J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Robust QTL fine mapping by applying a quantitative transmission disequilibrium test to the Mendelian sampling term

H. Simianer & E.C.G. Pimentel

Department of Animal Science, Animal Breeding and Genetics Group, Georg-August-University, Goettingen, Germany

Keywords
mendelian sampling; QTL fine mapping; quantitative TDT.

Correspondence
Henner Simianer, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany.
Tel: +49 551 39 5604;
Fax: +49 551 39 5587;
E-mail: hsimian@gwdg.de

Received: 26 September 2008;
accepted: 22 March 2009

Summary

In many farm animal populations, high-density single nucleotide polymorphism (SNP) genotypes are becoming available on a large scale, and routine estimation of breeding values is implemented for a multiplicity of traits. We propose to apply the basic principle of the quantitative transmission disequilibrium test (QTDT) to estimated Mendelian sampling terms. A two-step procedure is suggested, where in the first step additive breeding values are estimated with a mixed linear model and the Mendelian sampling terms are calculated from the estimated breeding values. In the second step, the QTDT is applied to these estimated Mendelian sampling terms. The resulting test is expected to yield significant results if the SNP is in sufficient linkage disequilibrium and linkage with quantitative trait loci (QTL). This principle is illustrated with a simulated data set comprising 4665 individuals genotyped for 6000 SNP and 15 true QTL. Thirteen of the fifteen QTL were significant on a genome-wide 0.1% error level. Results for the empirical power are derived from repeated samples of 1000 and 3000 genotyped individuals, respectively. General properties and potential extensions of the methodology are indicated. Owing to its computational simplicity and speed, the suggested procedure is well suited to scan whole genomes with high-density SNP coverage in samples of substantial size and for a multiplicity of different traits.

Introduction

The availability of high-density single nucleotide polymorphism (SNP) genotyping in farm animals has generated a need for novel tools in statistical genetics, especially for efficient and robust fine mapping of quantitative trait loci (QTL). The traditional mapping designs, i.e. the F_2-designs in multiparous species like pig or chicken, or (grand-)daughter designs in species with large paternal half-sib groups like cattle, are limited because the number of recombinations in the observed pedigree is low and thus the positional resolution is insufficient. Higher resolution can be achieved with approaches that capitalize on linkage disequilibrium (LD), which has been generated in the studied populations over longer time periods. Owing to factors like population subdivisions, large family structures caused by the extensive use of artificial insemination, smaller and continuously decreasing effective population size (Hayes *et al.* 2003), and intensive selection, LD in farm animal populations was found to be on a higher level than LD in large outbred populations such as the human population (Andersson 2001).

  doi:10.1111/j.1439-0388.2009.00812.x

In association studies based on historic LD, high statistical power exists only in the immediate vicinity of the QTL, so that high marker densities are required. Pure association tests are known to be highly sensitive to population stratification and admixtures (Freedman *et al.* 2004). The situation in farm animal study is expected to be similar or worse as in this case sporadic appearances of high LD were even reported for non-syntenic loci, caused by selection, population stratification or admixture and other mechanisms (Farnir *et al.* 2000).

To increase robustness of LD-based association tests, Spielman *et al.* (1993) proposed the transmission disequilibrium test (TDT), which tests the null hypothesis that a marker and a dichotomous phenotype (e.g. a disease incident) are both in linkage and in LD. In the original approach, this is achieved by considering the contrast between the transmitted versus the non-transmitted allele at a given locus in nuclear families (one affected offspring and its parents). The test is fully non-parametric and model-free, the only assumption made is that of Mendelian inheritance. The basic principle of the TDT has been extended to various more general cases like the use of haplotypes (Clayton 1999), general pedigrees (Rabinowitz & Laird 2000) or analysis of quantitative traits (Abecasis *et al.* 2000), the latter termed as quantitative TDT (QTDT). For a comprehensive review of the TDT methodology, see Laird & Lange (2006).

A general feature of the quantitative TDT is that it avoids false-positive results in the case of population stratification. The price for this robustness seems to be a reduced power compared with the straight association mapping approaches (Ewens *et al.* 2008). Applications of the QTDT to candidate gene regions are numerous in human genetics but can also be found for cattle (Kunej *et al.* 2007; Jiang *et al.* 2008) and pigs (Bink *et al.* 2000; Hernández-Sánchez *et al.* 2003; Wimmers *et al.* 2004). Aulchenko *et al.* (2007) compared different pedigree-based QTL association analysis methods in three different data structures, one of them reflecting an idealized pig population family structure. They also report a lack of power of different QTDT implementations compared with the competing approaches.

Meuwissen *et al.* (2002) suggested a population-based fine mapping approach that combines recombination fractions in the observed pedigree with LD information derived from earlier generations. This approach was shown to provide a high mapping resolution in the demonstrated example (twinning rate in cattle), but is computationally demanding as it involves the construction of a matrix of pair-wise identity by descent (IBD) probabilities between all haplotypes at the putative position of the QTL. It also requires the specification of typically unknown quantities, like the effective population size or the 'age' of the population. However, the approach was shown to be robust against misspecifications of those parameters (Meuwissen *et al.* 2002).

Data structures in farm animal studies differ from those in human genetics. Often large numbers of reliable phenotypes are routinely recorded, deep pedigrees are widely available, large families (e.g. half-sib groups) exist and targeted genotyping of informative individuals is possible. In many farm animal species, sophisticated polygenic breeding value estimation is routinely conducted for many economically and functionally important traits.

In this study, we suggest a two-step procedure for fine mapping of QTL, capitalizing on these structural advantages. In the first step, the complete phenotypic and pedigree information is used to estimate breeding values for the studied trait. From the estimated breeding values, estimates of the Mendelian sampling term of all animals are derived, which are readily corrected for all relevant non-genetic effects and reflect the deviation of an offspring from the parent mean owing to the sampling of parental alleles.

