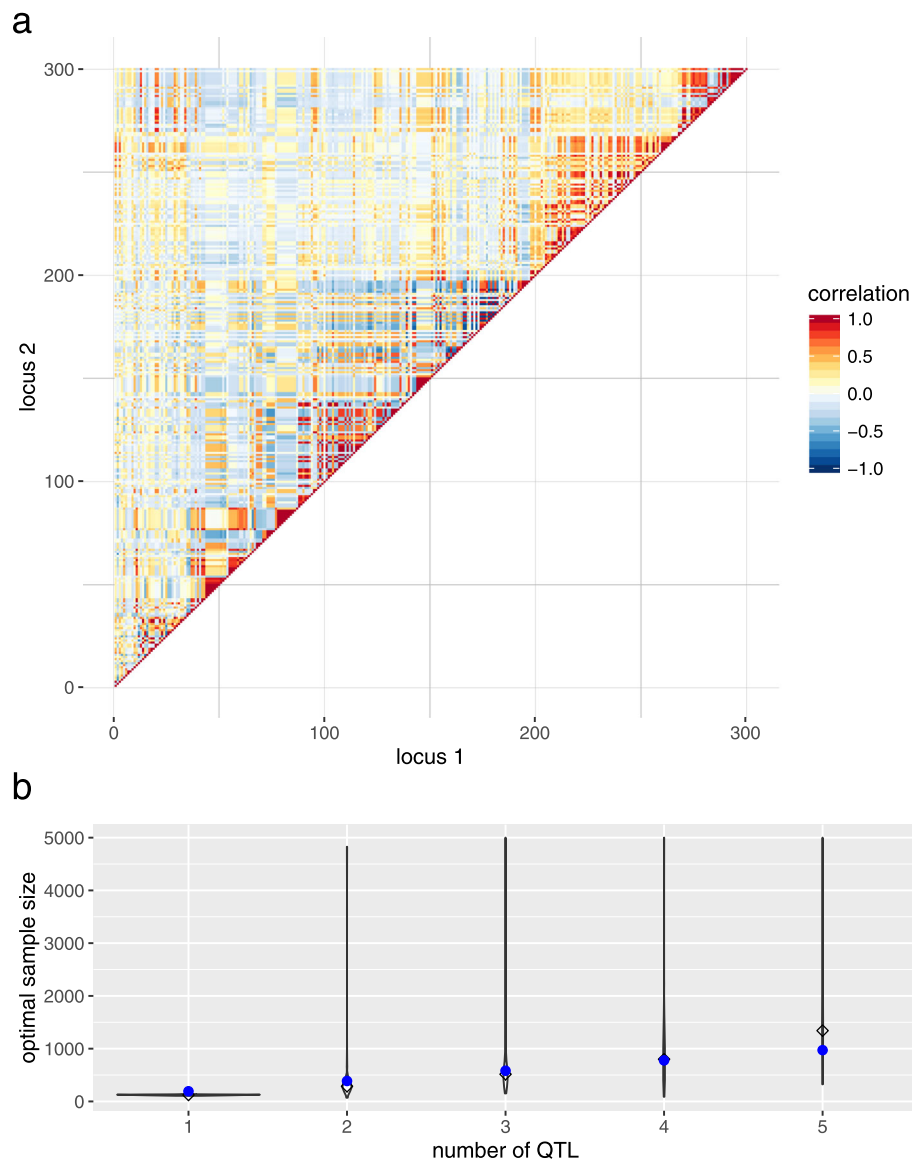
#### Population parameters

Our approach is applicable to any population structure. The matrix  $K$  of covariance between SNPs can be set up for any kind of family stratification by adapting the derivations of the Appendix or, in case of unrelated individuals, by using population LD in  $K$ .

Due to the way of model parametrization (columns  $X_k$  have been scaled), the dependence on allele frequency has been excluded. For instance, in a random mating population, the column-scaling term is  $\sqrt{2p_k(1-p_k)}$  with allele frequency  $p_k$  at SNP  $k$ . Likewise, a scaling term

can be derived for half-sib families as the square root of Eq. (7) in the Appendix by investigating maternally and paternally inherited SNP alleles separately. Results of our association analyses suggested that there was no clear relationship between high  $n_{\text{opt}}$  and maternal allele frequency or sire heterozygosity (Fig. S.6, Additional file 1). However, regions with large or low variation have to be taken into account when selecting sires for fine-mapping of QTL signals in a follow-up experiment. The lower sire heterozygosity or maternal minor allele frequency is, the lower the effect size  $\beta_k$  on the model scale will be and, consequently, higher  $n_{\text{opt}}^*$  is required in order to detect QTL signals. To investigate this, we employed the relationship  $X_k\beta_k = X_k^{(o)}\beta_k^{(o)}$  at SNP  $k$ . Here  $X_{j,k}^{(o)}$  is the allele count at SNP  $k$  for individual  $j$ , and  $\beta_k^{(o)}$  is the coefficient on the observed genotype scale, i.e.,  $\beta_k^{(o)}s_k = \beta_k$  with scaling term  $s_k = \sqrt{V(X_{j,k})}$ . The relationship between allele frequency and optimal sample size for detecting one QTL signal based on the single-SNP model is presented in Figure S.7 (Additional file 1). Extreme alleles, roughly





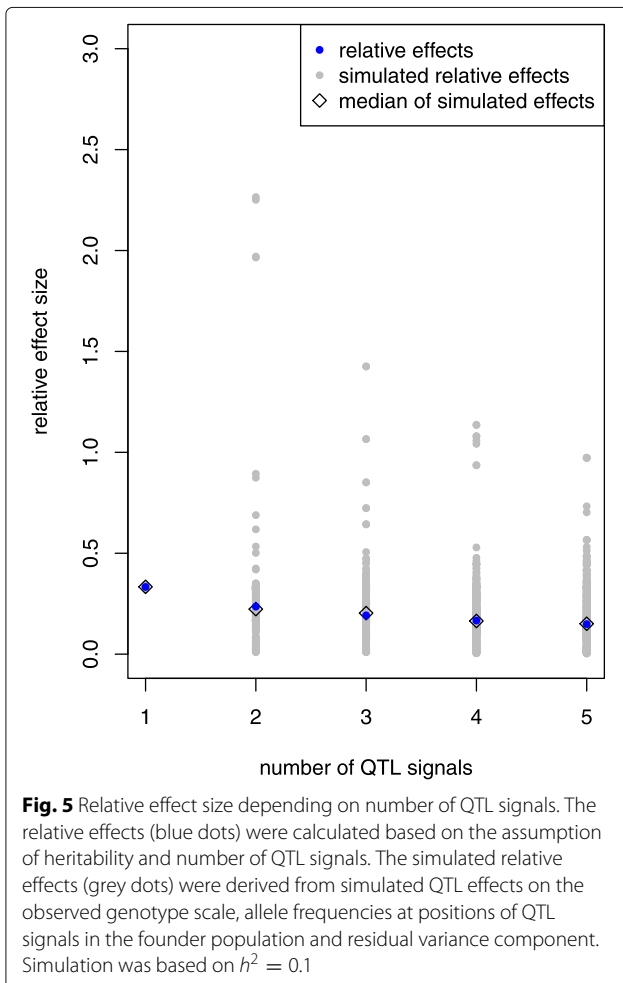
**Fig. 4** Empirical bovine HD SNP chip data. **a** Correlation matrix for a randomly selected window containing 300 SNPs on BTA7. **b** Violinplot of  $n_{\text{opt}}$  vs. number of QTL signals to be detected. The diamond indicates the median of  $n_{\text{opt}}$  and the blue dots mark the results based on a single-SNP model,  $N = 10$  and  $h^2 = 0.1$

spoken with major allele frequency  $> 0.95$ , require drastically increased sample size.

In the simulation study, equal effects of QTL signals were simulated on the observed genotype scale. For comparison with relative effect sizes derived from heritability (Eqn. 4), allele frequencies in the (random mating) founder population and the residual variance component were used to calculate the corresponding relative effect sizes:  $\beta_k^{(0)} \sqrt{2p_k(1-p_k)}/\sigma_e$ . Figure 5 shows the effect sizes of all repetitions of simulation with  $h^2 = 0.1$  separated by the number of simulated QTL signals ( $\kappa$ ). As expected, the

relative effect size decreased with increasing  $\kappa$  but a high fluctuation has been observed which was due to the high variation of allele frequencies in the simulated data. This observation underlines the difficulty of detecting multiple QTL signals at given  $h^2$  – the lower the effect size, the higher  $n_{\text{opt}}$  required.

The suggested optimal sample size is divided into  $N$  sires which are selected for most heterogeneity in the target region. The actual number of sires is of minor importance. The choice of individuals depends on the objectives of the follow-up study. Sires can be chosen independently



from the GWAS population in order to confirm and fine-map QTL signal(s). However, if the initial study indicated the presence of rare variants, sires under suspicion should be re-used. Selective genotyping is an option to increase power [11] but this might have negative impact on reproducibility of the study design [4]. In our investigation of paternal half-sib families, mothers are treated as random samples from a dam population. Thus, the choice of dams for future matings is not addressed here but is definitely an issue for other family designs.

Being equally important for fine-mapping of QTL signals is the positive correlation between SNPs. Positive correlatedness is a matter of genotype coding. Coding has to be consistent throughout the target region to avoid unnecessary sign changes in correlation. We employed coding in terms of counting the major allele in the population. But in regions of intermediate frequency, the coding might not be appropriate and hence a dynamic approach of coding the SNP alleles can circumvent negative correlations. A strategy on this is worth further investigation.

Power calculations are needed to quantify and judge the prospects of identifying causative variants with a hypothesized effect size in a particular population. In practice, however, experiments are usually not planned to obtain maximum power but data are regularly collected for purposes of breeding as a standard routine. Thus, experimental designs being theoretically optimal could be compared with available field data to understand the possible shortcomings of such data and to understand differences between theoretical/expected and actually achieved power. Based on the results, decisions can be made whether the amount of data is sufficient or, in case of underpowered experiments, more data should be acquired.

#### Necessity of fine-mapping of QTL signals using an appropriate design

The QTL databases of livestock species [33] contain information on several thousands QTL for a wide range of traits. This shows that the variability of most of the traits studied has a polygenic origin, with multiple QTL contributing to the overall genetic variance. Despite the number of QTL, only a handful of causal mutations could be detected and verified in the different livestock species [34]. This is partly due to the fact that GWAS show considerable weaknesses in the fine-mapping of QTL signals which are related to the SNP panel requirements for a genomewide distribution and high LD to neighboring markers [5]. Accordingly, these SNPs are usually indicative of a large genomic region that likely comprises the unmeasured causal SNP but does not provide information about the causal variant itself. Statistical methods for fine-mapping have been designed to overcome these issues and perform fine-mapping using the available SNP information from a SNP-chip or GBS (summarized by [5]). However, even these methods require a high SNP density in the region of interest, which favors a targeted sequencing strategy that enable the dissection of QTL regions and increase the chance of detecting causal variants [35]. Major factors to be considered for designing a targeted sequencing study are effect size, the number of causal SNPs, local LD structure and sample size [5]. The approach proposed in this study incorporates information on  $\kappa$ ,  $h^2$  and  $R$  derived from the data to estimate the optimal sample size ( $n_{opt}^*$ ) and thus provides all the information needed to design a fine-mapping experiment.

Currently, several fine-mapping studies are based on imputation strategies or the integration of results with functional enrichment analysis to identify promising candidate genes and QTNs (e.g., [36–38]). These approaches largely depend on imputation accuracy and the status of genome annotation, thus limiting the ability to detect causal variants, especially those with a low minor allele frequency [39]. Specific examples for the fine-mapping of



important genomic regions with a resequencing strategy are still rare nowadays. Fraser et al. [40] focused in their study on collagenous lectins in horses by resequencing 658 kb DNA consisting of different candidate genes and regulatory regions. Therefore, a case-control design with pooled samples was used and with this approach 113 variants were identified, which differed between the groups. Although the results are promising, the authors concluded that a large-scale genotyping of individual samples is necessary for deeper insights. In this context, and considering that targeted sequencing for a reasonable set of samples is becoming increasingly affordable, an accurate estimate of sample size is advisable.

**Other random effects**

Association analysis of empirical data with certain pedigree structure requires an additional model term to account for genetic effects beyond the target window ( $Zu$ ). Then  $u = (u_1, \dots, u_n)$ ,  $u \sim N(0, G\sigma_u^2)$ , is the vector of individual genetic effects with suitable relationship matrix  $G$ . The calculation of optimal sample size should consider the presence of additional random effects (genetic or environmental) for the design of experiments. For instance, the coefficients of the single-SNP model could be estimated via BLUE. This affects sample size calculation because the variance of the estimator  $\hat{\beta}_k$ ,

$$V(\hat{\beta}_k) = (X_k' V^{-1} X_k)^{-1} \quad \text{with} \quad V = ZZ' \sigma_u^2 + I_n \sigma_e^2,$$

has impact on the distribution of the test statistic. Accounting for  $V$  in the denominator of test statistic increases the denominator of non-centrality parameter. However, in order to keep it simple, it would be sufficient to increase  $\sigma_e^2$  or reduce  $h^2$  appropriately without any other alterations.

It is possible to consider other kinds of genetic effects with the proposed methods. For instance, exploring dominance genetic effects requires only one modification. Instead of coding SNP genotypes for additive effects via  $X_{j,k} \in \{1, 0, -1\}$ , genotypes can be coded as  $X_{j,k} \in \{0, 1, 0\}$  to account for dominance effects. The covariance between dominance effects has been worked out by Bonk et al. [41]. Feeding the Equation (3) with the corresponding dominance correlation matrix will provide estimates of optimal sample size to fine-map QTL signals with dominance effect.

**Conclusion**

For planning the design of experiment, we recommend a multi-SNP approach which considers the expected dependence among SNPs. Compared to a conventional approach, this leads to a reduced estimate of sample size and thus promises a more efficient use of animal resources. The benefit depends strongly on heritability: the lower heritability, the more resources can be saved.

In general, optimal sample size increases almost linearly with the number of QTL signals to be detected. This study constitutes a frequentist framework for the design of experiments in specific populations that may be characterized by family stratification. It will help differentiating independent signals in QTL regions that can be further examined for cellular and molecular properties.

**Appendix: Derivation of correlation matrix**

We study the dependence between pairs of SNPs, each with two alleles A and B, in a population consisting of  $N$  paternal half-sib families. Let  $X_{j,k}$  be the genotype code at SNP  $k \in \{1, \dots, p\}$  of individual  $j \in \{1, \dots, n\}$  being progeny of sire  $s$  and dam  $d$ . Homozygous genotypes A/A and B/B are coded as 1 and -1, respectively, and the heterozygous genotype A/B is indicated as 0. The family-specific (i.e., sire-specific) covariance between SNPs  $k$  and  $l$  of individual  $j$  is, according to Bonk et al. [41] and Wittenburg et al. [42],

$$\begin{aligned} K_{k,l}^s &= E(X_{j,k} X_{j,l} | \mathcal{S}_s) - E(X_{j,k} | \mathcal{S}_s) E(X_{j,l} | \mathcal{S}_s) \\ &= D_{k,l}^d + D_{k,l}^s \end{aligned}$$

a function of maternal and paternal contribution and depends on the sire diplototype  $\mathcal{S}_s$ . The  $D_{k,l}^d$  denotes the LD of maternal gametes in a dam population. The sire term depends on the phase of paternal haplotypes and recombination rate ( $\theta_{k,l}$ ). It is determined as

$$D_{k,l}^s = \begin{cases} \frac{1}{4}(1 - 2\theta_{k,l}), & \text{for sire with haplotypes A-A and B-B} \\ -\frac{1}{4}(1 - 2\theta_{k,l}), & \text{for sire with haplotypes A-B and B-A} \\ 0, & \text{else.} \end{cases} \quad (5)$$

To achieve the covariance between a pair of SNPs, we employ conditioning on families,

$$\begin{aligned} E(X_{j,k} X_{j,l}) &= \sum_{s=1}^N \Pr(\mathcal{S}_s) E(X_{j,k} X_{j,l} | \mathcal{S}_s) \\ E(X_{j,k}) &= \sum_{s=1}^N \Pr(\mathcal{S}_s) E(X_{j,k} | \mathcal{S}_s) \\ \text{cov}(X_{j,k}, X_{j,l}) &= \sum_{s=1}^N w_s E(X_{j,k} X_{j,l} | \mathcal{S}_s) \\ &\quad - \sum_{s=1}^N w_s E(X_{j,k} | \mathcal{S}_s) \sum_{s=1}^N w_s E(X_{j,l} | \mathcal{S}_s) \end{aligned}$$

and approximate  $\Pr(\mathcal{S}_s)$  by family weights  $w_s = \frac{n_s}{n}$  with  $\sum_{s=1}^N n_s = n$ . The aim is now to derive an expression that depends on already known terms. For instance, using

$$E(X_{j,k} X_{j,l} | \mathcal{S}_s) = K_{k,l}^s + E(X_{j,k} | \mathcal{S}_s) E(X_{j,l} | \mathcal{S}_s)$$

yields

$$E(X_{j,k}X_{j,l}) = \sum_{s=1}^N w_s (K_{k,l}^s + E(X_{j,k}|\mathcal{S}_s)E(X_{j,l}|\mathcal{S}_s)).$$

We exploit the separation into independently inherited maternal and paternal SNP alleles:  $X_{j,k} = X_{j,k,s} + X_{j,k,d}$ , where  $X_{j,k,s}$  and  $X_{j,k,d}$  take a value of  $\frac{1}{2}$  if the A allele was inherited but  $-\frac{1}{2}$  otherwise. Then

$$\begin{aligned} E(X_{j,k}|\mathcal{S}_s) &= E(X_{j,k,d}|\mathcal{S}_s) + E(X_{j,k,s}|\mathcal{S}_s) \\ E(X_{j,k,d}|\mathcal{S}_s) &= E(X_{j,k,d}) = p_k - \frac{1}{2}, \end{aligned}$$

where  $p_k$  denotes the maternal allele frequency at SNP  $k$ . Furthermore,

$$E(X_{j,k,s}|\mathcal{S}_s) = \begin{cases} \frac{1}{2}, & \text{for sire genotype A/A} \\ 0, & \text{for sire genotype A/B} \\ -\frac{1}{2}, & \text{for sire genotype B/B.} \end{cases} \quad (6)$$

Putting it all together,

$$\begin{aligned} K_{k,l} &= \sum_{s=1}^N w_s (D_{k,l}^d + D_{k,l}^s) \\ &+ \sum_{s=1}^N w_s \left[ \left( p_k - \frac{1}{2} \right) + E(X_{j,k,s}|\mathcal{S}_s) \right] \\ &\quad \left[ \left( p_l - \frac{1}{2} \right) + E(X_{j,l,s}|\mathcal{S}_s) \right] \\ &- \sum_{s=1}^N w_s \left[ \left( p_k - \frac{1}{2} \right) + E(X_{j,k,s}|\mathcal{S}_s) \right] \\ &\quad \sum_{s=1}^N w_s \left[ \left( p_l - \frac{1}{2} \right) + E(X_{j,l,s}|\mathcal{S}_s) \right]. \end{aligned}$$

This reduces to

$$\begin{aligned} K_{k,l} &= D_{k,l}^d + \sum_{s=1}^N w_s D_{k,l}^s + \sum_{s=1}^N w_s E(X_{j,k,s}|\mathcal{S}_s)E(X_{j,l,s}|\mathcal{S}_s) \\ &- \sum_{s=1}^N w_s E(X_{j,k,s}|\mathcal{S}_s) \sum_{s=1}^N w_s E(X_{j,l,s}|\mathcal{S}_s), \end{aligned}$$

and this is evaluated using the sire-specific terms in (5) and (6).

Now the variance of genotype codes at SNP  $k$  is derived explicitly – it also serves as a scaling term in a regression model for association analysis. The second moment of the paternally inherited SNP allele is constant  $E(X_{j,k,s}^2|\mathcal{S}_s) = \frac{1}{4}$  for all  $s$ . Hence

$$V(X_{j,k,s}) = \frac{1}{4} - \left( \sum_{s=1}^N w_s E(X_{j,k,s}|\mathcal{S}_s) \right)^2.$$

Then, the variance at SNP  $k$  is

$$\begin{aligned} V(X_{j,k}) &= V(X_{j,k,d}) + V(X_{j,k,s}) \\ &= p_k(1-p_k) + \frac{1}{4} - \left( \sum_{s=1}^N w_s E(X_{j,k,s}|\mathcal{S}_s) \right)^2 = K_{k,k}. \end{aligned} \quad (7)$$

Finally, the correlation matrix  $R = \{R_{k,l}\}_{k,l=1,\dots,p}$  is calculated by scaling the entries correspondingly,

$$R_{k,l} = \frac{K_{k,l}}{\sqrt{K_{k,k}K_{l,l}}}.$$

Note that the covariance based on non-centered genotype codes (as derived above) is identical to the one based on centered genotype codes (as used in [Methods](#) section). Centering is used to study within-family genetic effects, and it allows the direct estimation of allele substitution effects [43].

The R package `hscovar` for the calculation of  $K$  and  $R$  is provided at CRAN repository.

### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12863-020-00871-1>.

**Additional file 1:** PDF file containing additional figures.

**Additional file 2:** R script for data simulation and analysis. Running this script performs the complete data analysis and evaluation. It calls functions from Additional file 3.

**Additional file 3:** R script including functions used for data simulation and analysis.

### Abbreviations

ANOVA: Analysis of variance; BLUE: Best linear unbiased estimation; BTA: Bos taurus autosome; cM: CentiMorgan; DNA: Deoxyribonucleic acid; FPR: False positive rate; (G)BLUP: (Genomic) best linear unbiased prediction; GBS: Genotyping by sequencing; GWAS: Genomewide association study; HD: High density; LD: Linkage disequilibrium; Mbp: Mega base pairs; QTL: Quantitative trait locus; QTN: Quantitative trait nucleotide; ROC: Receiver operating characteristic; SNP: Single nucleotide polymorphism; TPR: True positive rate

### Acknowledgments

Special thanks are given to N. Reinsch (Leibniz Institute for Farm Animal Biology, Dummerstorf), J. Hartung (University of Hohenheim, Stuttgart) and V. Liebscher (University of Greifswald), who contributed invaluable ideas to the project, and to C. Gaynor (Roslin Institute, Edinburgh) for giving handy insight into `AlphaSimR`. We also thank the reviewers for their helpful comments.

### Authors' contributions

DW developed the theory, implemented the statistical methods, performed the analysis, and wrote the manuscript. SB contributed to the research on covariance between SNPs, MD was involved in theoretical investigations. HR contributed to the discussion. All authors have read and approved the final manuscript.

### Funding

The project was funded by the German Research Foundation (DFG, WI 4450/1-1). The funder had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### Availability of data and materials

R scripts used to simulate and analyze data are available in Additional files 2 and 3. However, an exact reproduction of the results of our simulation study is not possible because `AlphaSimR` relies on random number generation

which cannot be initialized yet. The R package `hscovar` version 0.2.1 is available at CRAN; it provides tools for setting up the covariance or correlation matrix as well as for performing power calculations. The empirical bovine HD SNP chip data are accessible through the Dryad repository <https://doi.org/10.5061/dryad.519bm> [31].

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Leibniz Institute for Farm Animal Biology, Institute of Genetics and Biometry, 18196 Dummerstorf, Germany. <sup>2</sup>University Medicine Greifswald, Department of Psychiatry and Psychotherapy, 17475 Greifswald, Germany. <sup>3</sup>Leibniz Institute for Farm Animal Biology, Institute of Genome Biology, 18196 Dummerstorf, Germany.

Received: 21 February 2020 Accepted: 9 June 2020

Published online: 29 June 2020

#### References

- Reyer H, Hawken R, Murani E, Ponsuksili S, Wimmers K. The genetics of feed conversion efficiency traits in a commercial broiler line. *Sci Rep*. 2015;5:16387.
- Sahana G, Guldbrandtsen B, Thomsen B, Holm LE, Panitz F, Brøndum RF, et al. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. *J Dairy Sci*. 2014;97(11):7258–75.
- Hampel A, Teuscher F, Gomez-Raya L, Doschoris M, Wittenburg D. Estimation of recombination rate and maternal linkage disequilibrium in half-sibs. *Front Genet*. 2018;9:186.
- Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5–23.
- Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*. 2018;19(8):491–504.
- Gauderman J, Morrison J. QUANTO Version 1.2. 2007. Retrieved June 10, 2015. Available from: <http://biostats.usc.edu/Quanto.html>.
- Schnabel R. ARS-UCD1.2 Cow Genome Assembly: Mapping of all existing variants. 2018. Retrieved Sep 21, 2018. Available from: [https://www.animalgenome.org/repository/cattle/UMC\\_bovine\\_coordinates/](https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/).
- Luo Z. Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity*. 1998;80(2):198.
- Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*. 2001;69(1):1–14.
- Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JA, Barris W, et al. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics*. 2008;9(1):187.
- Weller J. Quantitative trait loci analysis in animals: CABI Publishing; 2001. <https://doi.org/10.1079/9781845934675.0000>.
- Gualdrón Duarte JL, Cantet RJ, Bates RO, Ernst CW, Raney NE, Steibel JP. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinf*. 2014;15(1):246. Available from: <https://doi.org/10.1186/1471-2105-15-246>.
- Koivula M, Strandén I, Su G, Mäntysaari EA. Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *J Dairy Sci*. 2012;95(7):4065–73.
- Mucha S, Mrode R, MacLaren-Lee I, Coffey M, Conington J. Estimation of genomic breeding values for milk yield in UK dairy goats. *J Dairy Sci*. 2015;98(11):8201–8.
- Maier R, Moser G, Chen GB, Ripke S, Absher D, Agartz I, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet*. 2015;96(2):283–94.
- Kristensen PS, Jahoor A, Andersen JR, Cericola F, Orabi J, Janss LL, et al. Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines. *Front Plant Sci*. 2018;9:69.
- Taskinen M, Mäntysaari EA, Strandén I. Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. *Genet Sel Evol*. 2017;49(1):36.
- Aguilar I, Legarra A, Cardoso F, Masuda Y, Lourenco D, Misztal I. Frequentist p-values for large-scale single step genome-wide association, with an application to birth weight in American Angus cattle. *Genet Sel Evol*. 2019;51(1):28.
- Searle S. Linear models. New York: Wiley; 1971.
- Hoerl AE, Kennard RW, Baldwin KF. Ridge regression: some simulations. *Commun Stat Theor M*. 1975;4(2):105–23.
- Cohen J. Statistical power analysis for the social sciences. Hillsdale: Erlbaum; 1988.
- Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*. 2008 May;32:361–9.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna; 2019. Retrieved Dec 16, 2019. Available from: <https://www.R-project.org/>.
- Faux AM, Gorjanc G, Gaynor RC, Battagin M, Edwards SM, Wilson DL, et al. AlphaSim: software for breeding program simulation. *Plant Genome*. 2016;9(3):1–14. Available from: <https://doi.org/10.3835/plantgenome2016.02.0013>.
- Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. *Genome Res*. 2009;19(1):136–42.
- Butler D, Cullis BR, Gilmour A, Gogel B. ASReml-R reference manual. Brisbane: The State of Queensland, Department of Primary Industries and Fisheries; 2009.
- Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*. 2011;4(3):250–5.
- Cule E, Vineis P, De Iorio M. Significance testing in ridge regression for genetic data. *BMC Bioinf*. 2011;12:372.
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010;42(4):348.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*. 1995;57(1):289–300.
- Bermingham ML, Bishop SC, Woolliams JA, Pong-Wong R, Allen AR, McBride SH, et al. Data from: Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. Dryad, Dataset. 2013. Available from: <https://doi.org/10.5061/dryad.519bm>.
- Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JH. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet Sel Evol*. 2011;43(1):12.
- Hu ZL, Park CA, Wu XL, Reacy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res*. 2012;41(D1):D871–9.
- Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet*. 2004;5(3):202.
- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. *Nat Methods*. 2010;7(2):111.
- Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun Biol*. 2019;2(1):212.
- Cai Z, Guldbrandtsen B, Lund MS, Sahana G. Weighting sequence variants based on their annotation increases the power of genome-wide association studies in dairy cattle. *Genet Sel Evol*. 2019;51(1):20.
- Liu Z, Wang T, Pryce JE, MacLeod IM, Hayes BJ, Chamberlain AJ, et al. Fine-mapping sequence mutations with a major effect on oligosaccharide content in bovine milk. *Sci Rep*. 2019;9(1):2137.
- Dadaev T, Saunders EJ, Newcombe PJ, Anokian E, Leongamornlert DA, Brook MN, et al. Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nat Commun*. 2018;9(1):2256.

40. Fraser RS, Arroyo LG, Meyer A, Lillie BN. Identification of genetic variation in equine collagenous lectins using targeted resequencing. *Vet Immunol Immunopathol.* 2018;202:153–63.
41. Bonk S, Reichelt M, Teuscher F, Segelke D, Reinsch N. Mendelian sampling covariability of marker effects and genetic values. *Genet Sel Evol.* 2016;48(1):36.
42. Wittenburg D, Teuscher F, Klosa J, Reinsch N. Covariance between genotypic effects and its use for genomic inference in half-sib families. *G3 Genes Genom Genet.* 2016;6:2761–72.
43. Abecasis GR, Cardon LR, Cookson W. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet.* 2000;66(1):279–92.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)



## RESEARCH ARTICLE

## Open Access



# Grouping of genomic markers in populations with family structure

Dörte Wittenburg\* , Michael Doschoris and Jan Klosa

\*Correspondence:  
wittenburg@fhn-dummerstorf.de  
Institute of Genetics  
and Biometry, Leibniz Institute  
for Farm Animal Biology,  
18196 Dummerstorf, Germany

## Abstract

**Background:** Linkage and linkage disequilibrium (LD) between genome regions cause dependencies among genomic markers. Due to family stratification in populations with non-random mating in livestock or crop, the standard measures of population LD such as  $r^2$  may be biased. Grouping of markers according to their interdependence needs to account for the actual population structure in order to allow proper inference in genome-based evaluations.

