Novel use of derived genotype probabilities to discover significant dominance effects for milk production traits in dairy cattle

Teide-Jens Boysen*, Claas Heuer*, Jens Tetens*, Fritz Reinhardt§ and Georg Thaller*

*Institute of Animal Breeding and Husbandry, Christian-Albrechts-University Kiel
D-24098 Kiel, Germany

§Vereinigte Informationssysteme Tierhaltung w.V. (VIT), D-27283 Verden/Aller, Germany

These authors contributed equally to the work.
Running title: A novel approach to reveal dominance

To whom correspondence should be addressed:

Jens Tetens
Institute of Animal Breeding and Husbandry
Christian-Albrechts-University Kiel
Olshausenstraße 40
D-24098 Kiel
Germany
phone.: +49 431 880 5059
Fax: +49 431 880 2588
E-mail: jtetens@tierzucht.uni-kiel.de

Key words: dominance, genotype probabilities, genomic selection, dairy cattle
ABSTRACT

The estimation of dominance effects requires the availability of direct phenotypes, i.e. genotypes and phenotypes in the same individuals. In dairy cattle, classical QTL mapping approaches are, however, relying on genotyped sires and daughter based phenotypes like breeding values. Thus, dominance effects cannot be estimated. The number of dairy bulls genotyped for dense genomewide marker panels is steadily increasing in the context of Genomic Selection schemes. The availability of genotyped cows is, however, limited. Within the current study, the genotypes of male ancestors were applied to the calculation of genotype probabilities in cows. Together with the cows’ phenotypes, these probabilities were used to estimate dominance effects on a genomewide scale. The impact of sample size, the depth of pedigree used in deriving genotype probabilities, the linkage disequilibrium between QTL and marker, the fraction of variance explained by the QTL and the degree of dominance on the power to detect dominance was analyzed in simulation studies. The effect of relatedness among animals on the specificity of detection was addressed. Furthermore, the approach was applied to a real data set comprising 470,000 Holstein cows. In order to account for relatedness between animals a mixed model two-step approach was used to adjust phenotypes based on an additive genetic relationship matrix. Thereby, considerable dominance effects were identified for important milk production traits. The approach might serve as a powerful tool to dissect the genetic architecture of performance and functional traits in dairy cattle.
INTRODUCTION

In the context of genomic selection in dairy cattle, an abundance of bulls has been genotyped applying genome-wide dense marker panels. In 2010, the European reference population comprised more than 17,000 bulls representing more than 20 million daughters (LIU et al. 2011; LUND et al. 2010). In addition to their utilization in genomic prediction, these data are extensively used in genome-wide association studies to unravel the genetic factors affecting performance and functional traits. The expression of these traits is naturally limited to female individuals and thus, the phenotypes used in association studies are usually breeding values of sires based on performance data of many daughters. Such a structure of data only allows the estimation of allele substitution effects. There is no direct possibility to distinguish between additive and dominance effects. For the detection of these allelic interactions, genotypes and phenotypes had to be known in the same individuals. As compared to the bulls, the availability of genotype data for cows is limited. With the increasing number of genotyped bulls, genotypes of male ancestors become available for many cows enabling the derivation of genotype probabilities. Within the current study, these probabilities were converted to additive and dominance coefficients suitable for regression analysis analogous to the procedures commonly applied to QTL mapping in resource populations (HALEY and KNOTT 1992). The derivation of coefficients is not applied to inter-marker intervals based on recombination fractions, but to unknown genotypes at given SNP marker positions. At a specific marker locus with two alleles A and B, the probabilities of the possible genotypes AA, AB and BB can be deduced in cows based on the male ancestor’s genotypes and the allele frequencies in the population. The approach implies a loss of statistical power as compared to the utilization of real genotype data. A large
number of genotyped bulls and an extensive number of daughters per sire, however, should compensate these limitations.

Alternative methods to deduce genotype probabilities include “peeling” algorithms based on the ideas of Elston and Stewart (1971), Monte Carlo methods (e.g. Guo and Thompson 1992 or Henshall and Tier 2003) or Bayesian approaches. These methods are computationally infeasible for the very large data sets as used within the current study. Approaches to impute genotypes based on phase information are inapplicable because this would require at least partly genotyped females. We conducted a series of simulation studies to discover both the capabilities and limitations of the method and to evaluate the importance of factors like allele frequencies, sample size and size of daughter groups. Subsequently, the approach was validated in a large German Holstein data set. The method presented herein enhances the utilization of existing large data sets to dissect the genetic architecture of important performance and functional traits in dairy cattle.

**Material and Methods**

**Derivation of genotype probabilities**

To calculate the three genotype probabilities \( P(AA), P(AB) \) and \( P(BB) \) for a biallelic SNP marker with two alleles A and B and the respective allele frequencies \( p \) and \( q \) in a cow, we considered pedigrees including at least a genotyped sire and maternal grandsire (Figure 1). For any genotyped bull, the probability to transmit allele A is \( P(A) = 1 \) in homozygous \( (AA) \), \( P(A) = 0.5 \) for the heterozygous \( (AB) \) and \( P(A) = 0 \) for the alternative homozygous state \( (BB) \). The probability to transmit allele B is \( P(B) = 1 - P(A) \). This information is unavailable for the female individuals.
Therefore, the respective population frequency of the allele was used as an approximation for the transmission probability from the maternal granddam to the dam of the cow under consideration. The genotype probabilities for the dam \( P_d(AA, AB, BB) \) are the conditional probabilities emerging from this approximation and the known genotypes of the maternal grandsires as summarized in Table 1. The allele frequencies were estimated from the bull data set assuming that the allele frequencies in the cows are not considerably different from those observed in the bulls. For a given cow, there exist nine possible scenarios conditional on the genotypes of the sire and the maternal grandsire (Table 2). The respective genotype probabilities can be calculated once and assigned to the cows by looking up the precalculated values. However, based on the genotype probabilities presented in Table 1, the transmission probability for allele A of the dam conditional on the genotype of the maternal grandsire \( P(A_d | GT_{GS}) \) can be calculated as the sum of the probability of being homozygous AA and half the probability of being heterozygous:

\[
P(A_d | AA_{GS}) = p + \frac{q}{2} = \frac{2p}{2} + \frac{1-p}{2} = \frac{p + 1}{2},
\]

\[
P(A_d | AB_{GS}) = \frac{p}{2} + \frac{1}{4} = \frac{2p}{4} + \frac{1}{4} = \frac{1 + 2p}{4} = \frac{p + 0.5}{2},
\]

\[
P(A_d | BB_{GS}) = \frac{p}{2} = \frac{p + 0}{2}.
\]

