Archiv Tierzucht 56 (2013) 38, 380-398

Original study

Investigating a complex genotype-phenotype map for the development of methods to predict genetic values based on genome-wide marker data – a simulation study for the livestock perspective

Nina Melzer, Dörte Wittenburg and Dirk Repsilber

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

Abstract

Phenotypic variation can partly be explained by genetic variation, such as variation in single nucleotide polymorphism (SNP) genotypes. Genomic selection methods seek to predict genetic values (breeding values) based on SNP genotypes. To develop and to optimize these methods, simulated data are often used, which follow a rather simple genotype-phenotype map. Is the conventional approach for data simulation in this field an appropriate basis to optimize such methods in view of experimental data? Here, we present an alternative approach, striving to simulate more realistic data based on a genotype-phenotype map which includes a simulated metabolome level. This level was used to simulate genetic values, implicitly including additive and non-additive genetic effects, whereas in a conventional approach additive and dominance effects were explicitly simulated and assembled to genetic values. For both simulation approaches, different scenarios regarding numbers of quantitative trait loci (QTLs) and SNPs were analysed using fastBayesB as prediction method. We observed that our alternative map showed a smaller prediction precision (at least 3.75 %) compared to the conventional approach in all investigated scenarios. The observed degree of linearity is at least 94.12 % of the conventional approach or less. Additionally, we present results for different simulated data and experimental data to allow a comparison on a purely conceptual level. Concluding, simulating a more complex genotype-phenotype map including a molecular level, allows to study processing of variation from the genetic to the phenotype level in more detail and may prepare the ground for modern methods of genomic selection.

Archiv Tierzucht 56 (2013) 37, 380-398
doi: 10.7482/0003-9438-56-037

Received: 16 October 2012
Accepted: 20 December 2012
Online: 22 March 2013

Corresponding author:
Dirk Repsilber; email: dirk.repsilber@oru.se
Institute of Genetics and Biometry, Leibniz Institute for Farm Animal Biology, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany

Keywords: genotype-phenotype map, metabolome, genetic value prediction

Abbreviations: HWE: Hardy-Weinberg equilibrium, LD: linkage disequilibrium, MAF: minor allele frequency, ODE: ordinary differential equation, QTL: quantitative trait locus, SNP: single nucleotide polymorphism

Introduction

In the field of genomic selection, data are frequently simulated to compare different methods of genetic evaluation and to optimize methods. These studies have in common that the involved genotype-phenotype map based on a simple linear function. We term this simulation approach »conventional approach«. It is not known to which degree simulated data following the conventional approach are realistically mirroring the biology of real traits, sufficient for the development of methods for genetic value or phenotype prediction. In genomic selection, it is common to simulate SNP genotypes involving several hundreds or thousands of generations using a mutation-drift model, which leads to a more or less realistic linkage disequilibrium (LD) between the simulated marker loci. This is usually applied to equally sized chromosomes (e.g. Meuwissen *et al.* 2001, Calus & Veerkamp 2007, Habier *et al.* 2007). Calus *et al.* (2008) have shown that different spacing of markers has an influence on LD, which in turn has an impact on the precision of genetic value prediction. To obtain a more realistic dataset in simulations, we implement actual lengths of chromosomes and use SNP marker positions from an existing SNP annotation of the bovine genome. Genetic effects are randomly drawn at predefined marker loci, which simulate quantitative trait loci (QTLs). Various types of genetic effects are discussed to simulate a genetic value considering an additive and/or non-additive (dominance and/or epistasis) mode of gene action (Long *et al.* 2010, Ober *et al.* 2011). Note that epistasis is considered in a statistical sense in these contributions, based on the definition of Fisher (1918). The individual genetic value is calculated as sum of these effects depending on the realized QTL genotypes. Hill *et al.* (2008) reviewed that findings based on experimental data seem to point to prevailing importance of additive genetic variance, explaining more than 50 % and in most cases close to 100 % of the genetic variance. Molecular biology, however, proved that gene action is organized in interactive pathways, regulatory networks, which imply non-additive gene interactions and probably non-additive genotype-phenotype mapping (Moore 2005). Here, epistasis is considered in the biological sense (Cordell 2002). The importance of epistasis for mechanisms which underlie the genotype-phenotype map is not yet known (Moore 2005, Carlborg *et al.* 2006). It is suspected, however, that epistatic mechanisms may account for much of the causal genetic determination currently unexplained (e.g. Zuk *et al.* 2012).

We drafted an alternative simulation approach designed to be more realistic with respect to the complexity of the genotype-phenotype map: a simulated metabolome level is integrated on top of the conventional genotype-phenotype map to model biological epistasis. Towards this objective, we adopt an approach from the field of systems biology. Mendes *et al.* (2003) inspired us to model a metabolite level, determining enzyme parameters by marker status at specified marker positions. Mendes *et al.* simulated different gene-expression datasets based on artificial gene regulatory networks. These network models are composed of coupled ordinary differential equations (ODEs), where each equation describes

the production and degradation dynamics of a specified gene product. Biological variation is realized by adding random values to the kinetic parameters. Liu *et al.* (2008) adopted this approach and followed Mendes *et al.* (2003) by incorporating QTL variation to influence the kinetic parameters in their gene regulatory network. Based on these two approaches, we make use of a curated and ready parameterized SBML model (Systems Biology Markup Language; Hucka *et al.* 2003) of the central carbohydrate metabolism (Holzhütter 2004), which contains enzymes also found in cattle, to realize our simulated metabolome level (download from http://biomodels.org/, Le Novère *et al.* 2006), in the following termed »SBML approach«. The SMBL model was selected and considered to be adequate, because among the few existing curated metabolic network models it belongs to the few larger alternatives, especially as no curated metabolic SBML model for cattle exists. Our SBML approach allows investigating a more complex genotype-phenotype map, considering an additional level of gene expression in a broader sense. Additive and non-additive genetic effects are implicitly simulated. That means that varying one parameter of an enzyme has an effect on the interactions within the simulated system, affecting diverse metabolite concentrations, not only those catalysed by the respective enzyme. This offers the opportunity to investigate to which extent a change on the genotypic level leads to a different outcome of the metabolic level. We compare our SBML approach, which is also clearly artificial, with the conventional approach, where additive and dominance genetic effects are explicitly simulated.

Regarding the choice of a model for analysis, methods known from the field of genomic selection include genetic effects modelled with a purely additive model (e.g. Meuwissen *et al.* 2001, Daetwyler *et al.* 2010 and Zhang *et al.* 2010). However, Lee *et al.* (2008) as well as Toro & Varona (2010) have shown that the prediction precision of genetic values increased if an additive-dominance model is used compared to a purely additive model. It has become more and more common to extend existing genomic selection methods to include non-additive effects or to use non-parametric methods (Long *et al.* 2010, Ober *et al.* 2011). The fast Bayesian algorithm was proposed by Meuwissen *et al.* (2009) and extended to include non-additive effects by Wittenburg *et al.* (2011). In our study, we apply the extended fast Bayesian algorithm (fastBayesB), modelling additive and dominance effects.