In the second step, the basic idea of the TDT is applied to all genotyped parent–offspring pairs or triplets, and a combined test of all contrasts of the average Mendelian sampling between transmitted and non-transmitted alleles is suggested. As Mendelian sampling terms between individuals are independent, the test statistic is easily accumulated over related and non-related pairs and triplets, and a simple statistic based on the *t*-test is suggested. The resulting approach is shown to produce robust signals of LD and linkage, and at the same time it is fast and easy to implement.

The structure of this paper is as follows. First, we present the methodology and specify the necessary computations in different scenarios. The method then is illustrated with a simulated data set to demonstrate its resolution and numerical feasibility, and the empirical power is studied by analysing repeated sub-samples of the simulated data set of different sizes. In the *Discussion*, the general properties of the suggested approach are discussed and potential extensions of the methodology are indicated.

Methods

We consider a population in which a quantitative trait y (like milk yield or daily gain) is routinely

recorded for many individuals. Additive breeding values are estimated for all individuals in the pedigree using best linear unbiased prediction (BLUP; Henderson 1973). The estimated breeding value $\hat{a}_i$ of an individual i with parents j and k can be decomposed in

$$\hat{a}_i = 0.5(\hat{a}_j + \hat{a}_k) + \hat{m}_i, \quad (1)$$

where $\hat{a}_j$ and $\hat{a}_k$ are the estimated parental breeding values and $\hat{m}_i$ is the estimated Mendelian sampling term of individual i; $\hat{m}_i$ represents the deviation of an individual's additive polygenic effect from the expectation, after correction for all environmental effects and given the parents' genotype. Note that $\hat{m}_i$ is independent of information of possible half- or full-sibs of individual i. $\hat{m}_i$ is expected to be > 0 if the average of the parental QTL alleles obtained was greater than the parental average, and $E(\hat{m}_i) < 0$ otherwise. Following the concept of Avendano *et al.* (2005), the accuracy of the Mendelian sampling term of individual i is expressed as correlation between the true and the estimated Mendelian sampling terms $0 \le \rho_i \le 1$. According to the selection index theory, $\rho_i = \sqrt{\mathrm{Var}(\hat{m}_i)/\mathrm{Var}(m_i)}$, where $\mathrm{Var}(\hat{m}_i)$ and$\mathrm{Var}(m_i)$ are the variances of the estimated and the true Mendelian sampling terms. $\mathrm{Var}(m_i) = (0.5 - 0.25(F_j + F_k))\sigma_a^2$, where F_j and F_k are the inbreeding coefficients of the parents j and k. Hence, $\mathrm{Var}(m_i) = 0.5\sigma_a^2$ only if the parents are not inbred. $\mathrm{Var}(\hat{m}_i)$ can be calculated from the breeding value estimation. It follows from Equation (1) that

$$\begin{aligned}\mathrm{Var}(\hat{m}_i) = {} & \mathrm{Var}(\hat{a}_i) + 0.25(\mathrm{Var}(\hat{a}_j) + \mathrm{Var}(\hat{a}_k)) \\ & - \mathrm{Cov}(\hat{a}_i, \hat{a}_j) - \mathrm{Cov}(\hat{a}_i, \hat{a}_k) + 0.5\mathrm{Cov}(\hat{a}_j, \hat{a}_k).\end{aligned}$$

The variances and covariances of the estimated breeding values can be extracted from the variance–covariance matrix

$$\mathrm{VCV}(\hat{a}) = \mathbf{G} - \mathbf{C}_{22},$$

where $\hat{a}$ is the vector of all estimated breeding values, $\mathbf{G}$ is the additive-genetic variance–covariance matrix and $\mathbf{C}_{22}$ is the lower right part of the inverted coefficient matrix of the mixed model equations (Henderson 1984), pertaining to the random breeding values. If the exact inverse coefficient matrix is not available, the required elements of the covariance matrix can be derived by approximate approaches as suggested, e.g. by Tier & Meyer (2004). If no information on the offspring or its progeny is available, $\rho_i = 0$. In this case, the estimated breeding value of an offspring is the average of the breeding values of its parents and hence the estimated Mendelian sampling term is invariably 0. If an offspring has a very reliable estimated breeding value, e.g. if its progeny is tested, then ρ_i approaches 1. We now assume that individuals i, j and k are genotyped for a biallelic SNP, with alleles coded as '1' or '2'. We adopt the idea of TDT by contrasting average Mendelian sampling terms of individuals that have obtained opposite paternal alleles across the population. The basic concept is illustrated with the following example: consider a heterozygous sire j with genotype $\{1,2\}$. The offspring of this sire can be grouped in a transmission class that has inherited allele 1 but not allele 2 (indicated as $1\backslash 2$, read 'one and not two'), with average Mendelian sampling $\bar{m}_{1\backslash 2,j}$ and a second transmission class that has inherited allele 2 and not 1 (indicated as $2\backslash 1$), with average Mendelian sampling $\bar{m}_{2\backslash 1,j}$. We now accumulate the contrast over all J heterozygous sires $\delta = \sum_{j=1}^{J} (\bar{m}_{1\backslash 2,j} - \bar{m}_{2\backslash 1,j})$. Under the null hypothesis that the marker is not linked to a QTL affecting the trait or the marker and the QTL are in linkage equilibrium, δ is expected to be zero. If the marker is linked to the QTL, but not in LD, we will find non-zero contrasts within sires, but they will average out over all sire families, because in some families the positive allele is associated with marker allele 1, while in other families the positive allele is associated with marker allele 2. If, on the other hand, LD exists without linkage, we will not find any contrast within the sire families. Only if the marker is linked to the QTL, and marker and QTL are in LD, we expect to find a non-random deviation of δ from zero. This principle is now extended to the more general case of trios. Given complete genotypes of the sire, the dam and the offspring, it can in all but one case be fully determined which parental alleles are transmitted to the offspring and which ones are not (Table 1). Only in the case where all three individuals are heterozygous $\{1,2\}$, transmission is ambiguous and cannot be included in the statistic.