The second summand within the numerator in this case is equivalent to the transmission probability of the maternal grandsire \( P(A_{GS}) \). Thus, the three equations can be jointly expressed as
\[ P(A_D) = \frac{1}{2}p + \frac{1}{2}P(A_{GS}). \]

Equivalently, the transmission probability for allele B \( P(B_D) \) can be calculated as

\[ P(B_D) = \frac{1}{2}q + \frac{1}{2}P(B_{GS}). \]

The probabilities for the homozygous genotypes in the cow are finally calculated from the sire’s and dam’s transmission probabilities as follows:

\[ P(AA_{Cow}) = P(A_s) \cdot P(A_d) = P(A_s) \cdot \left( \frac{1}{2}p + \frac{1}{2}P(A_{GS}) \right), \]

\[ P(BB_{Cow}) = P(B_s) \cdot P(B_d) = P(B_s) \cdot \left( \frac{1}{2}q + \frac{1}{2}P(B_{GS}) \right). \]

The probability of being heterozygous is simply calculated as \( P(AB_{Cow}) = 1 - P(AA_{Cow}) - P(BB_{Cow}) \). It is also possible to directly calculate the cow’s probability of being heterozygous as:

\[ P(AB_{Cow}) = P(A_s) \cdot P(B_d) + P(B_s) \cdot P(A_d). \]
Referring to Table 1 and considering the example of a sire with genotype AA and a maternal grandsire with genotype BB, the probability can be calculated as:

\[
P(AB_{\text{cow}}|GT_s = AA, GT_{GS} = BB)
\]

\[
= P(A_s|GT_s = AA) \cdot P(B_d|GT_{GS} = BB) + P(B_s|GT_s = AA) \cdot P(A_d|GT_{GS} = BB)
\]

\[
= 1 \cdot \left( q + \frac{1}{2}p \right) + 0 \cdot \frac{1}{2}p = q + \frac{1}{2}p = 1 - p + \frac{1}{2}p = 1 - \frac{1}{2}p \text{ (see Table 2).}
\]

This calculation would allow distinguishing between different parental origins of the alleles resulting in the two probabilities \( P(A_{\text{pat}}B_{\text{mat}}) \) and \( P(B_{\text{pat}}A_{\text{mat}}) \), which could be used to analyze imprinting effects. This was, however, out of the scope of the current study.

For practical reasons, all calculations were based only on the paternal transmission probability \( P(A_s) \) and the frequency of the rare allele \( f(A) = p \) representing the minor allele frequency (MAF). \( P(B_s) \) and \( f(B) = q \) were thus replaced by \( 1 - P(A_s) \) and \( 1 - p \), respectively. The approach can readily be adapted for deeper pedigrees. With the inclusion of a genotyped maternal greatgrandsire (GGS), i.e. three ancestral generations, the probabilities for the homozygous genotypes of the cow under consideration can be calculated similarly to the procedure described above. The transmission probabilities in the granddam are \( P(A_{GD}) = \frac{1}{2}p + \frac{1}{2}P(A_{GGS}) \) and \( P(B_{GD}) = \frac{1}{2}q + \frac{1}{2}P(B_{GGS}) \). Multiplication by the transmission probabilities of the grandsire gives the genotype probabilities of the dam, which can then be used to calculate the transmission probabilities. Analogous to the initial approach with two ancestral generations, the probabilities are:
Thus, the probabilities for the homozygous genotypes in the cow under consideration can be calculated as:

\[
P(A_d) = \frac{1}{2} p + \frac{1}{2} P(A_{GGS}) + P(A_{GS}) = \frac{1}{4} p + \frac{1}{4} p(A_{GGS}) + \frac{1}{2} p(A_{GS}),
\]

\[
P(B_d) = \frac{1}{2} q + \frac{1}{2} P(B_{GGS}) + P(B_{GS}) = \frac{1}{4} q + \frac{1}{4} p(B_{GGS}) + \frac{1}{2} p(B_{GS}).
\]

The inclusion of a further generation (great-greatgrandsire = GGGS) consequently leads to the following transmission probabilities in the dam:

\[
P(A_d) = \frac{1}{8} p + \frac{1}{8} p(A_{GGGS}) + \frac{1}{4} p(A_{GGS}) + \frac{1}{2} p(A_{GS}),
\]

\[
P(B_d) = \frac{1}{8} q + \frac{1}{8} p(B_{GGGS}) + \frac{1}{4} p(B_{GGS}) + \frac{1}{2} p(B_{GS}).
\]
To generalize the approach for any number $N$ of ancestral generations, the generations were numbered such that the cow under consideration represents generation ‘0’ (Figure 1). The parents of the cow are generation 1 ($SIRE_1 =$ Sire and $DAM_1 =$ Dam), the maternal grandparents are generation 2 ($SIRE_2 =$Grandsire, GS and $DAM_2$), the parents of $DAM_2$ are generation 3. Defining $N$ as the number of ancestral generations, the genotype probabilities can be calculated as follows:

\[
P(\text{AA}_{COW}) = P(A_{SIRE_1}) \left( \frac{1}{2^{N-1}} p + \sum_{i=2}^{N} \frac{1}{2^{N-i+1}} P(A_{SIRE_{N-i+2}}) \right),
\]

\[
P(\text{BB}_{COW}) = P(B_{SIRE_1}) \left( \frac{1}{2^{N-1}} q + \sum_{i=2}^{N} \frac{1}{2^{N-i+1}} P(B_{SIRE_{N-i+2}}) \right).
\]