Our focus of interest is to compare precision of prediction among the two simulation approaches and also to evaluate goodness of model fit. To illustrate the respective analysis results also for an example of experimental data, we investigated an experimental dataset for three different milk traits of 1 307 SNP-genotyped Holstein Friesian cows.

Material and methods

Genome: SNPs selection and positions

A bovine genome-wide SNP dataset was modelled in the style of Illumina Bovine SNP 50K SNP chip (Illumina Inc., San Diego, CA, USA). From this chip, we used all SNPs with annotated position according to Btau4.0 (The Bovine Genome Sequencing and Analysis Consortium *et al.* 2009) resulting in 52 276 SNPs. Chromosome lengths were retrieved from database Ensembl cattle (Ensembl 2008) to check the plausibility of SNP positions. Three SNPs were omitted because they were outside the corresponding chromosome. Single nucleotide polymorphism positions were linearly converted from the physical map (given in physical

unit base pair, bp) to the genetic map (genetic unit centiMorgan, cM) using a chromosome-wise scaling factor based on chromosome lengths in cM from the database »marc-USDA cattle« (United States Department of Agriculture 2008). On the basis of the genetic distance between two adjacent loci, the recombination rate between them was determined, using the Haldane mapping function (Haldane 1919).

Population: mutation-drift model

Four hundred generations of a mutation-drift model with a constant effective population size N_e=100 (50 sires, 50 dams) were simulated employing random mating. In the founder generation, all alleles were denoted as zero and from one generation to the next generation each locus had the chance to mutate once with a mutation rate of 2.5×10^{-3}. A mutated allele was labelled as one. The mutation rate and the number of simulated generations were determined in a preliminary study (Melzer *et al.* 2010b) to obtain an adequate simulated LD (Hill & Robertson 1968). Following the 400 initial generations, four additional generations were simulated without mutation and the population size was increased from 100 to 1 000 animals. Here, a 50 half-sib mating design was applied (1 sire mated with 20 dams). Generations 401 and 402 were used as training set (first offspring generation) and generations 403 and 404 as test set (second offspring generation).

Simulation and analyses set-up for simulation approaches

The following simulation steps were applied. The number of QTLs (n_{QTL}) was determined based on our metabolome-level model (Holzhütter 2004). It is an erythrocyte metabolism non-linear ODE system model for human which includes glycolysis (including the 2, 3-bisphosylglycerate shunt) and pentose phosphate pathway. For all involved enzymes, it was verified whether they occur in cattle, using the databases KEGG cattle (Kanehisa & Goto 2000) and Ensembl cattle. While all 38 enzymes of this metabolome-level were simulated numerically, 23 enzymes covering parts of glycolysis, gluthathione and pentose phosphate pathways in cattle were selected to be influenced by 23 QTLs. In addition, to work with larger numbers of QTLs, we used the 10-fold quantity of QTLs (n_{QTL}=230, see details below). The positions of QTLs were chosen randomly from all simulated SNPs with a minor allele frequency (MAF) of at least 0.02 in generation 400. Furthermore, a reduced SNP dataset was created from the complete SNP dataset (n_{SNP}=52 273), where every 10th SNP was taken, but QTL positions were retained (n_{SNP}=5 227). Combining the different numbers of SNPs and QTLs resulted in four simulation scenarios. Phenotypes were simulated based on different choices for broad-sense heritability $H^2 \in \{0.1, 0.3, 0.5\}$. Values of H^2 were similar to narrow-sense heritability (h^2) estimated for different milk traits (Gregory & Grandin 2007). For each scenario and heritability the set-up was replicated 100 times.

The prediction precision (ρ) is defined as correlation between simulated (test set) and predicted genetic values. We also investigated the impact of all 23 QTLs on each metabolic outcome via regression analysis. In addition, the goodness of model fit was evaluated for all training datasets for all scenarios and heritabilities, where the correlation between fitted values and residuals were determined using the function cor.test in R (R Development Core Team 2010).

Simulating genetic value and phenotype - conventional approach

Following the conventional approach, the phenotype for an animal was simulated as:

$$y_i = \sum_{j=1}^{n_{QTL}} (X_{ij} a_j + D_{ij} d_j) + e_i \quad (1)$$

where $i \in \{1, ..., n\}$ is the animal index. X_{ij} represents the design matrix for the additive effect a_j and D_{ij} is the design matrix for the dominance effect d_j. Entries in the design matrices depend on the marker genotypes: $X_{ij}=\pm 1$ and $D_{ij}=0$ for homozygous (+1 means homozygous for the mutated alleles) and $X_{ij}=0$ and $D_{ij}=1$ for heterozygous individuals at locus j. The simulated additive effect was drawn from a gamma distribution with shape parameter $\alpha=0.42$ and scale parameter $\beta=2.619$ in case of 23 QTLs and $\beta=8.282$ in case of 230 QTLs, following Meuwissen *et al.* (2001). The sign of the additive effect was randomly drawn with equal chance. The dominance effect was calculated as product of the additive effect and the degree of dominance, which was drawn from a normal distribution with mean m=0.193 and variance $\tau^2=0.097$ (Bennewitz & Meuwissen 2010). The genetic value for an animal was composed as sum of locus-specific genotypic effects as in equation (1). Furthermore, genetic values of the training set, also for the test set, were separately standardized to obtain a simulated genetic variance $\sigma_g^2=1$. The phenotype for an animal was obtained by adding an error e_i to the genetic value. The error was drawn from a normal distribution $N(0, \sigma_e^2)$, for which the variance was determined according to the chosen $H^2 \in \{0.1, 0.3, 0.5\}$.