**Results:** Given a matrix reflecting the strength of association between markers, groups are built successively using a greedy algorithm; largest groups are built at first. As an option, a representative marker is selected for each group. We provide an implementation of the grouping approach as a new function to the R package `hscovar`. This package enables the calculation of the theoretical covariance between biallelic markers for half- or full-sib families and the derivation of representative markers. In case studies, we have shown that the number of groups comprising dependent markers was smaller and representative SNPs were spread more uniformly over the investigated chromosome region when the family stratification was respected compared to a population-LD approach. In a simulation study, we observed that sensitivity and specificity of a genome-based association study improved if selection of representative markers took family structure into account.

**Conclusions:** Chromosome segments which frequently recombine in the underlying population can be identified from the matrix of pairwise dependence between markers. Representative markers can be exploited, for instance, for dimension reduction prior to a genome-based association study or the grouping structure itself can be employed in a grouped penalization approach.

**Keywords:** Single nucleotide polymorphism, Covariance matrix, Clustering, TagSNP, Group lasso, SNP-BLUP

## Background

Genomic markers are an invaluable source for characterizing genetic variety and to elucidate the relationship between genetic and phenotypic variation in breeding populations. Dependencies among genomic markers are caused by linkage and linkage disequilibrium (LD) between genome regions. Though this condition complicates investigations on which genetic variants are truly associated with trait expression



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

[1], dependencies can be advantageous for grouping of markers. For example, clustering based on a greedy algorithm [2], hierarchical clustering (e.g., [3]) or grouping via interval-graph modeling [4] exploit the presence of LD blocks which are regions of particularly high correlation. To allow for proper inferences of such approaches, a suitable measure for the strength of dependence is needed. For instance, measuring LD in terms of  $r^2$  [5] is a natural choice but it is meaningful only for populations without stratification. In livestock and crop breeding, however, populations are often characterized by strong family stratification due to non-random mating of selected individuals. As examples, large paternal half-sib families are typical for cattle populations whereas chicken or fish populations consist of full-sib families. In plant breeding, maternal half-sib families are often produced in, for instance, wheat and clover. Then, linkage between markers within family leads to haplotype frequencies among progeny that are not conclusive for estimating  $r^2$ . Hence, there is need to promote measures of marker dependence which takes into account the particular family structure.

Especially in situations of ultra-dense panels of single nucleotide polymorphisms (SNPs), it is often sufficient to investigate representative SNPs (“tagSNPs”) out of each cluster. This subset can help identifying trait-associated genome regions in genome-wide association studies and allows comparing genome characteristics between ethnics/species/breeds (e.g., [2]). As the choice of tagSNPs is a consequence of grouping, it is also influenced by the underlying population structure.

The objective of this paper is to exploit the family structure of a population for specifying groups of associated markers. We generalize the grouping approach of Carlson et al. [2] in order to allow binning of markers given a correlation matrix or any kind of similarity matrix with scaled entries in  $[0, 1]$ . We investigate three case studies and a simulation study. For each case study, we visually inspect the correlation matrix and link to the outcome of grouping. Usability for genome-based association studies is shown as one possible field of application. Results were compared to the commonly used population-LD approach which ignores family structure. We provide a new function to the R package `hscovar` (available at CRAN) that enables grouping of markers and selection of representative markers.

## Methods

The dependence between pairs of SNPs, each with two alleles A and B, can be expressed in terms of a covariance or correlation matrix. It has already been shown in the literature how to calculate the theoretical covariance between markers in a population consisting of half-sib families [1]. It requires a genetic map, haplotypes of the common parent and LD information (or haplotype frequencies) of the population the individual parent comes from. This approach can be extended to be applicable to full-sib families by adding the paternal and maternal contribution into a single covariance matrix; the derivation is summarized in Additional file 1. Hence, a covariance matrix can be derived for any family structure, and this constitutes the input of the following grouping approach.



**Grouping of markers**

We generalized the strategy of Carlson et al. [2] for binning markers and selecting representatives to be applicable to any symmetric matrix which reflects a measure of dependence between markers and has entries scaled in [0, 1]. In particular, we considered the correlation matrix  $R$ . The idea is that SNPs which are associated to each other are assigned to groups. Groups are built one after the other—the largest group at first. For  $b = 1, 2, \dots$ , the  $b$ -th group is identified by searching for the SNP that has most occurrences of absolute correlation to other SNPs larger than a given threshold  $t$ . More precisely, let  $\mathcal{S}_b$  denote the set of SNP indices which have not been binned yet. Then, for each SNP  $k \in \mathcal{S}_b$ , the set of highly associated SNPs is determined as (Step A)

$$\mathcal{C}_k = \{l \mid l \in \mathcal{S}_b : |R_{k,l}| > t\},$$

and a set with highest cardinality (operator #) is chosen,

$$c \in \arg \max_k \#\mathcal{C}_k.$$

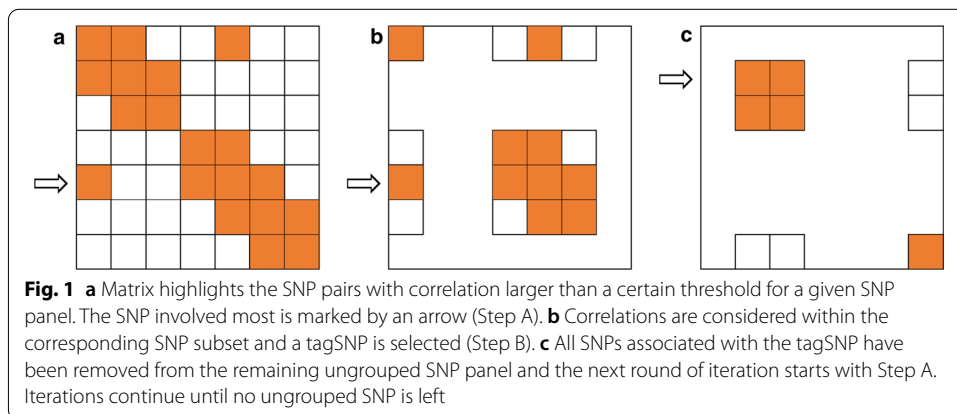
Thus,  $\mathcal{C}_c$  constitutes the  $b$ -th group, and a tagSNP is selected from this group. SNP  $c$  has strong correlation with any other SNP in  $\mathcal{C}_c$  but it can happen that also other SNPs of  $\mathcal{C}_c$  fulfill this criterion. Hence, a set of candidates is given by (Step B)

$$\mathcal{T} = \{k \mid k \in \mathcal{C}_c \wedge \forall l \in \mathcal{C}_c : |R_{k,l}| > t\}.$$

If more than one candidate remains, then the  $\lceil \#\mathcal{T}/2 \rceil$ -th SNP becomes the representative of group  $b$ . A next round of iteration is started using  $\mathcal{S}_{b+1} = \mathcal{S}_b \setminus \mathcal{C}_c$  until  $\mathcal{S}_{b+1} = \emptyset$ . The number of groups only depends on the threshold  $t$ . Similar to [2],  $t = 0.8$  is a suitable value. In an extreme case (with  $t$  approaching 1), each SNP builds a single group, yielding a complexity of this algorithm of  $\mathcal{O}(p^2)$ , with  $p$  the total number of SNPs. This approach is implemented as function tagSNP in the R package hscovar. A graphical representation of this algorithm is shown in Fig. 1.

**Evaluation**

It is an obvious choice to compare the family approach with a population-LD approach. Such an approach requires the population frequency of the different haplotypes ( $f_{A-A}, f_{A-B}$



$f_{B-A}$ ,  $f_{B-B}$ ) from which LD between markers is computed in terms of  $r^2$  according to [5]. For any marker pair  $k, l$  with allele frequencies  $f_k = f_{A-A} + f_{A-B}$  and  $f_l = f_{A-A} + f_{B-A}$ , we have

$$r_{k,l}^2 = \frac{(f_{A-A}f_{B-B} - f_{A-B}f_{B-A})^2}{f_k(1-f_k)f_l(1-f_l)}.$$

The LD matrix can be computed from progeny genotypes using the function `ld` from the R package `snpStats` version 1.38.0 [6]. This function undertakes phasing of genotypes using a maximum-likelihood approach [7]. Based on the LD matrix containing  $r^2$ , representative markers can be derived as described above.

Family and population-LD approach were compared using the Calinski–Harabasz (CH) index [8] which measures the cluster quality with respect to inter- and intra-cluster distances. The method with higher CH index performed better. For this, the function `calin-hara` from the R package `fpc` version 2.2-9 was applied to the groups obtained; the quality referred to distances based on the genotype matrix centered within family and scaled (as described below). Furthermore, we present the pure number of groups and highlight those groups with group size of at least three. This also helped visualizing the location of corresponding representative markers.

Selecting a representative set of markers is a natural tool for dimension reduction prior to genomic evaluations in order to reduce the impact of multicollinearity among predictor variables. Representative SNPs capture cumulative effects of the corresponding LD blocks on trait expression. We investigated a SNP-BLUP approach, which is widely used in genomic evaluations (e.g., [9]), and thereby demonstrate one possible application of the suggested approach. Representative markers selected from the family or from the population-LD approach were employed as predictor variables in a regression model

$$y = X\beta + e,$$

with  $y = (y_1, \dots, y_n)^\top$  the phenotype vector,  $\beta = (\beta_1, \dots, \beta_\tau)^\top$  the vector of genetic effects captured by  $\tau$  tagSNPs, the corresponding design matrix  $X$  with dimensions  $n \times \tau$  including the genotype codes in terms of major allele counts. The columns of  $X$  and the vector  $y$  were centered within family and scaled to obtain an empirical variance of one. The residual errors were assumed to be independently and normally distributed. For convenience, no other effects were assumed. We used the R package `asreml` version 3.0 [10] to estimate the vector of regression coefficients as

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y$$

where the shrinkage parameter  $\lambda$  was estimated via AI-REML. Significance of the  $k$ -th SNP effect was tested by a  $t$ -like test statistic as in [1],

$$T_k = \frac{\hat{\beta}_k}{SD(\hat{\beta}_k)}.$$

Significance was reported if  $T_k \geq q_{1-\alpha/2}$  or  $T_k < q_{\alpha/2}$  using the  $1 - \alpha/2$  and  $\alpha/2$  quantile of the standard normal distribution. The SNP-BLUP approach was evaluated in terms of sensitivity (i.e., true-positive rate) and specificity (i.e.,  $1 - \text{false-positive rate}$ ) over a range



of type-I error  $\alpha$ . We additionally verified the impact of threshold  $t \in \{0.5, 0.6, 0.7, 0.8\}$  on grouping and its consequences on the performance of the SNP-BLUP approach.

### Data

The data sets used for studying dependencies between SNP markers differed in SNP density and family structure. They covered a range of mean inter-marker distances from 0.003 cM to 0.23 cM. The case study of mouse data was based on low-density genotypes of full-sib families; progeny and parents were genotyped. The case study of cattle data comprised medium-density genotypes of half-sib families with genotyped progeny only. Furthermore, medium-density SNP data were available for full-sib families in maize. High-density genotype data of half-sib families were generated in simulations. For evaluation, the SNP-BLUP approach was applied to simulated data only. Unless otherwise stated, computations were done using R version 4.0.3 [11] and  $t = 0.8$ ; all scripts are included as Additional files 2–6.

### Mouse data

Genotype data of a heterogeneous stock of mice were available from <https://wp.cs.ucl.ac.uk/outbredmice/heterogeneous-stock-mice/>. We investigated chromosome 17 because it harbors the highly recombining MHC region which affects several immunological traits [12]. The chromosome data consisted of genotypes at 394 SNPs of 2002 individuals. After filtering for individual call rate  $\geq 90\%$ , 1998 genotyped individuals remained comprising 1759 progeny, 120 fathers and 119 mothers. In total, 138 full-sib families (family size ranged from 1 to 47) could be identified. The SNP call rate was  $\geq 90\%$ . All genotype data were phased with Beagle version 5.1 [13] and parental haplotypes were selected to set up the correlation matrix. Assuming a 1:1 relationship between physical (Build37 genome assembly) and genetic distance of adjacent markers, the genetic length was 91 cM. SNP alleles were coded in terms of the major allele in the given sample. The population-LD matrix was calculated from progeny genotypes.

### Cattle data

Genotype data of Holstein cattle were available from RADAR <https://dx.doi.org/10.22000/280>. The data comprised 50K SNP-chip data of five half-sib families with  $n = 265$  progeny in total; the family size ranged from 32 to 106. A chromosome window containing 300 SNPs was selected from BTA1. Based on the physical ordering of markers according to the genome assembly ARS-UCD1.2, this region corresponded to 20.59–39.44 Mbp. The haplotypes of sires were imputed from progeny genotypes using the R package *hsphase* version 2.0.2 [14]. Maternal LD and paternal recombination rates between SNP pairs were estimated according to Hampel et al. [15]. However, we used a 1:1 relationship between physical (Mbp) and genetic positions (cM) for convenience; the genetic length of this window was 19 cM. Sire haplotypes and maternal LD were also part of the RADAR data set. SNP alleles were coded in terms of the major maternal allele among progeny.

### **Maize data**

Raw marker data were available from NCBI GEO database under Accession Number GSE50558, accompanied with physical coordinates corresponding to the genome assembly B73. The data set contained two maize panels, Flint and Dent, for which about 50K SNPs have been assessed in order to estimate recombination activity in different maize populations [16]. We arbitrarily chose the Flint panel and chromosome 2 for further analysis. In this panel, 13 full-sib families have been obtained by crossing an inbred “central” line and several inbred “founder” lines. Double haploid (DH) lines have been derived from the F1 plants. This procedure allowed for studying maternal meioses only. A cM:Mbp ratio of 0.80 was reported for Flint [16]; the genetic length of chromosome 2 was approximately 188 cM. SNP genotypes of DH progeny being heterozygous were set to missing value. After filtering the data for SNP and individual call rate  $\geq 90\%$ ,  $n = 1248$  out of 1262 DH progeny and 1447 out of 2030 SNPs remained. Rarely missing marker information of DH progeny were imputed by sampling the homozygous genotypes according to their frequencies. Afterwards loci with minor allele frequency less than 5% were discarded, yielding 956 SNPs. As haplotypes of DH progeny were given with certainty, the population-LD matrix was set up directly using the squared Spearman correlation between SNPs based on haploid data. The family approach solely considered the female part of the covariance between SNPs. The haplotypes of F1 individuals were inferred from the marker data of inbred lines. SNP alleles were coded according to central-line origin.

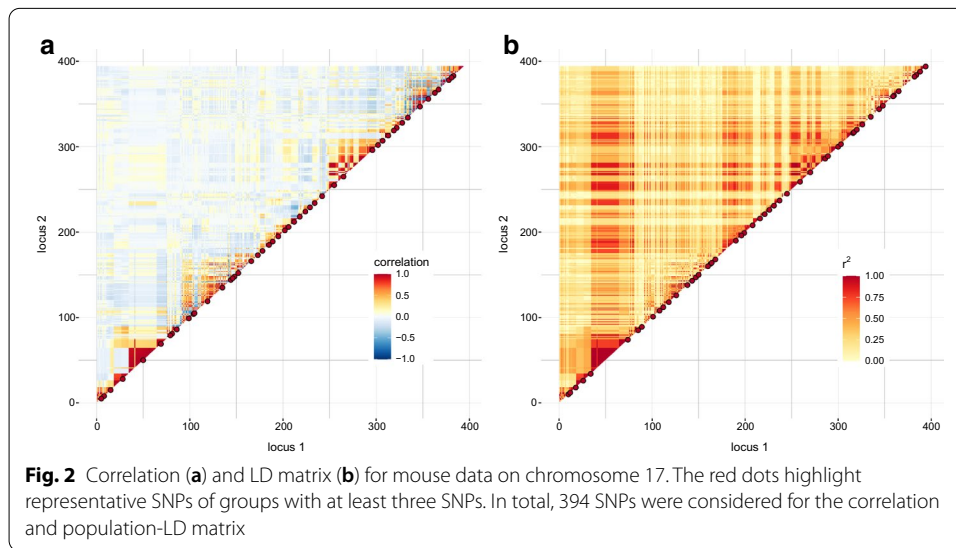
### **Simulated data**

The simulation study resembled the population structure of a dairy cattle population. The setup of simulation design is fully described in [1]. Briefly, we considered  $N = 1, 5, 10$  sires of half siblings. The overall number of progeny was  $n = 1000$  equally partitioned into half-sib families. Quantitative traits were simulated which were influenced by 2 and 5 QTLs with equal effect sizes. QTLs contributed 30% to the trait variation (i.e., heritability 0.3). In total, 300 SNPs were simulated on a chunk of DNA with 1 cM length. The data were generated using the R package `AlphaSimR` version 0.13.0 [17]. SNP alleles were recoded in terms of the major allele in the founder population. The simulation was repeated 100 times. For assessing the SNP-BLUP approach, a window of 0.05 cM to both sides of a simulated QTL was accepted as true-positive result.

## **Results**

### **Case studies**

For the mouse data consisting of 138 genotyped full-sib families, the population-LD approach exposed a wide region (13.82–21.13 Mbp) that had a strong association with the entire chromosome 17 shown as a red band in Fig. 2b. With the family approach, this region revealed only high positive interdependence with almost no impact on the remaining chromosome, see Fig. 2a. Moreover, a highly fragmented region appeared in the range of 27.59–45.88 Mbp that overlaps MHC regions 1 and 2 ([https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001635.18/](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.18/)). This was caused by high variation of parental haplotypes. Though the total number of groups was smaller with the family



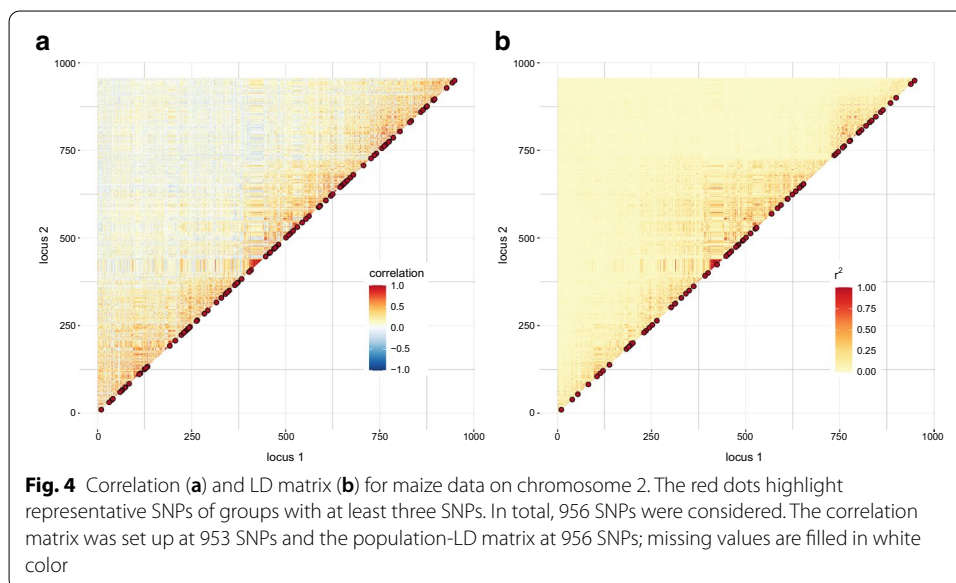
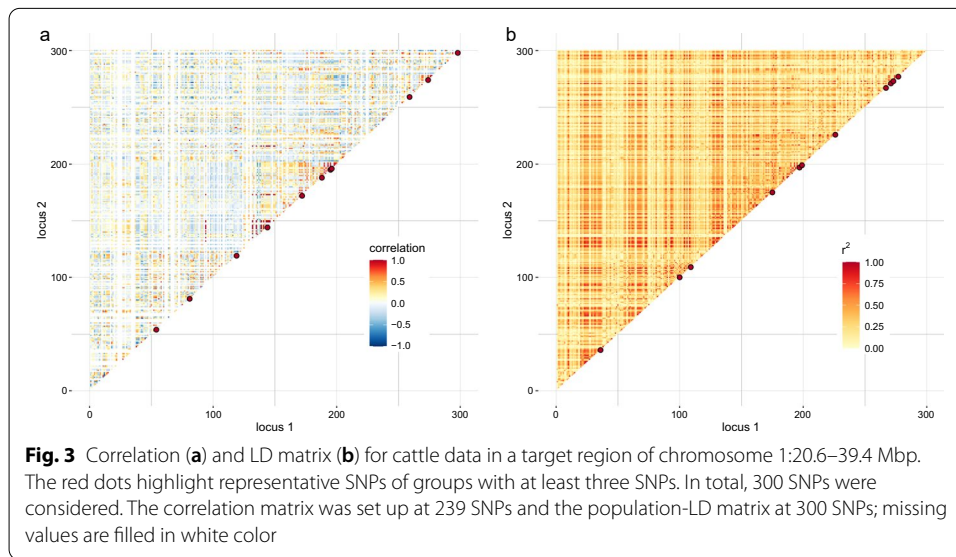
**Table 1** Number of groups, number of groups with at least three SNPs and Calinski-Harabasz index (CH)

	Families	QTLs	Family approach				Population-LD approach			
			SNPs	Groups	$\geq 3$	CH	SNPs	Groups	$\geq 3$	CH
Simulation	10	2	283	59	10	80.7	300	21	9	40.0
	10	5	282	61	10	75.0	300	17	9	38.9
	5	2	282	61	10	71.9	300	29	8	32.6
	5	5	281	64	10	85.6	300	24	9	35.0
	1	2	281	59	9	61.4	300	56	6	20.5
	1	5	282	59	9	83.0	300	49	7	25.0
Mouse	138		394	83	49	38.8	394	98	51	20.6
Cattle	5		237	172	11	2.9	300	210	11	2.4
Maize	13		953	426	93	10.4	956	576	70	21.9

Number of SNPs corresponds to the method applied (family or population LD). Average values of 100 repetitions are presented for simulated data and  $t = 0.8$ . Computing time for grouping markers was up to 0.2 s except for the maize data which required 2 s

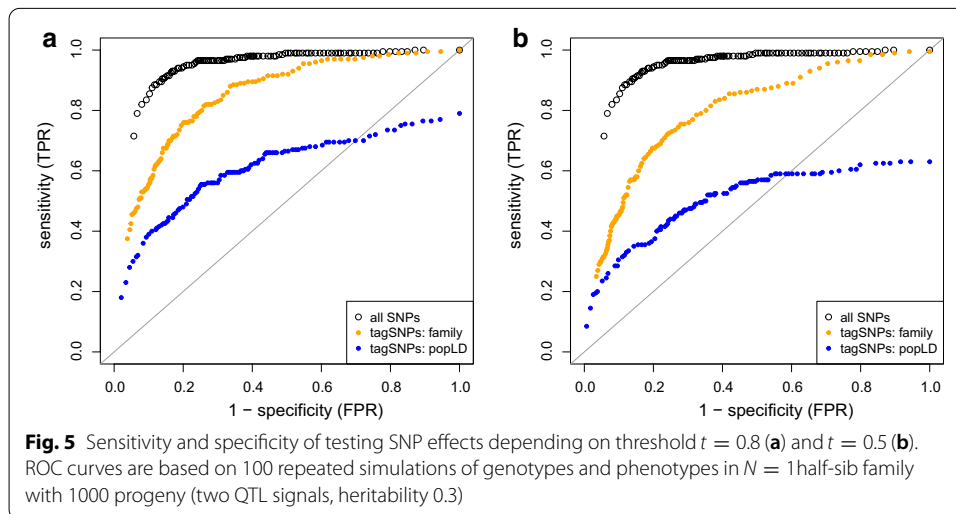
approach than with the population-LD approach (83 vs. 98), the number of large groups, i.e., with group size  $\geq 3$ , was almost equal (49 vs. 51), see Table 1. The CH index was about two times higher with the family approach. Figure 2 shows that tagSNPs of groups containing at least three markers were distributed more uniformly over the chromosome with the family approach than with the population-LD approach. The median of distances between all tagSNPs was 0.70 cM and 0.52 cM in the family and population-LD approach, respectively.

For the cattle data consisting of five half-sib families, 239 SNPs were taken into account in the family approach. At 61 out of 300 SNPs, all sires were homozygous for the major allele, leading to zeros on the diagonal of the paternal covariance part. Thus, these loci were discarded when setting up the correlation matrix  $R$ . A clear distinction of regions with particularly high interdependence was not possible for any of the approaches (Fig. 3). In total, 11 groups with size  $\geq 3$  were found with both the family and population-LD approach (Table 1) but the representative SNPs of these



groups were distributed more evenly over the chromosome window based on the family approach. The median of distances between all tagSNPs was 0.08 cM and 0.07 cM in the family and population-LD approach, respectively. When only those SNPs were used in the population-LD approach that were considered in the family approach, the outcome of grouping differed: 239 SNPs were binned into 194 groups (CH index 6.2); 8 out of them had group size of at least three SNPs vs. 300 SNPs were binned into 210 groups (CH index 2.4); 11 groups with group size of at least three SNPs.

The correlation matrix corresponding to 13 full-sib families in maize is shown in Fig. 4. In three out of 956 SNPs, maternal SNP alleles of F1 plants were missing; these loci were discarded in the family approach. In total, 953 SNPs have been binned into 426 groups based on the correlation matrix but the CH index was about 50% lower



than with the population-LD approach where 956 SNPs were grouped into 576 bins (Table 1). With both approaches, tagSNPs corresponding to the bins of at least three SNPs were similarly distributed over the chromosome (family approach: 93 bins, population-LD approach: 70 bins) except a gap between SNP index 650 and 750 which was better covered with tagSNPs from the family approach. The median distance between all representative SNPs was 0.29 cM with the family and 0.16 cM with the population-LD approach. With both methods, a block of strong (positive) association among SNPs appeared in the region of 85.04 to 95.31 Mbp which is in the vicinity of the functional centromere [18]. Two F1 plants were the driving factor: they were completely heterozygous in this window.

### Simulation study

The average number of groups over all repetitions, and groups with at least three SNPs are listed in Table 1; results are given for  $t = 0.8$  and varying number of half-sib families and QTLs. Grouping of markers appeared rather robust based on the family approach, about 60 groups have been built, but the number of groups strongly varied with the population-LD approach suggesting a dependence on the number of families. Few families needed more groups. Note that the number of rows in  $X$  was fixed ( $n = 1000$ ). The number of groups containing at least three markers was rather constant between methods and for different family sizes and QTLs. The CH index was at least two times larger with the family approach than with the population-LD approach. The population-LD approach becomes competitive with decreasing threshold  $t$  and performed better than the family approach with respect to the CH index when  $t \leq 0.6$  (see Additional file 7).

A SNP-BLUP approach was applied to simulated data in order to evaluate sensitivity and specificity if only tagSNPs were used as predictor variables in linear regression. As an example, results based on one half-sib family and two simulated QTLs are shown as ROC curve in Fig. 5; the number of groups was almost equal for this scenario. As expected, considering all SNPs simultaneously yielded highest accuracy in terms of true-positive and false-positive rate. However, if the aim was to use filtered data in SNP-BLUP,

then tagSNPs obtained from the family approach was the second best choice. Or in other words, using only one fifth of available genotypic information led to almost the same accuracy of genome-based association studies as using all genotypic information. The choice of  $t$  had no influence on which method performed best, see Fig. 5a for  $t = 0.8$  and Fig. 5b for  $t = 0.5$ . ROC curves looked very similar for all investigated scenarios of simulation though the number of groups obtained from the population-LD approach increased with decreasing number of families. Hence, a direct relationship between number of groups and sensitivity/specificity seems not to exist.

## Discussion

We have shown applicability of the suggested software tool to empirical data. Especially for the mouse data consisting of many genotyped full-sib families, the correlation matrix gave a clear representation of genomic regions with high or low interdependence. In contrast to a population-LD approach, which did not account for family stratification, it was also possible to identify regions with positive or negative relationship. Spurious dependencies with the population-LD approach disappeared with the family approach. Also Carlson et al. [2] reported that population stratification may generate artifactual LD and hence makes an LD-selection algorithm sensitive.

Instead of selecting representative markers for genome-based evaluations, employing the grouping structure itself can be a beneficial option. For instance, the group assignment derived from the family approach can directly be considered in a group lasso approach [3, 19]. Then the effects of markers in a group of highly dependent markers will jointly be shrunk towards zero or enlarged with respect to the relevance of this group for trait expression. Additional sparsity within group can be achieved with a sparse-group lasso approach [20]. Grouped approaches shall be investigated in more detail in future because they hold potential to cope with high multicollinearity. Possible benefits will likely depend on characteristics of the sample, such as the number of families, SNP density, and population-genetic parameters, e.g., heritability and heterozygosity.

In future research, the functionality of our package should be extended by grouping methods based on LD blocks which can optionally put restrictions to the physical distance between SNPs (similar to [21]). Other options for selecting tagSNPs (e.g., depending on allele frequency; [22]) will be verified.

## Conclusions

The extent of dependence among genomic markers is affected by the underlying population structure. Representative markers can be selected more efficiently if the corresponding matrix of pairwise dependencies takes this structure into account. The correlation matrix for half- or full-sib families highlights regions of high dependence between markers more precisely than the population-LD matrix. Additionally, it reveals regions of positive or negative association among markers.