The probabilities were transformed to regressors for additive and dominance effects with $add = P(\text{AA}) - P(\text{BB})$ and $dom = P(\text{AB})$, which can be used in a simple linear regression model of the form

\[
y_i = \mu + \beta_1 \cdot \text{add}_i + \beta_2 \cdot \text{dom}_i + e_i,
\]

where $\mu$ is the mean, $y_i$ represents the phenotype of the $i^{th}$ animal, $\text{add}_i$ and $\text{dom}_i$ are the additive and dominance coefficients calculated from the genotype probabilities, $\beta_1$ and $\beta_2$ are the respective regression coefficients and $e_i$ is a random residual term with $e \sim N(0, \sigma^2_e)$.
Simulation studies

To verify the functionality of our approach and to determine the necessary sample size to sufficiently detect dominance effects, populations of completely unrelated cows were simulated. This was achieved by assigning an individual sire and grandsire per cow exclusively. A single biallelic marker in complete LD with a nearby QTL was simulated. The sire genotypes for that marker were sampled from a population in Hardy-Weinberg equilibrium given a minor allele frequency (MAF) of 0.3. By utilizing the presented method for deriving genotype probabilities, the two coefficients $add$ and $dom$ were calculated for each cow based on the genotypes of sire, damsire and the allele frequency. Furthermore, we assigned a true genotype for every cow by sampling genotypes given the probabilities presented in Table 2. This ‘true’ genotype was used as a reference for the assignment of genetic effects to the simulated QTL. Phenotypes were calculated by assuming a defined phenotypic variance multiplied by a given heritability of $h^2 = 0.25$ leading to the additive genetic variance. The sizes of the simulated QTL effects were determined as a multiple of the additive genetic standard deviations ($\sigma_A$), with the additive effect being a half of the genetic standard deviation and the dominance effect making up half the size of the additive effect. Finally, a random residual effect was assigned to every phenotype.

Subsequently, the influence of various properties of the created population and the features of the simulated QTL on the power of detection was tested by separately varying the sample size, the number of considered generations, the minor allele frequency, the LD between QTL and marker, the fraction of variance explained by the QTL, and the degree of dominance. The effect of sample size was analyzed by a stepwise increase of up to $10^6$ cows. In order to achieve a better resolution across
the entire range of N, a reduced heritability of 0.2 and a smaller QTL effect of 0.2 additive genetic standard deviations were applied. The latter scenario was carried out with ‘true’ genotypes as well as with genotype probabilities. The different scenarios are described in Table 3. For each scenario, 1,000 iterations were performed and the power was defined as the rate of positives in the total number of iterations. The significance criterion was fixed to 5% probability of error after a Bonferroni correction to account for multiple testing. For this correction, 40,000 informative markers were assumed, which is a realistic scenario for the application of the Illumina 50k SNP Chip in Holstein cattle (see e.g. HABIER et al. 2010).

A second set of simulations was conducted to determine the influence of relatedness among animals and resulting stratification, which might lead to a high rate of false-positive signals. To this end, we repeated the simulations in a multi-marker scenario including related ancestors of the cows. A sample of 100,000 cows was created descending from initially 100 sires and 100 maternal grandsires. The phenotypes of the cows were assumed to consist of the combination of marker-effects and additional polygenic effects received from the ancestors. A trait of midrange heritability ($h^2 = 0.3$) was simulated with the genetic variance in the cows equally explained by the bull’s breeding values and the marker effects. The sires contributed half and the maternal grandsires a quarter of their respective breeding values. A total of 150 markers with minor allele frequencies ranging from 0.05 to 0.5 were defined. A tenth of the variance explained by markers was equally assigned to ten of the markers, therefore each simulating a strong QTL. The dominance ratio was randomly assigned within these ten markers. The remainder of markers was left without any effect. Subsequently, the effect of a varying family size was tested. For this purpose, the number of descendants per sire was varied from 50 to 5,000 daughters, leading
to different proportions of cows with identical sires and maternal grandsires, thus representing groups of animals with uniform genotype probabilities. For each family size, the specificity was calculated as the number of true negatives divided by the sum of true negatives and false positives.

To correct for relationship among animals as a source of stratification, especially due to sires with a very large number of daughters, a single marker linear mixed model regression using Residual Maximum Likelihood (REML) was applied. Therefore, a relationship matrix based on dam, sire and damsire was used with the following model.

\[
y = Xb + Za + e,
\]

where \( y \) is a vector of phenotypes, \( b \) is a vector of fixed effects consisting of the intercept and the coefficients \( add \) and \( dom \) and \( X \) is the incidence matrix for the fixed effects. The vector \( a \) represents the random effects with \( a \sim N(0, \sigma_a^2) \), where \( A \) is the additive genetic relationship matrix. \( Z \) is the incidence matrix for the random effects and \( e \) is a vector of residuals with \( e \sim N(0, \sigma_e^2) \), \( I \) representing an identity matrix. To test for significance, a Wald t-test was applied, where the denominator degrees of freedom were numerically estimated (KENWARD and ROGER 1997).