Simulating genetic value and phenotype - SBML approach

For the SBML approach, we simulated a metabolome level between genotype and phenotype. The transition from the genotype to the metabolome level was realized as follows: a QTL influenced a specific kinetic parameter k_{ij}, in our case mostly the maximum reaction rate (V_{max}) of a specific enzyme. This means that the kinetic parameter changed depending on the genotype of the QTL coded in X_{ij}. In detail, $k_{ij} \in \{\psi - 50\,\%, \psi, \psi + 50\,\%\}$, following Holzhütter (2004), corresponded to $X_{ij} \in \{-1, 0, 1\}$, if the sign of the additive effect was positive in the conventional approach. In the other case, the order of the values of the kinetic parameter was reversed, $k_{ij} \in \{\psi + 50\,\%, \psi, \psi - 50\,\%\}$. Here, ψ was the default value of the kinetic parameter in the originally parameterized SBML model. The following enzyme kinetics were affected: vRibPepi_Vmaxv21, vPK_Vmaxv12, vHEX_Vmax1v1, vBPGM_kDPGMv8, vTPI_Vmaxv5, GAPDH_Vmaxv6, vPFK_Vmaxv3, vLDHNADH_Vmaxv13, vPPRPPS_Vmaxv25, vRibPiso_Vmaxv22, vAK_Vmaxv16, vENO_Vmaxv11, vGPI_Vmaxv2, vALD_Vmaxv4, vBPGP_Vmaxv9, vPGK_Vmaxv7, vATPase_kATPasev15, vPGM_Vmaxv10, vGSSGRD_Vmaxv19, vTrAld_Vmaxv24, vTrKet1_Vmaxv23 and vG6PDH_Vmaxv17. All other parameters remained unaffected. The SBML model was implemented as numeric simulation of the ODE system for the respective kinetic parameter settings using Matlab R2009b (MATLAB, Release 7.9.0.529 (R2009b), The MathWorks Inc., Natick, MA, USA) and the Matlab toolbox SimBiology R2009b. The SBML model was simulated until the metabolite concentrations reached the steady-state. After test runs, the maximum number of iterations (time) was set to 500. On the basis of the standardized equilibrium metabolite concentrations, we simulated the phenotype for an animal as:

$$y_i = \sum_{i=1}^{q} (P_{il}) + f_i \quad (2)$$

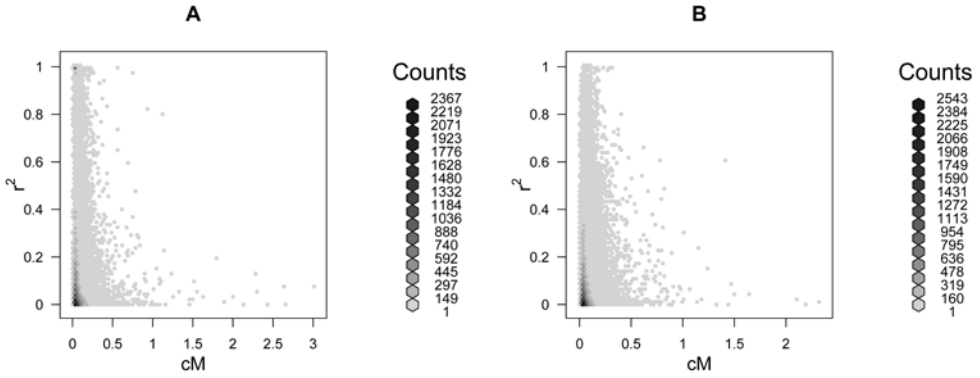
Here, P_{il} depicts the matrix of equilibrium metabolite concentrations, where $i \in \{1, ..., n\}$ is the animal index and $l \in \{1, ..., q\}$ denotes the index for equilibrium metabolite concentrations belonging to specific enzymes influenced by simulated QTLs. The equilibrium concentrations belonging to those metabolites catalysed by a specific enzyme were summed up, resulting in the specific metabolic outcome for simulated QTLs. Further, q represents the total number of metabolite concentrations; in some cases two concentrations were influenced by one enzyme. Also in some cases more than one column of *P* belongs to the same metabolite, if this metabolite is catalysed by more than one of the investigated enzymes. The sum over all equilibrium metabolite concentrations or the sum of all 23 metabolic outcomes results in the genetic value for an animal. Note that for this simulation approach of a genotype-phenotype map, the step from genotype to the metabolite concentrations is non-additive, whereas from the metabolite concentrations to the genetic value a purely additive step is implemented. Similar to the conventional approach, the genetic values for training set and test set were standardized separately. The phenotype was obtained by adding an error f_i, which was drawn from a normal distribution with mean zero and residual variance σ_f^2. The residual variance was again determined according to the chosen $H^2 \in \{0.1, 0.3, 0.5\}$.

Two sizes of SBML models were implemented, a 23-QTL model and a 230-QTL model. For the 230-QTL model, the original model was replicated 10 times, yielding 230 independent QTLs. For each replicate, the 23 enzymes available in cattle were simulated as QTLs as outlined above.

Experimental data and analyses set-up

Our experimental dataset compromises 1 307 Holstein Friesian cows from 18 agricultural holdings in Mecklenburg-Western Pomerania. From these cows, we obtained genome-wide SNP genotypes (Illumina Bovine SNP 50K). Milk traits were measured during the standard milk performance test at State Control Association for Quality Inspection (LKV, Güstrow, Germany). Milk samples were taken between the 20th and 120th day of the first lactation. More details can be found in Melzer *et al.* (2010a). Each cow had less than 10 % missing SNP genotypes. Single nucleotide polymorphism data pre-processing included several steps: First, SNPs with unknown positions were deleted. Thus, 52 255 SNPs were delivered to further quality checks. Second, SNPs were excluded if MAF<1 % and if Hardy-Weinberg equilibrium (HWE) was not fulfilled (P-value<10^{-4}, Samani *et al.* 2007) or if a SNP locus had more than 10 % missing values over all cows. In our experimental dataset, 263 SNPs were not in HWE, 7 233 SNPs did not fulfil the MAF criterion, 900 SNPs had more than 10 % missing values and 780 SNPs showed a combination of these properties, such that 43 079 SNPs were kept. The rarely missing SNP genotypes were imputed using Beagle v3.2 (Browning & Browning 2007). Phased SNP data were also obtained via Beagle and implied an average LD between adjacent markers of r^2=0.21. Comparative investigations of simulations regarding the LD in training sets, excluding SNPs with MAF less than 1 % (in average 5 688 SNPs), showed an average r^2=0.14. The average LD in test sets was r^2=0.15 after discarding SNPs with MAF less than

1 % (in average 5 826 SNPs). Figure 1 shows the LD for the experimental dataset and for an example of a simulated dataset after filtering the SNPs.



The graphic shows the distribution of LD as correlations between adjacent loci for (A) experimental dataset, (B) example simulated dataset.

Figure 1
Comparison of linkage disequilibria between experimental and simulated data

Three milk traits were chosen: fat (%), casein (%) and pH value. Milk traits were standardized and corrected for following systematic effects: agricultural holdings (ah), milk test day (stp) and linear and quadratic regression on lactation time-point (ltp). The following linear model was fitted:

$$y = ah \times stp + ltp + ltp^2 + \in \quad (3)$$

and the obtained residuals were used for further analyses. For comparative purposes, h^2 was estimated based on the sire-model using the R package nlme (R Development Core Team 2010, Pinheiro *et al.* 2009). The sire-model included the same effects as model (3) plus a random effect for sires to account for similarities among half-sibs. Estimation for fat (%) h^2=0.23, casein (%) h^2=0.24 and pH value h^2=0.38.