Using the order of the 15 possible constellations of unordered genotypes as given in Table 1, it should be noted that the transmission $1\backslash 2$ is observed on the paternal (maternal) side in cases 5, 7, 10 (2, 7, 13), while $2\backslash 1$ is observed in cases 6, 9, 11 (3, 9, 14). Note that in cases 7 and 9, informative transmissions are observed both on the paternal and the maternal sides, so these cases need to be counted twice. If $\bar{m}_c$ is the mean Mendelian sampling in case c based on n_c observations, the contrast reflects the mean difference of all

Table 1 Fifteen possible combinations of unordered genotype triplets (cases 1–15) and pairs (sire-offspring cases, s1 and s2; dam-offspring cases, d1 and d2) with transmitted and non-transmitted parental alleles and weight of the observation λ_i

	Unordered genotype			Paternal allele		Maternal allele		
Case	Sire	Dam	Offspring	Transmitted	Not transmitted	Transmitted	Not transmitted	λ_i
1	{1,1}	{1,1}	{1,1}	1	1	1	1	0
2	{1,1}	{1,2}	{1,1}	1	1	1	2	1
3	{1,1}	{1,2}	{1,2}	1	1	2	1	−1
4	{1,1}	{2,2}	{1,2}	1	1	2	2	0
5	{1,2}	{1,1}	{1,1}	1	2	1	1	1
6	{1,2}	{1,1}	{1,2}	2	1	1	1	−1
7	{1,2}	{1,2}	{1,1}	1	2	1	2	2
8	{1,2}	{1,2}	{1,2}	Unknown				0
9	{1,2}	{1,2}	{2,2}	2	1	2	1	−2
10	{1,2}	{2,2}	{1,2}	1	2	2	2	1
11	{1,2}	{2,2}	{2,2}	2	1	2	2	−1
12	{2,2}	{1,1}	{1,2}	2	2	1	1	0
13	{2,2}	{1,2}	{1,2}	2	2	1	2	1
14	{2,2}	{1,2}	{2,2}	2	2	2	1	−1
15	{2,2}	{2,2}	{2,2}	2	2	2	2	0
s1	{1,2}	–	{1,1}	1	2	1	Unknown	1
s2	{1,2}	–	{2,2}	2	1	2	Unknown	−1
d1	–	{1,2}	{1,1}	1	Unknown	1	2	1
d2	–	{1,2}	{2,2}	2	Unknown	2	1	−1

cases $1\backslash 2$ versus $2\backslash 1$ over all genotyped triplets and both on the paternal and maternal sides.

$$\delta = \frac{n_2\bar{m}_2 + n_5\bar{m}_5 + 2n_7\bar{m}_7 + n_{10}\bar{m}_{10} + n_{13}\bar{m}_{13}}{n_2 + n_5 + 2n_7 + n_{10} + n_{13}} - \frac{n_3\bar{m}_3 + n_6\bar{m}_6 + 2n_9\bar{m}_9 + n_{11}\bar{m}_{11} + n_{14}\bar{m}_{14}}{n_3 + n_6 + 2n_9 + n_{11} + n_{14}} \quad (2)$$

In real data, only a subset of all individuals is genotyped. Starting from a genotyped individual (note that only complete genotypes are considered, i.e. animals with just one allele known are ignored), the following four cases are relevant: (i) both parents are not genotyped; (ii) only the sire is genotyped; (iii) only the dam is genotyped; and (iv) both parents are genotyped. Case (i) is non-informative and can be discarded and case (iv) is the triplet situation described previously. Cases (ii) and (iii), however, contain relevant information and can be included in the analysis. Consider case (ii) that only the sire is genotyped and has genotype $\{1,2\}$. In this case, we can only infer the allele transmission pattern if the offspring is homozygous. If, say, the offspring genotype is $\{1,1\}$, transmission is clearly $1\backslash 2$, and the information content of this observation is equivalent to case 5 in the triplet situation. The four different informative single parent–offspring cases are listed in Table 1. We suggest merging case s1 with the triplet case 5, s2 with triplet case 11, d1 with triplet case 2 and d2 with triplet case 14. As δ is a contrast of sample means, we apply the *t*-test (Snedecor & Cochran 1956), using

$$\hat{\sigma}_\delta = \sqrt{\frac{(n_{1\backslash 2}-1)\hat{\sigma}^2_{1\backslash 2} + (n_{2\backslash 1}-1)\hat{\sigma}^2_{2\backslash 1}}{(n_{1\backslash 2}-1)+(n_{2\backslash 1}-1)}\left(\frac{1}{n_{1\backslash 2}}+\frac{1}{n_{2\backslash 1}}\right)},$$

where

$$n_{1\backslash 2} = n_2 + n_5 + 2n_7 + n_{10} + n_{13}$$

$$n_{2\backslash 1} = n_3 + n_6 + 2n_9 + n_{11} + n_{14}$$

and $\hat{\sigma}^2_{1\backslash 2}$ and $\hat{\sigma}^2_{2\backslash 1}$ are the estimated variances of the Mendelian sampling terms in the two transmission groups that can be computed from the observed data. Under the null hypothesis, the test statistic

$$t = \delta/\hat{\sigma}_\delta$$

approximately follows a *t*-distribution with $n_{1\backslash 2}+n_{2\backslash 1}-2$ degrees of freedom. In many cases, the number of degrees of freedom will exceed 50, so that the standard normal distribution can be used as a sufficient approximation to the null distribution.

Simulated data

To study the properties of the suggested methodology, we used a simulated data set that was distributed for the 12th QTL-MAS Workshop in May 2008 in Uppsala, Sweden (Crooks *et al.* 2009). A genome

of six chromosomes each 1 M (M indicates Morgan) long was generated. In the base population, 6000 SNP were generated uniformly spaced at 0.1 cM. For convenience, SNP are numbered across the genome, so that SNP 4215 is on chromosome four at position 21.5 cM from the start. In addition to the SNP, 48 QTL of variable sizes were assigned to chromosomes one to five. All QTL were fully additive and one pair of QTL (positioned at 30 cM on chromosome 1 and positioned at 60 cM on chromosome 2) interacted epistatically.