**Application to a real data set**

In order to validate the procedures with real data, phenotypic and pedigree data of 847,000 German Holstein cows were obtained from the vit - Vereinigte
Informationssysteme Tierhaltung w.V. (Verden, Germany). The phenotypic data consisted of yield deviations (YDs) for the traits milk, protein and fat yield as well as Somatic Cell Score (SCS) in the first lactation representing the deviations of the cow’s yields from the population mean adjusted for non-genetic fixed and random effects. The phenotypic data is summarized in Table 4. Genotype information was available for 3,200 Holstein bulls, genotyped with the Illumina BovineSNP50® BeadChip (Illumina, Inc. San Diego, California, USA) featuring a total of 54,001 SNPs. The genotyping was done in the context of the GenoTrack research programme funded by the German Federal Ministry of Education and Research. SNPs were filtered for their minor allele frequency and missing genotypes per locus. The thresholds were set to 0.05 and 0.1, respectively. Based on available pedigree information, a sample of ~470,000 cows per trait with complete phenotypic data and genotyped sires and maternal grandsires was assembled (Table 4). The cows descend from 2,081 bulls of which 1,916 do occur as sire and 1,981 as grandsire. A total of 86,254 unique sire-dam-sire combinations were observed. The data set consisting of the bull’s genotypes, the pedigree and the yield deviations of the cows is available in the supplemental material. For this sample, the add and dom coefficients were calculated according to the procedure developed within this study and phenotypes were regressed onto the coefficients applying a simple linear regression model as outlined above.

Furthermore, the mixed model approach as described above was applied to correct for family based stratification. This approach is computationally challenging on a genomewide scale. Thus, a two-step approach was applied, which is equivalent to the previously described genomewide rapid association using mixed model and regression (GRAMMAR) approach (AULCHENKO et al. 2007). It offers a way to
efficiently reduce the time-consuming computations required for extensive pedigrees. In the first step, phenotypes were adjusted for a polygenic effect by applying the mixed model, but omitting the effects for *add* and *dom*. The residuals coming from this step were used as phenotypes in a subsequent linear regression. With this two-step approach, the traits Somatic Cell Score (SCS), Milk yield (MY), Fat yield (FY) and Protein Yield (PY) were analyzed on a genomewide scale.

To validate the two-step approach, chromosome 14 (BTA14) was analyzed applying the full linear mixed model including a polygenic term as well as the coefficients for *add* and *dom*. BTA14 was chosen for this purpose, because the centromeric region of this chromosome contains a major QTL for milk production traits caused by polymorphisms in the *acylCoA-diacylglycerol-acyltransferase* (*DGAT1*) gene (Grisart *et al.* 2002; Riquet *et al.* 1999; Spelman *et al.* 2002; Thaller *et al.* 2003; Weller *et al.* 2003; Winter *et al.* 2002). This QTL segregates within the analyzed population and was thus used as a reference to evaluate the two-step approach as compared to a full model. Furthermore, those SNPs considered as genomewide significant at a threshold of $p \leq 0.01$ after Bonferroni-correction, were reanalyzed applying a full model including a polygenic effect and the coefficients for *add* and *dom* at the same time.

**Implementation**

All data handling procedures including genotype integrity checks, recoding and allele frequency calculation were performed using unix shell scripts together with the genome analysis toolset *plink* (Purcell *et al.* 2007). The derivation of genotype probabilities and simulation routines as well as regression models and test statistics
were realized using the \textit{R statistical environment} (R DEVELOPMENT CORE TEAM 2008). The R-script with the simulation routines as well as the R- and Shell-scripts used for data analysis can be found in Supplemental File 1. The linear mixed models were fitted using REML as implemented in ASReml 3.0 (VSN International Ltd., GILMOUR \textit{et al.} 1995).

\textbf{RESULTS AND DISCUSSION}

\textbf{Simulations}

The results of the single marker simulation scenarios are summarized in Figure 2, illustrating the major impacts on the power to detect additive and dominance effects using genotype probabilities. All parameters under consideration substantially influenced the power of detection. Under a scenario with a heritability of 0.2 and an effect size of 0.2 $\sigma_A$ with $d/|a|=0.5$ (scenario 1, Table 3), a sample size of 800,000 cows was needed to achieve a power of 0.7 for the detection of dominance effects (Figure 2A.). With an effect size of 0.5 $\sigma_A$ and a heritability of 0.25 the same power can be achieved with 100,000 animals (Figure 2B.). Approximately half of the sample size is needed to detect the corresponding additive effects (Figure 2A.). When using the ‘true’ genotypes, a power of 0.94 and 0.69 for additive and dominance effects, respectively, is achieved with 50,000 cows. A sample size of approximately 100,000 animals results in a power of one for both effects under these assumptions (not shown in Figure 2). Using genotype probabilities, the necessary sample size is ten times as large as needed with real genotype data. This clearly reflects the decrease in power arising from the impreciseness of genotype probabilities and the need to compensate this with a larger sample size. Under realistic conditions the power is even lower, because the simulated marker was assumed to be in complete LD with
the QTL and the MAF was fixed to 0.3. From Figure 2D. it arises that with the reduction of $r^2$ from 0.8 to 0.6 the power to detect dominance effects drops from 1 to approximately 0.5 (scenario 4, Table 3) when assuming a MAF of 0.5. This scenario was also carried out using the standard MAF of 0.3, but the power was considerably lower (not shown in Figure 2D.). The substantial dependency between power and LD is reflecting the fact that only parts of the QTL-variance can be explained by markers in incomplete LD. A marker showing an LD of $r^2=0.5$ can capture approximately a half of the QTL effect, meaning that the sample size must be doubled to capture an effect of given size. Regarding the MAF, it can be seen from Figure 2C. that a frequency of 0.5 results in the highest power to detect dominance effects (scenario 3, Table 3). This ideal situation cannot be expected under practical conditions. Thus, we set the MAF to 0.3 in the other simulations. For the additive effect, the MAF is not a limiting factor as in a sample of 100,000 cows the power equals one even for a MAF of 0.1. However, there might be a limitation as to the strong interdependency between MAF and LD shown above. When there is incomplete LD between QTL and marker, the results are very sensitive to low minor allele frequencies. This has to be elucidated in further studies.