To allow a conceptual comparison between milk traits and simulated datasets, 1 307 of the 2 000 simulated animals (training set) were randomly selected for each simulation approach. The following criteria were chosen: H^2=0.3, n_{SNP}=52 273 and n_{QTL}={23, 230}. Investigations were limited to H^2=0.3, because chosen milk traits had similar values of h^2. For all datasets, the prediction precision was obtained by using a 10-fold cross-validation (Hastie *et al.* 2009), for which the dataset was split into 10 equally-sized training sets and corresponding test sets. This implementation was followed for the sake of comparability, because no separate experimental test set was available.

Here, the prediction precision (ρ) is defined as correlation between estimated genetic values and phenotypes. The goodness of model fit was evaluated visually for the whole dataset involving all investigated traits (simulated and experimental).

Predicting genetic values using fastBayesB

Prediction of genetic values was based on using the genotypes and phenotypes from the training set to estimate the genetic effect sizes. These genetic effect sizes were combined with the genotype from the test set to estimate the genetic value. As appropriate choice for our studies, we considered the fastBayesB method (Meuwissen *et al.* 2009), which is an iterative fast Bayesian approach to estimate additive effects. An extended version of this method, including non-additive effects, is described in Wittenburg *et al.* (2011). We implemented a fastBayesB analysis considering additive and dominance effects. The additive and dominance effects are re-parameterized to prevent the estimation of covariances between them. For this orthogonal decomposition of the genetic values, the method of Álvarez-Castro & Carlborg (2007) is applied as in Wittenburg *et al.* (2011). The fastBayesB algorithm involves prior assumptions for genetic effects. The prior distribution of a genetic effect is a mixture of the double exponential distribution and the point mass at zero. The probability of having a zero genetic effect at some locus $j \in \{1, ..., n\}$ is 1-γ. Hence, γ represents the proportion of QTLs to SNPs and the algorithm requires a specification of this parameter. As the true number of QTLs for a given trait is unknown in principle, the following set of plausible values for $\gamma \in \{0.1, 0.05, 0.025, 0.01, 0.005, 0.001, 10^{-4}, 10^{-5}\}$ was tested for each run of fastBayesB. The resulting variation of prediction precision in the sets was evaluated to mirror the sensitivity of the algorithm to different choices of γ. The optimal γ was determined over the corresponding replicates, resulting in largest mean prediction precisions. The genetic variance σ_g^2 was determined as additive variance σ_a^2 plus dominance variance σ_d^2.

The maximum number of fastBayesB iterations was set to 1 000. Also, SNP alleles with MAF<0.01 were excluded from the analysis, but SNP alleles which were not in HWE were kept.

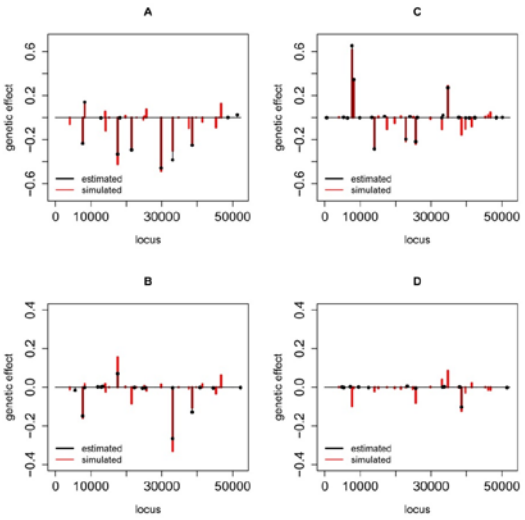
Results

Comparison of conventional approach versus SBML approach

To obtain an optimal choice of the parameter γ, which was required for the fastBayesB estimation algorithm, different γ-values were implemented to analyse the four simulation scenarios and for three values of H^2. In general, it was observed that not every γ-value is appropriate for each scenario and heritability in both conventional and SBML approach. For extreme choices of data or parameters, the fastBayesB algorithm aborts, e.g. for n_{SNP}=52 273 and n_{QTL}=23, each value of H^2 and γ=0.1, more than 74 % of the replicate runs aborted for both approaches.

In Table 1, we list prediction precisions, simulated and estimated variance components and corresponding standard deviations for all scenarios and heritabilities regarding the optimal γ-value for both approaches. In general, n_{SNP}=5 227 showed a larger prediction precision than n_{SNP}=52 273. In addition, n_{QTL}=23 showed a larger prediction precision than n_{QTL}=230. The quantity of QTLs had more influence on the prediction precision than the quantity of SNPs. In more detail, in all investigated scenarios it was observed that the mean prediction precision was at least 3.75 % lower for the SBML approach compared to the conventional approach. Estimated genetic variance components approached the true values for increasing values of simulated heritability. The estimated proportions of additive variance to total genetic variance

(σ_a^2/σ_g^2) were high compared to the proportion of dominance to total genetic variance. The estimated additive genetic variance σ_a^2 can be used to evaluate the degree of the linearity of both simulation approaches, which is at least 5.88 % lower for the SBML approach compared to the conventional approach for all investigated scenarios. Figures 2 A-B show the simulated and estimated additive and dominance effects for an example dataset in the conventional approach based on H^2=0.3, n_{QTL}=23 and n_{SNP}=52 273. It was observed that large simulated genetic effects were better detected than small genetic effects by the fastBayesB method. In comparison, Figures 2 C-D show the estimated additive and dominance effects for the comparable SBML approach. Here, sizes of the simulated genetic effects were unknown. Hence, in an additional analysis involving only the 23 simulated QTLs and genetic values of the simulated trait, we obtained estimates for the implicitly simulated genetic effects for the SBML approach.

