Aus dem Forschungsbereich Genetik und Biometrie

des Leibniz-Instituts für Nutztierbiologie, FBN Dummerstorf

MODIFIED RANDOM MODELS FOR QTL DISCOVERY IN F_2 POPULATIONS

Dissertation

zur Erlangung des Doktorgrades

der Agrar- und Ernährungswissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Dipl.-Biomath. (FH) Daisy Zimmer

aus Räckelwitz

Dekanin:	Prof. Dr. Karin Schwarz
1. Berichterstatter:	Prof. Dr. Norbert Reinsch
2. Berichterstatter:	Prof. Dr. Georg Thaller

Tag der mündlichen Prüfung: 14.07.2011

Gedruckt mit Genehmigung des Dekans der Agrar- und Ernährungswissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel

Contents

GENERAL INTRODUCTION

Quantitative trait and QTL: Quantitative traits, like blood pressure, plant height or milk yield, are continuously distributed and typically affected by multiple loci as well as the environment to varying proportions. The identification of regions in the genome which have significant influence on complex traits is important in animal science, plan breeding, biomedicine and human genetics. These regions are called quantitative trait loci (QTL; Geldermann 1975). In quantitative genetics an infinitesimal model was suggested by Fisher (1918), where an infinite number of unlinked QTL each with an infinitely small effect is assumed. Most quantitative (polygenic) traits are influenced by few QTL which have large or moderate effects and many QTL which have small effects (e. g. Shrimpton and Robertson 1988; Hayes and Goddard 2001).

Multiple and possibly linked QTL may act in an additive and nonadditive manner (dominance, epistasis), whereby the exact number of segregating QTL is unknown. QTL detection and estimation of genetic parameters are important in order to detect individual loci responsible for quantitative genetic variation and for differences between the phenotypes of diverging strains. The nature of quantitative genetic variation should be understood to utilize these variations in selection programs in plant and livestock populations.

For QTL mapping in segregating populations it is beneficial to use orthogonal models, because estimated genetic effects and genetic variance components are consistent (Zeng *et al.* 2005). In an orthogonal model the definition of the genetic effects are independent for all loci, especially in populations with Hardy-Weinberg and linkage equilibrium. Tests for different genetic effects or variance components are independent (Zeng *et al.* 2005).

Half of the difference between the homozygous genotypic values is the additive genetic effect (Falconer and Mackay 1996). Nonadditive genetic effects are interactions within a locus (dominance) or between two or more loci (epistasis), which may contribute markedly to the genetic variation in quantitative traits (e. g. Jannink and Jansen 2001; Carlborg and Haley 2004). The dominance effect is defined as the heterozygous genotypic value minus the mean of both homozygous genotypic values (Falconer and Mackay 1996).

Using the orthogonal partition of the genotypic effects as suggested by Cockerham (1954) the coefficients of the additive genetic effects are 1 for QQ, 0 for Qq and -1 for

qq, where the QTL alleles are denoted by Q and q. The respective coefficients of the dominance effects are $-\frac{1}{2}$, $\frac{1}{2}$ and $-\frac{1}{2}$. Coefficients of pairwise epistatic effects (between two loci) are determined as product of the coefficients of the single effects (see KAO and ZENG 2002; ZENG *et al.* 2005).

For quantitative traits, nonadditive genetic effects are an important source of genetic variation (e. g. BROCKMANN *et al.* 2000; CARLBORG *et al.* 2004, 2005; PALUCCI *et al.* 2007). Considering nonadditive genetic effects, the precision of estimated QTL positions and their effects may be improved as well as the understanding of complex genetic traits (e. g. KAO 2000; CARLBORG and HALEY 2004; CARLBORG *et al.* 2005; MEUWISSEN and GODDARD 2004; HU *et al.* 2011). Some QTL can only be identified if their interactions are considered (e. g. JANNINK and JANSEN 2001; CORDELL 2002; CARLBORG and HALEY 2004).