Another important parameter is the proportion of variance explained by the QTL (scenario 5, Table 3). For a given degree of dominance of $d/|a|= 1$, a QTL explaining ~5% of the additive genetic variance, would be detectable with a power of ~0.6 in a sample of 100,000 cows (Figure 2E.). Assuming a heritability of 0.25, this would correspond to 1.25% of the phenotypic variance. That is a realistic value in dairy cattle (HAYES and GODDARD 2001; KHATKAR et al. 2004) leading to the conclusion that the approach should work well even for QTL with smaller effects. The power curves for additive and dominance effects as shown in Figure 2E. are congruent, because
the degree of dominance was set to one. Figure 2F. shows the dependency of the power to detect dominance under different degrees of dominance for a given total QTL effect (scenario 6, Table 3). From the figure it can be seen that genotype probabilities calculated for a sample of 100,000 cows are sufficient to detect a degree of dominance of ~0.33 at the given total QTL effect with a power of ~0.8. Regarding the sample sizes discussed above, the detection of dominance effects even at a low degree of dominance is possible with the approach. Furthermore, Figure 2F. impressively illustrates the inferiority of genotype probabilities in small sample sizes. In a sample of 10,000 cows, a power of one can be achieved at $d/|a|=1$ with true genotypes, while the application of genotype probabilities results in a power close to zero for this $d/|a|$.

A part of the decrease in power arising from the use of genotype probabilities can be compensated by the utilization of deeper pedigrees in the estimation of these probabilities. When using $N$ ancestral generations, a proportion of $1/2^N$ of the available information arises from the allele frequency and is thus equal for all animals. With the inclusion of two ancestral generations, i.e. a genotyped sire and damsire, 1/4 of the information comes from the allele frequency. When including a genotyped greatgrandsire, only 1/8 does arise from the frequency. Figure 2B. illustrates the impact of 2 vs. 3 genotyped male ancestors on the power to detect dominance effects reflecting this dependency (scenario 2, Table 3). The inclusion of the third sire increases the power by 0.1 and 0.16 for 50,000 and 100,000 cows, respectively.

From the results of these single marker simulation scenarios it can be concluded that for traits of midrange heritability, at least strong QTL effects can be detected with markers being in LD with the QTL in the range of $r^2>0.4$. However, the necessary
sample size for detecting dominance effects is substantially larger than 100,000 cows. A reliable power to detect smaller dominance effects might under practical conditions even require a sample size of >1,000,000 cows. This is, however, not limiting. Phenotypic data sets of this size are available on the base of national evaluation and increasing reference samples in genomic selection schemes provide the necessary genotypic information. Furthermore, the simple regression approach presented herein is computationally not challenging. In reality, however, the finite number of sire-damsire combinations, representing the source of information for the genotype probabilities, will be the limiting factor in the sense of information content.

The second simulation was conducted as a multi marker scenario and focused on a more realistic population with family-based stratification. Expectedly, this led to the occurrence of false positives in addition to those arising at random as type I error. This is reflected by the specificity to detect dominance as depicted in Figure 3. When increasing the number of daughters per sire from 50 to 5,000, the specificity drops from ~1 to ~0.85. To obtain reliable results, a correction for this stratification is necessary. A rather simple approach would be the inclusion of the fixed effect of the sire and possibly also of the grandsire. This is hampered by the fact that deduced genotypes are a direct function of ancestral genotypes. When using two ancestral generations, all cows descending from the same combination of sire and grandsire have equal genotype probabilities. Correcting for the fixed effects of sire and grandsire is thus inadequate. A valid correction is inevitably linked to the availability of a more independent source of information. A mixed model using a pedigree based relationship matrix to include a polygenic effect meets this criterion. Figure 3 outlines the impact of family sizes and correction on the specificity to detect dominance effects. Even though there is no substantial decrease in specificity with increasing
family size in simulated data, the correction using the linear mixed effects model including the relationship matrix noticeably improves specificity. This is much more pronounced in real data as will be shown below, emphasizing the importance of properly accounting for stratification.

**Real Data analyses**

In a first attempt to validate our approach we analyzed cattle chromosome 14 (BTA14) for the trait *milk yield* without accounting for stratification. This chromosome was chosen, because *DGAT1* located in the centromeric region is the underlying gene of a major QTL, which we attempted to detect based on genotype probabilities. The additive effect estimators for the markers in close proximity to *DGAT1* showed results correctly reflecting the strong additive effect known to originate from this locus. However, the attempt was hampered by a striking lack of specificity due to stratification. The results are characterized by extremely low p-values along the entire chromosome, especially for the additive coefficients while the dominance part is less influenced. These results are depicted in the upper part of Figure 4. To analyze a possible effect of sample size, we started with 10,000 cows and sequentially increased the number up to 470,000 representing the entire data set (data not shown). Larger samples did not lead to an improvement; the specificity substantially decreased with an increasing number of cows. When reaching the maximum sample size, false positives were completely covering the association signal. The subsequent correction applying a mixed model resulted in an effective adjustment. The lower part of Figure 4 shows the corrected results for additive and dominance signals using the complete data set of 470,000 cows with genotype probabilities. The direct comparison of the results without and with correction impressively illustrates the
effectiveness of the approach. After adjustment, the aforementioned \textit{DGAT}-QTL can easily be detected at a position of approximately 0.5 Mb.