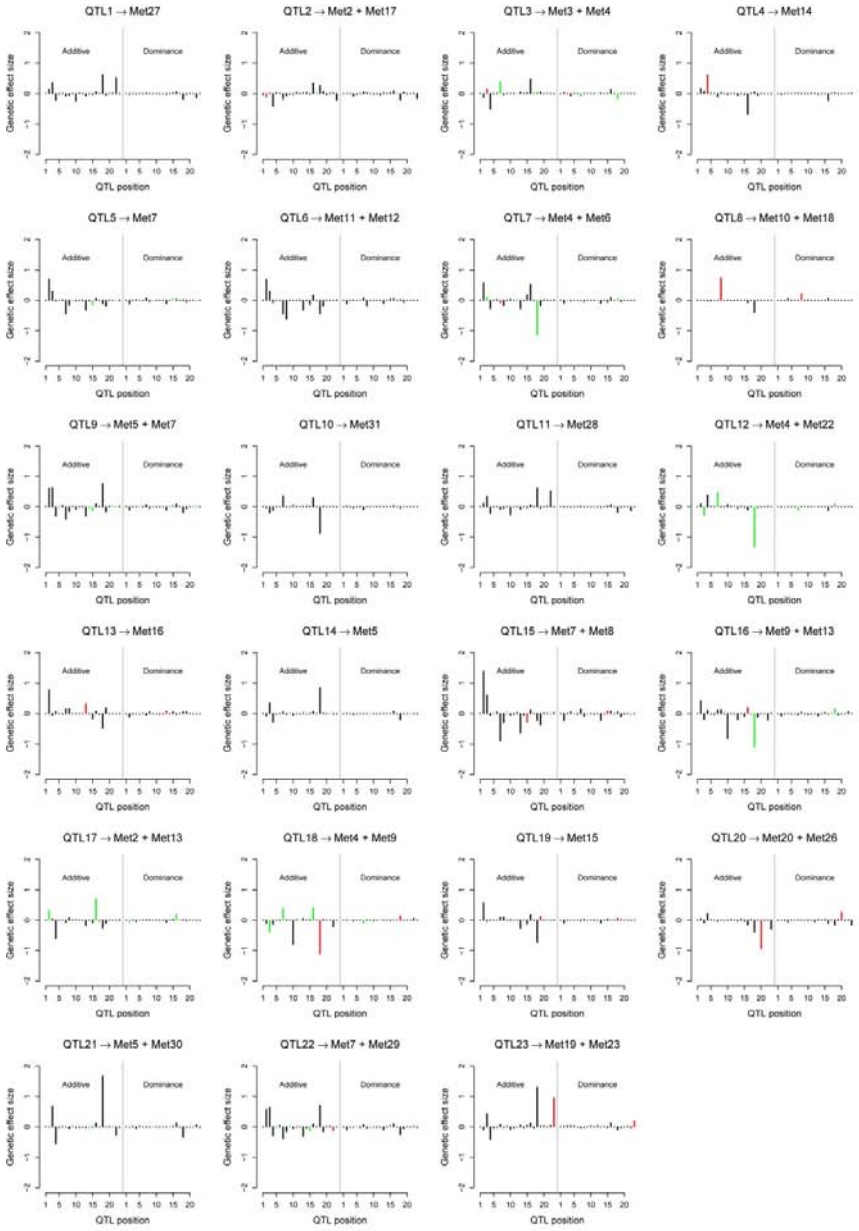


All figures are based on an example dataset with n_{SNP}=52 273, n_{QTL}=23, H^2=0.3 and the optimal γ-value. Estimated (A) additive and (B) dominance effects in the conventional approach. A filled circle was plotted for each genetic effect >10^{-4}. In this approach, sizes of the simulated genetic effects were known and additionally plotted in red. In comparison, estimated (C) additive and (D) dominance effects in the SBML approach. Here, the implicitly simulated main genetic effect sizes were estimated using the 23 QTLs to predict the corresponding genetic values. The observed estimated genetic effect sizes were plotted in red.

Figure 2
Estimated main genetic effects for conventional (left) and SBML (right) approach

To characterize possible deviations from the linearity in the SBML approach, we estimated the genetic effect sizes of all 23 simulated QTLs on the observed metabolic outcome of each single QTL-influenced enzymatic reaction. For an example dataset, which was also the basis for Figures 2 C-D, the impact of all QTLs on different metabolic outcomes is presented in Figure 3. Analysing all 100 datasets, two different kinds of genotype-phenotype mapping were observed on the level of metabolite concentrations: First, the QTL had no clear impact on its belonging metabolic outcome, whereas other QTL positions had. For example, QTL1 had no specific impact on the corresponding metabolic outcome, whereas, e.g. QTL18 and QTL22 clearly had an impact on the metabolic outcome belonging to the enzymatic reaction parameterized by QTL1. Second, the QTL had a clear impact on its belonging metabolic outcome as also on other QTL positions. This is the case for e.g. QTL18 and QTL23.

For all 100 training datasets, all scenarios, heritabilities and for the corresponding optimal γ-values, we investigated how good the linear model fitted the simulated data (results not shown). We observed, except of one case, that the linear model fitted all simulated datasets similarly and no significant difference for both simulation approaches was found.



Figures are based on the same example dataset as in Figure 2. All 23 QTL positions were used to estimate for each QTL the impact on each specific metabolic outcome. Each QTL is numbered and its specific metabolic outcome is presented. The Moreover, each specific metabolic outcome is split into the participating metabolites (M), where each metabolite is signed with the specific number as it is used in the SBML model (Holzhütter, 2004). The QTL positions which share the same metabolite for their belonging enzymes are marked in green and the corresponding QTL position is marked in red.

Figure 3
Estimated main genetic effect sizes for all QTLs for all metabolic outcomes

Table 1
For the optimal γ-value, the average estimated variance components, prediction precision and in brackets the corresponding standard deviations are given for 100 replicates and compared with the simulated variance components (italic) for the different scenarios