We contributed a new function `tagSNP` to the R package `hscovar` which is suited to samples from livestock and crop populations with typical family stratification. The covariance matrix can be set up in a piecewise manner, either separately for each chromosome or based on other meaningful information. The resulting grouping structure



can be exploited in genome-based evaluations to handle the problem of high multicollinearity between markers.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-04010-0>.

**Additional file 1.** Given the population structure, half-sib families or full-sib families, the covariance matrix is analytically retrieved. Its computation using the R package `hscovar` is shown.

**Additional file 2.** Raw mouse data are processed and the matrix of correlation between markers is derived.

**Additional file 3.** Cattle data are processed and the matrix of correlation between markers is derived.

**Additional file 4.** Raw maize data are processed and the matrix of correlation between markers is derived.

**Additional file 5.** With this script, genotype and phenotype data of half-sib families are simulated and a genome-based association analysis is carried out.

**Additional file 6.** Plots of correlation matrices and population-LD matrices are produced based on results with Additional files 2–5.

**Additional file 7.** Number of groups, number of groups with at least three SNPs and Calinski-Harabasz index for different simulation scenarios and varying threshold.

### Abbreviations

AI-REML: Average information restricted maximum likelihood; BLUP: Best linear unbiased prediction; BTA: Bos taurus autosome; cM: CentiMorgan; DNA: Deoxyribonucleic acid; FPR: False positive rate; GWAS: Genomewide association study; LD: Linkage disequilibrium; Mbp: Mega base pairs; MHC: Major histocompatibility complex; QTL: Quantitative trait locus; ROC: Receiver operating characteristic; SNP: Single nucleotide polymorphism; TPR: True positive rate.

### Acknowledgements

We thank the Reviewers for their helpful comments.

### Authors' contributions

DW developed the theory, implemented the statistical methods, performed the analysis, and wrote the manuscript. MD and JK contributed to software development and improved the manuscript. All authors have read and approved the final manuscript.

### Funding

Open Access funding enabled and organized by Projekt DEAL. The project was funded by the German Research Foundation (DFG, WI 4450/2-1). The funder had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### Availability of data and materials

All R scripts used are provided in Additional files 2–6. The R package `hscovar` version 0.4.0 is available at CRAN. The cattle data are accessible through RADAR <https://dx.doi.org/10.22000/280>. The mouse data are obtainable from <https://wp.cs.ucl.ac.uk/outbredmice/heterogeneous-stock-mice/>. Maize data are available at NCBI GEO under Accession Number GSE50558.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 5 August 2020 Accepted: 8 February 2021

Published online: 19 February 2021

### References

1. Wittenburg D, Bonk S, Doschoris M, Reyer H. Design of experiments for fine-mapping quantitative trait loci in live-stock populations. *BMC Genet.* 2020;21:66.
2. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Human Genet.* 2004;74(1):106–20.
3. Dehman A, Ambroise C, Neuvial P. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinform.* 2015;16:148.
4. Kim SA, Cho CS, Kim SR, Bull SB, Yoo YJ. A new Haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics.* 2018;34(3):388–97.

5. Hill W, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet.* 1968;38(6):226–31.
6. Clayton D. *snpStats: SnpMatrix and XsnpMatrix Classes and Methods*; 2017. R package version 1.34.0. <http://biocductor.org/packages/release/bioc/html/snpStats.html>.
7. Clayton D, Leung HT. An R package for analysis of whole-genome association studies. *Hum Hered.* 2007;64(1):45–51.
8. Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Theory Methods.* 1974;3(1):1–27.
9. Koivula M, Strandén I, Su G, Mäntysaari EA. Different methods to calculate genomic predictions-comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *J Dairy Sci.* 2012;95(7):4065–73.
10. Butler D, Cullis BR, Gilmour A, Gogel B. *ASReml-R Reference Manual*. The State of Queensland, Department of Primary Industries and Fisheries, Brisbane. 2009; <https://asreml.kb.vsnr.co.uk/wp-content/uploads/sites/3/2018/02/ASReml-R-3-Reference-Manual.pdf>.
11. R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2020. <https://www.R-project.org/>.
12. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet.* 2006;38(8):879–87.
13. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;81(5):1084–97.
14. Ferdosi M, Kinghorn B, van der Werf J, Lee S, Gondro C. *hsphase: an R package for pedigree reconstruction, detection of recombination events, phasing and imputation of half-sib family groups*. *BMC Bioinform.* 2014;15(1):172.
15. Hampel A, Teuscher F, Gomez-Raya L, Doschoris M, Wittenburg D. Estimation of recombination rate and maternal linkage disequilibrium in half-sibs. *Front Genet.* 2018;9:186.
16. Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, et al. Intraspecific variation of recombination rate in maize. *Genome Biol.* 2013;14:R103.
17. Gaynor RC, Gorjanc G, Hickey JM. *AlphaSimR: An R-package for Breeding Program Simulations*. G3: Genes Genomes Genetics. 2020.
18. Schneider KL, Xie Z, Wolfgruber TK, Presting GG. Inbreeding drives maize centromere evolution. *Proc Nat Acad Sci.* 2016;113(8):E987–96.
19. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc B.* 2006;68(1):49–67.
20. Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. *J Comput Graph Stat.* 2013;22(2):231–45.
21. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. *Science.* 2002;296(5576):2225–9.
22. Wang S, He S, Yuan F, Zhu X. Tagging SNP-set selection with maximum information based on linkage disequilibrium structure in genome-wide association studies. *Bioinformatics.* 2017;33(14):2078–81.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)







# Estimation of Recombination Rate and Maternal Linkage Disequilibrium in Half-Sibs

Alexander Hampel<sup>1</sup>, Friedrich Teuscher<sup>1</sup>, Luis Gomez-Raya<sup>2</sup>, Michael Doschoris<sup>1</sup> and Dörte Wittenburg<sup>1\*</sup>

<sup>1</sup> Leibniz Institute for Farm Animal Biology (FBN), Institute of Genetics and Biometry, Dummerstorf, Germany, <sup>2</sup> Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain

## OPEN ACCESS

### Edited by:

Han Mulder,  
Wageningen University & Research,  
Netherlands

### Reviewed by:

Steve Kachman,  
University of Nebraska-Lincoln,  
United States  
Hans D. Daetwyler,  
La Trobe University, Australia

### \*Correspondence:

Dörte Wittenburg  
wittenburg@fbn-dummerstorf.de

### Specialty section:

This article was submitted to  
Livestock Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 08 January 2018

**Accepted:** 07 May 2018

**Published:** 05 June 2018

### Citation:

Hampel A, Teuscher F, Gomez-Raya L, Doschoris M and Wittenburg D (2018) Estimation of Recombination Rate and Maternal Linkage Disequilibrium in Half-Sibs. *Front. Genet.* 9:186. doi: 10.3389/fgene.2018.00186

A livestock population can be characterized by different population genetic parameters, such as linkage disequilibrium and recombination rate between pairs of genetic markers. The population structure, which may be caused by family stratification, has an influence on the estimates of these parameters. An expectation maximization algorithm has been proposed for estimating these parameters in half-sibs without phasing the progeny. It, however, overlooks the fact that the underlying likelihood function may have two maxima. The magnitudes of the maxima depend on the maternal allele frequencies at the investigated marker pair. Which maximum the algorithm converges to depends on the chosen start values. We present a stepwise procedure in which the relationship between the two modes is exploited. The expectation maximization algorithm for the parameter estimation is applied twice using different start values, followed by a decision process to assess the most likely estimate. This approach was validated using simulated genotypes of half-sibs. It was also applied to a dairy cattle dataset consisting of multiple half-sib families and 39,780 marker genotypes, leading to estimates for 12,759,713 intrachromosomal marker pairs. Furthermore, the proper order of markers was verified by studying the mean of estimated recombination rates in a window adjacent to the investigated locus as well as in a window at its most distant chromosome end. Putatively misplaced markers or marker clusters were detected by comparing the results with the revised bovine genome assembly UMD 3.1.1. In total, 40 markers were identified as candidates of misplacement. This outcome may help improving the physical order of markers which is also required for refining the bovine genetic map.

**Keywords:** allele frequency, expectation maximization algorithm, genome assembly, likelihood function, linkage analysis

## INTRODUCTION

Population genetic parameters, such as linkage disequilibrium (LD) and recombination rate, are relevant parameters for characterizing a livestock population. Methods for genomic selection (e.g., Meuwissen et al., 2001) implicitly exploit the LD between quantitative trait loci (QTL) and genetic markers for estimating breeding values or genetic effects captured by the markers in a population. The accuracy of estimated effects depends on the extent of LD (e.g., Hayes et al., 2009), which is affected by several factors (e.g., Ardlie et al., 2002). For instance, recombination breaks the physical

linkage between chromosomal segments. Hence a recombination event creates new combinations of alleles in the next generation. Furthermore, population structure influences the extent of LD differing not only between populations and breeds (De Roos et al., 2008) but also between families within a breed (Dekkers, 2004).

The estimation of recombination rate requires genotypic and pedigree information—most often parent-offspring data are used. Different approaches are available enabling the parameter estimation based on phased (e.g., “direct method”; Ott, 1991) or unphased genotype data (e.g., likelihood approaches such as the LOD score; Ott, 1991). Such approaches are applicable to natural populations and experimental crosses (e.g., F2 or backcross) but typically not to livestock populations. Non-random mating influences the population structure and the way of data collection in livestock. For instance, in paternal half-sib families, the sire and its ancestors are genotyped but not necessarily the dam. Paternal half-sib families are a typical family structure in dairy cattle. In this context, the paternal recombination rate ( $\theta$ ) and the LD of maternal gametes ( $D^{dam}$ ) are of particular interest, for instance, to map QTL in such breeds. Neglecting the population structure may lead to biased parameter estimates.

Maps for cattle breeds are available in low dimensions, and they have mostly been based on the analysis of microsatellites (e.g., Kappes et al., 1997; Thomsen et al., 2001). Due to advances in molecular genetics, single nucleotide polymorphisms (SNPs) are available to cover animal genomes at high density. With the exception of some errors, the order of SNPs is known from the genome sequence (Bovine HapMap Consortium, 2009). In a recent effort, estimates of recombination rates considering more than 50 K SNPs have been determined (Ma et al., 2015) which opens up the way to a more comprehensive map.

Previous studies (Gomez-Raya, 2012; Gomez-Raya et al., 2013) explicitly exploited the family structure of paternal half-sibs and proposed a likelihood approach for the estimation of  $\theta$  and  $D^{dam}$  which was based on an expectation maximization (EM) algorithm. However, a special characteristic of the underlying likelihood function (LF) is its bimodality (Figure 2 in Gomez-Raya et al., 2013) and thus the possibility of having two maxima. Depending on the maternal allele frequencies at the two loci, the modes can have equal height. EM converges to a local maximum—which one depends on the start values (Dempster et al., 1977). Specifying alternative start values allows identifying the two modes and the selection of the most likely estimate. However, in case of equal-height modes, a unique estimation of  $D^{dam}$  and  $\theta$  is actually not possible. Therefore, the first objective of this study was to elucidate the relationship between the two modes of LF which is incorporated in determining alternative start values. In case of equally high modes, the most likely estimate of  $D^{dam}$  and  $\theta$  is selected based on additional information from the neighborhood. A stepwise procedure is proposed which was validated using simulated genotypes of half-sibs. The second objective addresses the proper order of SNPs according to the underlying genome assembly. As recombination rates can be estimated between all pairs of SNPs on a chromosome, and not only between adjacent SNPs, the pattern of estimates of  $\theta$  will be examined in depth to identify

SNPs which are putatively misplaced in the underlying genome assembly. Empirical data consisting of multiple half-sib families of Holstein-Friesian cows were analyzed, and candidates for misplacement in the genome assembly are presented.

## METHODS

At first, the LF for estimating  $\theta$  and  $D^{dam}$  is inspected to differentiate its bimodality. Then, the unknown parameters are estimated using an EM algorithm corresponding to the LF and two different sets of start values. A decision process (DP) is employed to find the most probable final estimate when the LF has two modes.

### Likelihood Function

The SNP genotypes of progeny are used to estimate the unknown parameters; SNP alleles are denoted as A and B. The observed genotype frequencies at two loci, e.g., AA and AB, are denoted as  $n_{AA,AB}$ . It is assumed that paternal linkage phases are known without error, and only those loci at which a sire is heterozygous are considered. Otherwise a recombination rate cannot be inferred. The sire haplotypes are denoted as A-A and B-B in the coupling phase as well as A-B and B-A in the repulsion phase. Hence the frequency of the paternal A allele is 0.5 among offspring.

Based on the multinomial distribution and considering a full model, in which the maternal haplotype frequencies and  $\theta$  are the unknown parameters, the logarithmic LF can be expressed similarly to Gomez-Raya (2012) as:

$$\log LF(\pi_{AA,AA}, \pi_{AA,AB}, \dots, \pi_{BB,BB} | n_{AA,AA}, n_{AA,AB}, \dots, n_{BB,BB}) \\ = \sum_{i,k \in \{AA,AB,BB\}} n_{i,k} \log \pi_{i,k} + \text{constant},$$

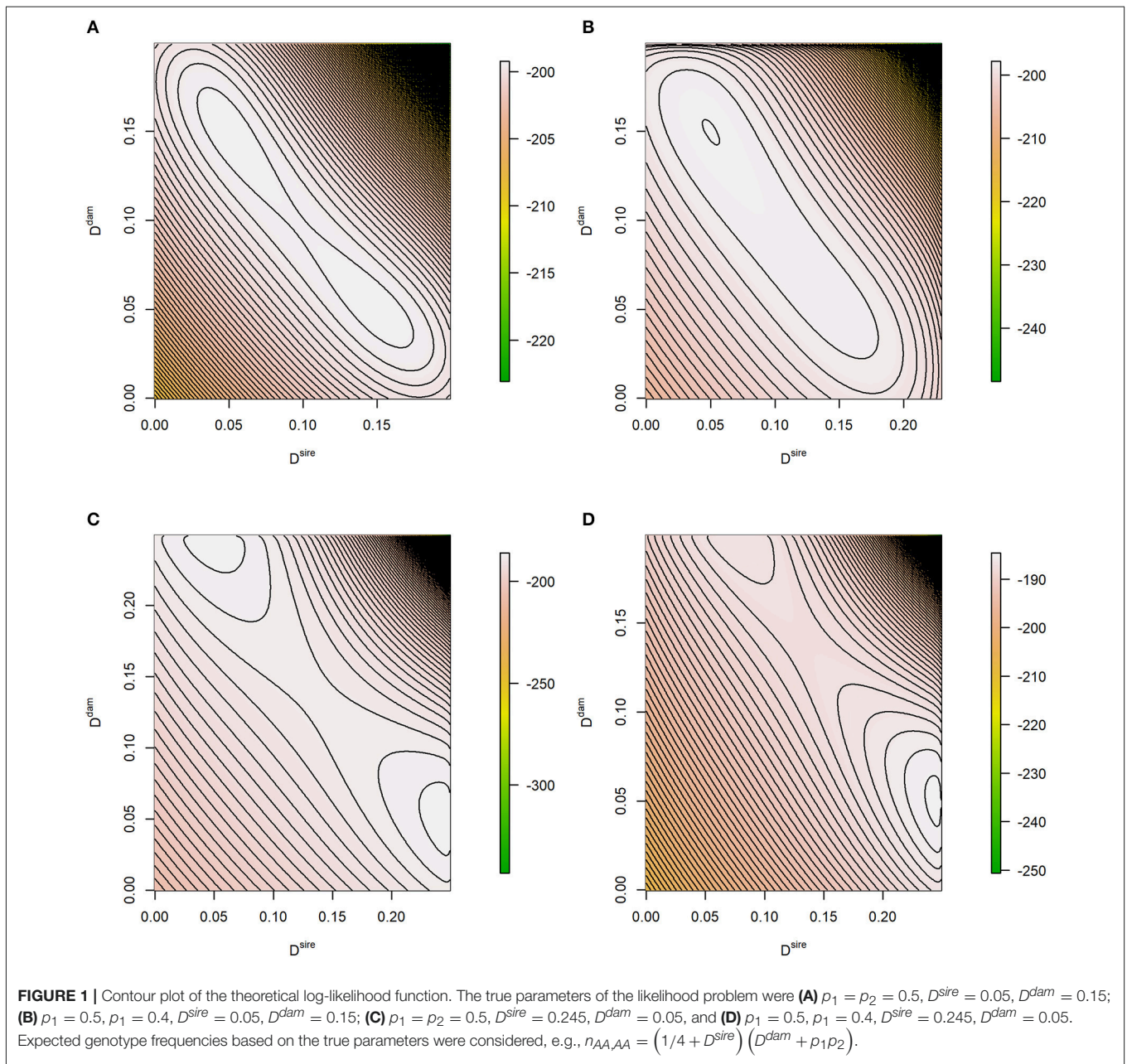
where  $\pi_{i,k}$  is the probability of the two-locus genotype  $i, k \in \{AA, AB, BB\}$ . For instance,  $\pi_{AA,AB} = \frac{1}{2}(1 - \theta)p_{A-B} + \frac{1}{2}\theta p_{A-A}$  in the coupling phase. If the sire haplotypes are in repulsion phase,  $\theta$  changes to  $1 - \theta$ . The  $\log LF$  depends on the frequency of maternal gametes (e.g.,  $p_{A-A}$  denotes the frequency of the maternal haplotype A-A) and  $\theta$ . It is now reparametrized and written in terms of LD of maternal and paternal gametes,  $D^{dam} = p_{A-A} - p_1 p_2$  and  $D^{sire} = \pm(1 - 2\theta)/4$  (the positive sign corresponds to the coupling phase), respectively, to better investigate the relationship between these parameters. Indices are used at frequencies ( $p_1$  and  $p_2$ ) of the maternal A allele to distinguish between the loci. Hence, using the computer algebra system *Maple*<sup>TM</sup> (Maplesoft, Waterloo, Ontario, Canada):

$$\log LF(D^{sire}, D^{dam}, p_1, p_2 | n_{AA,AA}, n_{AA,AB}, \dots, n_{BB,BB}) \\ = n_{BB,BB} \log \left( \frac{1}{4} (1 + 4D^{sire}) (D^{dam} + (-1 + p_1)(-1 + p_2)) \right) \\ + n_{BB,AB} \log \left( \frac{1 - p_1}{4} + D^{sire} (-2D^{dam} + (-1 + p_1)(1 - 2p_2)) \right) \\ + n_{BB,AA} \log \left( \frac{1}{4} (-1 + 4D^{sire}) (D^{dam} + (-1 + p_1)p_2) \right) \\ + n_{AB,BB} \log \left( \frac{1 - p_2}{4} + D^{sire} (-2D^{dam} + (1 - 2p_1)(-1 + p_2)) \right)$$

$$\begin{aligned}
 &+ n_{AB,AB} \log \left( \frac{1}{4} + D^{sire} (4D^{dam} + (-1 + 2p_1)(-1 + 2p_2)) \right) \\
 &+ n_{AB,AA} \log \left( \frac{p_2}{4} + D^{sire} (-2D^{dam} + p_2(1 - 2p_1)) \right) \\
 &+ n_{AA,BB} \log \left( \frac{1}{4} (-1 + 4D^{sire}) (D^{dam} + p_1(-1 + p_2)) \right) \\
 &+ n_{AA,AB} \log \left( \frac{p_1}{4} + D^{sire} (-2D^{dam} + p_1(1 - 2p_2)) \right) \\
 &+ n_{AA,AA} \log \left( \left( \frac{1}{4} + D^{sire} \right) (D^{dam} + p_1 p_2) \right) + \text{constant}.
 \end{aligned} \tag{1}$$

The shape of  $\log LF$  depends on the maternal allele frequencies at the two loci: they affect the number (one or two) and height of the

modes. Plotting this function also reveals its characteristics, see **Figure 1**. The examples consider maternal and paternal LD being inside (**Figures 1A,B**) or near the boundary (**Figures 1C,D**) of the parameter space which is  $D^{sire} \in [-0.25, 0.25]$  and  $D^{dam} \in [L_1, L_2]$  with  $L_1 = \max \{-p_1 p_2, -(1 - p_1)(1 - p_2)\}$  and  $L_2 = \min \{p_1(1 - p_2), (1 - p_1)p_2\}$  (Lewontin, 1964). In absence of distorted segregation,  $D^{sire}$  is not restricted by allele frequencies and has its extreme value of  $\pm 0.25$ . In case of maternal allele frequencies being 0.5 at both loci, the modes have equal height, and a symmetric pattern in terms of  $D^{dam}$  and  $D^{sire}$  is observed (**Figures 1A,C**). For frequencies beyond 0.5, the  $\log LF$  is either unimodal or the two modes have different heights.





## Start Values for the EM Algorithm

Although the form of  $\log LF$  is elementary, computing the maximum-likelihood (ML) estimation is not straightforward, see e.g., Hill (1974) and **File S1**. Gomez-Raya (2012) and Gomez-Raya et al. (2013) proposed an EM algorithm to estimate  $\theta$  and maternal haplotype frequencies based on the  $\log LF$ . For an EM algorithm, and for other commonly used optimization methods, it is known that it may converge to a local maximum if the underlying distribution is bimodal (Dempster et al., 1977). Which maximum is found depends on the start values of the unknown parameters. Thus, an EM algorithm shall be started multiple times with different start values to explore the modes of the likelihood surface (e.g., Biernacki et al., 2003). In our experience, there will be no more than two modes. Thus, we will use two sets of start values thereby employing the relationship between the two modes.

The theoretical derivations of Bonk et al. (2016) can be adapted and applied to estimate the statistical dependence in half-sib families. The covariance between genotype codes for additive effects  $\text{cov}_{add}$  (AA, AB and BB are coded as 1, 0, and  $-1$ , respectively) and dominance effects  $\text{cov}_{dom}$  (AA, AB, and BB are coded as  $-1$ , 1, and  $-1$ , respectively) require the paternal and maternal LD, and a system of two equations is achieved:

$$\begin{aligned}\text{cov}_{dom} &= 16D^{sire}D^{dam} + 4D^{sire}(1 - 2p_1)(1 - 2p_2), \\ \text{cov}_{add} &= D^{sire} + D^{dam}.\end{aligned}$$

Then two solutions for  $D^{dam}$  and  $D^{sire}$  (indices *I* and *II*) can be obtained. Given one solution, the other solution is derived as:

$$\begin{aligned}D_{II}^{sire} &= D_I^{dam} + \frac{1}{4}(1 - 2p_1)(1 - 2p_2) \text{ and} \\ D_{II}^{dam} &= D_I^{sire} - \frac{1}{4}(1 - 2p_1)(1 - 2p_2).\end{aligned}\quad (2)$$

In the particular case of  $p_1 = p_2 = 0.5$ , maternal and paternal LD are mutually exchangeable.

For the first run of EM, default start values are used:  $D^{sire} = 0.25$  ( $\theta = 0$ ) and  $D^{dam} = 0$ . Start values for the maternal haplotype frequencies are received from  $D^{dam}$  and ML estimates of the maternal A allele at the investigated loci, i.e.,  $\hat{p} = n_{AA}/(n_{AA} + n_{BB})$ , where  $n_{AA}$  and  $n_{BB}$  denote the number of offspring with AA and BB genotypes, respectively. Start values for the second run are obtained after using the EM estimates of the first run and employing the complementary relationship in Equation (2). If this relationship leads to start values outside the parameter space, we conclude that the underlying likelihood is unimodal and skip a second EM run. Otherwise, this “educated guess” shall lead the EM algorithm to the second mode.

## Decision Process for Non-critical Allele Frequencies

In our experience,  $\log LF$  can either be unimodal or bimodal with modes of different height when  $p_1$  and  $p_2$  differ from 0.5. Then, the most likely estimate of  $D^{dam}$  and  $\theta$  can be identified from comparing the value of  $\log LF$  at the two proposals. It is obvious to select the estimates of parameter values constituting the higher

value of  $\log LF$ . Because  $\log LF$  can be rather flat (e.g., **Figure 1** and **Figure 2** in Gomez-Raya et al., 2013), sometimes the height of the two modes cannot be differentiated even when  $p_1 \neq 0.5$  or  $p_2 \neq 0.5$ . Hence we introduce the interval  $[0.48, 0.52]$  for allele frequencies; this is denoted as the critical range. For allele frequencies outside this range,  $p_1 \vee p_2 \notin [0.48, 0.52]$ , the modes of  $\log LF$  differed clearly more than  $10^{-5}$  in prior investigations using simulated data with  $D^{sire}, D^{dam} \in \{0, 0.05, 0.15\}$  and 100 half-sibs.

In view of the empirical data analysis, the  $\log LF$  in Equation (1) was modified mildly by combining genotype frequencies. In a family with few progeny, not all genotype combinations may have been observed, i.e., some cells remained empty in a genotype-count table, and haplotype frequencies were then estimated to be close to zero. To level the influence of zero or almost-zero haplotype frequencies on the likelihood function, the none-observed genotypes and the least-observed genotype were combined into one cell. The corresponding set of index pairs was set up as  $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_1$  where  $\mathcal{J}_0 = \{i, k \in \{AA, AB, BB\} | n_{i,k} = 0\}$  collects the indices from the empty cells, and  $\mathcal{J}_1 = \arg \min_{i,k \in \{AA, AB, BB\}} \{n_{i,k} | n_{i,k} > 0\}$  refers to the genotype index with least observed frequency. This gave rise to a  $\log LF$  with a combined cell:

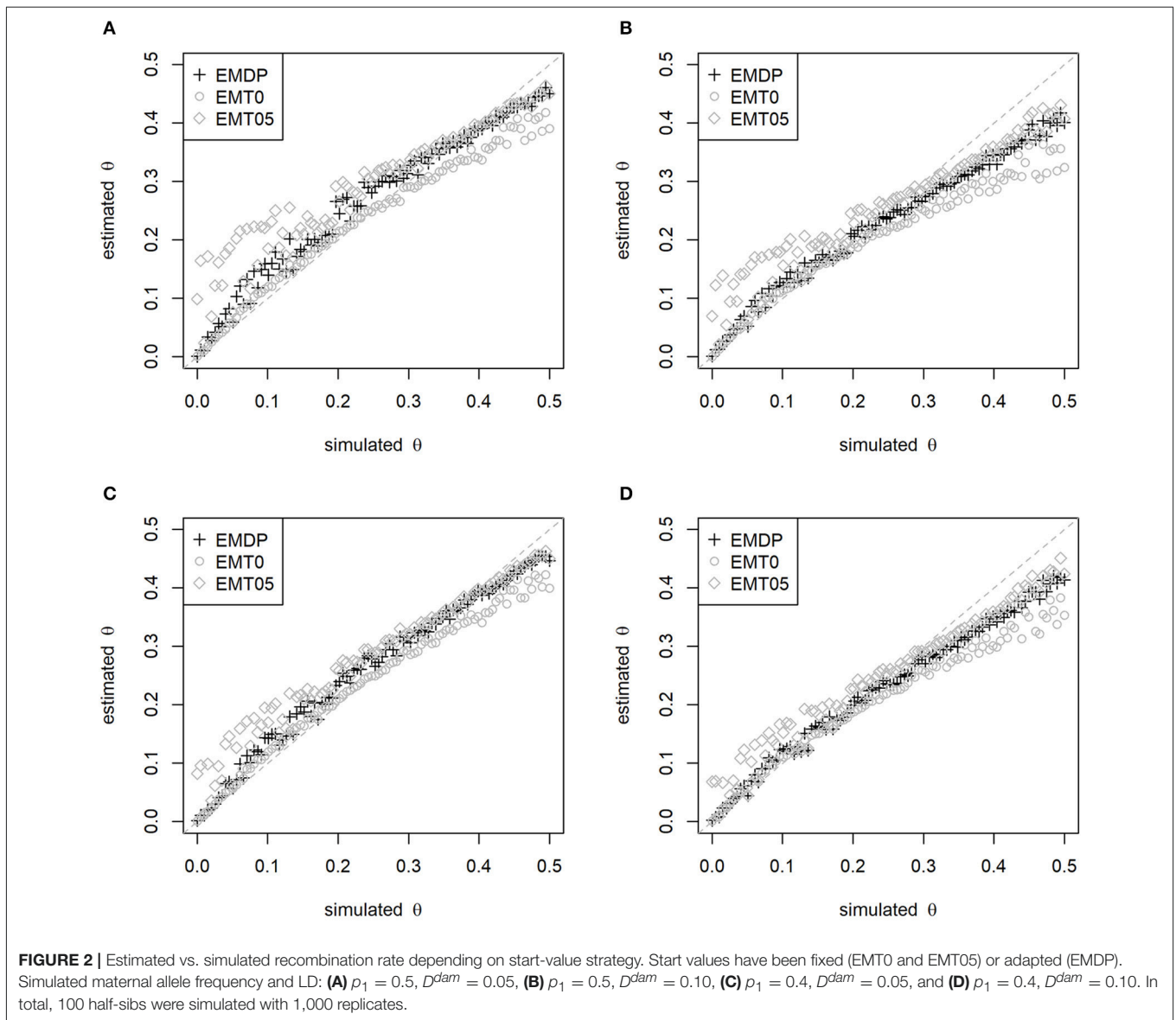
$$\begin{aligned}\log LF^*(\pi_{AA,AA}, \pi_{AA,AB}, \dots, \pi_{BB,BB} | n_{AA,AA}, \dots, n_{BB,BB}) \\ = \sum_{i,k \in \{AA, AB, BB\} \setminus \mathcal{J}} n_{i,k} \log \pi_{i,k} + \left( \sum_{i,k \in \mathcal{J}} n_{i,k} \right) \log \left( \sum_{i,k \in \mathcal{J}} \pi_{i,k} \right)\end{aligned}$$

with parameters as above. Thus, compared to a strategy of leaving empty cells out, the probability of the least observed genotype is marginally increased.