In a next step, we conducted analyses on a genomewide scale for the traits somatic cell score (SCS), milk yield (MY), fat yield (FY) and protein yield (PY) applying the substantially faster two-step approach. The direct comparison of the results for BTA14 revealed almost identical results for dominance effects. The preadjustment of phenotypes, however, dramatically increased the sensitivity for additive signals (data not shown). Therefore, it works as a quick scan for dominance. It might be worth investigating, whether this can be improved by the utilization of deeper pedigrees in the calculation of the relationship matrix. Figure 5 is an illustration of the result of a genome scan for dominance effects. Applying a significance threshold of \( p \leq 0.01 \) after Bonferroni-correction, there were no significant results for SCS. For MY, FY and PY, 29 SNPs on 15 chromosomes, 30 SNPs on 11 chromosomes and 59 SNPs on 17 chromosomes with significant dominance effects were detected, respectively (Supplemental Table 1). Some chromosomal regions, especially on BTA9 and BTA22 exhibit significant dominance effects for all analyzed yield traits. To further characterize these findings, additional analyses for FY and PY were performed including MY as a covariable in the preadjustment aiming to reflect the corresponding content traits. Notably, those regions on BTA9 and 22 affecting FY, also significantly affected fat content, while those affecting PY had no significant dominance effect for protein content (data not shown). Thus, it can be concluded that the effect in PY is due to an effect on MY, while fat content is directly affected. These two chromosomes seem to harbour QTL for milk yield and fat content displaying considerable dominance effects. QTL for yield traits have been previously described on BTA 9 (\textit{Wiener et al.} 2000) as well as on BTA22 (\textit{Ashwell et al.} 2004; \textit{Harder et
al. 2006). However, these studies used genotyped sires and daughter based phenotypes and did thus not identify any dominance effects. Furthermore, significant dominance effects were detected on BTA14 within the $DGAT1$ region. Applying the full model as described below, a dominance effect of 0.063 phenotypic standard deviations ($\sigma_p$) along with a significant additive effect of 0.202 $\sigma_p$ was found for marker ARS-BFGL-NGS-100480 at 4.36 Mb. This is in general accordance with the divergent effect of the heterozygous condition compared to the mean of the homozygous states found in a study using 1,035 Holstein cows genotyped for two $DGAT1$ polymorphisms (KUEHN et al. 2007).

As the dependant variables in this approach are residuals from the previous step, the effect estimators can not be interpreted readily. Thus, an analysis applying the full model is necessercary for the markers considered significant in order to obtain reliable effect estimators. This was conducted for the markers showing significant dominance effects with $p \leq 0.01$ for MY, FY and PY. The effect estimators obtained from this model are exemplarily shown for BTA9, 14 and 22 in Table 5 (see also Supplemental Table 1). The markers on BTA9 are distributed across a large chromosomal region ranging from 46.5 to 100.3 Mb. They are characterized by notably high degrees of dominance. For milk yield, there is a clear peak around 62 Mb. The highest dominance effect of 0.065 $\sigma_p$ for this trait was found at marker Hapmap35559-SCAFFOLD35632_12026 at 62.96 Mb. The degree of dominance ($d/|a|$) is 1.754 and no significant additive effect was detected applying the full model. This is even more pronounced for PY. For this trait, the same marker shows a dominance effect of 0.072 $\sigma_p$ and a $d/|a|$ value of 12.466. Such signals might be missed in studies using breeding values as phenotypes. In this case however, there are adjacent markers, for which considerable additive effects were detected (Table 5). The same applies for
BTA22, where the \(d/|a|\) values are generally smaller as compared to BTA9. In total, there are 10, 12 and 26 SNPs with \(d/|a|\) values \(>2\) for MY, FY and PY, respectively (Supplemetal Table 1). For none of these markers, a significant additive effect was observed when applying the full model. For each analyzed yield trait, there is at least one marker showing an additive effect of more than 0.1 \(\sigma_p\) with \(p \leq 0.01\). These effects would probably have been detectable in studies using daughter based breeding values, but the considerable dominance effects would be missed. Large scale studies using breeding values or daughter yield deviations exhibit high statistical power and have successfully been applied to the identification of numerous QTL in dairy cattle. However, our results emphasize the need to use direct phenotypes in order to better understand the genetic architecture of traits.

**CONCLUSION**

The detection of dominance effects relies on the utilization of direct phenotypes, while classical QTL mapping approaches in dairy cattle employed breeding values of sires based on daughter performance. The implementation of Genomic Selection schemes in dairy cattle breeding provides a large number of bulls genotyped for dense genomewide marker panels. Within the current study, these genotypes were used to derive genotype probabilities in female descendants of these bulls. Using these probabilities it is possible to detect significant dominance effects on a genomewide scale relying on a large sample of phenotyped cows. Together with a two-step mixed model approach to account for relationship among animals, the approach is computationally not challenging and applicable to datasets of more than \(10^6\) cows. The application of this approach could substantially enhance our knowledge about the genetic architecture of performance and functional traits in dairy cattle.
cattle. The introduced method could also be extended for the use in genomic prediction including the enlargement of reference populations or the prediction of individual performance accounting for non-additive genetic effects.

ACKNOWLEDGEMENT

This study was financially supported by the German Federal Ministry for Education and Research in the context of the FUGATO* project GenoTrack (grant 0315134A) and the competence networks Synbreed (grant 0315528F) and Phänomics (grant 0315536D).
REFERENCES


LUND, M. S., A. P. W. DE ROOS, A. G. DE VRIES, T. DRUET, V. DUCROOC et al., 2010 Improving genomic prediction by EuroGenomics collaboration, pp. in 9th World Congress on Genetics Applied Livestock Production, Leipzig, Germany.


TABLE 1. Genotype probabilities for the dam of the cow under consideration. The probabilities were calculated from the allele frequencies \( f(A) = p \) and \( f(B) = q \) and the probability of the grandsire (GS) to transmit allele A \( P(A)_{GS} \).

<table>
<thead>
<tr>
<th>Genotype GS</th>
<th>( P(A)_{GS} )</th>
<th>( P(AA)_{D} )</th>
<th>( P(AB)_{D} )</th>
<th>( P(BB)_{D} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>1</td>
<td>( p )</td>
<td>( q )</td>
<td>0</td>
</tr>
<tr>
<td>AB</td>
<td>( \frac{1}{2} )</td>
<td>( \frac{p}{2} )</td>
<td>( \frac{1}{2} )</td>
<td>( \frac{q}{2} )</td>
</tr>
<tr>
<td>BB</td>
<td>0</td>
<td>0</td>
<td>( p )</td>
<td>( q )</td>
</tr>
</tbody>
</table>
TABLE 2. Conditional genotype probabilities of a cow. The nine possible scenarios for the genotype probabilities of a cow conditional on the genotypes of the animal’s sire and maternal grandsire as well as the frequency of the rare allele $A$ $f(A) = p$ are shown. The probabilities were calculated by multiplying the respective transmission probabilities of the sire and dam of the cow under consideration.