Using marker information: Associations between polymorphic genetic markers and unobservable QTL affecting a quantitative trait are searched and used in QTL mapping in different population structures (e. g. GELDERMANN 1975; FLINT and MOTT 2001). Polymorphic genetic markers are markers which have at least two segregating alleles, where e. g. the minimum allele frequency is $> 2.5\,\%$ (HAYES *et al.* 2009). The availability of molecular genetic markers, like microsatellites, RFLP and SNPs, facilitates the construction of genetic linkage maps, i. e. positions of markers in the genome are known, in many species. A linkage disequilibrium (LD) between marker alleles and underlying QTL alleles of a quantitative trait is necessary in QTL analyses. LD can be artificially generated by crossbreeding and a maximum LD appears in F_2 populations derived from inbred lines, where only two QTL alleles are segregating. Using a limited number of genetic markers in linkage analyses (linkage mapping), the association between markers and QTL will, in general, exist only within families (family-specific LD). Due to recombination these LD will vanish after a number of generations. Association studies and linkage disequilibrium mapping requires a population-wide LD, therefore genetic markers and QTL must be closely linked. If the number of available genetic markers is dense (sufficiently high) and covering the whole genome, a population-wide LD is ensured. Mostly SNPs are located in smaller distance than 1 cM intervals (e. g. SCHAEFFER 2006) which are available due to advanced DNA chip technology. Costs of genotyping have decreased in the last years due to high-throughput genotyping technology. Furthermore, for more and more species chips with numerous SNPs are available (SCHAEFFER 2006).

Inbred line-derived populations are commonly used in plant breeding and laboratory

animals, as frequently described in the literature. A maximum LD appears in F_2 populations derived from inbred lines. In such populations the (joint) conditional QTL genotype probabilities can easily be inferred from flanking marker information, because markers are completely informative.

Application of QTL mapping results: Identified QTL which have a significant influence on one or more quantitative traits and estimated genetic effects are used e. g. in the marker assisted selection (MAS) in animal and plant breeding (e. g. SCHROOTEN *et al.* 2000; DEKKERS 2004). MAS is an indirect selection method, where individuals with advantageous traits (desirable gene variants) are selected based on markers which are linked to QTL of interest. The selection intensity (e. g. WHITTAKER *et al.* 2000) and the frequency of desired QTL alleles in populations increases due to MAS. The focus of MAS is on using QTL of large effects. MEUWISSEN *et al.* (2001) pointed out that the benefit from MAS depends on the heritability in the broad sense, i. e. the proportion of the genetic variance explained by QTL. Applications and limitations of MAS are further discussed by, e. g., DEKKERS (2004). A form of MAS is called genomic (assisted) selection (GS), which was suggested by MEUWISSEN *et al.* (2001). Genetic markers covering the whole genome, mostly SNPs in LD with the QTL, are used for prediction of total genetic values for quantitative traits simultaneously. The focus is on prediction of genetic differences between individuals and therefore multiple significance tests are unnecessary. In this way markers potentially explain all the genetic variance, i. e. parts of the genome which have no evidence for the presence of single QTL with large effects explain still something of the genetic variance (CALUS *et al.* 2008). Application and requirements for maximum benefits using GS are briefly given by GODDARD and HAYES (2007) and HAYES *et al.* (2009).

MAS and GS may lead to a substantial increase of genetic gain per year compared to using phenotypic data and kinship information alone. Predictions of breeding values may also become more accurate, especially for traits with low heritability (e. g. CALUS *et al.* 2008). The generation interval is shorted, because breeding values of sufficient accuracy are available early in the life of individuals without own phenotypes or progeny testing (MEUWISSEN *et al.* 2001; SCHAEFFER 2006).

Coarse QTL mapping is often aimed to encounter causal genes and mutations underlying the QTL effects and is followed by fine-scale mapping or high-resolution mapping, i. e. denser marker maps enclosing narrow regions around causal genes which are e. g. responsible for diseases (e. g. FLINT and MOTT 2001). In this way functional candidate genes suspected of being involved in the expression of a trait are identified (e. g.

Flint and Mott 2001). Positional cloning, i. e. identifying causal mutation in the genome became possible using results of QTL analyses (e. g. Grisart *et al.* 2002).

Statistical methods: Numerous statistical methods have been suggested to detect QTL and estimate their effects. Unbiasedness and accuracy for detecting positions and effects of the QTL are the most important issues. Among mapping procedures there are non-parametric methods (e. g. Kruglyak and Lander 1995), least-squares based methods (e. g. Haley and Knott 1992; Martínez and Curnow 1992; Haley *et al.* 1994; Carlborg *et al.* 2000; Knott and Haley 2000), maximum likelihood (ML) based methods (e. g. Lander and Botstein 1989; Kao and Zeng 1997; Zeng *et al.* 1999) and Bayesian methods (e. g. Thomas and Cortessis 1992; Hoeschele and VanRaden 1993a,b). QTL effects can be considered as random or fixed effects in a linear model. Accordingly, three categories of linear models can be differentiated: fixed, random and mixed models.