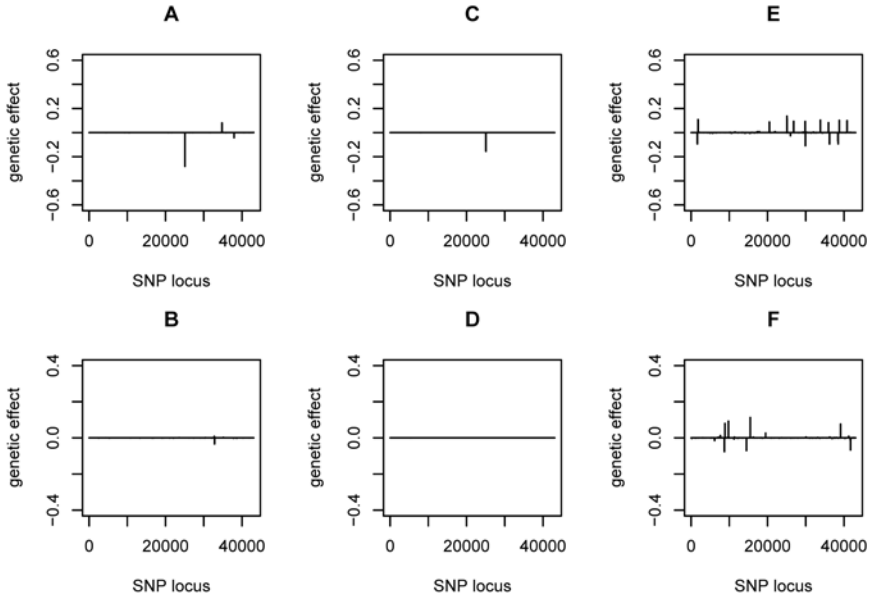
Approach	n_{QTL}	n_{SNP}	H^2	σ_g^2	σ_a^2	σ_d^2	σ_e^2	$\tilde{\sigma}_g^2$	$\tilde{\sigma}_a^2$	$\tilde{\sigma}_d^2$	$\tilde{\sigma}_e^2$	$\bar{H}^2$	ρ
conventional	23	5227	0.1	1	0.95	0.05	9.00	0.78 (0.17)	0.77 (0.17)	0.01 (0.03)	8.92 (0.28)	0.08	0.86 (0.06)
			0.3				2.33	0.93 (0.09)	0.89 (0.10)	0.03 (0.04)	2.33 (0.07)	0.29	0.96 (0.02)
			0.5				1.00	0.97 (0.07)	0.93 (0.08)	0.04 (0.04)	1.01 (0.03)	0.49	0.98 (0.01)
		52273	0.1				9.00	0.73 (0.18)	0.71 (0.18)	0.02 (0.03)	8.92 (0.29)	0.08	0.80 (0.10)
			0.3				2.33	0.90 (0.09)	0.87 (0.10)	0.03 (0.04)	2.34 (0.08)	0.28	0.94 (0.03)
			0.5				1.00	0.96 (0.07)	0.92 (0.08)	0.04 (0.04)	1.02 (0.04)	0.48	0.97 (0.01)
	230	5227	0.1	1	0.94	0.06	9.00	0.30 (0.11)	0.21 (0.09)	0.09 (0.03)	7.17 (0.29)	0.04	0.50 (0.10)
			0.3				2.33	0.71 (0.09)	0.68 (0.08)	0.03 (0.02)	2.32 (0.09)	0.23	0.75 (0.06)
			0.5				1.00	0.87 (0.08)	0.83 (0.08)	0.04 (0.03)	0.94 (0.04)	0.48	0.85 (0.03)
		52273	0.1				9.00	0.11 (0.08)	0.08 (0.07)	0.02 (0.02)	6.75 (0.30)	0.02	0.45 (0.11)
			0.3				2.33	0.59 (0.10)	0.58 (0.10)	0.01 (0.02)	2.59 (0.10)	0.19	0.66 (0.08)
			0.5				1.00	0.82 (0.08)	0.77 (0.08)	0.05 (0.03)	0.93 (0.04)	0.47	0.77 (0.05)
SBML	23	5227	0.1	1	-	-	9.00	0.73 (0.16)	0.70 (0.15)	0.03 (0.05)	9.04 (0.30)	0.07	0.82 (0.05)
			0.3				2.33	0.88 (0.08)	0.83 (0.08)	0.05 (0.03)	2.39 (0.07)	0.27	0.92 (0.02)
			0.5				1.00	0.93 (0.08)	0.86 (0.08)	0.06 (0.03)	1.00 (0.04)	0.48	0.94 (0.01)
		52273	0.1				9.00	0.68 (0.17)	0.66 (0.16)	0.02 (0.04)	9.05 (0.31)	0.07	0.77 (0.08)
			0.3				2.33	0.85 (0.09)	0.81 (0.09)	0.04 (0.03)	2.40 (0.08)	0.26	0.90 (0.03)
			0.5				1.00	0.90 (0.09)	0.85 (0.09)	0.05 (0.03)	1.07 (0.04)	0.46	0.93 (0.02)
	230	5227	0.1	1	-	-	9.00	0.22 (0.06)	0.13 (0.06)	0.08 (0.03)	7.30 (0.30)	0.03	0.37 (0.06)
			0.3				2.33	0.69 (0.09)	0.64 (0.09)	0.04 (0.02)	2.20 (0.07)	0.24	0.68 (0.04)
			0.5				1.00	0.76 (0.07)	0.73 (0.07)	0.02 (0.02)	1.10 (0.04)	0.41	0.78 (0.03)
		52273	0.1				9.00	0.04 (0.04)	0.02 (0.03)	0.02 (0.02)	6.87 (0.31)	0.01	0.32 (0.05)
			0.3				2.33	0.49 (0.09)	0.48 (0.09)	0.00 (0.01)	2.69 (0.10)	0.15	0.54 (0.07)
			0.5				1.00	0.73 (0.07)	0.69 (0.06)	0.04 (0.02)	1.00 (0.05)	0.42	0.69 (0.04)

σ_g^2 – simulated genetic variance, σ_a^2 – simulated additive variance, σ_d^2 – simulated dominance variance, σ_e^2 – simulated residual variance, H^2 – simulated broad-sense variance, $\tilde{\sigma}_g^2$ – estimated genetic variance, $\tilde{\sigma}_a^2$ – estimated additive variance, $\tilde{\sigma}_d^2$ – estimated dominance variance, $\tilde{\sigma}_e^2$ – estimated residual variance, $\bar{H}^2$ – estimated broad-sense heritability, ρ - prediction precision

Analysing experimental and simulated data

Two different approaches to simulate data were conceptually compared to experimental data by comparing the results of fastBayesB analyses. We observed that the optimal γ-value disagreed for the different investigated experimental datasets, e.g. pH value γ=0.001, fat (%) γ=10^{-4} and casein (%) γ=10^{-5}. The estimated variance components and prediction precisions for the optimal γ-value can be found in Table 2. In general, for experimental investigated milk traits, H^2 was underestimated by fastBayesB compared to estimated h^2 obtained with a sire-model for all investigated milk traits, e.g. fat (%) estimated h^2=0.23 with sire-model and H^2=0.105 with fastBayesB. Further, observed prediction precisions were scaled to 100 % to ease conceptual comparison between datasets. Therefore, the observed prediction precisions were divided by the square root of the estimated or simulated heritability (represents a possible upper bound for the prediction precision).

In Figure 4, the estimated genetic effects are presented for the different milk traits observed, using the whole dataset. In this figure, it is shown that casein content revealed only one intermediate additive effect. In comparison, beside one major additive effect, fat content showed further two intermediate additive and one dominance effects. Analyses for pH value revealed equally large genetic effects for additive and dominance effects. For simulated datasets the observed main genetic effect sizes were close to the observed, using the corresponding whole training sets (Figure 2).



Estimated (A) additive and (B) dominance genetic effects for fat content, (C) additive and (D) dominance for casein content, (E) additive and (F) dominance for pH value. The figures based on the whole dataset and analysed with fastBayesB for the optimal γ-value.

Figure 4
Estimated main genetic effects for different milk traits

Table 2
Estimated variance components and prediction precisions for simulated datasets and milk traits with fastBayesB.