<table>
<thead>
<tr>
<th>Genotype sire</th>
<th>AA</th>
<th>AB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AA</strong></td>
<td>$P(AA_{cow}) = 0.5 + 0.5p$</td>
<td>$P(AA_{cow}) = 0.25 + 0.25p$</td>
<td>$P(AA_{cow}) = 0$</td>
</tr>
<tr>
<td></td>
<td>$P(AB_{cow}) = 0.5 - 0.5p$</td>
<td>$P(AB_{cow}) = 0.5$</td>
<td>$P(AB_{cow}) = 0.5 + 0.5p$</td>
</tr>
<tr>
<td></td>
<td>$P(BB_{cow}) = 0$</td>
<td>$P(BB_{cow}) = 0.25 - 0.25p$</td>
<td>$P(BB_{cow}) = 0.5 - 0.5p$</td>
</tr>
<tr>
<td><strong>AB</strong></td>
<td>$P(AA_{cow}) = 0.25 + 0.5p$</td>
<td>$P(AA_{cow}) = 0.125 + 0.25p$</td>
<td>$P(AA_{cow}) = 0$</td>
</tr>
<tr>
<td></td>
<td>$P(AB_{cow}) = 0.75 - 0.5p$</td>
<td>$P(AB_{cow}) = 0.5$</td>
<td>$P(AB_{cow}) = 0.25 + 0.5p$</td>
</tr>
<tr>
<td></td>
<td>$P(BB_{cow}) = 0$</td>
<td>$P(BB_{cow}) = 0.375 - 0.25p$</td>
<td>$P(BB_{cow}) = 0.75 - 0.5p$</td>
</tr>
<tr>
<td><strong>BB</strong></td>
<td>$P(AA_{cow}) = 0.5p$</td>
<td>$P(AA_{cow}) = 0.25p$</td>
<td>$P(AA_{cow}) = 0$</td>
</tr>
<tr>
<td></td>
<td>$P(AB_{cow}) = 1 - 0.5p$</td>
<td>$P(AB_{cow}) = 0.5$</td>
<td>$P(AB_{cow}) = 0.5p$</td>
</tr>
<tr>
<td></td>
<td>$P(BB_{cow}) = 0$</td>
<td>$P(BB_{cow}) = 0.5 - 0.25p$</td>
<td>$P(BB_{cow}) = 1 - 0.5p$</td>
</tr>
</tbody>
</table>
**TABLE 3. Detailed description of the single marker simulation scenarios.** For each scenario, the range of the modified parameter is given together with the fixed parameters. The QTL effects are expressed as additive genetic standard deviations ($\sigma_A$), MAF and LD stand for minor allele frequency and linkage disequilibrium, respectively. The degree of dominance was calculated as the ratio of dominance effect and absolute additive effect ($d/|a|$). Sample sizes are given in multiples of 1,000 (k).

| Scenario | Tested parameter | $h^2$ | QTL effect | Sample size | MAF | LD | $d/|a|$ | Range of modified parameter | Explanatory notes |
|----------|------------------|-------|------------|-------------|-----|----|--------|----------------------------|------------------|
| 1 | Sample size | 0.2 | 0.2 $\sigma_A$ | -- | 0.3 | $r^2=1$ | 0.5 | N=50k-1000k (14 steps) | Heritability and effect size were reduced to achieve a higher resolution across the entire range of N. True genotypes were compared with genotype probabilities. |
| 2 | Number of considered generations | 0.25 | 0.5 $\sigma_A$ | N=5k-100k (19 steps) | 0.3 | $r^2=1$ | 0.5 | N=5k-100k (19 steps) | Genotype probabilities were calculated using a genotyped sire and grandsire only or including an additional greatgrandsire in an increasing sample size. |
| 3 | Minor allele frequency | 0.25 | 0.5 $\sigma_A$ | 10k vs. 100k | -- | $r^2=1$ | 0.5 | MAF=0.025-0.5 (19 steps) | MAF was increased stepwise in two different sample sizes. |
| 4 | LD between QTL and marker | 0.25 | 0.5 $\sigma_A$ | N=5k-100k (19 steps) | 0.3/0.5 | -- | 0.5 | $r^2=0.4/0.6/0.8$ | An increasing sample size was analyzed for three different $r^2$ values and two different MAFs. |
| 5 | Variance explained by the QTL | 0.25 | -- | 10k vs. 100k | 0.3 | $r^2=1$ | 1 | $\sigma_q^2/\sigma_e^2=0.01-0.2$ (14 steps) | $\sigma_q^2/\sigma_e^2$ was increased stepwise in two different sample sizes. |
| 6 | Degree of dominance | 0.25 | 0.5 $\sigma_A$ | 10k vs. 100k | 0.3 | $r^2=1$ | -- | $d/a=0.05-2$ (39 steps) | $d/|a|$ was increased stepwise in two different sample sizes. True genotypes were compared with genotype probabilities within these samples. |
TABLE 4. Description of the real data set. The upper part of the table summarizes
the mean (μ), minimum (min) and maximum (max) values of the YD as well as their
respective standard deviation (σP) and the number of cows with available
phenotypes. The lower part summarizes the family structure.