The DP based on the value of  $\log LF^*$  was applied to all locus pairs with allele frequencies outside the critical range.

## Decision Process for Critical Allele Frequencies

If the allele frequencies are within the critical range,  $p_1, p_2 \in [0.48, 0.52]$ , it is expected that the modes of  $\log LF$  have almost equal height, and a decision about the final estimates is actually not possible. Only additional information about the physical position and estimates of  $\theta$  in the neighborhood might allow selecting the most appropriate set of parameter estimates. If SNP  $i$  and  $k$  had estimates of allele frequency in the critical range, i.e.,  $\hat{p}_i, \hat{p}_k \in [0.48, 0.52]$ ,  $i, k \in \{1, \dots, m_c\}$ ,  $i < k$ ;  $m_c$  denotes the number of SNPs on a particular chromosome. Then the estimated recombination rates ( $\hat{\theta}_{i,j}$ ) between SNP  $i$  and all other SNPs  $j$ ,  $j = 1, \dots, m_c$ ,  $i \neq j$ , with non-critical allele frequency were used to determine a smoothing B-spline curve: The estimates  $\hat{\theta}_{i,j}$  were regressed onto the corresponding physical position of the SNPs  $j$ ,  $j = 1, \dots, m_c$ ,  $i \neq j$ , using the *R* function “smooth.spline” considering five degrees of freedom (R Core Team, 2014). The physical position was scaled according to the whole chromosome length. Then, for SNP  $k$  with critical allele frequency, the value was selected out of the two proposals  $\hat{\theta}_I$  and



$\hat{\theta}_{II}$  yielding the minimum squared deviation to  $\hat{\theta}$  the fitted curve at the relative physical position of SNP  $k$ .

The stepwise procedure using EM twice followed by a DP based on either  $\log LF^*$  or information from the neighborhood is denoted as EMDP. For comparison, EM estimates were also achieved based on fixed start values. In short, we applied three approaches to simulated and empirical data with the following start values:

- EMDP:  $\theta = 0$  and  $D^{dam} = 0$ , second set of start values was obtained from Equation (2),
- EMT0:  $\theta = 0$  and  $D^{dam} = 0$ ,
- EMT05:  $\theta = 0.5$  and  $D^{dam} = 0$ .

### Detection of Misplaced SNPs

The estimation of recombination rates based on empirical bovine data enables the identification of SNPs with wrong placement in

the underlying genome assembly. SNPs revealing high estimates of  $\theta$  with SNPs in close neighborhood or low estimates with distant SNPs are candidates for misplaced SNPs. In order to trace such SNPs, a statistic for unusually high estimates was calculated for each locus  $i \in \{1, \dots, m_c\}$  as the average of the nearby  $s$  markers:

$$\hat{\theta}_{i,c}^H = \begin{cases} \frac{1}{s} \sum_{j=i+1}^{i+s} \hat{\theta}_{i,j}; & \text{if } i < m_c - s, \\ \frac{1}{s} \sum_{j=i-s}^{i-1} \hat{\theta}_{i,j}; & \text{if } i \geq m_c - s. \end{cases}$$

The number of SNPs considered in a local environment of the target SNP was  $s = 30$ ; this choice will be discussed later. Similarly, for investigating the pattern of unusually low estimates, a statistic was computed for a window at the most distant

chromosome end:

$$\hat{\theta}_{i,c}^L = \begin{cases} \frac{1}{s} \sum_{j=m_c-s+1}^{m_c} \hat{\theta}_{i,j}; & i < \frac{m_c}{2}, \\ \frac{1}{s} \sum_{j=1}^s \hat{\theta}_{i,j}; & \text{if } i \geq \frac{m_c}{2}. \end{cases}$$

The 99% quantile of  $\hat{\theta}_{i,c}^H$  and 1% quantile of  $\hat{\theta}_{i,c}^L$  over the entire genome ( $i = 1, \dots, m_c$ ;  $c = 1, \dots, 29$ ) were taken as thresholds, and SNPs with  $\hat{\theta}_{i,c}^H$  ( $\hat{\theta}_{i,c}^L$ ) above (below) the corresponding threshold were candidates for misplacement. The subset of selected SNPs was visually inspected afterwards, and conspicuously misplaced SNPs were also matched to the improved reference assembly UMD 3.1.1 (Merchant et al., 2014).

## DATA

### Simulation

In order to validate the influence of start values on the parameter estimates, genotype data of a single half-sib family ( $N = 100$  or  $N = 1,000$  progeny) were simulated at two loci. The sire was double-heterozygous in coupling phase, and the paternal haplotypes of progeny were determined considering a random recombination event during meiosis. The maternal haplotypes of progeny were drawn by chance from a dam population with a given LD between the two loci.

In a first study, three LD scenarios were simulated with 10,000 replicates each:

- I) Small  $\theta$ :  $D^{sire} = 0.245$  ( $\theta = 0.01$ );  $D^{dam} = 0.05$ ,
- II) Medium  $\theta$ :  $D^{sire} = 0.15$  ( $\theta = 0.20$ );  $D^{dam} = 0.05$ ,
- III) Large  $\theta$ :  $D^{sire} = 0.05$  ( $\theta = 0.40$ );  $D^{dam} = 0.15$ .

Different maternal allele frequencies were considered for each scenario: (a)  $p_1 = p_2 = 0.5$ ; (b)  $p_1 = p_2 = 0.4$ ; (c)  $p_1 = 0.4$ ,  $p_2 = 0.6$ ; and (d)  $p_1 = p_2 = 0.8$ . Scenarios II (a) and III (a) are complementary settings. A balanced design, i.e.,  $D^{sire} = D^{dam}$ , leads to a unimodal log  $LF$  and is not considered here. As only a single SNP pair was simulated, the DP was reduced to comparing the log  $LF^*$  values only.

In a second simulation study,  $\theta$  was varied in a range from zero to 0.5 with step size 0.005. For each value, genotypes of  $N = 100$  or  $N = 1,000$  half-sibs and 1,000 replicates were simulated. Separate runs of the simulation were executed with fixed allele frequency at the first locus,  $p_1 = 0.4$  or  $p_1 = 0.5$ , but varying frequency at the second locus:  $p_2$  was randomly drawn from a uniform distribution on the interval  $[0.3, 0.7]$ . In each run  $D^{dam}$  was fixed at 0.05 or 0.10. Hence, 8 scenarios were studied in total.

Data simulation and all analyses were carried out with programs written in R (R Core Team, 2014).

### Empirical Dataset

The stepwise procedure was applied to an empirical dataset consisting of multiple families of Holstein-Friesian cows and 39,780 SNP genotypes on the autosomes; data and quality control were described in Melzer et al. (2013). SNPs were ordered according to the Btau 4.2 assembly (Bovine HapMap Consortium, 2009). Sires with at least 30 progeny were

selected; hence five paternal half-sib families comprising 265 cows remained. The phases of the non-genotyped sires were estimated using the R package *hsphase* version 2.0.0 (Ferdosi et al., 2014). Then, 8,512 loci were disregarded because all sires were homozygous. The number of intrachromosomal SNP pairs expected from the number of SNPs was reduced by the quantity of SNP pairs at which only sires with homozygous-heterozygous or heterozygous-homozygous haplotypes were observed (roughly one third of SNP pairs); 12,759,713 SNP pairs within chromosomes remained for the analysis. The estimate of  $D^{dam}$  and its standardized measure  $r^2$  were determined from the estimated haplotype frequencies.

The pattern search to detect misplaced SNPs was validated by rearranging SNPs within and between chromosomes. Firstly, a single SNP was moved on BTA4 from index position 3 to position 1,000 as well as a SNP cluster on BTA4 from index positions 10–15 to 1,602–1,607. Secondly, single SNPs were moved from BTA20 (index positions 1 and 1,000) to BTA4 (index positions 100 and 1,000) as well as to a SNP cluster from BTA1 (index positions 411–420) to BTA4 (index positions 411–420).

The empirical data are provided in **Files S2, S3**. Marker names and positions according to the genome assemblies Btau 4.2 and UMD 3.1.1 are listed in **File S4**.

## RESULTS

### Simulated Data

Using either one set of fixed start values (EMT0 or EMT05) or two proposals (EMDP) influenced the estimates of disequilibrium in sires and dams. The first simulation study showed that EMT0 and EMDP performed equally well for the small- $\theta$  setting in terms of bias and MSE of  $\hat{\theta}$  and  $\hat{D}^{dam}$ . For the medium- and large- $\theta$  setting, EMT0 and EMT05, respectively, had the least bias and MSE (**Tables 1, 2**) when  $N = 100$  half-sibs were considered. In such cases, EMDP was the second-best choice. If  $N = 1,000$ , bias and MSE reduced for all approaches and EMDP performed best overall in terms of MSE (**Tables S1, S2**). Furthermore, it was obvious from the two-locus investigation that a decision solely based on log  $LF^*$  is not sufficient if  $p_1 = p_2 = 0.5$ . Using EMDP in such a case and  $N = 1,000$ , both modes were selected at almost equal proportion in the medium- and large- $\theta$  scenario. However, if  $N = 100$ , in 58.0% and 57.9% of the repetitions in the medium- $\theta$  and small- $\theta$  scenarios, respectively, the decision was in favor of the mode derived from the default start values. In the small- $\theta$  scenario, the mode derived from the default start values yielded most often the higher value of log  $LF^*$ : in 95.5% of the repetitions if  $N = 100$  and 81.1% if  $N = 1,000$ .

The second simulation study confirmed the ranking of methods as described above. Moreover, an impact of the simulated maternal LD on the estimates of  $\theta$  was observed. Using  $N = 100$ , EMDP had the least bias of  $\hat{\theta}$  in the upper range ( $\theta > 0.30$ ) if  $D^{dam} = 0.05$  (**Figures 2A,C**) and the least bias in the lower range ( $\theta < 0.05$ ) if  $D^{dam} = 0.10$  (**Figures 2B,D**). The bias of  $\hat{\theta}$  was reduced for all approaches if  $p_1 = 0.4$  instead of  $p_1 = 0.5$  was simulated as well as when  $N = 1,000$  half-sibs were considered (**Figure S1**). In summary, EMDP has led to robust

**TABLE 1** | Bias of estimated paternal recombination rate and maternal linkage disequilibrium for simulated scenarios.

	bias $\hat{\theta}$						bias $\hat{D}^{dam}$					
	$p_1 = p_2 = 0.5$	$p_1 = p_2 = 0.4$	$p_1 = 0.4, p_2 = 0.6$	$p_1 = p_2 = 0.8$	$p_1 = p_2 = 0.5$	$p_1 = p_2 = 0.4$	$p_1 = 0.4, p_2 = 0.6$	$p_1 = p_2 = 0.8$	$p_1 = p_2 = 0.5$	$p_1 = p_2 = 0.4$	$p_1 = 0.4, p_2 = 0.6$	$p_1 = p_2 = 0.8$
$\theta = 0.01, D^{dam} = 0.05$	0.153	0.079	0.023	0.013	0.073	0.036	0.009	0.002	0.036	0.009	0.002	
	<b>0.002</b>	<b>0.003</b>	<b>0.001</b>	<b>0.002</b>	<b>-0.001</b>	<b>-0.001</b>	<b>-0.001</b>	<b>-0.001</b>	<b>-0.001</b>	<b>-0.001</b>	<b>-0.001</b>	
	0.016	0.009	0.002	<b>0.002</b>	0.006	0.002	0.002	0.006	0.002	<b>-0.001</b>	<b>-0.001</b>	
$\theta = 0.20, D^{dam} = 0.05$	0.097	0.059	0.043	<b>0.000</b>	0.051	0.031	0.022	0.051	0.031	0.022	0.003	
	<b>0.012</b>	<b>0.011</b>	<b>0.009</b>	-0.014	<b>0.008</b>	<b>0.007</b>	<b>0.006</b>	<b>0.008</b>	<b>0.007</b>	<b>0.006</b>	<b>-0.002</b>	
	0.069	0.042	0.018	-0.014	0.036	0.023	0.010	0.036	0.023	0.010	<b>-0.002</b>	
$\theta = 0.40, D^{dam} = 0.15$	<b>-0.106</b>	<b>-0.080</b>	<b>-0.053</b>	<b>-0.013</b>	<b>-0.051</b>	<b>-0.039</b>	<b>-0.025</b>	<b>-0.051</b>	<b>-0.039</b>	<b>-0.025</b>	<b>-0.010</b>	
	-0.187	-0.163	-0.123	-0.252	-0.092	-0.080	-0.060	-0.092	-0.080	-0.060	-0.095	
	-0.130	-0.109	-0.105	-0.248	-0.064	-0.053	-0.051	-0.064	-0.053	-0.051	-0.093	

Start values have been fixed (EMT0 and EMT05) or adapted (EMDP), 100 half-sibs were simulated with 10,000 replicates. The best outcome for each parameter setting is highlighted in bold.

**TABLE 2** | MSE of estimated paternal recombination rate and maternal linkage disequilibrium for simulated scenarios.

	MSE $\hat{\theta}$						MSE $\hat{D}^{dam}$					
	$p_1 = p_2 = 0.5$	$p_1 = p_2 = 0.4$	$p_1 = 0.4, p_2 = 0.6$	$p_1 = p_2 = 0.8$	$p_1 = p_2 = 0.5$	$p_1 = p_2 = 0.4$	$p_1 = 0.4, p_2 = 0.6$	$p_1 = p_2 = 0.8$	$p_1 = p_2 = 0.5$	$p_1 = p_2 = 0.4$	$p_1 = 0.4, p_2 = 0.6$	$p_1 = p_2 = 0.8$
$\theta = 0.01, D^{dam} = 0.05$	0.061	0.030	0.009	0.002	0.014	0.008	0.002	0.014	0.008	0.002	<b>0.001</b>	
	<b>0.001</b>	<b>0.002</b>	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	
	0.007	0.004	<b>0.001</b>	<b>0.001</b>	0.002	<b>0.001</b>	<b>0.001</b>	0.002	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>	
$\theta = 0.20, D^{dam} = 0.05$	0.030	0.022	0.018	<b>0.008</b>	0.008	0.006	0.004	0.008	0.006	0.004	<b>0.002</b>	
	<b>0.014</b>	<b>0.013</b>	<b>0.012</b>	0.010	<b>0.003</b>	<b>0.003</b>	<b>0.002</b>	<b>0.003</b>	<b>0.003</b>	<b>0.002</b>	<b>0.002</b>	
	0.026	0.020	0.014	0.010	0.006	0.005	0.003	0.006	0.005	0.003	<b>0.002</b>	
$\theta = 0.40, D^{dam} = 0.15$	<b>0.032</b>	<b>0.026</b>	<b>0.020</b>	<b>0.006</b>	<b>0.008</b>	<b>0.006</b>	<b>0.004</b>	<b>0.008</b>	<b>0.006</b>	<b>0.004</b>	<b>0.002</b>	
	0.048	0.045	0.035	0.099	0.011	0.010	0.007	0.011	0.010	0.007	0.013	
	0.038	0.034	0.032	0.097	0.009	0.007	0.007	0.009	0.007	0.007	0.012	

Start values have been fixed (EMT0 and EMT05) or adapted (EMDP), 100 half-sibs were simulated with 10,000 replicates. The best outcome for each parameter setting is highlighted in bold.



estimates of  $\theta$  and  $D^{dam}$  being least susceptible to changes in the parameter settings compared with EMT0 and EMT05.

## Parameter Estimation in Cattle

For the empirical data analysis, the results of estimating  $\theta$  and  $D^{dam}$  based on EMDP are described. Loci at which all sires were homozygous did not contribute to parameter estimations, and they are shown blank in all figures. Usually, the chance of recombination was lower in the immediate neighborhood than for more distant SNPs: lower estimates of  $\theta$  were found between the next 50 neighboring SNPs (about 3.2 Mb range, **Figure 3** and **Figure S2**), and  $\hat{\theta}$  increased with increasing distance between loci. Typical triangles with lower estimates showing regions of high paternal LD were also found for more distant SNPs.

The estimated  $D^{dam}$  between distant SNPs was generally close to zero but specific regions existed where  $\hat{D}^{dam} \pm 0.2$  (see **Figure S3**; the corresponding  $r^2$  is shown in **Figure S4**). High LD was estimated near the center on BTA10 (23,319.73 kb range, mean  $r^2 = 0.170$ ), BTA13 (16,326.63 kb range, mean  $r^2 = 0.140$ ) and BTA20 (13,870.19 kb range, mean  $r^2 = 0.179$ ). Smaller regions of high LD were also observed outside the center, e.g., on BTA14 (3,811.90 kb range, mean  $r^2 = 0.178$ ), BTA23 (474.85 kb range, mean  $r^2 = 0.699$ ), and BTA29 (3,474.42 kb range, mean  $r^2 = 0.176$ ). Such regions may indicate the presence of cold spots.

Among all SNP pairs, 37,040 (0.3%) estimates of  $\theta$  followed the rules of the DP for critical maternal allele frequencies: the most appropriate  $\hat{\theta}$  was determined using the cubic B-spline smoothing approach. For example, considering SNP 61 on BTA1, 25 mates with critical allele frequency existed:  $61 \times 188$ ,  $61 \times 424$ ,  $61 \times 510$ ,  $61 \times 604$ ,  $61 \times 644$ , ...,  $61 \times 2,539$ , and  $61$

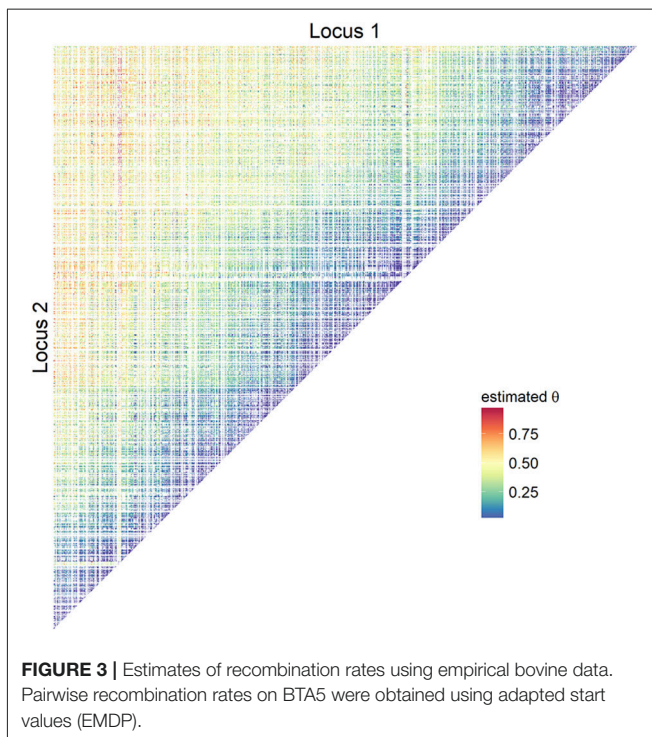
$\times 2,560$ . Estimates based on mates with SNP 61 having a non-critical allele frequency were employed to generate the B-spline curve (**Figure 4**). Five degrees of freedom were equivalent to 152 proper knots in this case. Then, for instance, the two proposals for SNP pair  $61 \times 604$  were  $\hat{\theta}_I = 0.428$  and  $\hat{\theta}_{II} = 0.572$ , and the first was closer to the fitted curve.

In 99.7% of pairwise investigations, a comparison of  $\log LF$  at the two proposals yielded the final estimate of genetic parameters. The decision was in favor of the first proposal, which was derived from default start values  $\theta = 0$  and  $D^{dam} = 0$ , in 72.13% regarding all chromosomes. On the other hand, in 27.87% of pairwise investigations, the proposal obtained from the complementary relationship (Equation 2) yielded the larger  $\log LF$ ; this makes EMDP superior to EMT0. Either way, the smaller  $\hat{\theta}$  out of the two proposals was related to the larger likelihood in about 89% of the comparisons.

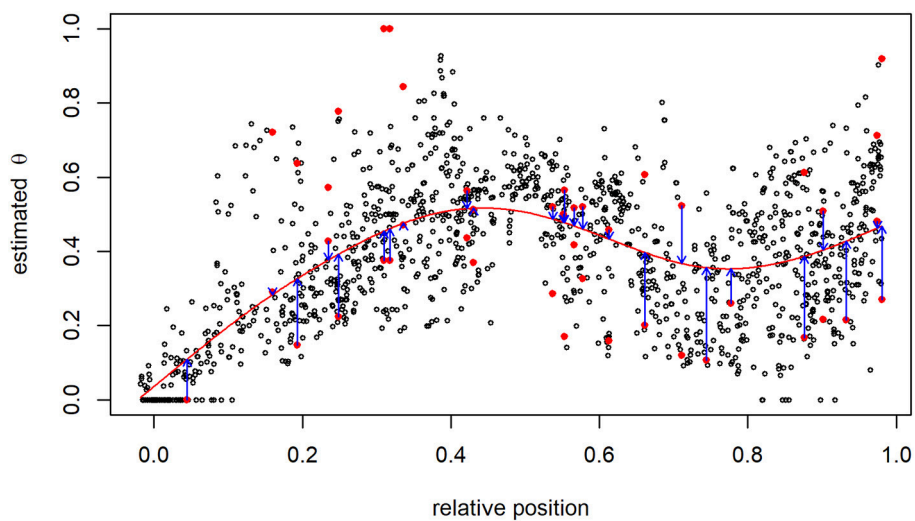
Taking the SNPs with statistics  $\hat{\theta}_{i,c}^H$  above and  $\hat{\theta}_{i,c}^L$  below the corresponding thresholds and looking at the pattern of  $\hat{\theta}$  in the rows or columns of the matrix representation, we concluded that individual SNPs or even clusters of SNPs were wrongly placed on some chromosomes. For instance, a SNP cluster was identified on BTA5 ranging from 10.11 Mb to 10.86 Mb (13 SNPs; rs110564524 to rs110035083, see **Table 3**) where  $\hat{\theta}$  was exceptionally high not only with very distant SNPs but also with adjacent SNPs. (If non-informative SNPs occurred within such a cluster, they were declared candidates as well.) We compared the chromosome assignment, the SNP order and the base pair position for all SNPs with putatively wrong placement to the revised *Bos taurus* assembly UMD 3.1.1. The aforementioned SNP cluster on BTA5 had the same chromosome assignment but the base pair position now ranged from 103.47 to 104.23 Mb based on UMD 3.1.1. The order of SNPs within this cluster remained. A second SNP cluster was found on BTA3 (rs109129646, rs41661711, rs41665543) which was now assigned to BTA2 based on UMD 3.1.1. Furthermore, the order of SNPs within this cluster was reversed. The rsID of a single SNP on BTA2 (BTA-45006; position: 10,760,381 bp) was unavailable. The DNA sequence of this SNP derived from Btau 4.2 was mapped to the sequence based on UMD 3.1.1 using Bowtie 2 (Langmead and Salzberg, 2012); the sequence fitted to BTA4 with position 33,987,225 bp. This result is similar to that of Khatkar et al. (2010), who determined this SNP on BTA4 at 33,987,224 bp based on UMD 3.0. Furthermore, the manually replaced SNPs and SNP clusters were detected as misplacements (see **Figure S5**). The complete list of 40 putatively misplaced SNPs, mostly confirmed with UMD 3.1.1, is given in **Table 3**.

## DISCUSSION

The recombination rate and maternal LD were estimated using an EM algorithm. Due to the bimodality of the underlying LF, two sets of start values were proposed to guide the iterative process into the right directions. With EMDP, the most likely estimate was selected out of the two proposals. A B-spline smoothing method was suggested for SNP pairs with critical maternal allele frequencies. The approach was validated using simulated







**FIGURE 4** | Recombination rate between the 61st and all other SNPs on BTA1. The pair of proposals at SNPs with critical allele frequency are shown as red points, and smoothing via B-splines (red line) was used to identify the most suitable estimate having the smallest distance to the B-spline curve (blue arrows).

genotypes of half-sibs, and it was applied to empirical dairy cattle data to identify putatively misplaced SNPs. The position of SNPs with unusually large estimates of  $\theta$  was compared to the improved bovine genome assembly UMD 3.1.1.

## Recombination Rate

The “direct method” of linkage analysis requires the haplotype phases of parents and offspring to be known; they might be inferred from genotypes ordered according to a given genome assembly (e.g., using LINKPHASE3 for half-sib families, Druet and Georges, 2015). As an option, approaches for long-range phasing of haplotypes (e.g., Daetwyler et al., 2011; Hickey et al., 2011; Ferdosi et al., 2014) enable determining the location of recombination events. Such an algorithm deduces blocks of haplotypes passed from parent to offspring. From this block structure, the number of break points can be counted and the recombination rate can be estimated. The advantage of the proposed method is that only the sire’s haplotypes and progeny genotypes are required for estimating the recombination rates and the error caused by phasing offspring genotypes are circumvented. A comparison of methods is worth further investigation.

Assuming that the maternal allele frequencies are known, the recombination rate and maternal LD can also be inferred from the reduced LF where  $\theta$  and  $D^{dam}$  are the unknown parameters. In this reduced LF case, the maximum of  $\log LF$  can quickly be obtained numerically. For instance, the function “optim” in R can consider the boundaries of the parameter space of  $\theta$  and  $D^{dam}$  using the option “L-BFGS-B”. Though the proposed sets of start values can also be subjected to the numerical optimization procedure, we have observed that some estimates have been stuck in one region of the parameter space and lead to the same maximum (results not shown). This outcome might be explained by the algorithm-internal step-width regularization. To circumvent this, the two local maxima can be sought via grid search. This procedure is feasible in the reduced LF case but

not recommended for a search in the whole parameter space considering not only  $\theta$  and  $D^{dam}$  but also the allele frequencies as unknown parameters (or, equivalently, treating  $\theta$  and the maternal haplotype frequencies as unknowns).

For various species it is known that the recombination rate varies between males and females (e.g., mouse, Liu et al., 2014; salmon, Moen et al., 2004). This was also observed for dairy cattle (e.g., Ma et al., 2015). Moreover, the recombination rate may also be different among individuals. Kadri et al. (2016) identified causal variants with effect on an individual’s number of recombination events in cattle. In general, the EM approach can be applied to each sire family separately, thus allowing the estimation of sire-specific recombination rates. Once the number of genome-wide or chromosome-wide recombination events would have been inferred from  $\hat{\theta}$ ’s of each sire, it could be further studied which loci are relevant to the variation of recombination events. This issue claims further investigation based on a larger number of sires.

The empirical data analysis required smoothing of estimates of  $\theta$  to select the final estimate in the ambiguous case where  $\hat{p}_1, \hat{p}_2 \in [0.48, 0.52]$ . The proposal bearing the least squared error with its predicted outcome on the curve was chosen. For convenience, a B-spline smoothing approach was employed with a fixed degree of smoothness. In contrast to fixing the degree of smoothness, as it was also done in Ma et al. (2015), the number and position of knots can be estimated using a Bayesian approach (Zhang et al., 2003). Preliminary investigations on the decision of knots revealed only minor differences in the shape of the smoothing curve (results not shown). Thus, it is expected that the DP is hardly influenced by this outcome. Other methods for selecting the more appropriate estimate are conceivable. They might be based on a measure of similarity. For instance, the sampling variance of estimates in a certain window around the investigated locus-pair mate can be considered, including either  $\hat{\theta}_I$  or  $\hat{\theta}_{II}$ . The higher measure of similarity, i.e., the lower sampling variance, would point to the more suitable estimate.