A ML based method of interval mapping (IM) was suggested by Lander and Botstein (1989) using flanking markers. IM is computationally demanding, especially if the number of estimated parameters increases. At the same time, Haley and Knott (1992) and Martínez and Curnow (1992) presented an approximation of IM, a simple regression method using flanking markers. This method is a least-squares method based on multiple regression of the quantitative trait on the conditional expected genotypic values and is called regression interval mapping (RIM). Extensions of the regression methods for multiple traits (Knott and Haley 2000), for multiple markers (Knott and Haley 1998) and for utilization of estimating equations based on both means and variances (Feenstra *et al.* 2006) were suggested. As Haley and Knott (1992) mentioned, RIM is a ML based method, when residuals are independently and normally distributed. A further ML procedure which considers the QTL as fixed effects is the multiple interval mapping (MIM; Kao and Zeng 1997; Zeng *et al.* 1999). Both methods, MIM and RIM, were proposed for mapping multiple QTL with additive genetic, dominance and epistatic effects using multiple marker intervals simultaneously. Composite interval mapping combines interval mapping with multiple regression analysis (Jansen 1993; Zeng 1993, 1994). This method considers, besides the single putative QTL, other markers as covariates to control variation due to additional QTL.

Methods which consider QTL effects as random in a linear mixed model are called variance component methods (VCM). The basic idea of the VCM is, that the variation of phenotypes is small if a pair of individuals share alleles which are identical by descent

(IBD). Instead of average effects of alternative genotypes, the VCM directly estimates and test the QTL variance components (Carlborg and Haley 2004). The VCM was introduced by Haseman and Elston (1972) for detection of QTL in multiple human families, usually of small size. Numerous QTL analyses were performed with the VCM by several authors (e. g. Fernando and Grossman 1989; Cantet and Smith 1991; Goddard 1992; van Arendonk *et al.* 1994; Grignola *et al.* 1996a,b; Xu 1996a; Xie *et al.* 1998; Lee and Van der Werf 2007; Zimmer *et al.* 2011). Different authors gave rules for setting up required QTL relationship matrices from marker information (Wang *et al.* 1995; Grignola *et al.* 1996a; Abdel-Azim and Freeman 2001; Pong-Wong *et al.* 2001). For noninbred populations Liu *et al.* (2002) suggested marker-based relationship matrices of additive and nonadditive genetic effects.

Firstly Quaas and Pollak (1980) used the concept of an equivalent reduced model to decrease the number of equations that needed to be solved. IBD matrices with reduced rank (only eigenvalues greater a given threshold are considered) were suggested for faster computing (Rönnegård *et al.* 2007). For an F_2 cross they showed that the results of the conventional and fast method were very similar.

Combining multiple line crosses for QTL mapping in experimental populations using the VCM was suggested by Xie *et al.* (1998). Their approach calculated the additive genetic and dominance relationship matrices from conditional QTL genotype probabilities and elementary covariance matrices. Since genetic effects for each individual are estimated, this approach is called individual random model (IRM). Crepieux *et al.* (2004) extended the VCM from Xie *et al.* (1998) for QTL mapping to any type of multicross populations obtained from inbred parents. They estimated the coefficients of coancestry between parents and used these coefficients to build the IBD matrices. Recently, Li and Cui (2009) developed a general VCM to map imprinted QTL underlying complex traits by combining different line crosses and backcrosses derived from inbred lines.

It is advisable to fit multiple linked QTL simultaneously in the genome using flanking markers to increase the accuracy and reliability. Large populations are necessary to identify multiple QTL and a multi-dimensional search should be done when epistasis is considered (Carlborg and Haley 2004), because the existence of multiple QTL in a linkage group can distort the identification of QTL if only a single QTL is modeled. There are more technical challenges and demands with the data caused by epistatic effects than individual QTL mapping (e.ġ. Jannink and Jansen 2001; Carlborg and Haley 2004). Step-wise selection by added or deleted QTL one by one in the model, as suggested e. g. by Kao *et al.* (1999), sometimes does not allow to detect

linked QTL. A multi-dimensional quest probably prevent such so-called "ghost" QTL phenomenon (a QTL was identified incorrectly between two true QTL; e. g. HALEY and KNOTT 1992; MARTÍNEZ and CURNOW 1992), but they are computationally demanding, because the number of combinations of QTL positions increases quickly if the number of considered QTL and their effects increases.