Data	Approach	$\hat{\sigma}_g^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_d^2$	$\hat{\sigma}_e^2$	$\hat{H}^2$	ρ	*scaled* ρ
n_{QTL}=23	conventional	0.714 (0.07)	0.627 (0.06)	0.087 (0.02)	2.561 (0.03)	0.218	0.454 (0.07)	82.36 %
	SBML	0.756 (0.09)	0.756 (0.09)	0.001 (0.00)	2.481 (0.06)	0.233	0.475 (0.09)	86.36 %
n_{QTL}=230	conventional	0.734 (0.08)	0.640 (0.05)	0.094 (0.05)	1.839 (0.07)	0.285	0.380 (0.08)	69.09 %
	SBML	0.556 (0.09)	0.554 (0.09)	0.002 (0.00)	2.676 (0.05)	0.172	0.263 (0.08)	47.82 %
Experimental	Fat, %	0.085 (0.01)	0.081 (0.01)	0.004 (0.01)	0.722 (0.02)	0.105	0.292 (0.07)	60.83 %
	Casein, %	0.023 (0.00)	0.023 (0.00)	0.000 (0.00)	0.674 (0.01)	0.034	0.186 (0.07)	37.96 %
	pH value	0.191 (0.03)	0.143 (0.02)	0.048 (0.02)	0.035 (0.02)	0.356	0.255 (0.10)	41.12 %

$\hat{\sigma}_g^2$ – estimated genetic variance, $\hat{\sigma}_a^2$ – estimated additive variance, $\hat{\sigma}_d^2$ – estimated dominance variance, $\hat{\sigma}_e^2$ – estimated residual variance, $\hat{H}^2$ – estimated broad-sense heritability, ρ – prediction precision

For comparison, an example training dataset was selected including 1 307 animals (settings: n_{SNP}=52 273 and H^2=0.3) for each kind of simulated dataset. The experimental dataset included 1 307 animals and different milk traits were studied. Ten-fold cross-validation was applied to determine prediction precision. The average estimated variance components and prediction precisions (in brackets the corresponding standard deviations are given) are presented for the optimal γ-value.

Discussion

Methodological developments for algorithms in the field of genomic selection are typically based on simulated data. In our contribution, as an alternative to the state-of-art simplistic simulation approach, we investigated consequences of using a more complex, partly non-additive genotype-phenotype map compared to a conventional genotype-phenotype map. Our comparisons revealed that the SBML approach produced lower prediction precisions and clearly less linear additivity compared to the conventional approach.

Comparison of conventional versus SBML approach

In the conventional approach, the contributions of additive and dominance effects were explicitly modelled and thus known. For the SBML approach instead, the influences of additive and non-additive effects and their specific impacts on the total genetic variance were unknown and genetic effects were estimated based on the simulated genetic values.

Our comparison of fastBayesB results showed that conventional and SBML approaches were not similar regarding prediction precision and mostly show clear differences in estimated variance components (Table 1). In general, however, the choice of heritability and simulated quantity of QTLs and/or SNPs had a similar influence on the prediction precision for both simulation approaches. The prediction precision decreased with increasing quantity of SNPs, because the larger SNP set only included additional non-informative SNPs without impact on the phenotypic variation. The estimated genetic effect of all these additional SNPs should be zero. The fastBayesB method estimated an effect size for each locus (iteratively) under the assumption of linkage equilibrium. Additionally, the LD (mean value of r^2=0.15 for neighbouring SNPs (Figure 1)) is weak between our simulated SNPs, such that we do not expect linkage influences on estimated genetic effects. Hence, estimation errors accumulated with increasing number of SNPs. The quantity of simulated QTLs has a major influence on the prediction precision, which is in agreement with the observation of Daetwyler *et al.* (2010) and Zhang *et al.* (2010). As the same amount of genetic variation in the simulation is now spread over more loci, most QTLs had small effect sizes. Smaller effects were more difficult to detect by fastBayesB. The details of results depend on the value of H^2.

The SBML approach enabled further research opportunities regarding the inner structure of the (simulated) genotype-phenotype map compared to the conventional approach. It was found that some genetic effects were negligible if the sum was taken over all specific QTL outcomes (Figure 3). In our case, investigations of the specific QTL outcomes revealed two different mappings. The first type involved QTL variation showing no impact on the metabolic outcome belonging to the enzymatic reaction parameterized by this QTL, indicating that changes at the corresponding QTL position had no direct influence on this metabolic outcome. For example, QTL1 position appears to have a negligible effect on the specific investigated metabolic outcome of all 23 investigated enzymatic reactions (Figure 3) consistently over all datasets. Genetic variation at QTL1, however, is not without importance for the main trait: If mutations or diseases would affect either the metabolome network model or the weights for the summation of single metabolites to yield the phenotype, variation at QTL1 would likely become measurable. The second type of observed mapping is the corresponding QTL position affecting its specific metabolic outcome as well as other QTL

positions. In this case, some of the QTL positions interacted. Comparing estimated genetic effects for an example dataset for the genetic value prediction based on the 23 QTLs in the SBML approach (Figure 2 CD) with those for the single metabolites (Figure 3), which are summed up to build these genetic values, it can be concluded that some genetic effect sizes, which exist on the metabolome level, are negligible on the level of the genetic value.

For conventional and SBML approach, the goodness of model fit was evaluated and it was observed that the used linear model explained both simulated datasets similar in almost all cases. Hence, the observation that the simulated data of the SBML approach can be well analysed with a conventional linear model, including additive and dominance effects can be traced back to the arbitrary simple genotype-phenotype mapping from the metabolome level to genetic value in the SBML approach. We conclude that, for our chosen simulation approach, the SBML approach involves both a non-additive genotype-phenotype mapping as well as an additive part (metabolome to genetic value). In this context, we hypothesize that the genetic effects of the non-additive part lead to possible small deviations from a clear additive genotype-phenotype map for the phenotype. However, to decipher the details of these interwoven influences is certainly a rewarding field for a future study and beyond our current contribution.

Analysing experimental and simulated data

Our approach to compare milk traits and different simulated datasets is not meant as direct comparison, as we do not fit any kind of simulation model parameters using our experimental data. The comparison is rather on a conceptual level via comparing the structure of fastBayesB results (Table 2), because of the unknown underlying number of QTLs for the experimental traits. The comparison of the composition of the genetic effect sizes for simulated and experimental data offers another perspective to compare experimental data and different alternatives of simulation. Different estimated genetic effect compositions were observed for simulated datasets (Figure 2) and also for milk traits (Figure 4), where casein (%) and fat (%) mainly depended on one major additive effect (which is known from the literature as DGAT1 (Grisart *et al.* 2004)). Also, the used linear model seems to be sufficient for experimental data and simulated data (visual residual analysis).

In this context, the SBML approach offers further investigation opportunities, e.g. studying the genotype-metabolite map or metabolite-phenotype map. Compared to the conventional approach, these additional possibilities make simulation approaches alike the proposed SBML approach eligible for improving the genetic value prediction for experimental data also with respect to non-additive genetic effects which are sought to be exploited by modern methods in the field of genomic selection.

The more realistic simulation approach

Our set-up of the SNP datasets was based on annotated SNP positions and we used the actual lengths of the bovine chromosomes. This is different from most approaches recently chosen, where chromosomes have equal size and mostly 3 to 10 chromosomes were simulated (e.g. Meuwissen *et al.* 2001, Calus & Veerkamp 2007). Our set-up generated a distribution of LD values for adjacent SNPs similar to the experimental data (Figure 1).