**TABLE 3** | List of misplaced SNPs in the genome assembly Btau 4.2 in comparison to the improved assembly UMD 3.1.1.

rsID	Synonym	Btau 4.2			UMD 3.1.1		
		Chr	Index Chr	Position bp	Chr	Index Chr	Position bp
41646101	BTA-49243	1	370	24,597,406	1	368	24,338,769
43237745	ARS-BFGL-NGS-101533	1	781	51,186,654	1	788	51,033,230
NA	BTA-45006	2	179	10,760,381	4	1,915	33,987,225
109129646	ARS-BFGL-NGS-31620	3	1,048	77,548,374	2	1,739	118,217,699
41661711	BTA-94835	3	1,049	77,635,918	2	1,738	118,130,312
41665543	BTA-98418	3	1,050	77,712,209	2	1,737	118,054,344
109560518	ARS-BFGL-NGS-16314	3	1,970	127,186,019	22	1,024	61,258,859
43504210	ARS-BFGL-NGS-105794	3	1,971	127,245,912	22	1,023	61,199,665
110389139	ARS-BFGL-NGS-92218	3	1,972	127,272,424	22	1,022	61,172,869
109112664	ARS-BFGL-NGS-67185	3	1,973	127,296,240	22	1,021	61,149,265
29018845	rs29018845	3	1,974	127,346,352	22	1,020	61,099,118
43505889	BTB-00297874	3	1,975	127,377,195	22	1,019	61,069,286
110104891	ARS-BFGL-NGS-4905	3	1,976	127,407,117	22	1,018	61,040,701
110739182	ARS-BFGL-NGS-25064	3	1,977	127,432,016	22	1,017	61,015,806
41586006	BTA-55057	3	1,978	127,462,142	22	1,016	60,985,682
109051741	ARS-BFGL-NGS-30266	3	1,979	127,516,161	22	1,015	60,930,725
110904383	ARS-BFGL-NGS-33950	3	1,980	127,552,306	22	1,014	60,897,444
29011651	BTA-03895	3	1,981	127,572,719	22	1,013	60,877,108
43506988	ARS-BFGL-NGS-90519	3	1,982	127,599,954	22	1,012	60,850,524
110144486	ARS-BFGL-NGS-70541	3	1,983	127,660,653	22	1,011	60,785,491
111007440	ARS-BFGL-NGS-58676	3	1,984	127,710,668	22	1,010	60,736,089
109871231	ARS-BFGL-NGS-114675	3	1,985	127,759,149	22	1,009	60,691,260
110218851	ARS-BFGL-NGS-42322	3	1,986	127,790,957	22	1,008	60,659,773
110564524	BTA-74753	5	178	10,108,998	5	1,330	103,472,289
110171642	ARS-BFGL-NGS-1422	5	179	10,175,658	5	1,331	103,539,042
110040410	ARS-BFGL-NGS-101402	5	180	10,211,997	5	1,332	103,583,249
109446437	ARS-BFGL-NGS-30033	5	181	10,288,674	5	1,333	103,659,884
41592968	BTA-74776	5	182	10,456,513	5	1,334	103,821,233
109366282	ARS-BFGL-NGS-678	5	183	10,495,942	5	1,335	103,860,658
109046936	ARS-BFGL-NGS-18320	5	184	10,545,228	5	1,336	103,911,258
109283161	BTA-143449	5	185	10,590,963	5	1,337	103,957,794
109955767	ARS-BFGL-NGS-118997	5	186	10,669,337	5	1,338	104,038,847
110253084	ARS-BFGL-NGS-52457	5	187	10,744,828	5	1,339	104,116,518
41654485	BTA-74821	5	188	10,769,105	5	1,340	104,140,809
110872738	ARS-BFGL-NGS-26592	5	189	10,815,342	5	1,341	104,188,186
110035083	ARS-USMARC-201	5	190	10,857,544	5	1,342	104,230,386
43521798	BTB-00316147	7	989	61,935,287	17	229	12,137,554
43696553	ARS-BFGL-NGS-69516	12	760	53,996,667	12	755	53,756,393
42103498	ARS-BFGL-NGS-77278	19	544	34,916,761	26	753	46,541,332
41255607	BTB-00780480	20	606	38,986,288	20	602	36,757,600

The misplaced SNPs confirmed by UMD 3.1.1 (based on discrepancy in SNP position or chromosome) are marked gray.

Estimates of recombination rate can reveal regions of hot and cold spots (e.g., Sandor et al., 2012; Weng et al., 2014). Our results based on LD estimates are in close agreement with those studies: cold spots were mainly found at the chromosome center or proximal chromosome end. Studies of human populations have shown the co-occurrence of runs of homozygosity (ROH) in regions with extended LD and low recombination rates (Gibson et al., 2006; Curtis et al., 2008). We did not investigate ROH

distribution in our material but our results support the existence of LD blocks in regions with low recombination rate.

### Order of SNPs

The physical order of SNPs has been updated regularly, and misplacements have been corrected successively (e.g., Milanesi et al., 2015). In this study, misplacement of SNPs was inferred from unusually large estimates of  $\theta$ : SNPs having low  $\hat{\theta}$  to distant

SNPs or large  $\hat{\theta}$  to close SNPs were candidates for misplacement. The inspection of  $\hat{\theta}$ 's followed heuristic rules, and it enabled the identification of the most serious candidates. The pattern search was based on a window size of 30 SNPs. Choosing, for instance,  $s = 300$  yielded 38 candidates of misplacement (rs43237745 on BTA1 and rs43696553 on BTA12 were not detected). Such a large window could be counteractive: it would definitely detect candidates being putatively placed on another chromosome but might fail to detect candidates which have a different position on the same chromosome.

The genome assembly UMD 3.1.1 was used to confirm candidates for misplacement in the older assembly Btau 4.2, hence providing evidence that the suggested pattern search is convenient. Additionally, the approach was validated based on manually misplaced SNPs. The comparison of genome assemblies leaves some candidates which shall be matched to the most recent assembly (<http://bovinegenome.org/>). Those candidates that cannot be matched may be further explored based on information about the raw data used for the current assembly. As, for instance, not only coverage but also read lengths of contigs and mate-pair distances have an impact on the quality of the genome build (Schatz et al., 2010), the value of such parameters has to be verified for suspicious SNPs. Eventually, an eligible strategy of replacement might consider the same criteria as those used for the identification of misplacements: in an iterative approach, a window measure seeks the position showing the least  $\theta$  estimates to neighboring SNPs and the maximum estimates to distant SNPs, which is similar to genetic mapping algorithms (e.g., Falk, 1989; see Cheema and Dicks, 2009 for a review of methods).

A more formal approach of identifying misplacements is testing the mean of  $\hat{\theta}$ 's in a local environment of the investigated SNP. This can be achieved by a statistical test, such as a  $t$ -test. Under the null hypothesis a recombination rate close to zero is expected. Such an approach bears a severe difficulty. As a recombination event likely follows a binomial distribution, the mean and variance approach zero under the null, and the null distribution of the test statistic cannot be determined. Alternatively, as the Fisher information  $F$  based on the  $\log LF$  (Equation 1) is available (see **File S1**), an asymptotic test based on the asymptotic normality of  $D^{sire}$  is conceivable. Because two-locus genotype frequencies—also the expected ones—may be zero in small samples, division by zero appears in terms of  $F$ , see Equations (S10)–(S16) in the **File S1**. Likewise, the term  $\log(0)$  appears in a likelihood ratio test at particular SNP pairs. Hence locus-pair-specific tests are not applicable to all locus combinations in the local environment.

Not only does the investigation of  $\hat{\theta}$  enable the identification of misplacements, but also that of the pattern of LD. Verifying the decay of maternal LD, Bohmanova et al. (2010) detected 223 misplaced SNPs in the Btau 4.0 assembly. Utsunomiya et al. (2016) even identified 2,906 SNPs with unexpected LD decay in the UMD 3.1 assembly; our set of candidates was not among this set. Furthermore, the presence of LD between unlinked markers (on different chromosomes) is an indication of misplacement. However, unusually large non-syntenic LD was not observed in the study of Bohmanova et al. (2010) though the underlying

assembly Btau 4.0 contained wrong chromosome assignments in comparison to the improved assembly UMD 3.1.1.

## Half-Sib Family

The proposed method is applicable not only to paternal but also to maternal half-sib families as they appear, for instance, in insects (Milne and Friars, 1984) or fish (Gjerde et al., 2004).

Half-sib families of dairy cattle typically contain many offspring. For instance, on average 828 offspring were reported for German Holstein bulls (median 172.5, minimum 12.0 and maximum 25,590; Vereinigte Informationssysteme Tierhaltung, 2017). A minimum family size of 30 was accepted in our study, allowing for phasing the sire haplotypes. Only the offspring of double-heterozygous sires were considered. To reduce the estimation error of the EM approach, the information from families of homozygous-heterozygous and double-homozygous sires may also be used in the estimation of  $\theta$  and  $D^{dam}$ . In such families, there is no means to estimate  $\theta$ , and haplotype frequencies can be estimated using simplified equations (Gomez-Raya, 2012). Haplotype frequencies are then treated to be known while the recombination rate is estimated for double-heterozygous sires.

## Chromatid Interference

In some cases, the smoothing curve adopts  $\hat{\theta}$ -values larger than 0.5. This may be a non-significant outcome and may occur by chance, but otherwise it can be explained by two reasons: (I) the uncertainty about the sire's estimated haplotype phase or (II) the presence of chromatid interference of recombination events. Additionally, a non-monotonic behavior of the smoothing curve was observed (see **Figure 4**) which is another and more severe indication of chromatid interference (Figure 6 in Stam, 1979; Zhao et al., 1995). The specification of the second set of start values does allow for chromatid interference. Excluding chromatid interferences requires some minor adjustment of the proposed method: in the start-value derivation step,  $\theta$  has to be restricted to lay in  $[0, 0.5]$  and hence  $D^{sire} \in [0, 0.25]$  in the coupling phase. The extent of genetic and chromatid interference shall be studied in more depth in future. It may help fitting a mapping function to the estimated  $\theta$ 's and improving the corresponding genetic map.

## CONCLUSIONS

The theoretical consideration of the likelihood function shows that the maternal allele frequencies have an influence on the position and height of its maxima. We employed an expectation maximization algorithm with two sets of start values to estimate recombination rate and maternal linkage disequilibrium near the two modes. Knowing the two modes gave a competitive edge on the parameter estimates, particularly if the maternal allele frequencies at the investigated SNP pair were about 0.5. In this case, a decision rule was proposed for selecting the most suitable proposal based on estimates in the neighborhood and the corresponding physical position. This study also showed that half-sib families can be used for the identification of misplaced SNPs in the genome assembly due to the information provided from linkage analyses and linkage disequilibrium coming from

maternal contribution. A correct physical order is helpful for further investigating the genetic distance between SNPs and refining the bovine genetic map.

## ETHICS STATEMENT

Animal Care and Use Committee approval was not obtained for this study because the blood samples were collected during a veterinary routine examination in 2009.

## AUTHOR CONTRIBUTIONS

AH implemented the theory, developed the simulation design, analyzed the data and drafted the manuscript. FT raised the initial question, contributed to the theoretical investigations and to the discussion of the results. LG-R contributed to the theoretical investigations and suggested improvements to the manuscript. MD contributed to the investigations of the likelihood function. DW supervised the project and contributed to the theoretical derivations, data analysis, discussion of results, and writing of the manuscript. All of the authors have read and approved the final manuscript.

## ACKNOWLEDGMENTS

We gratefully acknowledge the contribution of M. Stanke and V. Liebscher (University Greifswald, Germany) to this study. We also thank the reviewers for their helpful comments.

The empirical genotype data were obtained in the Fugato-plus project BovIBI funded by the German Federal Ministry of Education and Research (BMBF).

The publication of this article was funded by the Open Access Fund of the Leibniz Institute for Farm Animal Biology (FBN).

## REFERENCES

- Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 3, 299–309. doi: 10.1038/nrg777
- Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.* 41, 561–575. doi: 10.1016/S0167-9473(02)00163-9
- Bohmanova, J., Sargolzaei, M., and Schenkel, F. S. (2010). Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics* 11:421. doi: 10.1186/1471-2164-11-421
- Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., and Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. *Genet. Sel. Evol.* 48:36. doi: 10.1186/s12711-016-0214-0
- Bovine HapMap Consortium (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324, 528–532. doi: 10.1126/science.1167936
- Cheema, J., and Dicks, J. (2009). Computational approaches and software tools for genetic linkage map estimation in plants. *Brief. Bioinformatics* 10, 595–608. doi: 10.1093/bib/bbp045
- Curtis, D., Vine, A. E., and Knight, J. (2008). Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann. Hum. Genet.* 72, 261–278. doi: 10.1111/j.1469-1809.2007.00411.x

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00186/full#supplementary-material>

**Figure S1** | Estimated vs. simulated recombination rate depending on start-value strategy. Start values were fixed (EMT0 and EMT05) or adapted (EMDP). Simulated maternal allele frequency and LD: **(A)**  $\rho_1 = 0.5$ ,  $D^{dam} = 0.05$ , **(B)**  $\rho_1 = 0.5$ ,  $D^{dam} = 0.10$ , **(C)**  $\rho_1 = 0.4$ ,  $D^{dam} = 0.05$ , and **(D)**  $\rho_1 = 0.4$ ,  $D^{dam} = 0.10$ . In total, 1,000 half-sibs were simulated with 1,000 replicates.

**Figure S2** | Estimates of paternal recombination rate for all autosomes using empirical bovine data. Pairwise recombination rates were obtained using the stepwise procedure EMDP.

**Figure S3** | Estimates of maternal linkage disequilibrium for all autosomes using empirical bovine data. Pairwise linkage disequilibria were obtained using the stepwise procedure EMDP.

**Figure S4** | Estimates of  $r^2$  for all autosomes using empirical bovine data. Pairwise  $r^2$ 's were obtained using the stepwise procedure EMDP.

**Figure S5** | Validation of recombination-rate estimates using empirical bovine data on BTA4. Pairwise recombination rates were obtained using the stepwise procedure EMDP. BTA4 contained SNPs which were misplaced **(A)** within and **(B)** between chromosomes.

**Table S1** | Bias of estimated paternal recombination rate and maternal linkage disequilibrium for simulated scenarios. Start values were fixed (EMT0 and EMT05) or adapted (EMDP), 1,000 half-sibs were simulated with 10,000 replicates.

**Table S2** | MSE of estimated paternal recombination rate and maternal linkage disequilibrium for simulated scenarios. Start values were fixed (EMT0 and EMT05) or adapted (EMDP), 1,000 half-sibs were simulated with 10,000 replicates.

**File S1** | Investigations on the likelihood function and derivation of the Hessian matrix.

**File S2** | Genotype data consisting of 39,780 SNPs (columns), and 265 progeny (rows).

**File S3** | Assignment of genotypes to half-sib families.

**File S4** | Physical position of SNPs based on assembly Btau 4.2 and UMD 3.1.1.

- Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A., and Goddard, M. E. (2011). Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics* 189, 317–327. doi: 10.1534/genetics.111.128082
- De Roos, A. P. W., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179, 1503–1512. doi: 10.1534/genetics.107.084301
- Dekkers, J. C. (2004). Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.* 82(Suppl. E), E313–E328. doi: 10.2527/2004.8213\_supplE313x
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* 39, 1–38.
- Druet, T., and Georges, M. (2015). LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. *Bioinformatics* 31, 1677–1679. doi: 10.1093/bioinformatics/btu859
- Falk, C. T. (1989). A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. *Prog. Clin. Biol. Res.* 329, 17–22.
- Ferdosi, M. H., Kinghorn, B. P., van der Werf, J. H., Lee, S. H., and Gondro, C. (2014). hspbase: an R package for pedigree reconstruction, detection of recombination events, phasing and imputation of half-sib family groups. *BMC Bioinformatics* 15:172. doi: 10.1186/1471-2105-15-172
- Gibson, J., Morton, N. E., and Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* 15, 789–795. doi: 10.1093/hmg/ddi493



- Gjerde, B., Terjesen, B. F., Barr, Y., Lein, I., and Thorland, I. (2004). Genetic variation for juvenile growth and survival in Atlantic cod (*Gadus morhua*). *Aquaculture* 239, 531–531. doi: 10.1016/j.aquaculture.2004.06.010
- Gomez-Raya, L. (2012). Maximum likelihood estimation of linkage disequilibrium in half-sib families. *Genetics* 191, 195–213. doi: 10.1534/genetics.111.137521
- Gomez-Raya, L., Hulse, A. M., Thain, D., and Rauw, W. M. (2013). Haplotype phasing after joint estimation of recombination and linkage disequilibrium in breeding populations. *J. Anim. Sci. Biotechnol.* 4:30. doi: 10.1186/2049-1891-4-30
- Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646
- Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N., and van der Werf, J. H. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.* 43:12. doi: 10.1186/1297-9686-43-12
- Hill, W. G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 33, 229–239. doi: 10.1038/hdy.1974.89
- Kadri, N. K., Harland, C., Faux, P., Cambisano, N., Karim, L., Coppieters, W., et al. (2016). Coding and noncoding variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle. *Genome Res.* 26, 1323–1332. doi: 10.1101/gr.204214.116
- Kappes, S. M., Keele, J. W., Stone, R. T., McGraw, R. A., Sonstegard, T. S., Smith, T. P., et al. (1997). A second-generation linkage map of the bovine genome. *Genome Res.* 7, 235–249. doi: 10.1101/gr.7.3.235
- Khatkar, M. S., Hobbs, M., Neuditschko, M., Sölkner, J., Nicholas, F. W., and Raadsma, H. W. (2010). Assignment of chromosomal locations for unassigned SNPs/scaffolds based on pair-wise linkage disequilibrium estimates. *BMC Bioinformatics* 11:171. doi: 10.1186/1471-2105-11-171
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923
- Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49, 49–67.
- Liu, E. Y., Morgan, A. P., Chesler, E. J., Wang, W., Churchill, G. A., and de Villena, F. P. M. (2014). High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. *Genetics* 197, 91–106. doi: 10.1534/genetics.114.161653
- Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., et al. (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genet.* 11:e1005387. doi: 10.1371/journal.pgen.1005387
- Melzer, N., Wittenburg, D., and Reipsilber, D. (2013). Integrating milk metabolite profile information for the prediction of traditional milk traits based on SNP information for Holstein cows. *PLoS ONE* 8:e70256. doi: 10.1371/journal.pone.0070256
- Merchant, S., Wood, D. E., and Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675. doi: 10.7717/peerj.675
- Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.
- Milanesi, M., Vicario, D., Stella, A., Valentini, A., Ajmone-Marsan, P., Biffani, S., et al. (2015). Imputation accuracy is robust to cattle reference genome updates. *Anim. Genet.* 46, 69–72. doi: 10.1111/age.12251
- Milne, C. P., and Friars, G. W. (1984). An estimate of the heritability of honeybee pupal weight. *J. Hered.* 75, 509–510. doi: 10.1093/oxfordjournals.jhered.a110003
- Moen, T., Hoyheim, B., Munck, H., and Gomez-Raya, L. (2004). A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. *Anim. Genet.* 35, 81–92. doi: 10.1111/j.1365-2052.2004.01097.x
- Ott, J. (1991). *Analysis of Human Genetic Linkage*. London: John Hopkins.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available online at: <http://www.R-project.org/>
- Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., and Georges, M. (2012). Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genet.* 8:e1002854. doi: 10.1371/journal.pgen.1002854
- Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Res.* 20, 1165–1173. doi: 10.1101/gr.101360.109
- Stam, P. (1979). Interference in genetic crossing over and chromosome mapping. *Genetics* 92, 573–594.
- Thomsen, H., Reinsch, N., Xu, N., Bennewitz, J., Looft, C., Grupe, S., et al. (2001). A whole genome scan for differences in recombination rates among three *Bos taurus* breeds. *Mamm. Genome* 12, 724–728. doi: 10.1007/s00335-001-2068-0
- Utsunomiya, A. T., Santos, D. J., Boison, S. A., Utsunomiya, Y. T., Milanese, M., Bickhart, D. M., et al. (2016). Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. *BMC Genomics* 17:705. doi: 10.1186/s12864-016-3049-8
- Vereinigte Informationssysteme Tierhaltung w.V. (vit), (2017). *German Bull Top List (April 2017)*. Available online at: [http://www.vit.de/fileadmin/user\\_upload/vit-fuers-rind/zuchtwertschaetzung/milchrinder-zws-online/SchwarzBunt\\_RZG\\_aktiveBullen.pdf](http://www.vit.de/fileadmin/user_upload/vit-fuers-rind/zuchtwertschaetzung/milchrinder-zws-online/SchwarzBunt_RZG_aktiveBullen.pdf) (Accessed June 21, 2017).
- Weng, Z.-Q., Saatchi, M., Schnabel, R. D., Taylor, J. F., and Garrick, D. J. (2014). Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Gen. Sel. Evol.* 46:34. doi: 10.1186/1297-9686-46-34
- Zhang, X., Roeder, K., Wallstrom, G., and Devlin, B. (2003). Integration of association statistics over genomic regions using Bayesian adaptive regression splines. *Hum. Genomics* 1, 20–29. doi: 10.1186/1479-7364-1-1-20
- Zhao, H., McPeck, M. S., and Speed, T. P. (1995). Statistical analysis of chromatid interference. *Genetics* 139, 1057–1065.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Hampel, Teuscher, Gomez-Raya, Doschoris and Wittenburg. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## RESEARCH ARTICLE

## Open Access



# Male recombination map of the autosomal genome in German Holstein

Saber Qanbari\*  and Dörte Wittenburg

## Abstract

**Background:** Recombination is a process by which chromosomes are broken and recombine to generate new combinations of alleles, therefore playing a major role in shaping genome variation. Recombination frequencies ( $\theta$ ) between markers are used to construct genetic maps, which have important implications in genomic studies. Here, we report a recombination map for 44,696 autosomal single nucleotide polymorphisms (SNPs) according to the coordinates of the most recent bovine reference assembly. The recombination frequencies were estimated across 876 half-sib families with a minimum number of 39 and maximum number of 4236 progeny, comprising over 367 K genotyped German Holstein animals.

**Results:** Genome-wide, over 8.9 million paternal recombination events were identified by investigating adjacent markers. The recombination map spans 24.43 Morgan (M) for a chromosomal length of 2486 Mbp and an average of  $\sim 0.98$  cM/Mbp, which concords with the available pedigree-based linkage maps. Furthermore, we identified 971 putative recombination hotspot intervals (defined as  $\theta > 2.5$  standard deviations greater than the mean). The hotspot regions were non-uniformly distributed as sharp and narrow peaks, corresponding to  $\sim 5.8\%$  of the recombination that has taken place in only  $\sim 2.4\%$  of the genome. We verified genetic map length by applying a likelihood-based approach for the estimation of recombination rate between all intra-chromosomal marker pairs. This resulted in a longer autosomal genetic length for male cattle (25.35 cM) and in the localization of 51 putatively misplaced SNPs in the genome assembly.

**Conclusions:** Given the fact that this map is built on the coordinates of the ARS-UCD1.2 assembly, our results provide the most updated genetic map yet available for the cattle genome.

## Background

Recombination is a process by which chromosomes are broken and recombine to produce new combinations of alleles, so-called haplotypes. Haplotypes possess specific genetic features, and thus play a major role in shaping genome variation. Crossover events are not uniformly distributed and regional rates of crossovers vary considerably across individual genomes and populations mostly because of the combined effects of mutation, recombination, and demographic history [1, 2]. Recombination frequencies between markers are used to construct genetic

maps, which have important implications in genomic studies. High-resolution genetic maps are key elements of a successful fine-mapping program. Moreover, genetic linkage maps are valuable resources for the improvement of chromosome-level assemblies of whole-genome sequences and for comparative genome analyses to name just a few applications.

Genetic maps are built based either on tracing parent-offspring transmission [3, 4], sperm typing [5], or exploiting polymorphism data on a population scale [6, 7]. Given the controlled mating scheme in commercial animals, the primary strategy for the analysis of recombination has been through pedigree to benefit from the fully recorded genealogies. Such an approach traces transmission of haplotypes between pairs of loci

\*Correspondence: [qanbari@fhn-dummerstorf.de](mailto:qanbari@fhn-dummerstorf.de)  
Leibniz Institute for Farm Animal Biology (FBN), Institute of Genetics and Biometry, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany



© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.



from parents to offspring and infer genetic distance based on proportion of recombinant haplotypes.

Cattle have a vital role in the global food system and, given their economic importance, this species was among the first livestock to own a genetic map of recombination, which was built based on microsatellite markers [8, 9]. With the advent of single nucleotide polymorphism (SNP) arrays, studying recombination in farm animal genomes was accelerated with the motivation to assess accurate haplotype phasing and imputation that are required for implementing the genomic selection strategy. Subsequent recombination maps were then constructed at a higher resolution based on genotyping arrays in several beef [10, 11] and dairy breeds [12].

Holstein is the world's most significant cattle breed with a prominent role in producing dairy products. Two recombination studies on Holstein cattle using medium-density genotypes were recently reported. Sandor et al. [13] characterized male bovine meiotic recombinations using 10,192 bulls from the Netherlands and 3783 bulls from New Zealand with 19,487 SNPs in common between the two groups. Ma et al. [4] reported a cattle sex-specific recombination map in a large pedigree of Holstein in the United States.

The recent genetic maps of cattle are built based on the arrays of SNPs that are mapped to the genome assembly UMD3.1 [14]. The emerging advances in long-read sequencing technologies have enabled a better alignment of sequence reads in the ARS-UCD1.2 re-assembly and improved overall continuity by reducing both gaps and inversions by more than 250-fold [15, 16]. The improved marker coordinates in the new assembly facilitate reliable haplotype phasing and imputation, and thus provide appropriate estimation of population genetics parameters such as inter-marker linkage disequilibrium and recombination frequencies, and eventually contribute to the success of gene mapping or genomic prediction projects.

In this study, we take the advantage of 50 K genotypes from a large pedigree of German Holstein cattle to construct an up-to-date genetic map, locate hotspot regions of recombination, and identify candidate genes that contribute to recombination. The novelty of our findings is twofold: (1) given the fact that this study uses the coordinates of the ARS-UCD1.2 assembly, it presents the most updated genetic map yet available for the cattle genome; and (2) we evaluate estimates of recombination rate between intra-chromosomal SNP pairs to identify misplaced markers. Furthermore, we introduce an optimization approach to verify the genetic map length.

## Methods

### Half-sib families from a large pedigree

This study used a large pedigree that includes 367,056 German Holstein cattle, and a subset of the animals have been genotyped for the genomic selection program in Germany. Data were provided by the German Evaluation Center, VIT ([www.vit.de](http://www.vit.de)). The pedigree involved 1053 half-sib families with sires born between 1979 and 2017.

### Genetic material, quality control and imputation

Genetic data involved bi-allelic genotypes of 45,613 autosomal SNPs, which are mapped to the coordinates of the most recent ARS-UCD1.2 assembly (available at <https://bovinegenome.elsiklab.missouri.edu/downloads/ARS-UCD1.2>).

We used the PLINK v1.9 program [17] to clean the data for Mendelian inconsistencies both on the marker and individual levels. Markers that had a Mendelian inheritance error for more than 5% of the individual genotypes were removed. In total, 44,696 SNPs with a minor allelic frequency (MAF) higher than 0.01 and an average inter-marker distance of 55 kb were retained for the subsequent analyses. At the individual level, the Mendelian inconsistency threshold was set to 0.1. Genotypes with a Mendelian inheritance error were set to 'NA' and were imputed in a subsequent step. For the imputation of missing genotypes, we used the Eagle v2.3 software [18], which exploits available pedigree information and is capable of handling very large cohorts of individuals. Program parameters were set to the default values and were run chromosome-wise overnight in a multi-thread module.

### Recombination rates and genetic map positions

Recombination frequencies were estimated across 876 half-sib families, with sires having a minimum number of 39 progeny (see Additional file 1: Figure S1). Exploiting the genetic similarity between paternal half-sibs, the male recombination rate between marker pairs was assessed for each chromosome by the following methods.

### Deterministic approach

The deterministic approach of Ferdosi et al. [19] enabled inference of sire haplotypes from progeny genotypes, thus sire genotypes are not needed. The locations of recombination events and the most likely haplotype phases of a sire were reconstructed by grouping consecutive markers depending on the occurrence of opposite homozygous genotypes among the progeny. We used the implementation of this approach in the R package "hspase" [20] and counted the number of crossovers between adjacent markers in each half-sib

family. The proportion of recombinant haplotypes in a marker interval was then averaged over all the families to estimate recombination rate. Given the close proximity of markers and assuming no interference between successive crossovers, estimated recombination rates were directly converted into genetic distances in Morgan (M) units. The hspase method is limited to adjacent markers only. In addition, these estimates were considered for the evaluation of hotspot regions.

**Likelihood-based approach**

Let  $p$  denote the number of markers on a chromosome. The recombination rate  $\theta_{i,j}$  between each pair of markers  $i$  and  $j$ , ( $i, j = 1, \dots, p; i < j$ ) was estimated using an expectation-maximization approach which relies on likelihood theory [21–23]. This approach uses sire haplotypes that have been reconstructed within each half-sib family by hspase and progeny genotypes. Estimating recombination frequency across all intra-chromosomal marker pairs allowed the identification of markers that are misplaced in the current genome assembly. For this purpose, SNPs with a markedly high recombination rate with the neighboring markers were identified following Hampel et al. [23]. Briefly, the mean recombination rate of  $\theta_{i,j+1}, \dots, \theta_{i,j+30}$  was calculated for all SNPs  $i = 1, \dots, p - 30$ , and the mean of  $\theta_{i-30,i}, \dots, \theta_{i-1,i}$  was taken for  $i = p - 29, \dots, p$ . If the mean recombination rate exceeded the chromosome-wide 99% quantile, the SNP was considered as a misplaced candidate, which was confirmed through subsequent visualization of the increased recombination rate with the following SNPs on a heatmap.

In order to account for possible genotype errors, and to reduce the influence of statistical uncertainty on parameter estimates, we developed a smoothing approach to approximate genetic distances between adjacent markers. Instead of converting recombination rate between adjacent markers only, we considered all the estimates  $\hat{\theta}_{i,j} \leq 0.05$  in a quadratic optimization approach. Then, only a linear relationship between recombination rate and genetic distance was assumed to hold. Let  $d_k$  denote the genetic distance between markers  $k$  and  $k + 1$  in M units. As genetic distances are additive, e.g.  $d_1 + d_2 + d_3$  is the genetic distance that corresponds to  $\theta_{1,4} \leq 0.05$ , the optimization problem was specified in terms of squared deviations:

$$\min_{d_1, \dots, d_{p-1}} \left\{ \sum_{\substack{i, j = 1 \\ i < j \\ \hat{\theta}_{i,j} \leq 0.05}}^p \left( \hat{\theta}_{i,j} - \sum_{k=i}^{j-1} d_k \right)^2 \right\} \text{ s.t. } d_k \geq 0, k = 1, \dots, p - 1.$$

The genetic length of a chromosome was derived as the sum over interval lengths. All steps of the likelihood-based approach were implemented in the R package “hsrecombi” version 0.3.1 that is available at CRAN [24].