Advantages of random models: The VCM require no information about linkage phases and number of QTL alleles in the populations (XU and ATCHLEY 1995; XU 1996a). In some situations an "infinite" number of QTL alleles may be present, i. e. the number is mostly unknown and the degree of allele shared among siblings is assessed. The treatment of QTL effects as random effects causes shrinkage of the estimated QTL effects toward a prior mean (usually towards zero) if the proportion of the phenotypic variance explained by QTL is small or with a low number of individuals, e. g. population with small families (GRIGNOLA *et al.* 1996a).

The experimental power of a single-line cross depends on the genetic construction of the two parental lines (XIE *et al.* 1998). Such QTL mapping results may provide limited information to understand the architecture of complex traits, because the chosen inbred lines may not represent the population structure completely. Furthermore, statistical inference of the estimated QTL variance is not simply expandable to other crosses (CREPIEUX *et al.* 2004; LI and CUI 2009), because the genetic variances are generally determined for underlying populations. The non-detection of present QTL due to fixation of the same alleles in parental lines, the so-called "genetic drift error", is a type II error and can be avoided by using multiple line crosses (XU 1996b, 1998).

The VCM is advantageous if a large number of small families or complex family structures occur (XU 1998). Using multiple parental lines and therefore combined data of multiple line crosses and multiple families, the interference on QTL variances have a general character (LI and CUI 2009) and can be adopted for various crosses (CREPIEUX *et al.* 2004). Such approaches should be more powerful in QTL mapping (XIE *et al.* 1998) with estimated genetic variances consistent across different genetic backgrounds.

Thesis outline: The VCM for multiple QTL mapping in F_2 crossbred populations derived from inbred lines are considered. For marked chromosome segments such methods include random genetic effects of different kinds – additive genetic, dominance – for each individual and their corresponding marker derived genetic covariances. Extensions of the IRM as suggested by XIE *et al.* (1998) to pairwise epistatic effects are presented. The IRM is computationally demanding, especially if the number of F_2

individuals or the number of QTL is increased. Therefore, main emphasis in **Chapter One** is on models where genetic covariances are approximated by replacing individual genetic effects by average genetic effects for each marker class, the so-called reduced random model (RRM). The result is a substantial decrease of the dimensions of covariance matrices of genetic effects and, consequently, a remarkable gain in computing speed in estimating the variance components and evaluating the residual log-likelihood. It could be shown that the RRM is asymptotically equal to the IRM. The RRM provides a general framework for mapping multiple linked QTL from inbred line-derived F_2 populations, where additive, dominance and pairwise epistatic effects are taken into account. Considering limited numbers of simulations indicated that the RRM approximates the IRM very well in terms of experimental power and accuracy of estimated QTL positions. The RRM was also compared to two fixed models (RIM, MIM) and it was shown that the RRM was again competitive. Single and multiple families were simulated.

More comprehensive simulations were done in **Chapter Two**, where two linked QTL with additive genetic effects for each QTL and an additive-by-additive genetic effect were considered simultaneously. A single F_2 family was assumed for the purpose of comparison with the fixed models (RIM, MIM). The competitiveness of RRM with IRM, RIM and MIM was shown in terms of observed power to detect both simulated QTL, the accuracy of the estimated QTL positions and their effects. The application of the RRM to the analysis of multiple families was discussed.

Chapter Three deals with elementary covariance matrices which consider the perfect negative correlation between both QTL alleles with allele frequencies of a half. Using the restricted log-likelihood function, it could be shown that the additive and additive-by-additive genetic relationship matrices calculated with the elementary covariance matrices suggested in Chapter One lead to the same restricted log-likelihood and parameter estimates (positions and effects of QTL) applying the suggested elementary covariance matrices from this Chapter, but the standard errors of the estimated genetic effects are reduced with the new covariance matrices. The suggested relationship matrices require less computing time compared to the matrices from Chapter One, especially for the IRM. Furthermore, for a single additive genetic effect the equivalence to a random regression approach was shown, in case of identity of marker and QTL.