Fifty males and fifty females were generated in the base population, and subsequently 50 generations of random mating were performed. In generation 51, 15 males and 150 females were sampled at random as parents. Each female was mated to a random male and produced 10 offspring, so that generation 52 encompassed 1500 individuals. From these, again 165 animals were selected at random and mated to produce generation 53, and generation 54 was produced similarly. All 4665 individuals in generations 51–54 were fully genotyped and had phenotypes. The phenotype of an animal was generated by summing up the additive and epistatic effects over all QTL and adding a random noise $e \sim N(0, \sigma_e^2)$, where $\sigma_e^2 = 4.2$. This resulted in an approximate true heritability of 0.3. In generations 51–54, only 15 of the simulated 48 QTL explained more than 1% of the additive genetic variance (Table 2).

Table 2 Simulated quantitative trait loci that explain more than 1% of the total additive genetic variance

Position	Allele substitution effect (absolute value)	Minor allele frequency	Percentage of additive genetic variance explained
200	0.62	0.28	11.8
400	0.56	0.07	3.1
772	0.37	0.29	4.3
1274	0.35	0.44	4.6
1300	0.33	0.21	2.8
1486	0.37	0.40	5.0
1749	0.50	0.18	5.6
1935	0.25	0.32	2.1
2149	0.30	0.40	3.3
2600	0.68	0.07	4.6
3032	0.61	0.39	13.5
3369	0.34	0.24	3.2
3761	0.58	0.41	12.4
3965	0.29	0.19	2.0
4935	0.75	0.26	16.5

Results

We applied the method suggested before to test for the presence of QTL based on genotypes at 6000 SNP and phenotypes of the 4665 animals in generations 51–54. For these animals, the full pedigree information was available.

First, a polygenic animal model was used to estimate variance components, breeding values and ultimately the Mendelian sampling terms. The statistical model was:

$$y_i = \mu + a_i + e_i,$$

where y_i is the phenotype of individual i; μ is the overall fixed mean; a_i is the random additive genetic effect (or breeding value) of individual i; and e_i is the random residual term of individual i.

The random variables are distributed as $a \sim N(0; A\sigma_a^2)$ and $e \sim N(0; I\sigma_e^2)$, where σ_a^2 and σ_e^2 are the additive genetic and the residual variance, I is the identity matrix and A is the numerator relationship matrix (Henderson 1975). First, residual maximum likelihood (REML; Patterson & Thompson 1971) estimates of the variance components were obtained using the program VCE 4.2.5 (Groeneveld 1998), resulting in $\hat{\sigma}_a^2 = 1.36$ and $\hat{\sigma}_e^2 = 3.12$ giving the heritability $\hat{h}^2 = 0.304$. Using these variance components and the same statistical model, BLUP breeding values $\hat{a}_i$ were calculated for all $i = 1, \ldots, 4665$ animals. Then, for each animal i with known parents j and k (i.e. the 4500 animals in generations 52–54) the estimated Mendelian sampling term $\hat{m}_i$ was computed as:

$$\hat{m}_i = \hat{a}_i - 0.5(\hat{a}_j + \hat{a}_k). \quad (3)$$

Using these estimates of the Mendelian sampling terms, the contrast δ_l was calculated for all loci $l = 1, \ldots, 6000$ using Equation (2). Note that a complete triplet was available for all 4500 offspring belonging to generations 52–54. From the resulting t-values and the degrees of freedom, the error probability was calculated. In Figure 1, the logarithm to the base 10 of the error probabilities is plotted over the entire simulated genome. Significance was tested on the genome-wide 0.1% error level using a Bonferroni correction. As it is a two-sided t-test, the critical error probability was set to $p_c = 0.0005/6000$ with $\log_{10}(p_c) = -7.079$.

In the lower part of Figure 1, the chromosome segments in which the test statistic exceeds the critical threshold are indicated together with the positions of those 15 true QTL explaining more than 1% of the additive genetic variance. We define a QTL to