**Genome-wide association study for recombination frequency**

A linear mixed model for genome-wide association analysis (GWAS) implemented in the GCTA program [25] was used to identify loci that have large effects on recombination activity. GWAS was conducted on all sires for which the genotype was available. The phenotype for each sire was estimated by averaging the number of recombination events across progeny. We tested the association between each SNP and the phenotype “recombination frequency” using the following model equation:

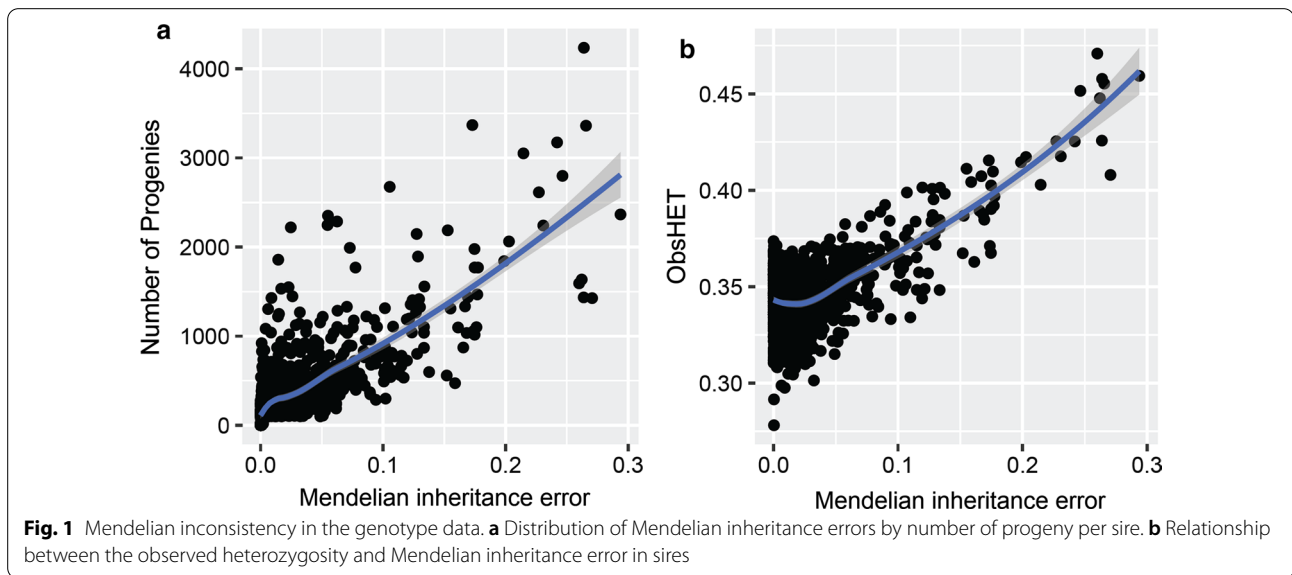
$$\mathbf{y} = \mathbf{X}\mathbf{g} + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

where  $\mathbf{y}$  is the vector of phenotypes for 875 sires,  $\mathbf{X}$  is the design matrix of fixed effects  $\mathbf{g}$ , including a population mean and the additive effect of the candidate SNP,  $\mathbf{Z}$  is the design matrix for a random animal effect  $\mathbf{a}$  with  $\mathbf{a} \sim N(0, \mathbf{G}\sigma_a^2)$  with  $\mathbf{G}$  the genomic relationship matrix of sires, and  $\mathbf{e}$  is the vector of independent and identically distributed residuals.

**Results and discussion**

**Mendelian inconsistency**

Checking genotype data for Mendelian inconsistency is a necessary step to estimate recombination frequencies. A Mendelian inheritance error is defined as the discrepancy between the genotype and pedigree data of two related animals (e.g., parents and offspring). This may result from an error in the recorded pedigree, from genotyping errors, or from mixing up DNA samples, and in very rare cases from mutations [26]. We conducted an exploratory analysis on Mendelian inconsistency in the marker dataset before investigating recombination. A subset of 69 sires showed a Mendelian inconsistency rate higher than 10% for the



genotypes and these were excluded from subsequent analyses. As expected, we observed a positive association of the Mendelian inheritance error with the number of progeny genotyped per sire, which is obviously explained by the number of assessments performed per sire to verify genotypes between sire-offspring (Fig. 1a). Mendelian inconsistency was also positively correlated with sire heterozygosity (Fig. 1b).

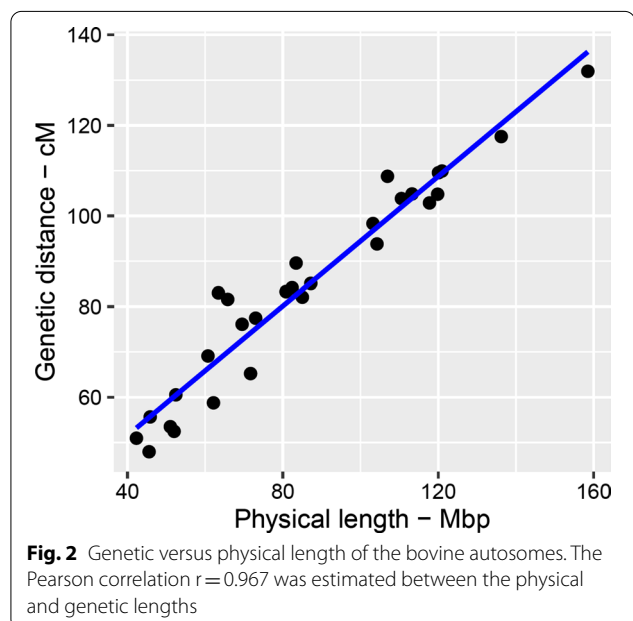
#### Construction of the male recombination map

We built the male recombination map based on genotypes for 44,696 autosomal SNPs with the coordinates derived from the most recent cattle genome assembly. To ensure accurate estimates of recombination frequencies, sires with more than 39 progenies were excluded. Recombination rates were estimated across 876 half-sib families with a maximum number of 4236 progenies (see Additional file 1: Figure S1).

#### Deterministic approach

By tracking paternal meiosis through sire/offspring genotypes, over 8.9 million recombination events were identified genome-wide in the pairwise comparison of adjacent markers. The recent genetic map in US Holstein cattle was constructed across 8.5 million paternal and maternal recombination events; on average, i.e. 36 recombination events per individual across the genome [4].

The recombination map spans 24.43 M on the autosomal genome (Fig. 2). Based on the bovine ARS-UCD1.2 assembly, the total physical length of the autosomes was 2.486 Gbp. The average recombination distance was approximately  $\sim 0.98$  cM per million bp (cM/Mbp). This is fairly consistent with the most recent linkage maps



built by Ma et al. [4] and Sander et al. [13] who reported autosomal genome lengths of 25.5 and 25.7 M, respectively, for male cattle. In cattle, the male recombination map has been reported to be 10% longer than the female map [12]. As a general trend genome-wide, we observed significantly higher recombination rates on short chromosomes than on long chromosomes ( $P$ -value = 0.0003 two-sample  $t$ -test). Accordingly, we found the longest genetic map for *Bos taurus* chromosome (BTA19), which spanned on average  $\sim 1.31$  cM/Mbp, versus  $\sim 0.83$  cM/Mbp for BTA1. The full list of recombination frequencies

between pairs of adjacent markers is in Additional file 2: Table S1 and local recombination rates chromosome-wise are in Additional file 3: Figure S2. The emerging picture is that the recombination activity increased across the middle part of most chromosomes and dropped towards both chromosome ends given the acrocentric nature of bovine autosomes and the underlying differences between the structure of centromeric and telomeric DNA.

#### Likelihood-based approach

The likelihood-based approach generated estimates of  $\theta$  for all intra-chromosomal marker pairs if at least one sire was double heterozygous. For instance, 2902 out of the 2911 SNPs considered on BTA1 yielded only 4,116,903

estimates of recombination rate. In general, the number of estimates was smaller than expected from the SNP number (e.g., 4,209,351 on BTA1) since double heterozygosity was observed only for 98% of all eligible SNP pairs on each chromosome. Note that sires with long runs of homozygosity can still be effective in the estimation of local recombination rates if heterozygous loci occasionally appear. Estimates of  $\theta$  were based on the genotypes of at least 39 or at most 212,823 progeny across families. For instance, on average 42,422 progeny were involved for BTA1. The total genetic map length estimated by using the likelihood-based approach was 25.35 cM, which is in perfect agreement with the most recent linkage maps built by Ma et al. [4] and Sander et al. [13]. On average, the genetic length was 1.05 times

**Table 1 A summary of the statistics of the genetic map for bovine autosomes**

Chr	nSNP	bp	Gap (bp)	Space (kb)	nRec	D (M)	cMMb <sup>-1</sup> (D)	L (M)	cMMb <sup>-1</sup> (L)
1	2911	158,517,589	497,531	54.16	483,569	1.319	0.832	1.269	0.801
2	2355	136,218,516	669,500	57.82	430,718	1.175	0.863	1.156	0.848
3	2190	120,957,517	799,833	55.15	402,820	1.099	0.909	1.152	0.952
4	2164	119,841,669	467,518	55.32	384,044	1.048	0.874	1.088	0.907
5	1868	120,055,511	727,739	64.24	401,460	1.095	0.912	1.142	0.951
6	2213	117,744,633	554,476	53.14	376,980	1.029	0.874	1.046	0.889
7	1945	110,528,375	870,522	56.66	380,527	1.038	0.939	1.062	0.961
8	2104	113,252,524	498,996	53.82	384,270	1.048	0.926	1.030	0.909
9	1761	104,228,150	663,781	59.18	343,786	0.938	0.900	0.973	0.934
10	1852	103,192,471	2,750,827	55.70	360,313	0.983	0.953	1.036	1.004
11	1937	106,932,443	700,224	55.17	398,538	1.087	1.017	1.067	0.998
12	1484	87,186,356	1,266,681	58.64	311,932	0.851	0.976	0.887	1.017
13	1522	83,402,661	726,077	54.53	328,307	0.896	1.074	0.940	1.127
14	1544	82,366,657	575,373	53.20	308,463	0.842	1.022	0.883	1.072
15	1498	85,007,180	727,620	56.38	300,883	0.821	0.966	0.862	1.014
16	1437	80,814,937	693,107	56.11	305,304	0.833	1.031	0.881	1.091
17	1390	72,986,398	779,268	52.45	282,258	0.775	1.061	0.802	1.099
18	1173	65,793,776	872,112	55.58	299,098	0.816	1.240	0.822	1.249
19	1189	63,394,562	674,383	53.01	304,320	0.830	1.310	0.914	1.441
20	1385	71,677,629	546,922	51.56	239,093	0.652	0.910	0.679	0.948
21	1192	69,498,436	737,425	57.95	278,990	0.761	1.095	0.779	1.121
22	1100	60,710,593	465,820	55.08	253,327	0.691	1.138	0.736	1.212
23	943	52,433,171	625,714	55.50	221,849	0.605	1.154	0.637	1.214
24	1091	62,127,707	427,626	56.72	215,397	0.588	0.946	0.654	1.052
25	865	42,292,572	234,159	48.89	186,752	0.510	1.205	0.566	1.338
26	944	51,990,348	367,274	54.36	192,294	0.525	1.009	0.578	1.113
27	862	45,553,866	1,148,867	52.75	175,767	0.480	1.053	0.553	1.214
28	840	45,834,413	338,885	54.34	203,987	0.557	1.214	0.553	1.206
29	937	51,028,789	1,374,496	54.06	196,068	0.535	1.048	0.607	1.190
#	44,696	2,485,569,449	2,750,827	55.46	8,951,114	24.426	0.983	25.354	1.020

bp: chromosome length in base pairs; Gap: maximum gap size between pairs of adjacent markers; Space: inter-marker space; nRec: number of cross-overs detected; D (M): genetic length in Morgan estimated based on deterministic approach; L (M): genetic length in Morgan estimated with the likelihood-based approach.

#Depending on the parameter, either mean (Space, cMMb<sup>-1</sup>), maximum (Gap) or sum (nSNP, BP, nRec, Morgan) is represented

longer with the likelihood-based approach than with the deterministic approach (see Table 1) in which only a fraction of the available information was exploited (i.e., 2910 estimates of genetic distances between SNPs on BTA1). The relationship between genetic and physical positions is shown chromosome-wise in Additional file 4: Figure S3. Whereas a linear relationship was obtained for some chromosomes (e.g., BTA25 and 27), an S-shaped curve was found for most of the other chromosomes, which is explained by the variation of local recombination rates as stated above. The visible gaps on BTA10, 27 and 29 are in Table 1.

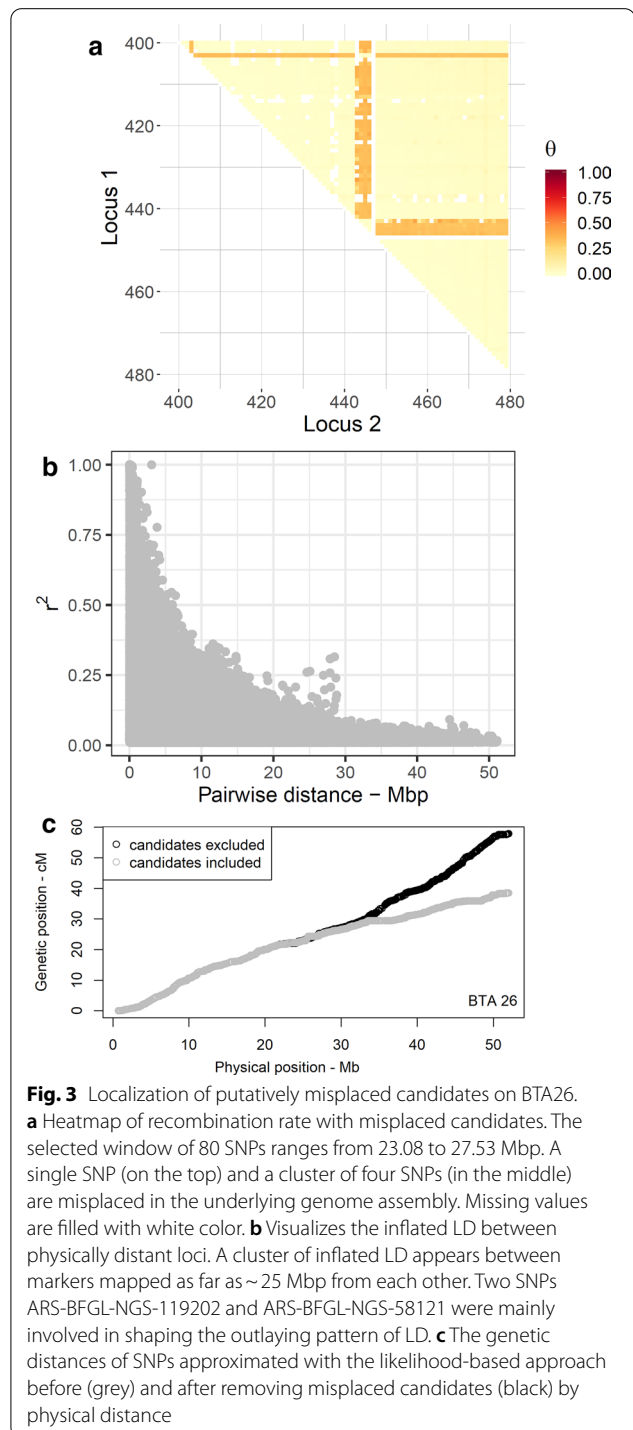
### Candidates of misplaced SNPs

Although we used coordinates of the most recent genome assembly, we draw attention to the fact that some remaining errors in the genome assembly such as misplaced markers may still lead to erroneous assessment of recombination frequencies and a spurious hotspots landscape [27]. This suggests that the identified hotspots could be targets for further investigations to correct the genome assembly. We followed two strategies in parallel to position misplaced markers and to circumvent the false-positive recombination assessments.

We searched for markers with markedly high recombination rates to the neighboring markers according to Hampel et al. [23]. In total, we found that 51 of the SNPs mapped on 18 chromosomes were putatively misplaced in the ARS-UCD1.2 assembly. As an example, on BTA26, single SNPs that mapped at positions 23.16 and 51.58 Mbp together with a cluster of four SNPs at 25.65–25.76 Mbp revealed increased recombination rates with all other SNPs (see Fig. 3a). The full list of misplaced candidates is in Additional file 5: Table S2.

Alternatively, we used the linkage disequilibrium (LD) between markers to verify putatively misplaced SNPs. To this end, sire haplotypes were reconstructed in hspbase, and the LD that was estimated as the allelic correlation ( $r^2$ ) between pairs of markers was plotted as a function of physical distance. The pattern of LD decay revealed clusters of inflated LD between loci that were physically mapped as far as several millions of bp from each other, which indicates misplaced markers even in the recent assembly (Fig. 3b). LD analysis successfully detected the misplaced candidates that were detected by the likelihood-based approach. The subsequent removal of misplaced SNPs resulted in a smooth decay of LD as a function of inter-marker distance, which provided evidence that the methodology used to detect these markers was appropriate.

Excluding SNPs with a putatively wrong physical position is also essential for a proper approximation of the genetic distances. For example, the genetic lengths



**Fig. 3** Localization of putatively misplaced candidates on BTA26.

**a** Heatmap of recombination rate with misplaced candidates. The selected window of 80 SNPs ranges from 23.08 to 27.53 Mbp. A single SNP (on the top) and a cluster of four SNPs (in the middle) are misplaced in the underlying genome assembly. Missing values are filled with white color. **b** Visualizes the inflated LD between physically distant loci. A cluster of inflated LD appears between markers mapped as far as ~25 Mbp from each other. Two SNPs ARS-BFGL-NGS-119202 and ARS-BFGL-NGS-58121 were mainly involved in shaping the outlying pattern of LD. **c** The genetic distances of SNPs approximated with the likelihood-based approach before (grey) and after removing misplaced candidates (black) by physical distance

of BTA26 and 23 were estimated to be respectively 49% and 2% longer when misplaced candidates were excluded (e.g., see Fig. 3c). In contrast, the estimated genetic lengths of BTA1 and 28 declined by 5% and 2%, respectively, after removing the misplaced candidates. The genetic length of the remaining chromosomes was



almost unaffected. Thus, we argue that the application of the likelihood-based approach followed by a verification step based on LD analysis can be efficiently used to screen marker panels of different densities for putatively misplaced SNPs. Improved map coordinates will eventually contribute to the success of gene mapping studies that are conducted based on available genotyping arrays in different species.

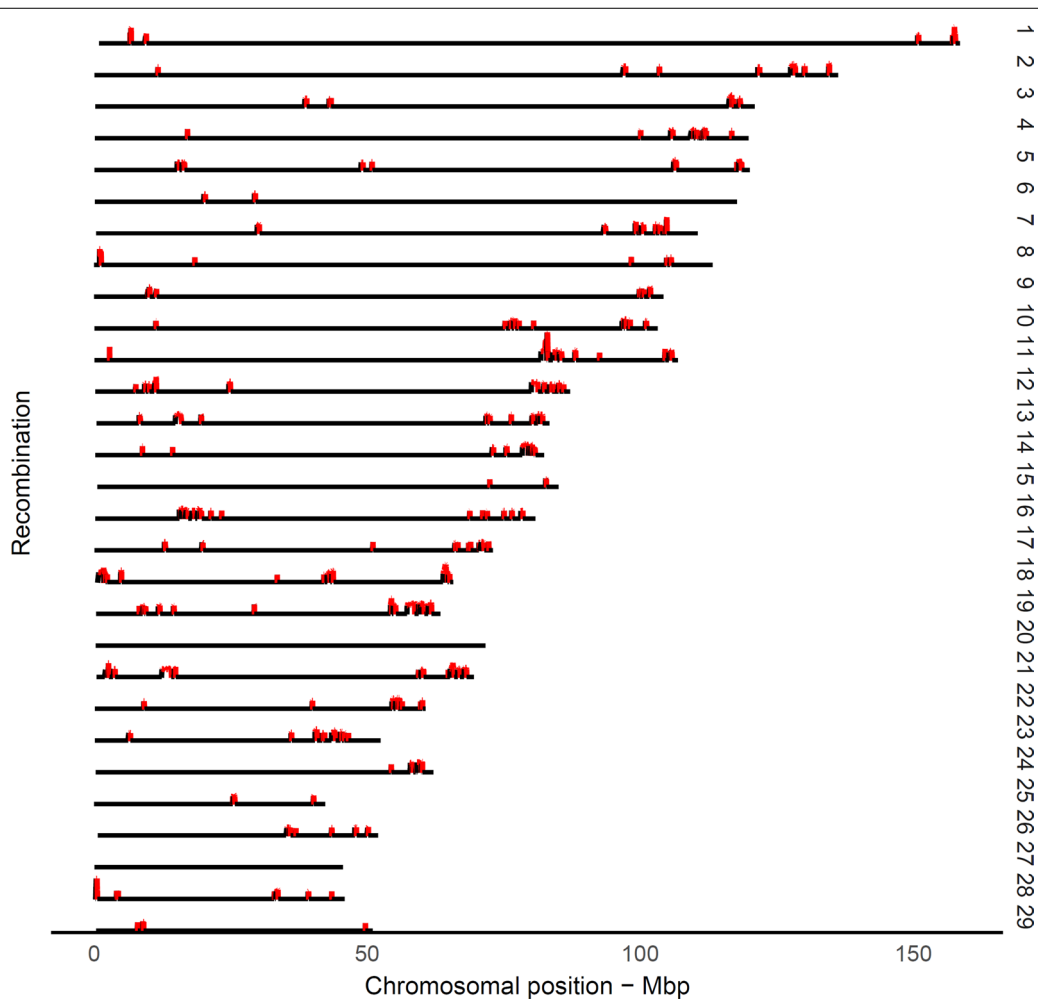
#### Deterministic versus likelihood-based approach

We applied two approaches to estimate genetic map length and found that the lengths obtained differed by about 4% (almost 1 M). A simulation study on the verification of the two approaches is provided in Additional file 6. The accuracy of the genetic distances that were obtained from the likelihood-based approach was higher than that of the deterministic estimates with a difference in total genetic length of the same order of magnitude

as in the real data analysis. However, both approaches underestimated the simulated genetic distances, and more research is needed to improve these estimation methods. Still, we decided to present both approaches since they possess different advantages: the deterministic approach allowed the elucidation of hotspot intervals and enabled identification of genome regions associated with recombination activity. Although, in principle, the likelihood-based approach is also applicable for the verification of hotspot regions, only this approach made it possible to clearly pinpoint putatively misplaced SNPs in the genome assembly.

#### Landscape of recombination hotspots

Following Ma et al. [4], we defined a hotspot region as a region with a recombination rate exceeding 2.5 standard deviations from the genome-wide average of recombination rates. The landscape of highly recombinant



**Fig. 4** Genome-wide landscape of recombination hotspot intervals. The putative hotspot interval was defined as having a recombination rate with more than 2.5 standard deviations greater than the genome-wide mean of recombination rates

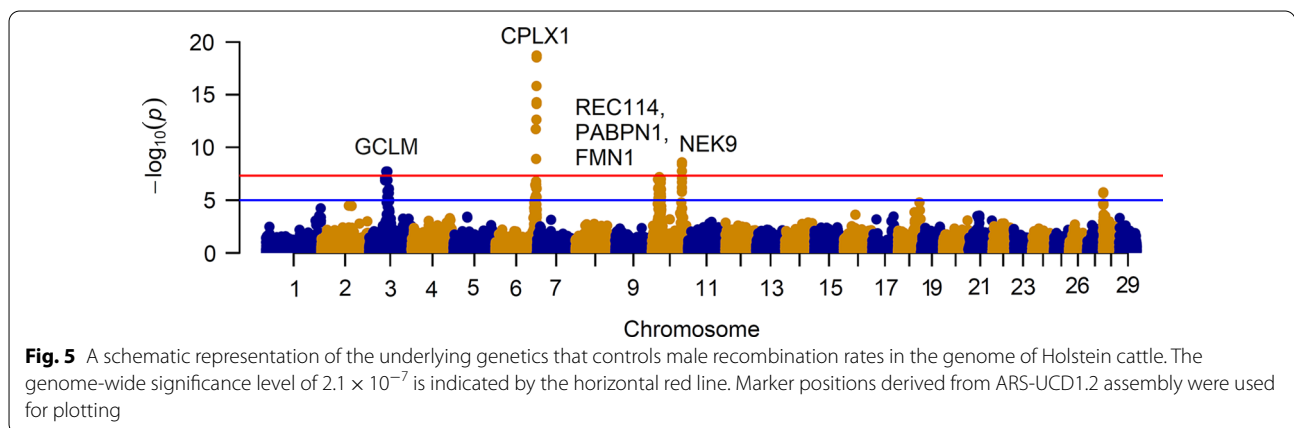
intervals or hotspot regions emerged as sharp and narrow peaks that occurred for a small proportion of the genome (Fig. 4). As expected, hotspot regions were non-uniformly distributed across the genome, which is consistent with previous observations in other mammals [2, 7]. After removing spurious hotspot intervals due to misplaced SNPs, a panel of 971 putative hotspot intervals were identified that represented ~5.8% of the recombination that occurred in only ~2.4% of the genome (see Additional file 1: Table S1). Previous studies in cattle based on medium-density SNP panels reported rather similar numbers of putative hotspot intervals. For example, Ma et al. [4] detected 1792 male putative hotspot regions that represented 3% of the genome. Another study identified 1378, 1295, and 1317 hotspot regions in Jersey, Brown Swiss, and Ayrshire breeds, respectively [12]. In contrast, studies on recombination in the human and mouse genomes that used full re-sequencing or very dense genotyping data, identified ~33,000 [7] and ~47,000 [2] hotspots, respectively, given that the genome size in mammals is comparable. In humans, ~80% of the crossovers map to ~10 to 20% of the genome, where the typical length of hotspots is less than 5 kb [28]. It is worth noting that, in our study, hotspots were localized by using a medium-density panel of SNPs with an average inter-marker space of 55 kb, thus they cannot be directly compared to the hotspots detected by using sequence data in the human or mouse genome. Therefore, we used the term “hotspot interval” instead of “hotspot” to report highly recombinant regions. In addition, it should be noted that the variation between physical inter-marker intervals (see Table 1) has not been taken into account in the definition of hotspot intervals, where the larger gaps between adjacent markers are expected to result in increased recombination frequencies.

#### Genomic regions associated with recombination frequency

We conducted a GWAS including 875 sires of half-sib families to detect genes that influence the trait

“recombination frequency”. The genome-wide number of crossovers that occur between adjacent SNP pairs of each sire was treated as the phenotype. Given the fact that all sires were genotyped on the same panel of markers, we ruled out the possible effect of SNP density on the number of crossovers counted for each sire. Over 8.9 million recombination events were detected across the half-sib families which was a sufficiently large number to result in accurate estimates of recombination frequencies.

Our results show that there are standing alone signals on a number of chromosomes and that the strongest candidates are on BTA3, 6 and 10 (Fig. 5), which is consistent with previous reports in cattle [4, 12, 13]. Applying a Bonferroni correction, a genome-wide threshold of  $2.1 \times 10^{-7}$  was set to identify significant signals. In total, 24 significant SNPs (corrected P-value  $\leq 0.01$ ), emerged with a strong effect on recombination activity (Table 2). The signal with the strongest effect ( $P = 1.89 \times 10^{-19}$ ) corresponded to SNP ARS-BFGL-NGS-110507 on BTA6, which co-localized with a region that contains three candidate genes, *CPLX1*, *GAK* and *PCGF3*. *CPLX1* encodes a protein that belongs to a family of cytosolic proteins, which have a role in synaptic vesicle exocytosis and are reported to be associated with sex-related variation of recombination frequency in sheep [29] and cattle [4, 12]. We also mapped two strong candidate SNPs on BTA10. The first signal had a maximum peak at SNP Hapmap47676-BTA-61231 ( $P = 6.75 \times 10^{-8}$ ) and was localized in the vicinity of several meiosis-related genes including *REC8*, *REC114*, and *FMN1*. *REC8* is a key component of the meiotic cohesion complex, and is associated with recombination activity in cattle [4, 12, 13], mouse [30] and Red Deer [31]. The second significant signal was associated with SNP BTA-78285-no-rs ( $P = 2.68 \times 10^{-9}$  on BTA10 and overlapped with the *NEK9* gene. *NEK9* mediates cell cycle progression that is essential for interphase progression during oocyte formation [32, 33] and is



**Fig. 5** A schematic representation of the underlying genetics that controls male recombination rates in the genome of Holstein cattle. The genome-wide significance level of  $2.1 \times 10^{-7}$  is indicated by the horizontal red line. Marker positions derived from ARS-UCD1.2 assembly were used for plotting



**Table 2 Summary of the statistics of SNPs associated with recombination frequency**

Chr	SNP	bp	Frequency	P-value	BF	Candidate gene
3	INRA-598	45,930,136	0.21	$8.67 \times 10^{-8}$	0.003	
3	ARS-BFGL-NGS-112152	45,978,363	0.21	$1.42 \times 10^{-7}$	0.006	
3	Hapmap59096-rs29024776	49,181,271	0.21	$1.97 \times 10^{-8}$	0.001	<i>GCLM</i>
3	Hapmap58808-rs29017431	52,452,892	0.20	$1.98 \times 10^{-8}$	0.001	
3	INRA-170	52,783,346	0.20	$1.32 \times 10^{-7}$	0.006	
6	ARS-BFGL-NGS-28350	114,972,434	0.26	$1.88 \times 10^{-12}$	0.000	
6	ARS-BFGL-NGS-18656	115,871,184	0.39	$1.28 \times 10^{-9}$	0.000	
6	ARS-BFGL-NGS-104112	115,942,196	0.30	$1.73 \times 10^{-7}$	0.007	
6	ARS-BFGL-NGS-61359	116,788,648	0.28	$2.36 \times 10^{-13}$	0.000	
6	ARS-BFGL-NGS-10037	117,015,280	0.27	$1.50 \times 10^{-16}$	0.000	
6	ARS-BFGL-NGS-112242	117,124,190	0.29	$7.68 \times 10^{-15}$	0.000	
6	BTB-00284077	117,271,685	0.27	$5.18 \times 10^{-15}$	0.000	
6	ARS-BFGL-NGS-117763	117,368,760	0.29	$2.89 \times 10^{-19}$	0.000	
6	ARS-BFGL-NGS-110507	117,390,034	0.29	$1.89 \times 10^{-19}$	0.000	<i>CPLX1, GAK, PCGF3</i>
10	ARS-BFGL-NGS-99693	17,886,463	0.61	$1.09 \times 10^{-7}$	0.004	
10	Hapmap47676-BTA-61231	21,768,228	0.15	$6.75 \times 10^{-8}$	0.003	
10	ARS-BFGL-NGS-19822	22,284,939	0.42	$2.09 \times 10^{-7}$	0.009	
10	ARS-BFGL-NGS-42815	25,998,000	0.59	$1.08 \times 10^{-7}$	0.004	<i>REC114</i>
10	ARS-BFGL-NGS-118433	26,023,168	0.59	$1.08 \times 10^{-7}$	0.004	
10	BTB-00438757	86,199,353	0.40	$1.85 \times 10^{-8}$	0.001	
10	Hapmap57084-ss46526565	86,260,186	0.40	$4.17 \times 10^{-9}$	0.000	
10	BTB-00438922	86,284,751	0.40	$2.05 \times 10^{-7}$	0.009	
10	BTA-78285-no-rs	86,322,591	0.55	$2.68 \times 10^{-9}$	0.000	<i>NEK9</i>
10	UA-IFASA-7857	86,379,951	0.56	$8.37 \times 10^{-8}$	0.003	

BF: Bonferroni adjusted P-value for multiplicity

associated with crossover interference levels [34] and recombination activity in mammals [4]. Another significant signal peaked at SNP Hapmap59096-rs29024776 ( $P = 1.97 \times 10^{-8}$ ) in a gene-rich region on BTA3. Although a statistical association with a QTL has already been reported [4], a biological association with the neighboring candidate genes needs to be established.