LITERATURE

Abdel-Azim, G. and A. E. Freeman, 2001 A rapid method for computing the inverse of the gametic covariance matrix between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **33**: 153–173.

Brockmann, G. A., J. Kratzsch, C. S. Haley, U. Renne, M. Schwerin, and S. Karle, 2000 Single QTL effects, epistasis, and pleiotropy account for two-thirds of the phenotypic F_2 variance of growth and obesity in DU6i x DBA/2 mice. Genome Res. **10**: 1941–1957.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics **178**: 553–561.

Cantet, R. J. C. and C. Smith, 1991 Reduced animal model for marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **23**: 221–233.

Carlborg, O., L. Andersson, and B. Kinghorn, 2000 The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics **155**: 2003–2010.

Carlborg, O., G. A. Brockmann, and C. S. Haley, 2005 Simultaneous mapping of epistatic QTL in DU6i x DBA/2 mice. Mamm. Genome **16**: 481–494.

Carlborg, Ö. and C. S. Haley, 2004 Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. **5**: 618–625.

Carlborg, O., P. M. Hocking, D. W. Burt, and C. S. Haley, 2004 Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth. Genet. Res. **83**: 197–209.

Cockerham, C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics **39**: 859–882.

Cordell, H. J., 2002 Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. **11**: 2463–2468.

Crepieux, S., C. Lebreton, B. Servin, and G. Charmet, 2004 Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. Genetics **168**: 1737–1749.

Dekkers, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci. **82 E-Suppl**: E313–E328.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman, Harlow, UK.

Feenstra, B., I. M. Skovgaard, and K. W. Broman, 2006 Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. Genetics **173**: 2269–2282.

Fernando, R. and M. Grossman, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21**: 467–477.

Fisher, R. A., 1918 The correlation between relatives on the supposition of mendelian inheritance. Trans. R. Soc. Edin. **52**: 399—433.

Flint, J. and R. Mott, 2001 Finding the molecular basis of quantitative traits: successes and pitfalls. Nat. Rev. Genet. **2**: 437–445.

Geldermann, H., 1975 Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. Theor. Appl. Genet. **46**: 319–330.

Goddard, M. E., 1992 A mixed model for analyses of data on multiple genetic markers. Theor. Appl. Genet. **83**: 878–886.

Goddard, M. E. and B. J. Hayes, 2007 Genomic selection. J. Anim. Breed. Genet. **124**: 323–330.

Grignola, F. E., I. Hoeschele, and B. Tier, 1996a Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. Methodology. Genet. Sel. Evol. **28**: 479–490.

Grignola, F. E., I. Hoeschele, Q. Zhang, and G. Thaller, 1996b Mapping quantitative trait loci in outcross populations via residual maximum likelihood. II. A simulation study. Genet. Sel. Evol. **28**: 491–504.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell, 2002 Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. **12**: 222–231.

Haley, C. S. and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–324.

Haley, C. S., S. A. Knott, and J. M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136**: 1195–1207.

Haseman, J. K. and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2**: 3–19.

Hayes, B. and M. E. Goddard, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. **33**: 209–229.

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

Hoeschele, I. and P. M. VanRaden, 1993a Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theor. Appl. Genet. **85**: 953–960.

Hoeschele, I. and P. M. VanRaden, 1993b Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor. Appl. Genet. **85**: 946–952.

Hu, Z., Y. Li, X. Song, Y. Han, X. Cai, S. Xu, and W. Li, 2011 Genomic value prediction for quantitative traits under the epistatic model. BMC Genet. **12**: 15.

Jannink, J.-L. and R. Jansen, 2001 Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **157**: 445–454.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135**: 205–211.

Kao, C.-H., 2000 On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics **156**: 855–865.

Kao, C.-H. and Z.-B. Zeng, 1997 General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53**: 653–665.

Kao, C.-H. and Z.-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160**: 1243–1261.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

Knott, S. A. and C. S. Haley, 1998 Simple multiple-marker sib-pair analysis for mapping quantitative trait loci. Heritity **81**: 48–54.

Knott, S. A. and C. S. Haley, 2000 Multitrait least squares for quantitative trait loci detection. Genetics **156**: 899–911.