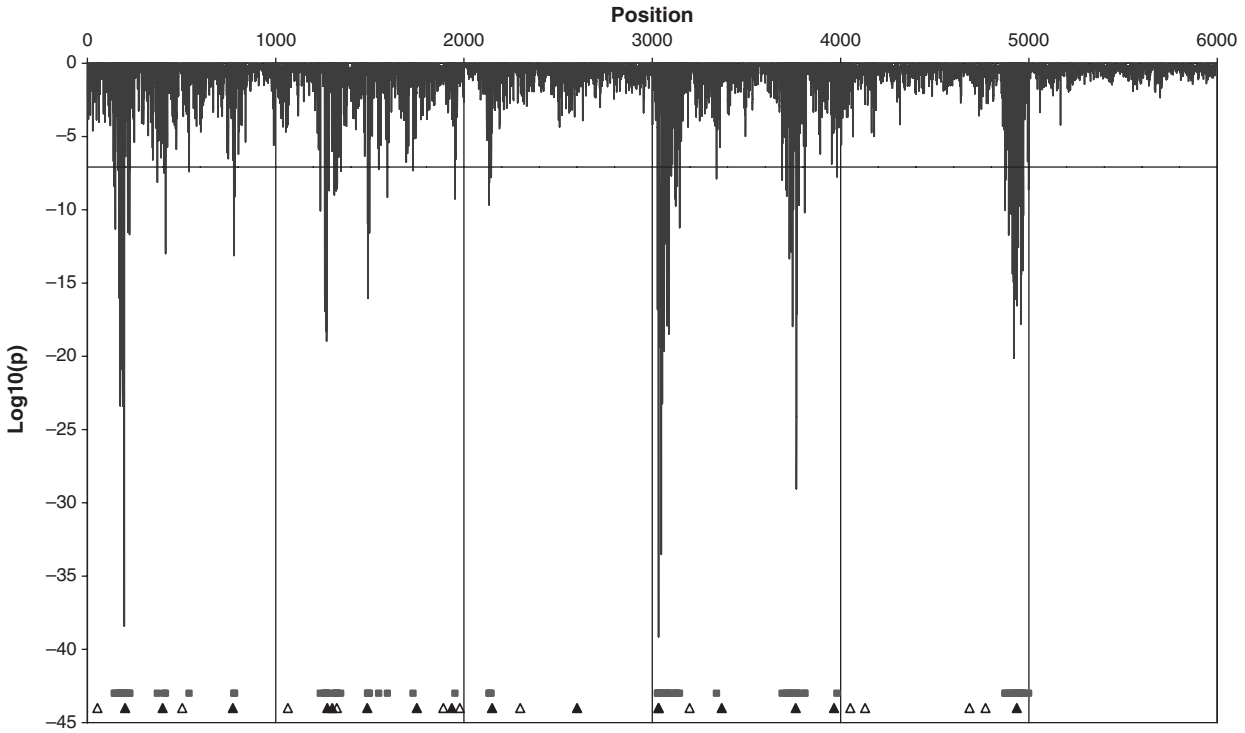

Figure 1 Log to the base 10 of error probabilities from the analysis of the simulated data set at the 6000 loci. Horizontal line indicates critical value for genome-wide p = 0.001 with Bonferroni correction; horizontal bars indicate genome segments with significant signals; filled triangles indicate position of simulated quantitative trait loci (QTL) explaining >1% of the genetic variance; and empty triangles indicate position of simulated QTL explaining <1% of the genetic variance.

be successfully mapped if there is a significant signal within ±1 cM of the true position. According to this criterion, 13 of the 15 QTL were successfully mapped. One QTL at position 3369 has a peak in the neighbourhood at position 3341, which is 2.8 cM away. The QTL at position 2600 is completely missed. No QTL was mapped on the last chromosome, which did not carry any true QTL. There are only spurious false-positives, and these effects may reflect some of the simulated QTL of minor effect (<1% of the additive genetic variance). This was tested by considering the 13 QTL explaining between 0.1% and 0.8% of the genetic variance (positions indicated in Figure 1). While at these positions the average error probability was significantly lower than in the regions carrying no QTL [average $\log_{10}(p) = -1.27$ at 13 minor QTL versus average $\log_{10}(p) = -0.41$ in the QTL-free regions], there was no association between QTL size and significance level within the group of minor alleles. In fact, minor alleles explaining a very low proportion of genetic variance tended to have the higher peaks in the test statistic. Therefore, we conclude that spurious signals are more likely because of random noise than caused by minor alleles.

For the 13 successfully mapped QTL, the mean absolute deviation of the true position and the adjacent highest peak of the test statistic was 0.89 cM. In general, the positional deviation decreased with increasing true QTL effects (results not shown).

We also tested the power of the approach to detect QTL when only a subset of n = 1000 or 3000 individuals from the complete simulated data set is genotyped. For this, we sampled at random one offspring from generations 52 to 54 and assigned this individual and its parents to be genotyped (if they were not already assigned to be genotyped in an earlier sample). This was repeated until a total of 1000 or 3000 animals to be genotyped were selected, respectively. In the analysis only these genotypes were used, but Mendelian sampling terms were calculated from the complete data set, reflecting that in real applications typically only a sub-sample of individuals having phenotypes and breeding values will be genotyped. Note that here incomplete triplets were included, because animals may have only one

genotyped parent. For each sample size, 1000 replicates were generated and analysed, using the same concept as before, but testing on the 1% genome-wide error level.

Figure 2 shows the proportion of the significant results plotted over the whole genome. In general, most peaks are associated with true QTL positions. We define the empirical power of the proportion of replicates in which the significant threshold was exceeded in a range of ±1 cM of the true position of a QTL. With n = 1000 individuals genotyped, only three QTL are significant in more than 20% of the replicates. With n = 3000 individuals genotyped, the empirical power is close to 100% for six QTL and most other QTL are detected quite frequently.

The power to detect a QTL depends on its effect. In Figure 3, the empirical power is plotted as a function of the QTL size (per cent of the genetic variance explained) for both sample sizes. With n = 3000, the power declines when a QTL explains only 5% or less of the genetic variance. With n = 1000, only QTL explaining more than 10% of the genetic variance can be detected. However, the power is not only a function of the QTL effect, but also of other systematic (like proximity to the next QTL) or unsystematic (like allele frequencies and LD pattern) effects, generating substantial variability especially concerning the power to map QTL of minor effects. Strikingly, the QTL at position 4935 explaining the highest proportion of the genetic variance was almost completely missed with sample size n = 1000, while for the other QTL with smaller effects the empirical power was up to 46.5% with the same sample size. A closer inspection of the QTL at position 4935 revealed that this locus is in remarkably low LD with the neighbouring SNP, the r^2 (Hill & Robertson 1968) with the two adjacent SNP being only 0.01 and 0.16, respectively. As a TDT only will find associations when both linkage and LD are present, a lack of LD may cause that some QTL are missed, which would have been mapped in a pure linkage study.