<table>
<thead>
<tr>
<th>Phenotype summary</th>
<th>μ</th>
<th>Min</th>
<th>Max</th>
<th>σP</th>
<th>N cows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk yield</td>
<td>7.61·10^{-9}</td>
<td>-1426.85</td>
<td>1284.05</td>
<td>213.54</td>
<td>469,454</td>
</tr>
<tr>
<td>Fat yield</td>
<td>7.26·10^{-10}</td>
<td>-125.86</td>
<td>114.04</td>
<td>16.75</td>
<td>470,260</td>
</tr>
<tr>
<td>Protein yield</td>
<td>-6.75·10^{-13}</td>
<td>-63.22</td>
<td>54.86</td>
<td>8.71</td>
<td>469,171</td>
</tr>
<tr>
<td>Somatic Cell Score</td>
<td>5.57·10^{-12}</td>
<td>-3.60</td>
<td>3.61</td>
<td>0.42</td>
<td>471,093</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary of Family structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daughters per sire</td>
</tr>
<tr>
<td>Granddaughters per sire</td>
</tr>
<tr>
<td>Cows per combination of sire and grandsire</td>
</tr>
</tbody>
</table>
TABLE 5. Results from applying the full mixed model to the real data set for selected markers and chromosomes. Given are the additive (Add) and dominance (Dom) effects as obtained from the full model as well as the degrees of dominance (d/|a|) for SNP markers showing significant ($p \leq 0.01$) dominance effects in the two-step genome scan. The table exemplarily summarizes the markers on BTA9, 14 and 22 (for a full list see Supplemental Table 1). Effects are given in phenotypic standard deviations ($\sigma_P$). Significance of additive effects is indicated by superscripts (a: $p \leq 0.01$, b: $p \leq 0.05$). CHR and BP are the cattle chromosome and the position in base pairs according to genome build UMD 3.1.

<table>
<thead>
<tr>
<th>SNP Name</th>
<th>CHR</th>
<th>BP</th>
<th>Milk yield</th>
<th>Fat yield</th>
<th>Protein yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hapmap33674-BTA-156367</td>
<td>9</td>
<td>46,531,368</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BTA-01878346</td>
<td>9</td>
<td>54,937,110</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BTA-01576221</td>
<td>9</td>
<td>56,786,292</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BTA-01573160</td>
<td>9</td>
<td>58,011,347</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BTA-01352997</td>
<td>9</td>
<td>61,883,731</td>
<td>0.158a</td>
<td>0.064</td>
<td>0.070b</td>
</tr>
<tr>
<td>Hapmap35559-SCAFFOLD35632_12026</td>
<td>9</td>
<td>62,961,279</td>
<td>0.037</td>
<td>0.065</td>
<td>1.754</td>
</tr>
<tr>
<td>BTA-01573160</td>
<td>9</td>
<td>61,883,731</td>
<td>0.032</td>
<td>0.055</td>
<td>1.703</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>14</td>
<td>2,054,457</td>
<td>0.094a</td>
<td>0.023</td>
<td>0.245</td>
</tr>
<tr>
<td>BTA-84397-100124</td>
<td>22</td>
<td>26,073,419</td>
<td>0.027</td>
<td>0.054</td>
<td>1.974</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>14</td>
<td>21,431,682</td>
<td>0.051</td>
<td>0.077</td>
<td>1.530</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>22</td>
<td>21,455,286</td>
<td>0.051</td>
<td>0.078</td>
<td>1.530</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>22</td>
<td>26,073,419</td>
<td>0.027</td>
<td>0.054</td>
<td>1.974</td>
</tr>
<tr>
<td>BTA-84397-100124</td>
<td>22</td>
<td>31,702,243</td>
<td>0.152a</td>
<td>0.086</td>
<td>0.565</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>22</td>
<td>32,234,944</td>
<td>0.150a</td>
<td>0.094</td>
<td>0.626</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>22</td>
<td>33,766,195</td>
<td>0.121a</td>
<td>0.089</td>
<td>0.737</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>22</td>
<td>38,727,002</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hapmap49345-BTA-85101</td>
<td>22</td>
<td>38,815,613</td>
<td>0.056</td>
<td>0.057</td>
<td>1.009</td>
</tr>
</tbody>
</table>
FIGURE 1. Graphical representation of the pedigree structure used in the calculation of genotype probabilities. N determines the generation in relation to the cow under consideration as used in the formulas to derive probabilities. P(Genotype) with genotypes AA, AB and BB is the genotype probability and P(Allele) with alleles A and B is the transmission probability for the respective allele.
FIGURE 2. Graphical representation of the results obtained from single marker simulation scenarios according to Table 3. A. Power to detect additive (add) and dominance (dom) effects conditional on sample size (scenario 1). B. Power to detect dominance effects conditional on sample size and the number of sires used in calculating genotype probabilities (scenario 2). C. Power to detect additive (add) and dominance (dom) effects conditional on the minor allele frequency (MAF) in two different sample sizes (scenario 3). D. Power to detect dominance effects conditional on sample size assuming different LD ($r^2$) between marker and QTL (scenario 4). E. Power to detect additive (add) and dominance (dom) effects conditional on the proportion of additive genetic variance explained by the QTL ($\sigma^2_{QTL}/\sigma^2_A$) in two different sample sizes (scenario 5). F. Power to detect dominance effects conditional on the degree of dominance (d/|a|) and compared between true genotypes (GT) and genotype probabilities (P(GT), scenario 6).
FIGURE 3. Graphical representation of the specificity to detect dominance effects as obtained from the multi marker simulation scenario. Compared are the specificities conditional on family size (daughters per sire) without any correction for relatedness (black) and after correction applying a linear mixed model (blue).
FIGURE 4. Analysis of BTA14 for the trait milk yield in a real data set comprising 469,454 cows. The y-axis represents the negative decadic logarithm of the p-values for the additive (left) and dominance (right) effects after Bonferroni correction. The x-axis gives the position on BTA14 in megabasepairs according to genome build UMD 3.1. The upper panel represents the results without correction. Especially for the additive effects, there is a striking lack of specificity. The lower panel depicts the results obtained by applying mixed model correction.
FIGURE 5. Manhattan plots representing the results of a genome scan for dominance applying correction for relatedness among animals in a two-step mixed model approach. Analyzed were the traits Somatic Cell Score (SCS), milk yield, fat yield and protein yield. Shown are the negative decadic logarithms of the raw p-values with a Bonferroni-corrected 1% genomewide significance level (dashed line).