Kruglyak, L. and E. S. Lander, 1995 A nonparametric approach for mapping quantitative trait loci. Genetics **139**: 1421–1428.

Lander, E. S. and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.

Lee, S. H. and J. H. J. Van der Werf, 2007 Fine mapping of multiple interacting quantitative trait loci using combined linkage disequilibrium and linkage information. Univ. Sci. B **8(11)**: 787–791.

Li, G. and Y. Cui, 2009 A statistical variance components framework for mapping imprinted quantitative trait locus in experimental crosses. J. Probab. Stat. **2009**: 1–27.

Liu, Y., G. B. Jansen, and C. Y. Lin, 2002 The covariance between relatives conditional on genetic markers. Genet. Sel. Evol. **34**: 657–678.

Martínez, O. and R. N. Curnow, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85**: 480–488.

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

Meuwissen, T. H. E. and M. E. Goddard, 2004 Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. **36**: 261–279.

Palucci, V., L. R. Schaeffer, F. Miglior, and V. Osborne, 2007 Non-additive genetic effects for fertility traits in canadian holstein cattle. Genet. Sel. Evol. **39**: 181–193.

Pong-Wong, R., A. W. George, J. A. Woolliams, and C. S. Haley, 2001 A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol. **33**: 453–471.

Quaas, R. L. and E. J. Pollak, 1980 Mixed model methodology for farm and ranch beef cattle testing programs. J. Anim. Sci. **51**: 1277—1287.

Rönnegård, L., K. Mischenko, S. Holmgren, and Ö. Carlborg, 2007 Increasing the efficiency of variance component quantitative trait loci analysis by using reduced-rank identity-by-descent matrices. Genetics **176**: 1935–1938.

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

Schrooten, C., H. Bovenhuis, W. Coppieters, and J. A. V. Arendonk, 2000 Whole genome scan to detect quantitative trait loci for conformation and functional traits in dairy cattle. J. Dairy Sci. **83**: 795–806.

Shrimpton, A. E. and A. Robertson, 1988 The isolation of polygenic factors controlling bristle score in drosophila melanogaster. II. Distribution of third chromosome bristle effects within chromosome sections. Genetics **118**: 445–459.

Thomas, D. C. and V. Cortessis, 1992 A gibbs sampling approach to linkage analysis. Hum. Hered. **42**: 63–76.

van Arendonk, J. A. M., B. Tier, and B. P. Kinghorn, 1994 Use of multiple genetic markers in prediction of breeding values. Genetics **137**: 319–329.

Wang, T., R. L. Fernando, S. van der Beek, M. Grossman, and J. A. M. van Arendonk, 1995 Covariance between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **27**: 251–274.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. **75**: 249–252.

Xie, C., D. D. Gessler, and S. Xu, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics **149**: 1139–1146.

Xu, S., 1996a Computation of the full likelihood function for estimating variance at a quantitative trait locus. Genetics **144**: 1951–1960.

Xu, S., 1996b Mapping quantitative trait loci using four-way crosses. Genet. Res. **68**: 175–181.

Xu, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics **148**: 517–524.

Xu, S. and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141**: 1189–1197.

Zeng, Z.-B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90**: 10972–10976.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136**: 1457–1468.

Zeng, Z.-B., C.-H. Kao, and C. J. Basten, 1999 Estimating the genetic architecture of quantitative traits. Genet. Res. **74**: 279–289.

Zeng, Z.-B., T. Wang, and W. Zou, 2005 Modeling quantitative trait loci and interpretation of models. Genetics **169**: 1711–1725.

Zimmer, D., M. Mayer, and N. Reinsch, 2011 Complex genetic effects in quantitative trait locus identification: A computationally tractable random model for use in F_2 populations. Genetics **187**: 261–270.