The above-mentioned regions are implicated in recombination variation at the individual level in humans, cattle and mice, which suggests a common genetic architecture of recombination activity in mammals. The variation observed in genome-wide recombination frequency among sires can be used as an opportunity to maintain the genomic diversity of intensively selected dairy cattle, which has been shrinking for decades.

## Conclusions

We present a bovine genetic map with a medium SNP density resolution based on a large pedigree of German Holstein animals. The deterministic approach used recombination frequencies between adjacent markers to construct the genetic map that spans 24.4 M with an average length of  $\sim 0.98$  cM/Mbp-1. We identified 971

highly recombinant marker intervals/hotspot regions that were non-uniformly distributed across  $\sim 2.4\%$  of the genome. The likelihood-based approach resulted in a genetic length of 25.3 M, which fits better with the available linkage map lengths. Taking benefit of all pairwise recombination estimates, the likelihood-based approach was able to localize 51 SNPs that were putatively wrongly assigned on the physical map. The genome-wide association study identified several candidate loci including *REC8*, *REC114*, *FMN1* and *CPLX1* that affect recombination frequency. Our results successfully validated those of previous reports on the genetics that underlies recombination activity in cattle. Given the fact that this map is built on the coordinates of the ARS-UCD1.2 assembly, our results provide the most updated genetic map yet available for the cattle genome. The map presented in this study will be useful for both breeders and researchers and will support further investigation of the genome of this economically important species. The R package and workflow provided will allow to estimate the length of the genetic map of other breeds and thus will facilitate future comparisons of the genome characteristics between breeds.

## Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12711-020-00593-z>.

**Additional file 1: Figure S1.** Number of progeny per sire of half-sib families.

**Additional file 2: Table S1.** Genetic-map coordinates. The table contains the genetic-map coordinates that were estimated from deterministic (cM\_deterministic) and likelihood-based (cM\_likelihood) approaches in Holstein cattle. Marker physical coordinates (Mbp\_position) are based on the ARS-UCD1.2 genome assembly. Furthermore, recombine\_adjacent\_deterministic denotes the recombination rate between adjacent markers based on the deterministic approach.

**Additional file 3: Figure S2.** Illustration of the recombination rate along the chromosomes. The figure shows the relationship between recombination rate between adjacent markers based on the deterministic approach and the relative physical position for each chromosome.

**Additional file 4: Figure S3.** Physical genetic maps for each chromosome. The figure shows the relationship between physical and genetic-map coordinates for each chromosome.

**Additional file 5: Table S2.** Panel of misplaced candidates in the ARS-UCD1.2 genome assembly. The table lists all the markers that are putatively misplaced in the underlying genome assembly, as revealed by the likelihood-based approach. Physical position (bp) corresponds to ARS-UCD1.2 genome assembly; the index refers to consecutive numbering of SNPs to facilitate the identification of clusters.

**Additional file 6.** Description of the simulation study to compare genetic map positions derived from the deterministic and likelihood-based approach [20–24, 35, 36].

### Acknowledgements

We gratefully acknowledge the generous support of the Association for Bioeconomy Research (FBF, Bonn) as representative of German cattle breeders for participating in this project and the German Evaluation Center (VIT, Verden) for providing the genotype data. Especially, Erik Pasman and Fritz Reinhardt (VIT, Verden) assisted and promoted this project. We further thank Friedrich Teuscher (FBN Dummerstorf) for his valuable contribution.

### Authors' contributions

DW conceived the idea and developed the project. Both authors performed the computations, drafted the manuscript and contributed to its final shape. Both authors read and approved the final manuscript.

### Funding

Open Access funding enabled and organized by Projekt DEAL. This research was conducted within the frame of the project CLARITY and financially supported by the grant from the German Federal Ministry of Education and Research (BMBF, FKZ 031L0166 CompLS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Availability of data and materials

The data supporting the findings of this study was provided by the German Evaluation Center (VIT, Verden) and their access is restricted by the provider to be publicly available. Permission for data access was granted by the Association for Bioeconomy Research (FBF, Bonn). The R package *hsrecombi* version 0.3.1 is available at CRAN; it provides tools for estimating recombination rates between marker pairs, identifying candidates of misplaced markers and approximating genetic-map positions.

### Ethics approval and consent to participate

Not applicable as no experimental study was carried out.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 10 July 2020 Accepted: 26 November 2020

Published online: 14 December 2020

### References

- Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*. 2010;467:1099–103.
- Brunschwig H, Levi L, Ben-David E, Williams RW, Yakir B, Shifman S. Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics*. 2012;191:757–64.
- Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans RPMA, et al. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res*. 2009;19:510–9.
- Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, et al. Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genet*. 2015;11:e1005387.
- Tiemann-Boege I, Calabrese P, Cochran DM, Sokol R, Arnheim N. High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. *PLoS Genet*. 2006;2:e70.
- McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science*. 2004;304:581–4.
- Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. *Science*. 2005;310:321–4.
- Bishop MD, Kappes SM, Keele JW, Stone RT, Sunden SL, Hawkins GA, et al. A genetic linkage map for cattle. *Genetics*. 1994;136:619–39.
- Ihara N, Takasuga A, Mizoshita K, Takeda H, Sugimoto M, Mizoguchi Y, et al. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Res*. 2004;14:1987–98.
- Weng Z-Q, Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genet Sel Evol*. 2014;46:34.
- Mouresan EF, González-Rodríguez A, Cañas-Álvarez JJ, Munilla S, Altarriba J, Díaz C, et al. Mapping recombination rate on the autosomal chromosomes based on the persistency of linkage disequilibrium phase among autochthonous beef cattle populations in Spain. *Front Genet*. 2019;10:1170.
- Shen B, Jiang J, Seroussi E, Liu GE, Ma L. Characterization of recombination features and the genetic basis in multiple cattle breeds. *BMC Genomics*. 2018;19:304.
- Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. Genetic variants in *RECB*, *RNF212*, and *PRDM9* influence male recombination in cattle. *PLoS Genet*. 2012;8:e1002854.
- Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow *Bos taurus*. *Genome Biol*. 2009;10:R42.
- Bickhart DM, McClure JC, Schnabel RD, Rosen BD, Medrano JF, Smith TPL. Symposium review: advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection. *J Dairy Sci*. 2020;103:5278–90.
- Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*. 2020;9:giaa021.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
- Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the haplotype reference consortium panel. *Nat Genet*. 2016;48:1443–8.
- Ferdosi MH, Kinghorn BP, van der Werf JHJ, Gondro C. Detection of recombination events, haplotype reconstruction and imputation of sires using half-sib SNP genotypes. *Genet Sel Evol*. 2014;46:11.

20. Ferdosi MH, Kinghorn BP, van der Werf JH, Lee SH, Gondro C. hspbase: an R package for pedigree reconstruction, detection of recombination events, phasing and imputation of half-sib family groups. *BMC Bioinformatics*. 2014;15:172.
21. Gomez-Raya L. Maximum likelihood estimation of linkage disequilibrium in half-sib families. *Genetics*. 2012;191:195–213.
22. Gomez-Raya L, Hulse AM, Thain D, Rauw WM. Haplotype phasing after joint estimation of recombination and linkage disequilibrium in breeding populations. *J Anim Sci Biotechnol*. 2013;4:30.
23. Hampel A, Teuscher F, Gomez-Raya L, Doschoris M, Wittenburg D. Estimation of recombination rate and maternal linkage disequilibrium in half-sibs. *Front Genet*. 2018;9:186.
24. Wittenburg D. hsrecombi: estimation of recombination rate and maternal LD in half-sibs. URL <https://cran.r-project.org/package=hsrecombi>. Accessed 18 Mar 2020.
25. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88:76–82.
26. Calus MPL, Mulder HA, Bastiaansen JWM. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. *Genet Sel Evol*. 2011;43:34.
27. Druet T, Georges M. Pedigree-based haplotype reconstruction, identification of cross-overs and detection of map and genotyping errors using PHASEBOOK. In: *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production: 17–22 August 2014; Vancouver*. 2014.
28. Paigen K, Petkov P. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet*. 2010;11:221–33.
29. Johnston SE, Bérénos C, Slate J, Pemberton JM. Conserved genetic architecture underlying individual recombination rate variation in a wild population of Soay sheep (*Ovis aries*). *Genetics*. 2016;203:583–98.
30. Bannister LA, Reinholdt LG, Munroe RJ, Schimenti JC. Positional cloning and characterization of mouse mei8, a disrupted allele of the meiotic cohesin Rec8. *Genesis*. 2004;40:184–94.
31. Johnston SE, Huisman J, Pemberton JM. A genomic region containing REC8 and RNF212B is associated with individual recombination rate variation in a wild population of Red deer (*Cervus elaphus*). *G3 (Bethesda)*. 2018;8:2265–76.
32. Tan BCM, Lee SC. Nek9, a novel FACT-associated protein, modulates interphase progression. *J Biol Chem*. 2004;279:9321–30.
33. Yang SW, Gao C, Chen L, Song YL, Zhu JL, Qi ST, et al. Nek9 regulates spindle organization and cell cycle progression during mouse oocyte meiosis and its location in early embryo mitosis. *Cell Cycle*. 2012;11:4366–77.
34. Wang Z, Shen B, Jiang J, Li J, Ma L. Effect of sex, age and genetics on crossover interference in cattle. *Sci Rep*. 2016;6:37698.
35. Turlach BA. quadprog: functions to solve quadratic programming problems. 2019. <https://cran.r-project.org/package=quadprog>. Accessed 3 Nov 2020.
36. Haldane JBS. The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*. 1919;8:299–309.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)





# Chapter 4

## General discussion

In this thesis, I explored two perspectives of dependence among SNPs. (i) Interactions of SNP variants within locus and between loci can have distinct impact in trait expression (Chapter 2). (ii) Dependencies between SNP genotypes occur due to linkage and linkage disequilibrium (LD) among markers and lead to multicollinearity in a genome-wide regression model (Chapter 3). Relationships between these issues are highlighted in the following discussion.

The investigations concentrate on data collected in half-sib families which is a typical family structure in livestock. However, an extension to consider full-sib families has been worked out, providing a general framework for future studies in various livestock species.

### 4.1 Model parameterisation for non-additive genetic effects

Statistical approaches are being studied to elucidate sources of genetic variation, to evaluate the association between genetic and phenotypic variation and to better understand the genetic architecture of livestock populations. Genetic architecture is not only characterised by the size of the genetic effect captured by a molecular marker (e.g. SNP) and the position on the genome but also by the mode of inheritance, which might be additive or dominant, and any kind of interaction. Additive  $a = (a_1, \dots, a_p)'$  and dominance effects  $d = (d_1, \dots, d_p)'$  at  $p$  SNPs and interaction effects  $e = (e_{1,1}, e_{1,2}, \dots, e_{p,p})'$  between  $p(p-1)/2$  pairs of SNPs can be jointly considered in a linear regression model fitted to a trait  $y = (y_1, \dots, y_n)'$ ,

$$y = Xa + Dd + We + \varepsilon, \quad (4.1)$$

with design matrices  $X$ ,  $D$  and  $W$ , respectively, of corresponding dimensions. The  $j$ -th column of  $X$  contains the individual genotype codes as counts of the alternate allele at SNP  $j$ , i.e.  $X_{t,j} \in \{0, 1, 2\}$  for animal  $t = 1, \dots, n$ . In  $D$ , an individual's genotype is coded as 1 if a heterozygous SNP genotype has been observed and 0 otherwise. A column of  $W$  is obtained from element-wise multiplication of columns of the design matrices for main effects. For instance, it is  $W_k = X_i X_j$  for an additive  $\times$  additive interaction between loci  $i$  and  $j$  or  $W_k = D_i X_j$  for a dominance  $\times$  additive interaction with index  $k = (i-1)p - \sum_{l=1}^{i-1} l + j$ . The index  $k$  reflects the ordering of interaction effects. To begin with, interactions between the first SNP  $i = 1$  and any other second SNP  $j = i + 1, \dots, p$  are built consecutively. Then, index  $i$  of the first SNP increases by one and all combinations with subsequent

SNPs  $j = i + 1, \dots, p$  follow and so on. In general, it is possible to model all four kinds of epistatic effects simultaneously with design matrices  $W_{a \times a}$ ,  $W_{a \times d}$ ,  $W_{d \times a}$  and  $W_{d \times d}$  (instead of  $W$ ) leading to  $2p^2$  genetic effects in total.

The parameterisation (i.e. the genotype codes in  $X$ ,  $D$ , and  $W$ ) according to model (4.1) reflects genetic effects at the genotype level – this in turn shall resemble their biological meaning. However, this coding harbours limitations for the statistical approach. Genotype codes for additive and dominance effects are dependent on each other because the heterozygous genotype is considered twice, i.e.  $D_{t,j} = 1$  only if  $X_{t,j} = 1$  for individual  $t$  at SNP  $j$ . This dependence carries over to epistatic effects by construction but can be circumvented by introducing an orthogonal decomposition of genetic effects that relies on allele frequencies of the underlying population (Cockerham, 1954; Vitezica et al., 2013; Zeng et al., 2005). The decomposition of effects can be generalised to consider genotype instead of allele frequencies proving useful if the population is not in Hardy-Weinberg equilibrium, denoted as the “natural and orthogonal interactions” approach (NOIA; Álvarez-Castro and Carlborg, 2007; Vitezica et al., 2017).

When searching for interacting loci, estimating the effect size and identifying the significant marker pairs increases the computational burden in terms of speed and memory allocation dramatically. In Paper 2.1, I selected an approximate Bayesian approach “fastbayes” (Meuwissen et al., 2009) and extended it to include dominance and all kinds of pairwise epistatic effects. In this approach, the analytically derived posterior mean of the marker-effect distribution immediately provides an estimate of genetic effect. Time-consuming Gibbs sampling or other Markov-chain Monte Carlo methods as typically used in other Bayesian approaches for genomic evaluations are avoided (as reviewed by de los Campos et al., 2013). The clue was that the association between each marker and the phenotype  $y$  was considered autonomously as in a single-marker model but  $y$  was corrected by all other (previously) estimated genetic effects. This approach also required iterating over estimates of genetic effects until convergence but the number of iterations was typically low (a few dozens were required if 0.4 % of markers had causal effects but the number of iterations was increased by one order of magnitude if 4.4 % of markers were causative). It was shown in a simulation study resembling a dairy cattle population that the accuracy of prediction of the total genetic values improved if dominance effects were included but epistasis complicated matters (Tab. 4 in Paper 2.1). Only if broad-sense heritability was high ( $H^2 = 0.5$ ) and epistatic effects were present, modelling epistatic effects truly increased the accuracy of predicting total genetic values. In complement to the published results in Paper 2.1, accuracy of breeding value estimation, measured as correlation between estimated additive genetic values and simulated breeding values, was almost unaffected by modelling non-additive effects in general, see Table 4.1. The NOIA parameterisation was used here but this hardly affected the outcome because the simulated population was almost always in HWE.

No matter which parameterisation is chosen, it leads to a decomposition of effects being orthogonal only in the absence of LD (Cockerham, 1954; Hill and Mäki-Tanila, 2015). Especially in dense marker panels, a certain amount of linkage between SNPs exists, introducing multicollinearity in  $X$  and  $D$  and hence in  $W$ . Thus, estimates of genetic effects and corresponding variance components are likely biased. This is critical from the population geneticists’ perspective because the additive genetic variance is particularly crucial for breeding decisions. Hill and Mäki-Tanila (2015) analytically derived formulas for genetic variation including LD between pairs of markers and observed “interaction

Broad-sense heritability	Model	Accuracy (w/o)	Accuracy (w)
0.1	a	0.881	0.737
0.1	a + d	0.878	0.736
0.1	a + d + e	0.833	0.590
0.3	a	0.967	0.905
0.3	a + d	0.967	0.905
0.3	a + d + e	0.960	0.895
0.5	a	0.985	0.949
0.5	a + d	0.985	0.951
0.5	a + d + e	0.983	0.953

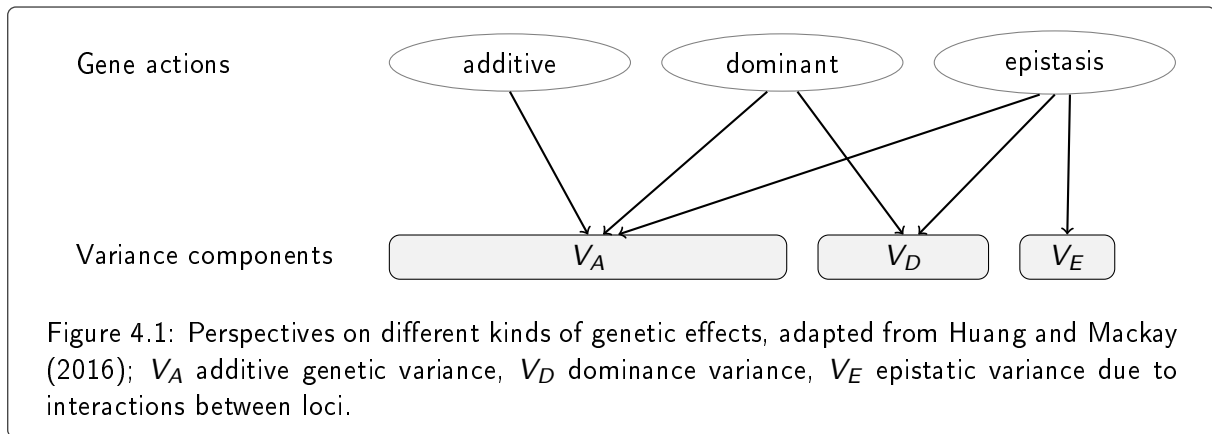
Table 4.1: Accuracy of breeding value prediction in a simulation study without (w/o) or with (w) epistatic effects. Statistical model includes additive (a), dominance (d) and all kinds of pairwise epistatic (e) effects.

variance to be higher in the presence of LD [...] but not dramatically so". This conclusion, however, is dependent on the model parameters chosen in their investigation. In Paper 2.1, it was observed that assuming LE led to almost unbiased estimates of all kinds of genetic variance components if only few loci (i.e. 0.4 % of the markers; average  $r^2 = 0.07$  between markers) had causal effects. If 4.4 % of the markers were causative (average  $r^2 = 0.12$  between markers), interaction variance were biased up to 22 % but additive genetic and dominance variance were slightly biased (up to 7 and 1 %, respectively).

Figure 4.1 summarises the relationship between the two levels of handling genetic effects – the genotype level describing the gene action and the (reparameterised) statistical level giving feedback on genetic variation – corresponding to Huang and Mackay (2016). It visualises that all kinds of genetic effects contribute to the additive genetic variance. However, there is no direct link between the size of genetic variance component and the importance of gene action. It is worth mentioning, that the "importance" of non-additive effects also depends on the way of parameterisation (Martini et al., 2019). Particular attention has to be paid to the epistatic variance, ( $V_E$  in Fig. 4.1) which "is the residual genetic variance after  $V_A$  has been maximized and bears no genetic meaning" (Huang and Mackay, 2016). Thus, variance components do not really help tracing back the actual kind of effect (i.e. gene action) (Huang and Mackay, 2016). However, marker pairs with large interaction effects give good reason for a follow-up study on the biological relevance based on specific experimental designs (such as backcross, Miller et al., 2020).

Hence, not only quantifying genetic effect sizes but also verifying its significance is decisive in genomic evaluations. I continued working with the rapid Bayesian approach "fastbayes" and derived a measure of evidence to select markers with significant effect on a trait (Paper 2.2). It is based on the credibility of the highest posterior density interval next to zero. This methodology was applied to simulated data in order to verify sensitivity of testing. Moreover, it was compared to a variational Bayes approach which is comparative in speed and sensitivity (Logsdon et al., 2012). Sensitivity and specificity of fastbayes were similar to the variational Bayesian method, and a further reduction of computing time could be achieved. Both approaches identified up to one third of the simulated causal variants with only few false-positive detections. The validation study showed that both approaches





are powerful dual-purpose tools for genomic inferences – they are applicable to predict future outcome of not-yet phenotyped individuals with high precision as well as to estimate and test single-marker effects – but only fastbays allowed the estimation of billions of interaction effects.

A real data application complemented the study above. The same data and methodology as in Paper 2.1 were used but the treatment of a hyperparameter which was responsible for the proportion on non-zero genetic effects changed. This might have improved the statistical approach at the price that epistatic effects were no longer present. A previously found strong dominance $\times$ additive interaction effect between loci on two chromosomes probably split into separate dominance and additive effects on the initially found chromosomes. Apart from this technically reasonable explanation, sometimes epistatic effects can be attributed to “phantom epistasis” (de los Campos et al., 2019): incomplete LD between markers and causal variant may lead to apparent epistatic effects (Wood et al., 2014). Hence, also main effects can be captured by epistatic effects. Furthermore, with increasing marker density, an advantageous predictive ability of a complex model including epistatic effects may become weaker if an interaction was not biologically sound (Schrauf et al., 2020). It remains challenging to identify epistatic effects that can explain functional mechanisms. This remains a controversial discussion. “The discussion [on epistasis] is of theoretical interest and will continue being so but it has been in stasis for decades” (D. Gianola, personal communication at ICQG6). It is not questioned if epistasis matters (clearly yes; Barton, 2020) but interaction effects are typically neglected in practical applications because parameters are imprecisely estimated and epistasis does not noticeably improve genetic value prediction (D. Gianola, personal communication at ICQG6). This attitude seems to drive opponents because case studies continue appearing, for instance, in laboratory mice. Recently, Miller et al. (2020) detected strong inter- and intrachromosomal interactions on blood-related traits (e.g. percentage of lymphocytes) detached from any main effects.

Papers 2.1 and 2.2 have been of theoretical nature in order to evaluate benefits from modelling non-additive genetic effects. In a genomic evaluation of milk metabolites (Paper 2.3), traditional milk production and novel milk-component traits (Paper 2.4) of  $n = 1295$  dairy cows, I concentrated on the practically most decisive genetic components: additive and dominance effects expressed in terms of breeding values and dominance deviations. Analyses were performed with GBLUP which became a standard approach in genomic evaluations (e.g. Koivula et al., 2012). Cumulative genetic effects exploit the similarity between animals by means of genomic relationship matrices that were based on Cockerham’s classical decomposition of genetic effects. Additionally, to consider the dependence

between markers due to linkage and LD, a kind of genetic background effect was taken into account. It was modelled as a polygenic effect with a pedigree-based covariance matrix. Presence of each genetic effect was tested by a likelihood ratio test of the corresponding variance component. In total, 55 out of 190 milk metabolites revealed significant additive genetic variation (broad-sense heritability ranged from 0.11 to 0.69) but dominance deviation and background genetics were not significantly present. Furthermore, predictive ability measured as correlation between estimated total genetic value and milk metabolite was fairly low (at most 0.28 for highly heritable metabolites). This is in line with the fact that milk metabolites are sensitive to the current nutritional and physiological state of cow (e.g. Billa et al., 2020). When milk production traits were studied, only protein and casein content showed significant contribution of dominance but none of the milk-component traits. In any case, presence of dominance needs to be studied and monitored because it is a fundamental driver of inbreeding depression in populations exposed to intense selection (Howard et al., 2017).

As a final option to be mentioned, Xiang et al. (2018) proposed a statistical model which allows for a non-zero correlation between additive and dominance effects. The model can be implemented in standard software used for GBLUP. A consistent coding of marker genotypes with regard to the major allele and the explicit estimation of the covariance component between additive and dominance effects yielded a slight improvement of predicted total genetic values of 1.5 % compared to a classical approach based on simulated data.

In summary and sharply speaking, groundbreaking advances since Cockerham's classical method of decomposition have not been achieved owing to the complexity of genome biology that obviously cannot be separated into a sum of genetic terms. Furthermore, as parameterisation methods for genotype codes rely on allele or genotype frequencies, attention has to be paid to estimating these quantities with respect to the underlying population. If any population structure exists, the application of population-genetics approaches can lead to biased estimates of frequencies and hence genetic parameters, especially with small sample sizes. Thus, to achieve progress in genomic evaluations in populations with family structure, I continued working on statistical models that explicitly considered the dependence among SNP genotypes and the underlying family design.

## 4.2 Covariance between genotype codes

The dependence between genotypes at a pair of SNPs was derived theoretically for populations characterised by family stratification. First investigations considered a single paternal half-sib family (Paper 3.1) and were 1:1 transferable to maternal half-sib families. These investigations required knowledge of LD and recombination rates. Later on the theory was extended to cover also multiple half-sib families (Paper 3.2) and full sib-families (Paper 3.3).

In general, the covariance  $K_{i,j} = \text{cov}(X_{t,i}, X_{t,j})$  between genotype codes at SNPs  $i$  and  $j$  can be separated into paternal ( $D_{i,j}^{\sigma}$ ) and maternal ( $D_{i,j}^{\phi}$ ) LD as  $K_{i,j} = D_{i,j}^{\sigma} + D_{i,j}^{\phi}$  because parental gametes are transmitted independently from parent to progeny. Starting from a single paternal half-sib family, LD of paternal haplotypes among progeny was derived conditionally on the sire haplotypes. Furthermore, maternal LD was derived from gamete frequencies corresponding to the dam population. Then a  $p \times p$  covariance matrix  $K$  can be set up that considers all SNP pairs on a chromosome or in a specific region.

Several fields of application of  $K$  exist. As a first option, it has been included in a statistical model for analysing a quantitative trait (Paper 3.1). For this, it was obvious to choose a Bayesian approach. The covariance matrix was directly used to specify different prior assumptions on the distribution of SNP effects. The assumptions differed in the choice of penalisation parameters leading to local or global shrinkage of genetic effects. Approaches were applied to a single half-sib family using simulated data in different scenarios (few versus many causal variants) and to semi-real data from dairy cattle (real genotypes but simulated phenotypes). In particular, it was evaluated how different prior specifications influenced the predictive ability and the power of identifying genome segments associated with performance traits. Compared with a prior specification that does not explicitly consider any relationship among predictor variables, the accuracy of genetic value prediction was improved by 10 to 22 % when the dependence between SNPs was included and when sample sizes was small (100 progeny). Although the accuracy of genetic value estimation was higher with genetically founded prior information, the individual SNP effects still were not estimated with more strength. Increasing family size and the number of families would be convenient. A second and more extensive simulation study was conducted to account for multiple families ( $N$ ), different marker densities and varying genetic parameters (Klosa and Wittenburg, 2017). The statistical model considered either family-specific ( $Np$  unknowns) or population-average ( $p$  unknowns) effects of each marker. Covariance matrices were set up separately for each half-sib family or as a kind of weighted average over all families; priors were specified as in Paper 3.1. However, modelling family-specific instead of average effects was in general not a competitive approach due to the  $N$ -fold increased number of parameters to be estimated. Moreover, due to the varying linkage phases of sire haplotypes, genetic-effect estimates cancelled out based on the average approach, making an identification of trait-associated genomic regions almost intractable. Thus, a revision of the prior specification remains necessary. Benefits of jointly modelling linkage and LD have been reported by X. Sun et al. (2016). Their methodology requires that genotypes of all individuals have been phased. Imputing haplotypes, however, introduces a severe bias in genetic value prediction that heavily levels off the profits of their novel method. The covariance between SNPs also requires haplotypes but only of the parent animals which can be imputed with high accuracy from progeny genotypes.