To simulate more realistic genetic values, several further opportunities exist. We decided to integrate the level of the metabolome between genotype and phenotype and kept the construction of the genetic value as simple as possible: each QTL influenced only one kinetic parameter per enzyme. Further, taking the sum of equilibrium enzyme products is a simple strategy to simulate genetic values of a «complex trait». The following alternative example approaches might be conceived: (1) Genetic variation at a specific QTL may influence more than one enzyme parameter at a time. This would allow for concrete pleiotropy. Also explicit epistasis could be a possible extension, as proposed by Long *et al.* (2010) or Ober *et al.* (2011), but these authors employed statistical epistasis. (2) Genetic values could be directly constructed from multiple metabolite concentrations in various other ways. (3) The most advanced possibility of simulating phenotypes would certainly be to implement a systems biology model including cell, organ and physiology levels, which could lead to more realistic, implicit genotype-phenotype mappings (e.g. Nomura 2010).

Sensitivity of fastBayesB

The parameter γ of the fastBayesB method often has a significant influence on the results of the analyses, especially on the prediction precision. Therefore, different γ-values were tested to study the influence of the fastBayesB method on performance with respect to different simulation approaches as well as experimental data. The optimal γ-value was determined by using the γ-value with the largest prediction precision covering a certain set of γ-values. For conventional and SBML approach, it can be summarized that the optimal γ-value with respect to cross-validation prediction accuracy was mostly lower than the simulated proportion of QTL to SNPs. The range of γ-values, which was appropriate for n_{SNP}=5 227, was in the interval [10^{-4}; 0.05] and for n_{SNP}=52 273 the range was [10^{-5}; 0.001]. In other cases, the fastBayesB algorithm did not converge or it aborted. If the algorithm did not converge, the optimum was not reached within the 1 000 iteration steps. There are several possible reasons for abortion, which were discussed in Wittenburg *et al.* (2011).

In conclusion, the SBML approach was simulated using a more complex genotype-phenotype map than the conventional approach including the metabolome level. A deeper investigation of the simulated genetic effect sizes for the SBML approach revealed that some genetic effect sizes were negligible after the additive second step of the simulated composite mapping. Also, some simulated QTLs do not seem to be important for the phenotype, because their genetic impact on the investigated metabolic outcome is very low. We could show that our proposed SBML approach offers various further investigation opportunities compared to the conventional approach.

Our conceptual comparison revealed that similarities can be found between simulated and experimental datasets regarding genetic architecture, with trait-specific details. For further investigations, we propose to simulate a genotype-phenotype map including the molecular level to explore the importance of the genetic variation on this intermediate level and its transformation through molecular networks.

Acknowledgements

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

The State Control Association for Quality Inspection (LKV, Güstrow, Germany) provided us with information from standard milk performance tests. Information Technology-Solutions for Animal Production (VIT, Verden, Germany) kindly provided pedigree information. The working group of Prof. Dr. T. Meitinger and Dr. P. Lichtner (Helmholtz Zentrum München, Germany) measured the SNP genotypes. Special thank deserve colleagues at the Leibniz Institute for Farm Animal Biology (Dummerstorf, Germany): R. Grahl collected blood and milk samples, C. Reiko helped to prepare blood samples for DNA extraction, the working group of PD Dr. J. Vanselow provided technical facilities for DNA preparation and especially M. Nimz, M. Anders and M. Spitschak gave assistance in preparation and A. Rief built the SNP marker map.

References

Álvarez-Castro JM, Carlborg Ö (2007) A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. Genetics 176, 1151-1167

Bennewitz J, Meuwissen THE (2010) The distribution of QTL additive and dominance effects in porcine F2 crosses. J Anim Breed Genet 127, 171-179

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

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178, 553-561

Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Anim Breed Genet 124, 362-368

Carlborg Ö, Jacobsson L, Åhgren P, Siegel P, Andersson L (2006) Epistasis and the release of genetic variation during long-term selection. Nat Genet 38, 418-420

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11, 2463-2468

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

Ensembl (2008) http://www.ensembl.org/Bos_taurus [last accessed 09.11.2013]

Fisher RA (1919) XV. – The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans R Soc Edinb 52, 399-433

Gregory N, Grandin T (2007) Animal welfare and meat production. 2nd ed., CABI Pub, Wallingford, Oxfordshire, UK

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

Habier D, Fernando RL, Dekkers JCM (2007) The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177, 2389-2397

Haldane JBS (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8, 299-309

Hastie TJ, Tibshirani RJ, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer, New York, USA

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38, 226-331

Hill WG, Goddard ME, Visscher PM (2008) Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLoS Genet 4, e1000008

Holzhütter HG (2004) The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. Eur J Biochem 271, 2905-2922

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, and the rest of the SBML Forum: Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novère N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19, 524-531

Kanehisa M, Goto S (2000) Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27-30, http://www.genome.jp/kegg/pathway.html [last accesssed 09.11.2013]

Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res 34 (Suppl.), D689-D691

Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. PLoS Genet 4, e1000231

Liu B, de la Fuente A, Hoeschele I (2008) Gene Network Inference via Structural Equation Modeling in Genetical Genomics Experiments. Genetics 178, 1763-1776

Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, González-Recio O (2010) Radial basis function regression methods for predicting quantitative traits using snp markers. Genet Res Camb 92, 209-225

Melzer N, Jakubowski S, Hartwig S, Kesting U, Wolf S, Nürnberg G, Reinsch N, Repsilber D (2010a) Design, Infrastructure And Database Structure For A Study On Predicting Of Milk Phenotypes From Genome-Wide SNP Markers And Metabolite Profiles. In: Proc 9th World Congr Genet Appl Livest Prod, Leipzig, Germany, 427

Melzer N, Wittenburg D, Repsilber D (2010b) Simulating SNP data: influence of simulation design on the extent of inkage disequilibrium. In: Schriftenreihe Leibniz-Institut für Nutztierbiologie 16, Dummerstorf, Germany

Mendes P, Sha W, Ye K (2003) Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics 19 (Suppl.), ii122-ii129

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819-1829

Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA (2009) A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol 41, 2

Moore JH (2005) A global view of epistasis. Nat Genet 37, 13-14

Nomura T (2010) Toward Integration of Biological and Physiological Functions at Multiple Levels. Front Physiol 1,164

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

Pinheiro J, Bates D, DebRoy S, Sarkar D, the R Core team (2009) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-96. R Foundation for Statistical Computing, Vienna, Austria

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org [last accessed 09.11.2013]

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, König IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H,

Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H (2007) Genomewide Association Analysis of Coronary Artery Disease. N Engl J Med 357, 443-453

The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC (2009) The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324, 522-528

Toro MA, Varona L (2010) A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol 42, 33

United States Department of Agriculture (2008) http://www.marc.usda.gov/genome/cattle/cattle.html [last accessed 09.11.2013]

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

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

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109, 1193-1198