CHAPTER ONE

COMPLEX GENETIC EFFECTS IN QUANTITATIVE TRAIT LOCUS IDENTIFICATION: A COMPUTATIONALLY TRACTABLE RANDOM MODEL FOR USE IN F_2 POPULATIONS

Daisy Zimmer, Manfred Mayer, Norbert Reinsch

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

Published in Genetics 187: 261–270 (January 2011)

ABSTRACT

Methodology for mapping quantitative trait loci (QTL) has focused primarily on treating the QTL as a fixed effect. These methods differ from the usual models of genetic variation that treat genetic effects as random. Computationally expensive methods that allow QTL to be treated as random have been explicitly developed for additive genetic and dominance effects. By extending these methods with a variance component method (VCM), multiple QTL can be mapped. We focused on an F_2 crossbred population derived from inbred lines and estimated effects for each individual and their corresponding marker-derived genetic covariances. We present extensions to pairwise epistatic effects, which are computationally intensive because a great many individual effects must be estimated. But by replacing individual genetic effects with average genetic effects for each marker class, genetic covariances are approximated. This substantially reduces the computational burden by reducing the dimensions of covariance matrices of genetic effects, resulting in a remarkable gain in the speed of estimating the variance components and evaluating the residual log-likelihood. Preliminary results from simulations indicate competitiveness of the reduced model with multiple interval mapping, regression interval mapping, and VCM with individual genetic effects in its estimated QTL positions and experimental power.

INTRODUCTION

Mapping procedures often treat the effects of quantitative trait loci (QTL) as fixed, in particular the maximum likelihood based method of interval mapping (IM) of Lander and Botstein (1989) and the least-squares regression interval mapping (RIM) of Haley and Knott (1992) and Martínez and Curnow (1992).

Single-QTL approaches with fixed effects were later extended to multiple QTL to avoid the so-called "ghost-QTL" phenomenon (e. g. Haley and Knott 1992) and to improve the power to detect linked QTL in repulsion (e. g. Kao 2000) as well as epistatic QTL (e. g. Jannink and Jansen 2001; Carlborg and Haley 2004). The multiple interval mapping (MIM) approach of Kao and Zeng (1997) and Kao *et al.* (1999) as an extension of IM considers fixed additive genetic, dominance, and epistatic QTL effects as parts of the likelihood function for a mixture model in experimental populations. Both MIM and RIM are known to be powerful and well suited to identifying multiple, possibly interacting QTL in mapping experiments. However, the accuracy of the estimates of the positions and effects of the QTL from RIM is less compared with MIM in some situations [e. g. QTL in repulsion; Kao (2000); Mayer *et al.* (2004); Mayer (2005)].

Considering QTL effects as random in a linear mixed model (LMM) leads to the variance component method (VCM) for QTL mapping. This is often applied in scenarios with a large number of small families as is frequently found in humans (e. g. Haseman and Elston 1972; Xu and Atchley 1995) or in livestock (e. g. Grignola *et al.* 1996), where a mixture of families with parents of different QTL genotypes is expected to occur. Experiments with multiple line crosses, e. g. F_2, are often advocated because of their potential to avoid non-detection of QTL by representing genetic variability of a population by only a few lines – the so-called "genetic drift error" (Xu 1996). Although fixed effect approaches are equivalent in power, at least in situations with a single QTL, VCM are easier to implement and have computational advantages in this context (Xu 1998). Rules for setting up the required QTL allelic relationship matrices from marker data were given by Wang *et al.* (1995) and Abdel-Azim and Freeman (2001). Marker-based relationship matrices for QTL with additive genetic and nonadditive genetic (dominance, epistasis) gene action in noninbred populations were applied by Liu *et al.* (2002).

The focus of Xie *et al.* (1998) was on backcross (BC) and F_2 designs descending from inbred lines. For these types of experiments additive genetic and dominance relationship matrices can be calculated from conditional QTL genotype probabilities (given the flanking marker genotypes) for all individuals of the mapping population (as used as regressor variables in RIM). Crepieux *et al.* (2004) provided a general extension to any type of multicross designs from inbred parents. Furthermore, Li and Cui (2009) demonstrated how VCM can be employed for mapping imprinted QTL in a combination of different BC populations derived from inbred lines.

In this article we first propose extensions of the variance component approach of Xie *et al.* (1998) to multiple interacting QTL with pairwise epistatic effects. Then, maintaining the focus on inbred line-derived F_2 populations, a reduced model is suggested, in which individual genetic effects are replaced by average genetic effects for different marker classes. The covariance matrix of the phenotypes is approximated in different ways, leading to less computational effort.