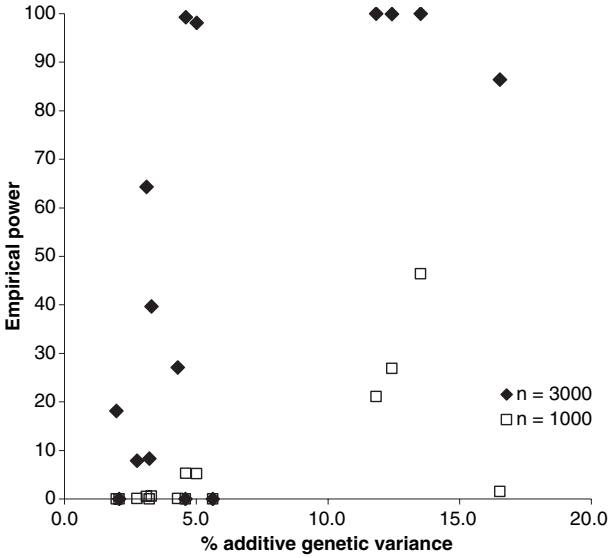

Figure 3 Empirical power to detect a quantitative trait locus (QTL) on the genome-wide 1% error level with sample size n = 1000 and n = 3000, respectively, as a function of the proportion of the additive genetic variance explained by the simulated QTL.

Discussion

The results presented here demonstrate that the suggested methodology is highly efficient in detecting QTL for a complex trait in large pedigrees with dense marker coverage. Through the combination of linkage and LD mapping, the positional resolution is

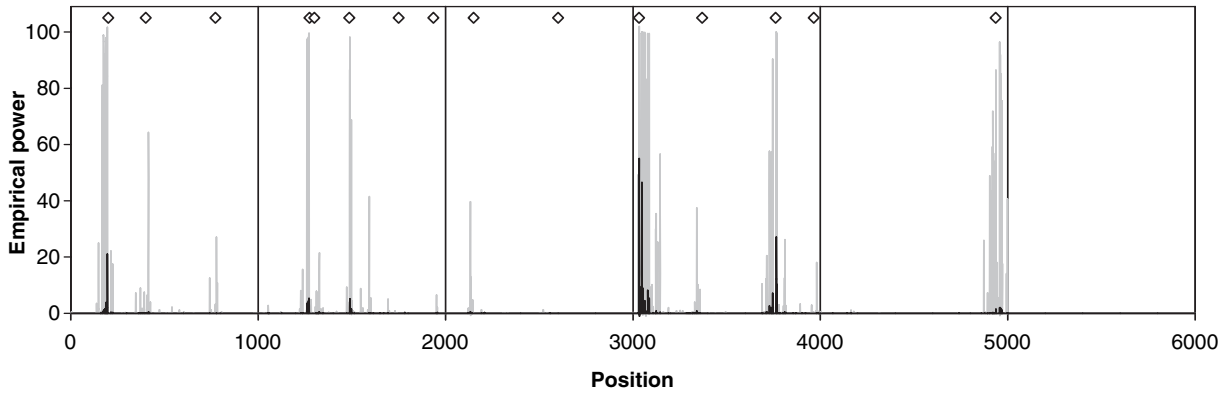

Figure 2 Empirical power to detect a quantitative trait locus (QTL) on the genome-wide 1% error level with sample size n = 1000 (black bars) and n = 3000 (grey bars), respectively (diamonds indicate the position of the simulated QTL explaining >1% of the genetic variance).

very good, leading to a mapping precision of less than 1 cM as demonstrated before.

The suggested approach is computationally efficient. The estimation of breeding values and Mendelian sampling terms in the first step is a routine procedure in most animal breeding populations, as comprehensive sets of high-quality estimates of breeding values are readily available for most production and functional traits. This is not restricted to normally distributed quantitative traits, but includes many other more complex traits like binary or multicategorical traits such as fertility or disease resistance (Gianola & Foulley 1983), longevity data using survival analysis (Ducrocq & Sölkner 1998) or analyses assuming other unconventional distributions (see, e.g. Gianola & Simianer 2006; Rodrigues-Motta *et al.* 2007). All these methods provide estimated breeding values on an underlying additive scale and thus the suggested method can be applied for QTL detection based on the estimated Mendelian sampling terms of the genotyped offspring.

In the second step, the data of the genotyped animals are processed linearly, i.e. one marker after another and one family (triplet or pair) after another. So the computing time is proportional to the number of genotyped offspring times the number of markers (times the number of traits). In the present study, the analysis of the full data set (4500 offspring with 6000 markers) took 67.9 s on a Digital Ultimate Workstation with a 533 MHz double processor.

Using the Mendelian sampling term in this procedure has many advantages. When considering nuclear families, the Mendelian sampling reflects the deviation of the offspring from the parent average and pedigree information is fully accounted for. Thus, complex calculations like setting up IBD or gametic relationship matrices, as in the mapping approach suggested by Meuwissen *et al.* (2002), are not required.

Morsci *et al.* (2006) suggested using the Mendelian sampling term in a study on the effects of polymorphisms in two genes on cattle chromosome BTA1 and various beef traits. The main intention of using estimated Mendelian sampling terms rather than estimated breeding values in that study was to avoid the impact of the time trend owing to selection (Bullock *et al.* 2000). However, the study was based on association only, and implications like the heterogeneous variances of estimated Mendelian sampling terms were not addressed.

Aulchenko *et al.* (2007) also suggested a two-step approach, called 'GRAMMAR', in which in the first step a mixed model with a polygenic effect is fitted to the data. In the second step, estimated residuals are analysed for association with the observed polymorphisms. They show with different simulated data sets that the power of this approach is similar to a one-step procedure testing association while fully accounting for the pedigree, called 'measured genotype' (MG; George & Elston 1987), but the computing time of their approach was only a fraction of the MG approach. A direct comparison of the empirical power observed in the study of Aulchenko *et al.* (2007) with the present study is difficult, because other data structures and significance levels (5% genome-wise in their study versus 1% genome-wise in our study) were used.