As a second option, the theoretical covariance matrix can be utilized for planning experimental designs. Particularly for fine-mapping studies given dense marker data, it is unavoidable to consider the strong dependence among markers. I proposed a ridge regression approach for which the empirical correlation matrix  $\frac{1}{n}X'X$  (each column of  $X$  was centered and scaled) is needed to calculate power and to estimate sample size to reach a certain power threshold. Because the empirical correlation would have been known only *after* an experiment had taken place, I employed the correlation matrix obtained from  $K$  as an approximation to it (Paper 3.2). Knowing the haplotypes of the putative common parents and LD from the population the individual parents come from, it is now possible to determine the required number of progeny in a future fine-mapping experiment. A simulation study was conducted and estimates of sample size based on the novel approach were compared with sample size estimates from a classical single-SNP model. Estimates of sample size were at least 17 % lower with the novel multi-SNP approach than with the single-SNP model; the lower heritability was or the less causal variants were assumed the higher the discrepancy between methods turned out. Negative correlations between SNPs, however, induced essentially inflated estimates of sample size

with the novel approach. It may be argued that planning experimental designs should be based on selection and shrinkage approaches which are more frequently used in genomic evaluations due to their superiority to ridge regression in many situations (e.g. Ogutu et al., 2012). For instance, the lasso enables selection of loci with strongest effects, while the effect estimates of other predictor variables remain zero. First and second moment of an estimator for effect size are needed to derive the test statistic and to approximate its distribution. As only approximated moments are available for the lasso estimator (van Wieringen, 2020), a straightforward derivation is not feasible.

Planning experimental designs requires assumptions on the breeding population: heritability, number of families as well as number of SNPs and causal variants. Moreover, in a population under selection, allele and haplotype frequencies of SNPs change over time. The covariance matrix of dependencies between SNP genotypes can theoretically be adapted to account for such dynamics of parameters. Because experiments are designed for one or two future generations and allele frequencies change moderately during that period, including dynamics is practically not relevant.

A pipeline was implemented as R package *hscovar* (Wittenburg et al., 2020); it includes the set up of covariance matrix, power and sample size calculation and, as a third option of using  $K$ , grouping of markers. The extent of dependence measured as correlation between SNP genotypes can be exploited for grouping (Paper 3.3). Then, for genomic evaluations, groups of highly associated markers can be utilised in grouped penalised regression approaches which are outlined in Section 4.4.

Though the covariance matrix was theoretically derived, few challenges remain for practical applications. For genetic-effect estimation as in Paper 3.1, the *inverse* covariance matrix is actually required but  $K$  is not necessarily positive definite due to the close proximity between markers. Hence, before inversion,  $K$  needs to be forced to be positive definite, either by bending or by replacing it by the nearest positive definite matrix (Higham, 2002). Furthermore, missing values may appear in  $K$  due to unknown (maternal) LD. For instance, in a study of five paternal half-sib families in Paper 3.4, on average 32 % missing values were observed in  $D^{\varnothing}$  over all chromosomes. Investigating 876 paternal half-sib families in Paper 3.5 led to 2 % missing estimates in  $D^{\varnothing}$ . (For SNP pairs, where no sire was double heterozygous, paternal recombination rate could not be assessed and estimation of maternal LD was skipped in these studies. The absence of double heterozygosity does not affect  $D^{\sigma}$  because the corresponding entries are zero.) Missing entries in  $D^{\varnothing}$  should be imputed to include as many markers as possible in the analysis; this is particularly important if sample size is small. Advanced tools are available that ensure the outcome to be a symmetric matrix (e.g. Ryan et al., 2010) which can then be used in the subsequent steps of genomic evaluation.

The covariance matrix between SNP genotypes requires a genetic map to determine the recombination rate between pairs of SNPs. For most species, genetic maps are available or can be approximated from physical maps. If these parameters are not available or are needed for a particular breed, they can be estimated from SNP genotypes as discussed below.

### 4.3 Inferences from recombination rates

Gomez-Raya et al. (2013) proposed a likelihood-based approach for estimating recombination rates in half-sib families. This method does not require the phase of progeny genotypes as it would be the case for a conventional linkage analysis. The likelihood function can be written in terms of the unknown

parameters  $D^{\sigma}$ ,  $D^{\varphi}$  and maternal allele frequencies and the observed progeny genotype frequencies. The likelihood may have two maxima – a point that has been ignored in the past. Furthermore, the magnitudes of the maxima are influenced by the maternal allele frequencies at the investigated marker pair. Because parameter estimates are obtained using an expectation-maximisation (EM) algorithm, the choice of start values has strong impact on which maximum the algorithm converges to. Knowing the covariance between genotype codes for additive effects  $\text{cov}(X_{t,i}, X_{t,j})$  and the covariance between genotype codes for dominance effects  $\text{cov}(D_{t,i}, D_{t,j})$ , corresponding to Bonk et al. (2016), was key for exploiting the relationship between  $D^{\sigma}$  and  $D^{\varphi}$  and for providing start values for the EM algorithm. The likelihood-based approach has been advanced to deal with flatness and bimodality of the log-likelihood surface in Paper 3.4. A stepwise procedure was presented in which the relationship between the two modes was exploited. The EM algorithm for parameter estimation was applied twice using reasonable start values from the relationship between LD and the most likely solution was selected. This approach was validated using simulated genotype data of half sibs. It was also applied to German Holstein data consisting of five half-sib families (with minimum family size of 30 progeny) and 39,780 marker genotypes on 29 chromosomes, leading to estimates for 12,759,713 intrachromosomal marker pairs.

Unlike previous studies in cattle (Druet and Georges, 2015; Ma et al., 2015; Sandor et al., 2012), recombination rates were estimated between all intra-chromosomal SNP pairs leading to an exciting by-product in Paper 3.4. Single SNPs and SNP clusters that were putatively misplaced in the underlying genome assembly Btau 4.2 were found by inspecting the mean of recombination rates in a local environment. A pattern search for unusually high estimates of recombination rate in the near neighbourhood of a SNP, or low recombination rates for distant SNPs, revealed such candidates for misplacement. In total, 40 misplaced SNPs were reported, most of them have been confirmed and assigned to different chromosomes or were located in different regions on the same chromosome regarding the revised genome assembly UMD 3.1.1. Similarly, Utsunomiya et al. (2016) also suggested a heuristic approach based on unexpected LD decay and filtered those SNPs with pairwise  $r^2 > 0.5$  but there was no overlap of candidates.

In Paper 3.5, the likelihood-based methodology was applied to large scale data in German Holstein cattle comprising more than 367 000 genotyped animals, thus supporting high certainty of parameter estimates. A quadratic optimisation approach was developed that incorporated all estimates of recombination rate between any SNP pairs which were less than 0.05 for the approximation of genetic distances between adjacent SNPs. Eventually, genetic-map positions were determined based on the ordering of markers according to the most recent genome assembly ARS-UCD1.2. The outcome was compared to a deterministic approach of Ferdosi et al. (2014) working with recombination rate between adjacent SNPs and a 1:1 transformation of recombination rate into genetic distances in Morgan units. Although there was a good coincidence between both approaches in general, estimates of genetic-map positions were 5 % larger with the likelihood-based approach than with the deterministic approach. Additionally, supplemented to Paper 3.5, a simulation study was performed showing that both methods underestimate the true genetic distances. Therefore, more research is still needed to reduce this bias. The methodology was implemented as an R package *hsrecombi* (Wittenburg, 2020). The pipeline is suited to half-sib families, thus it is applicable to genotypic data in dairy, beef, dual-purpose cattle and also in fish, honey bee and others.

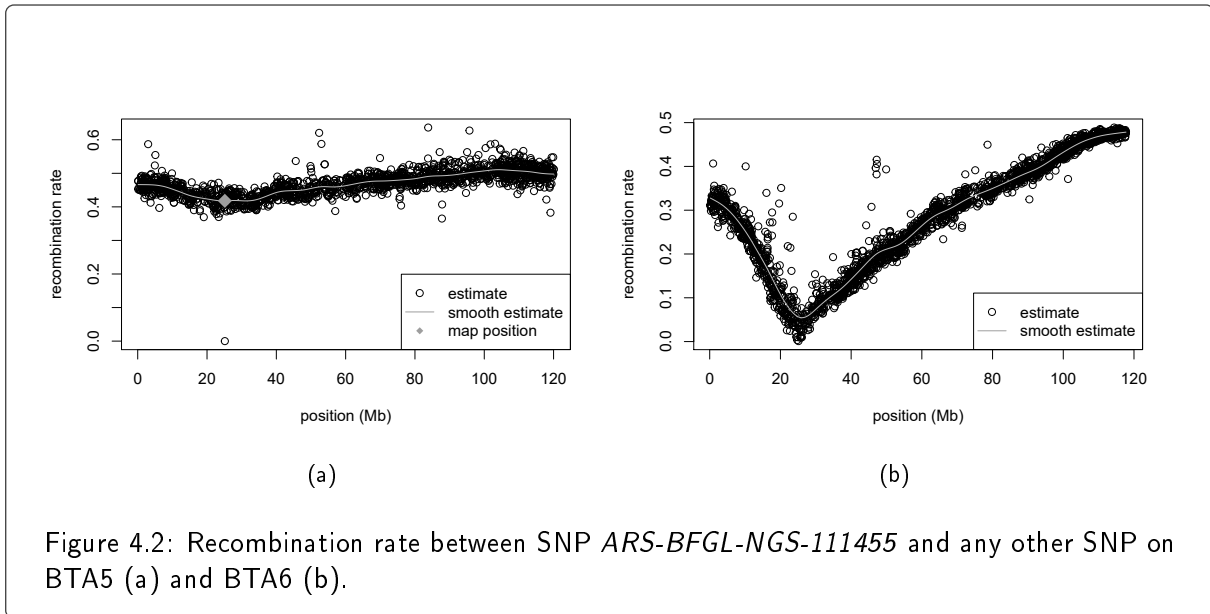


Figure 4.2: Recombination rate between SNP *ARS-BFGL-NGS-111455* and any other SNP on BTA5 (a) and BTA6 (b).

Despite improved technologies for assembling the bovine genome, 51 SNP markers with unusually large recombination rate to close SNPs turned out, indicating a wrong placement (supplemented to Paper 3.5). The majority of these SNPs (34 out of 51) are located in very small “problematic” regions of the genome and they represent either assembly errors or structural variants (R. Schnabel, personal communication). Another four single SNPs, which are inside problematic regions, did not show up in the analyses of Paper 3.5 because all 876 sires were homozygous and no inference was possible. These SNPs can likely be added to the list of misplaced candidates. Some problematic regions did not contain any candidate of misplacement. Another 6 candidates, which were not located in problematic regions, have a clear indication of being better placed somewhere else on the same chromosome or on another chromosome after inspecting the recombination rate between the candidate SNP and any other SNP on that chromosome. For instance, Figure 4.2 shows the pattern of recombination rate between SNP *ARS-BFGL-NGS-111455* and any other SNP on the initial chromosome BTA5 and on BTA6. The position of recombination rate close to zero on BTA6 suggests a more likely placement. Eventually, a solid list of 44 SNPs was assembled. As this study revealed only a handful SNPs based on the 50K-SNP panel and imputation accuracy of SNPs in problematic regions is known to be poor (C. Kühn, personal communication; Utsunomiya et al., 2016), the pragmatic solution is to exclude them from further evaluations, otherwise they might lead to biased inferences.

Due to the acrocentric nature of the bovine autosomes, recombination rate was generally low at the chromosome ends and elevated towards the middle of chromosome (supplemented to Paper 3.5). Moreover, studying patterns of recombination rate in more detail along the chromosomes pointed to various hotspots intervals. Hotspots are genomic regions of increased frequency of recombination events (e.g. Sandor et al., 2012); they can reveal gene regulations or DNA-protein interactions. For instance, the zinc-finger motif of the *PRDM9* protein is known to influence recombination activity (Paigen and Petkov, 2018). Shifting a recombination hotspot towards a causative variant – for instance, due to changes in the zinc-finger motifs of the *PRDM9* protein – causes new combinations of causative variants to occur. This can create new genetic variation that levels the intentional loss of genetic variation due to selection (Gonen et al., 2017). As another option, breeding for an increased

recombination rate is possible, reducing the loss of genetic variance (Battagin et al., 2016).

To develop breeding schemes for a proper management of genetic diversity and inbreeding depression, population-genetic parameters such as the rate of inbreeding and effective population size  $N_e$  have to be monitored (e.g. Howard et al., 2017).  $N_e$  is directly related to the rate of inbreeding (Falconer and Mackay, 1996) or it can be approximated from genetic distances between markers and corresponding  $r^2$  (Sved, 1971). Moreover, breeding strategies pay attention to the known non-favourable marker variants which may cause an individual disease, increased loss of progeny or may lead to other undesirable side effects. The association (i.e. the genetic distance) between favourable and non-favourable marker variants can be considered in mate allocation, leading to improved animal health and production in future generations.

## 4.4 Accounting for multicollinearity

Tens of thousands of molecular markers are often used as predictor variables in a statistical model for genomic evaluations leading to a high dimensional model. The high dimensionality not only increases the computational burden, it can also negatively influence the accuracy of effect estimates. Though the first issue can be solved by using specifically designed algorithms as in Paper 2.2, the latter remains challenging. Two fundamental options are available which (i) process or filter the predictor variables with no regard to the investigated trait or (ii) account for dependencies among predictors in a linear model approach. For aspect (ii), it is possible, among others, to exploit dependencies among predictors implicitly by employing latent variables as in partial least squares or explicitly by working with autocorrelated variables in a Bayesian context (Yang and Tempelman, 2012). A locally adaptive specification of an autocorrelated prior was proven to be advantageous in Paper 3.1. Referring to aspect (i), the genotype code matrix  $X$  can be preprocessed and input variables, such as principal components, can be derived from columns of  $X$ . The derived variables are then used as predictors in a linear model. Alternatively, filtering for SNPs bearing highest LD to SNPs in a defined neighbourhood (i.e. tagging SNPs; Carlson et al., 2004) is a suitable tool for reducing the model dimension and thus for weakening multicollinearity. The depth of reduction can be adapted by specifying an  $r^2$  threshold. Similarly, SNPs can be grouped according to LD among each other.

Grouping of predictors needs to rely on the study design and can account for different sources of data such as SNP genotypes, expression levels and other omics data (Boulesteix et al., 2017). In that sense, if it is relevant to evaluate what kind of effect is prevailing, different kinds of genetic effects may be treated as separate groups. Or, considering whether a SNP is influential or not, and no matter what kind of effect is ruling, additive and dominance effect build a group at each SNP, as suggested by Sabourin et al. (2015). In that study, grouping yielded up to 8% improvement for selecting the influential loci on the genome compared to an approach without grouping additive and dominance effects (measured as AUC of sensitivity vs. specificity). Noteworthy, that the inclusion of dominance resulted in an up to 2-fold larger AUC than a purely additive model if dominance effects have been simulated (Sabourin et al., 2015). The options for grouping are various and depend on the scope of study.

SNPs can be clustered according to the extent of dependence among them. For instance, Dehman et al. (2015) proposed grouping SNPs based on the population-LD measure  $r^2$ . The authors have



shown that using groups of markers in a penalised regression approach was superior to an ungrouped approach in detecting the important loci if the number of causal variants was at least three (out of 2048 loci) and thus LD between causal variants clearly mattered. In Paper 3.3, it has been suggested to bin SNPs upon the extent of correlation between them. Though the correlation matrix highlighted regions of high dependence between markers more precisely in populations with family structure than the population-LD matrix, there was no general trend in the number of groups obtained with either approach when different empirical data sets were analysed. A simulation study, however, could show that the groups obtained from the correlation matrix had better cluster quality by means of inter- and intra-cluster distance. If the aim was to reduce the model dimension and to use only a representative marker of each group in a ridge regression approach, then representative SNPs based on the correlation matrix had better sensitivity and specificity to reveal genome regions harbouring causal variants than the population-LD approach. Thus, using a measure of dependence suited to the population under investigation proved useful in this context.

Note that grouping based on population LD or correlation between SNPs actually does not rely on the physical order of SNPs but on the frequency of haplotypes. Based on a proper physical order, a grouping approach may be restricted to a certain physical proximity between SNPs which helps accelerate the approach.

A forthcoming statistical model can comprise additive and non-additive genetic effects while markers are grouped for each kind of genetic effect with respect to the correlation between the corresponding genotype codes. A grouped penalised regression approach can handle this adequately. The technical framework for incorporating groups of predictor variables in high dimensions was furnished in an efficient sparse-group lasso approach (Klosa et al., 2020; Simon et al., 2013) and other methods (e.g. IPF lasso; Boulesteix et al., 2017). I expect that such an approach with population-specific grouping will enable to distinguish genomic regions associated to a trait from the rest of the genome. That kind of evaluation will provide further insight into the genetic architecture in livestock animals.

## 4.5 Concluding remark

The first part of this thesis (Chapter 2) contributes to the broad field of research on non-additive genetic effects. A statistically proper treatment of different kinds of genetic effects requires an orthogonal decomposition of sources of genetic variation. But a truly orthogonal decomposition is actually not possible in the presence of LD between markers – this is not an exceptional but normal circumstance in genomic evaluations. This drawback gives a smooth transition to the second part of this habilitation thesis (Chapter 3) in which dependencies between loci due to linkage and LD have been studied, theoretically and empirically. During these studies, several points for further improvement became obvious which particularly have practical relevance and were discussed above (Chapter 4).

## 4.6 Literature cited

- Álvarez-Castro, J. M., & Carlborg, Ö. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*, *176*(2), 1151–1167. <https://doi.org/10.1534/genetics.106.067348>
- Barton, N. Does epistasis matter? In: *lcqg6 abstracts*. 2020, ID 291.
- Battagin, M., Gorjanc, G., Faux, A.-M., Johnston, S. E., & Hickey, J. M. (2016). Effect of manipulating recombination rates on response to selection in livestock breeding programs. *Genet. Sel. Evol.*, *48*(1), 44.
- Billa, P.-A., Faulconnier, Y., Larsen, T., Leroux, C., & Pires, J. (2020). Milk metabolites as noninvasive indicators of nutritional status of mid-lactation holstein and montbéliarde cows. *J. Dairy Sci.*, *103*(4), 3133–3146.
- Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., & Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. *Genet. Sel. Evol.*, *48*(1), 36. <https://doi.org/10.1186/s12711-016-0214-0>
- Boulesteix, A.-L., De Bin, R., Jiang, X., & Fuchs, M. (2017). IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Comput. Math. Methods Med.*, *2017*, 7691937. <https://doi.org/10.1155/2017/7691937>
- Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., & Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, *74*(1), 106–120. <https://doi.org/10.1086/381000>
- Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, *39*(6), 859–882.
- 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*(2), 327–345. <https://doi.org/10.1534/genetics.112.143313>
- de los Campos, G., Sorensen, D. A., & Toro, M. A. (2019). Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data). *G3 Genes Genom. Genet.*, *9*(5), 1429–1436.
- Dehman, A., Ambrose, C., & Neuvial, P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinf.*, *16*, 148. <https://doi.org/10.1186/s12859-015-0556-6>
- Druet, T., & Georges, M. (2015). Linkphase3: An improved pedigree-based phasing algorithm robust to genotyping and map errors. *Bioinformatics*, *31*(10), 1677–1679. <https://doi.org/10.1093/bioinformatics/btu859>
- Falconer, D. S., & Mackay, T. F. C. (1996). *Quantitative genetics*. Longman.
- Ferdosi, M. H., Kinghorn, B. P., van der Werf, J. H., & Gondro, C. (2014). Detection of recombination events, haplotype reconstruction and imputation of sires using half-sib snp genotypes. *Genet. Sel. Evol.*, *46*(1), 11.
- Gomez-Raya, L., Hulse, A. M., Thain, D., & Rauw, W. M. (2013). Haplotype phasing after joint estimation of recombination and linkage disequilibrium in breeding populations. *J. Anim. Sci. Biotechnol.*, *4*(1), 30. <https://doi.org/10.1186/2049-1891-4-30>

- Gonen, S., Battagin, M., Johnston, S. E., Gorjanc, G., & Hickey, J. M. (2017). The potential of shifting recombination hotspots to increase genetic gain in livestock breeding. *Genet. Sel. Evol.*, *49*(1), 55. <https://doi.org/10.1186/s12711-017-0330-5>
- Higham, N. J. (2002). Computing the nearest correlation matrix – a problem from finance. *IMA J. Numer. Anal.*, *22*(3), 329–343.
- Hill, W., & Mäki-Tanila, A. (2015). Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. *J. Anim. Breed. Genet.*, *132*(2), 176–186. <https://doi.org/10.1111/jbg.12140>
- Howard, J. T., Pryce, J. E., Baes, C., & Maltecca, C. (2017). Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. *J. Dairy Sci.*, *100*(8), 6009–6024. <https://doi.org/10.3168/jds.2017-12787>
- Huang, W., & Mackay, T. F. C. (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS Genet.*, *12*(11), e1006421. <https://doi.org/10.1371/journal.pgen.1006421>
- Klosa, J., Simon, N., Westermark, P. O., Liebscher, V., & Wittenburg, D. (2020). Seagull: Lasso, group lasso and sparse-group lasso regularisation for linear regression models via proximal gradient descent. *BMC Bioinf.*, *21*, 407. <https://doi.org/10.1186/s12859-020-03725-w>
- Klosa, J., & Wittenburg, D. Family-specific analysis of whole-genome regression models in half-sibs. In: 68th Annual Meeting of the EAAP in Tallinn, Estland, 2017, August.
- Koivula, M., Strandén, I., Su, G., & Mäntysaari, E. A. (2012). Different methods to calculate genomic predictions – comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *J. Dairy Sci.*, *95*(7), 4065–4073.
- Logsdon, B. A., Carty, C. L., Reiner, A. P., Dai, J. Y., & Kooperberg, C. (2012). A novel variational bayes multiple locus z-statistic for genome-wide association studies with bayesian model averaging. *Bioinformatics*, *28*(13), 1738–1744.
- Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., Bickhart, D. M., Cole, J. B., Null, D. J., Liu, G. E. et al. (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genet.*, *11*(11), e1005387.
- Martini, J. W., Rosales, F., Ha, N.-T., Heise, J., Wimmer, V., & Kneib, T. (2019). Lost in translation: On the problem of data coding in penalized whole genome regression with interactions. *G3 Genes Genom. Genet.*, *9*(4), 1117–1129.
- Meuwissen, T. H. E., Solberg, T. R., Shepherd, R., & Woolliams, J. A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.*, *41*, 2. <https://doi.org/10.1186/1297-9686-41-2>
- Miller, A. K., Chen, A., Bartlett, J., Wang, L., Williams, S. M., & Buchner, D. A. (2020). A novel mapping strategy utilizing mouse chromosome substitution strains identifies multiple epistatic interactions that regulate complex traits. *G3 Genes Genom. Genet.*, *10*(12), 4553–4563. <https://doi.org/10.1534/g3.120.401824>
- Ogut, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, *6*(S2), S10. <https://doi.org/10.1186/1753-6561-6-S2-S10>

- Paigen, K., & Petkov, P. M. (2018). Prdm9 and its role in genetic recombination. *Trends Genet.*, *34*(4), 291–300.
- Ryan, C., Greene, D., Cagney, G., & Cunningham, P. (2010). Missing value imputation for epistatic maps. *BMC Bioinf.*, *11*(1), 197.
- Sabourin, J., Nobel, A. B., & Valdar, W. (2015). Fine-mapping additive and dominant SNP effects using group-LASSO and fractional resample model averaging. *Genet. Epidemiol.*, *39*(2), 77–88.
- Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., & Georges, M. (2012). Genetic variants in rec8, rnf212, and prdm9 influence male recombination in cattle. *PLoS Genet.*, *8*(7), e1002854. <https://doi.org/10.1371/journal.pgen.1002854>
- Schrauf, M. F., Martini, J. W., Simianer, H., de Los Campos, G., Cantet, R., Freudenthal, J., Korte, A., & Munilla, S. (2020). Phantom epistasis in genomic selection: On the predictive ability of epistatic models. *G3 Genes Genom. Genet.*, *10*(9), 3137–3145.
- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *J. Comp. Graph. Stat.*, *22*(2), 231–245.
- Sun, X., Fernando, R., & Dekkers, J. (2016). Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. *Genet. Sel. Evol.*, *48*(1), 77. <https://doi.org/10.1186/s12711-016-0255-4>
- Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.*, *2*(2), 125–141.
- Utsunomiya, A. T. H., Santos, D. J. A., Boison, S. A., Utsunomiya, Y. T., Milanese, M., Bickhart, D. M., Ajmone-Marsan, P., Sölkner, J., Garcia, J. F., da Fonseca, R., & da Silva, M. V. G. B. (2016). Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. *BMC Genomics*, *17*(1), 705.
- van Wieringen, W. N. (2020). Lecture notes on ridge regression [Retrieved Jan 14, 2021]. <https://arxiv.org/pdf/1509.09169>
- Vitezica, Z. G., Legarra, A., Toro, M. A., & Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics*, *206*(3), 1297–1307. <https://doi.org/10.1534/genetics.116.199406>
- Vitezica, Z. G., Varona, L., & Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, *195*(4), 1223–1230. <https://doi.org/10.1534/genetics.113.155176>
- Wittenburg, D. (2020). Hsrecombi: Estimation of recombination rate and maternal ld in half-sibs. <https://cran.r-project.org/package=hsrecombi>
- Wittenburg, D., Doschoris, M., & Klosa, J. (2020). Hscovar: Calculation of covariance between markers for half-sib families. <https://cran.r-project.org/package=hscovar>
- Wood, A. R., Tuke, M. A., Nalls, M. A., Hernandez, D. G., Bandinelli, S., Singleton, A. B., Melzer, D., Ferrucci, L., Frayling, T. M., & Weedon, M. N. (2014). Another explanation for apparent epistasis. *Nature*, *514*(7520), E3–E5.
- Xiang, T., Christensen, O. F., Vitezica, Z. G., & Legarra, A. (2018). Genomic model with correlation between additive and dominance effects. *Genetics*, *209*(3), 711–723. <https://doi.org/10.1534/genetics.118.301015>

- Yang, W., & Tempelman, R. J. (2012). A Bayesian antedependence model for whole genome prediction. *Genetics*, *190*(4), 1491–1501. <https://doi.org/10.1534/genetics.111.131540>
- Zeng, Z.-B., Wang, T., & Zou, W. (2005). Modeling quantitative trait loci and interpretation of models. *Genetics*, *169*(3), 1711–1725. <https://doi.org/10.1534/genetics.104.035857>



# Chapter 5

## Summary

My thesis is dedicated to statistical methods which have proved useful for high dimensional data analysis in genomic evaluations. The intention of such evaluations is to explain the association between genetic and phenotypic variation or to explore associations among molecular markers.

Molecular markers are used as predictor variables in a statistical model to investigate the genetic impact on a quantitative trait. Genetic effects are assumed to be attributable to different kinds of gene action which can be additive or non-additive. For a proper statistical treatment, the genotype codes for different kinds of genetic effects need to be statistically parameterised in such a way that the different sources of genetic variation constitute an orthogonal decomposition of the total genetic variation. In my studies, the accuracy of breeding value estimation was hardly affected by inclusion or exclusion of non-additive genetic effects though the accuracy of total genetic values could improve with non-additive effects considered. The identification of causal variants or genome regions associated with trait expression was difficult when interaction effects were included. I developed a testing approach to identify at least those markers with strongest additive, dominant or any kind of pairwise interaction effect in a computationally rapid manner. Though there is no direct link between statistical effect and gene action, the resulting significant markers give rise to be further studied for their biological relevance using specific breeding designs.

The orthogonal decomposition of genetic variation is actually only possible in the absence of linkage disequilibrium between markers which is not a realistic assumption. I studied the dependence between markers in populations with family stratification, starting with half-sib families which is a typical family structure in livestock and generalising the methodology to full-sib families. The covariance between SNPs was analytically derived with respect to the genotype codes for the additive effect of SNP loci. The resulting covariance matrix deserved special attention due to its multi-functionality in several fields of genomic evaluations. First, it was used to incorporate prior knowledge about the distribution of marker effects in regression models for a genome-wide association analysis. Second, this matrix was key to design future experiments for fine-mapping loci that are associated with trait expression based on dense marker data. Knowing the genetic information of selected mates, the number of progeny could be determined to guarantee a certain power of association analysis later on. Third, markers were grouped depending on the extent of covariance/correlation. Grouping is a universal strategy to cope with multicollinearity in genomic evaluations due to the (partly) tight linkage of markers. Groups of highly associated markers may be employed in a penalised regression approach that allows to differentiate genomic regions with impact on trait expression from neutral



regions. Such an approach can be extended to also include non-additive genetic effects.

Additionally knowing the covariance between genotype codes for the dominance effect of SNP loci enabled the improvement of a likelihood-based approach for estimating paternal recombination rates and maternal linkage disequilibrium. This was a very thankful approach as it allowed the estimation of recombination rate between any SNP pair. Not only were recombination rates required to set up the covariance matrix as outlined above, they were also important for the derivation of a genetic map. The genetic distance between markers was estimated from recombination rates that were not restricted to neighbouring SNPs. Furthermore, SNPs that were putatively misplaced in the underlying genome assembly could be identified from unusually large recombination rates to other SNPs in close proximity. The likelihood-based approach was applied to large-scale cattle data and thus the outcome contributed to further improving the bovine genome assembly.

This thesis clearly showed that the dependence among SNPs can be controlled from a statistical point of view and exploited for beneficial use in genomic evaluations.

**Erklärung nach §3 (2) der Habilitationsordnung**

Hiermit versichere ich, dass ich diese Habilitationsschrift selbständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen als solche gekennzeichnet habe.

Ferner erkläre ich, dass diese Arbeit von mir weder an der Agrar- und Umweltwissenschaftlichen Fakultät der Universität Rostock noch einer anderen wissenschaftlichen Einrichtung zum Zwecke der Habilitation eingereicht wurde.