The most comparable result likely is the large-scale analysis in the complex Erasmus Rucphen family (ERF) pedigree (Pardo *et al.* 2005) comprising 1010 phenotype- and genotype-related individuals trying to map a QTL explaining 10% of the additive genetic variation with a total heritability of 0.3 (table 5 in Aulchenko *et al.* 2007). With classical TDT approaches (QTDT and FBAT), the genome-wise power ($\alpha = 0.05$, 100 000 tested SNP) was below 5% while it was 33% with GRAMMAR and 62% with MG, the latter requiring however 76 days of computing time for a single chromosome. In our study, the genome-wise power ($\alpha = 0.01$, 6000 tested SNP) to find a QTL explaining more than 10% of the additive genetic variation with 1000 genotyped individuals ranges from 1.5% to 46.4% and was on average 24% (see Figure 3). This is merely an indication that the suggested approach may have a comparable power as GRAMMAR and performs better than a phenotype-based TDT, but extended simulations need to be performed to verify this.

A direct comparison is possible with six different approaches based on linkage, LD or a combination of both applied to the same data set, as reviewed by Crooks *et al.* (2009). Of the 14 QTL considered in both studies as major ones (all but the QTL at position 1935 in our study), the alternative methods detected between 7 and 11 QTL, while our method detected 12 QTL. Note that the definitions of a successful detection and the applied significance thresholds differ between the studies. Interestingly, the QTL at position 2600 was missed by all approaches except a haplotype-based LD mapping approach. Only four QTL (at positions 200, 400, 1486 and 3032) were detected by all approaches.

The approach suggested in our study is nonparametric in the sense that the estimated contrast δ is not interpreted as an estimate of an underlying QTL effect like, e.g. the allele substitution effect.

Even under a simple genetic model like complete additivity δ reflects a complex mixture of effects.

While true Mendelian sampling terms of different individuals are uncorrelated, this is not true for estimated Mendelian sampling terms. Strictly speaking, the straight summation of weighted Mendelian sampling terms as suggested before violates the assumptions underlying the *t*-test, which therefore is only approximately valid. More reliable genome-wide critical values again can be derived by applying a permutation-based test (Churchill & Doerge 1994).

If for each Mendelian sampling term, an accuracy ρ_i is available and individuals differ substantially in this parameter, estimated accuracies can be used as weights in the test statistic in the form

$$\delta' = 1/n \frac{\sum_{i=1}^{n} \rho_i \lambda_i \hat{m}_i}{\sum_{i=1}^{n} \rho_i |\lambda_i|},$$

where n is the number of individuals contributing a Mendelian sampling term to either of the transmission classes $1\backslash 2$ or $2\backslash 1$, and the indicator variable λ_i is +1 or +2 if the Mendelian sampling term of individual i is in the transmission class $1\backslash 2$ or −1 or −2 if the Mendelian sampling term of individual i is in the transmission class $2\backslash 1$. The values for λ_i for the possible parent–child combinations are given in the last column of Table 1. For the test statistic δ', the assumption of a *t*-distribution is no longer valid, and appropriate threshold values should be derived empirically by applying a permutation test (Churchill & Doerge 1994).

The assumed data structure, comprising complete triplets for all families, is somewhat idealized. In many applications to farm animal populations, other family structures like extended paternal half-sib families, will prevail, and maternal genotypes will often be missing. The presented method can also be applied if only one parent is genotyped, but in this case only a fraction of the meioses are informative and power will be reduced. In many cases, it will be possible to derive missing genotypes from the observed ones, but, as noted by Curtis (1997), it is erroneous to treat these reconstructed families as if parental genotypes have been typed. Knapp (1999) has suggested an approach to correct for this bias and has shown that including reconstructed genotypes properly may increase the power to detect linkage.

Regarding the power, data structures with multiple offspring like extended half- or full-sib families may be advantageous compared to triplets with a single offspring only. Primarily, this is because of the amount of information relative to the number of genotyped animals. In the triplet case, one bit of information requires three genotyped individuals, so the ratio or informative Mendelian sampling terms versus genotypes is 1/3. In a half-sib structure where a sire has one offspring each from mating with n dams the ratio is $n/(2n + 1)$ and approaches ½ for large values of n. In a full-sib structure with pairs of parents with n offspring each, this ratio is $n/(n + 2)$ and approaches 1 for large values of n. While for large families the amount of information on linkage, i.e. within family segregation, increases, it is important that sufficient numbers of families remain to ensure that the between-family information reflecting the LD in the population is accounted for.

Owing to the computational simplicity and the speed of the suggested procedure, it is well suited to scan whole genomes with high-density SNP coverage in samples of substantial size and for a multiplicity of different traits. Applying permutation-based testing procedures is computationally demanding and will cause the loss of the claimed computational advantage of the suggested method. Therefore, we propose using tabulated values of the test statistic known to be only approximately valid for the genome screening step. Taking into account that the applied Bonferroni procedure is known to be systematically over-conservative (Benjamini & Yekutieli 2001), the risk of getting false-positives will be acceptable in most cases, nevertheless the results should be interpreted with caution and regions showing an indication of significant results should be further analysed with more rigorous statistical approaches. Also, the observed signals will only reflect a qualitative indication that the SNP at the respective position is linked to a pertinent genetic effect of unknown size and nature. The method will not provide any unbiased effect estimates. Further exploration of the underlying genetic mechanisms in the small sub-sample of significant SNP positions or regions then can be performed via model-based parametric approaches, such as the one suggested by Meuwissen *et al.* (2002) in combination with reliable empirical testing procedures.

Acknowledgements

The authors thank two anonymous referees for constructive suggestions on an earlier version of the manuscript. This study was part of the project FUGATO-Plus GenoTrack and was financially supported by the German Ministry of Education and

Research, BMBF, the Förderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven.

References

Abecasis G.R., Cardon L.R., Cookson W.O. (2000) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.*, **66**, 279–292.

Andersson L. (2001) Genetic dissection of phenotypic diversity in farm animals. *Nat. Rev. Genet.*, **2**, 130–138.

Aulchenko Y.S., de Koning D.J., Haley C. (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577–585.

Avendano S., Woolliams J.A., Villanueva B. (2005) Prediction of accuracy of estimated Mendelian sampling terms. *J. Anim. Breed. Genet.*, **122**, 302–308.

Benjamini Y., Yekutieli D. (2001) The control of false discovery rate under dependency. *Ann. Statist.*, **29**, 1165–1188.

Bink M.C.A.M., te Pas M.F.W., Harders F.L., Janss L.L.G. (2000) A transmission/disequilibrium test approach to screen for quantitative trait loci in two selected lines of Large White pigs. *Genet. Res. Camb.*, **75**, 115–121.

Bullock K.D., Thrift F.A., Aaron D.K., Bertrand J.K. (2000) Relationship between a bull's parental genetic merit difference and subsequent progeny trait variability in Angus, Gelbvieh, and Limousin cattle. *J. Anim. Sci.*, **78**, 2540–2546.

Churchill G.A., Doerge R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Clayton D. (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am. J. Hum. Genet.*, **65**, 1170–1177.

Crooks L., Sahana G., de Koning D.J., Lund M.S., Carlborg Ö. (2009) Comparison of analyses of the QTLMAS XII common dataset. II: genome-wide association and fine mapping. *BMC Proc.*, **3**, S2.

Curtis D. (1997) Use of siblings as controls in case-control association studies. *Ann. Hum. Genet.*, **61**, 319–333.

Ducrocq V., Sölkner J. (1998) The Survival Kit-V3.0: a package for large analyses of survival data. In: International Committee for World Congresses on Genetics Applied to Livestock Production (ed.) Sixth World Congress on Genetics Applied to Livestock Production, Armidale, Australia. University of New England, Armidale, New England, **27**, pp. 447–448.

Ewens W.J., Li M., Spielman R.S. (2008) A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. *PLoS Genet.*, **4**, e1000180.

Farnir F., Coppietersen W., Arranz J.-J., Berzi P., Cambiano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., Georges M. (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.*, **10**, 220–227.

Freedman M.L., Reich D., Penney K.L., McDonald G.J., Mignault A.A., Patterson N., Gabriel S.B., Topol E.J., Smoller J.W., Pato C.N., Pato M.T., Petryshen T.L., Kolonel L.N., Lander E.S., Sklar P., Henderson B., Hirschhorn J.N., Altshuler D. (2004) Assessing the impact of population stratification on genetic association studies. *Nat. Genet.*, **36**, 388–393.

George V.T., Elston R.C. (1987) Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet. Epidemiol.*, **4**, 193–201.

Gianola D., Foulley J.L. (1983) Sire evaluation for ordered categorical data with a threshold model. *Genet. Sel. Evol.*, **15**, 201–224.

Gianola D., Simianer H. (2006) A Thurstonian model for quantitative genetic analysis of ranks: a Bayesian approach. *Genetics*, **174**, 1613–1624.

Groeneveld E. (1998) VCE User's Manual, Version 4.2.5. Institute of Animal Breeding and Animal Behavior, Federal Research Institute for Agriculture, Mariensee, Germany.

Hayes B.J., Visscher P.M., McPartlan H.C., Goddard M.E. (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.*, **13**, 635–643.

Henderson C.R. (1973) Sire evaluation and genetic trends. In: Proceedings of Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush, Champaign, Illinois, USA, pp. 10–41; Copyright American Society of Animal Science (ASAS).

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

Henderson C.R. (1984) Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Canada.

Hernández-Sánchez J., Visscher P., Plastow G., Haley C. (2003) Candidate gene analysis for quantitative traits using the transmission disequilibrium test: the example of the melanocortin 4 receptor in pigs. *Genetics*, **164**, 637–644.

Hill W.G., Robertson A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38**, 226–231.

Jiang Z., Michal J.J., Tobey D.J., Daniels T.F., Rule D.C., MacNeil M.D. (2008) Significant associations of stearoyl-CoA desaturase (SCD1) gene with fat deposition and composition in skeletal muscle. *Int. J. Biol. Sci.*, **4**, 345–351.

Knapp M. (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am. J. Hum. Genet.*, **64**, 861–870.

Kunej T., Wang Z., Michal J.J., Daniels T.F., Magnuson N.S., Jiang Z. (2007) Functional UQCRC1 polymorphisms affect promoter activity and body lipid accumulation. *Obesity*, **15**, 2896–2901.

Laird N.M., Lange C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.*, **7**, 385–394.

Meuwissen T.H.E., Karlsen A., Lien S., Olsaker I., Goddard M.E. (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics*, **161**, 373–379.

Morsci N.S., Schnabel R.D., Taylor J.F. (2006) Association analysis of adiponectin and somatostatin polymorphisms on BTA1 with growth and carcass traits in Angus cattle. *Anim. Genet.*, **37**, 554–562.

Pardo L.M., MacKay I., Oostra B., van Duijn C.M., Aulchenko Y.S. (2005) The effect of genetic drift in a young genetically isolated population. *Ann. Hum. Genet.*, **69**, 288–295.

Patterson H.D., Thompson R. (1971) Recovery of interblock information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Rabinowitz D., Laird N. (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.*, **50**, 211–223.

Rodrigues-Motta M., Gianola D., Heringstad B., Rosa G.J., Chang Y.M. (2007) A zero-inflated Poisson model for genetic analysis of the number of mastitis cases in Norwegian Red cows. *J. Dairy Sci.*, **90**, 5306–5315.

Snedecor G.W., Cochran W.G. (1956) Statistical Methods Applied to Experiments in Agriculture and Biology, 5th edn. Iowa State University Press, Ames, Iowa, USA.

Spielman R.S., McGinnis R.E., Ewens W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **52**, 506–516.

Tier B., Meyer K. (2004) Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *J. Anim. Breed. Genet.*, **121**, 77–89.

Wimmers K., Schellander K., Ponsuksili S. (2004) BF, HP, DQB and DRB are associated with haemolytic complement activity, acute phase protein reaction and antibody response in the pig. *Vet. Immunol. Immunopathol.*, **99**, 215–228.