THEORY

Linear mixed model: From an F_2 generation derived from a cross between inbred lines, one observation per individual is considered. The vector of phenotypes $\boldsymbol{Y}$ (length n) is modeled with respect to additive genetic, dominance, and pairwise epistatic effects of the QTL, whose total number is ν. A pair of QTL is indexed by ℓ and k. The LMM

in matrix notation is given as

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \sum_{\ell=1}^{\nu} \boldsymbol{Z}_\ell(\boldsymbol{u}_{a_\ell} + \boldsymbol{u}_{d_\ell}) \\ &\quad + \sum_{\ell=1}^{\nu-1} \sum_{k=\ell+1}^{\nu} \boldsymbol{Z}_{\ell k}(\boldsymbol{u}_{aa_{\ell k}} + \boldsymbol{u}_{ad_{\ell k}} + \boldsymbol{u}_{da_{\ell k}} + \boldsymbol{u}_{dd_{\ell k}}) + \boldsymbol{e}\,. \end{aligned} \tag{2.1}$$

The vector of fixed effects $\boldsymbol{\beta}$ has the related design matrix $\boldsymbol{X}$. The random vectors $\boldsymbol{u}_\tau$ with $\tau \in \{a_\ell, d_\ell, aa_{\ell k}, ad_{\ell k}, da_{\ell k}, dd_{\ell k}\}$ denote the additive genetic, the dominance, and the four pairwise epistatic effects (first-order interactions) at QTL ℓ and k. For each τ the length of $\boldsymbol{u}_\tau$ equals the number of F_2 individuals n, i. e. all QTL effects differ between individuals. The incidence matrices $\boldsymbol{Z}_\ell$ and $\boldsymbol{Z}_{\ell k}$ with $\dim(\boldsymbol{Z}_\ell) = \dim(\boldsymbol{Z}_{\ell k}) = n \times n$ relate the observations to genetic effects. The residuals are assumed to be independently and identically normally distributed with $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$, where $\boldsymbol{I}$ is the identity matrix of order n and σ_e^2 is the residual variance. The covariances between normally distributed random genetic effects $\boldsymbol{u}_\tau$ and the residuals $\boldsymbol{e}$ are assumed zero as well as the covariances between different types of genetic effects $\boldsymbol{u}_\tau$. The expectations of the QTL effects are $E(\boldsymbol{u}_\tau) = \boldsymbol{0}$ and the variances are $\mathrm{Var}(\boldsymbol{u}_\tau) = \boldsymbol{V}_\tau\sigma_\tau^2$, where σ_τ^2 is the related QTL variance and $\boldsymbol{V}_\tau$ is the corresponding expected QTL relationship matrix conditional on the marker genotypes. The phenotypic vector therefore follows a multivariate normal distribution with $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$. The covariance matrix $\boldsymbol{V}$ is derived conditional on the observed marker genotypes and can be written as

$$\begin{aligned} \boldsymbol{V} &= \sum_{\ell=1}^{\nu} \boldsymbol{Z}_\ell \left(\boldsymbol{V}_{a_\ell}\sigma_{a_\ell}^2 + \boldsymbol{V}_{d_\ell}\sigma_{d_\ell}^2\right) \boldsymbol{Z}_\ell' + \sum_{\ell=1}^{\nu-1} \sum_{k=\ell+1}^{\nu} \boldsymbol{Z}_{\ell k} \left(\boldsymbol{V}_{aa_{\ell k}}\sigma_{aa_{\ell k}}^2 + \boldsymbol{V}_{ad_{\ell k}}\sigma_{ad_{\ell k}}^2\right. \\ &\quad \left. + \boldsymbol{V}_{da_{\ell k}}\sigma_{da_{\ell k}}^2 + \boldsymbol{V}_{dd_{\ell k}}\sigma_{dd_{\ell k}}^2\right) \boldsymbol{Z}_{\ell k}' + \boldsymbol{I}\sigma_e^2\,. \end{aligned} \tag{2.2}$$

Calculation of covariance matrices: We follow the approach of Xie *et al.* (1998) and derive the required genetic covariance matrices of (2.2) from conditional QTL genotype probabilities and elementary covariance matrices.

Conditional QTL genotype probabilities: For a particular QTL the F_2 generation can be partitioned into nine different marker classes (see Table 2.2 column headings) conditional on the observed genotype of the flanking markers. QTL alleles originating from the first line are denoted by uppercase letter indexes (Q, H) and those from the second line by lowercase indexes (q, h), and for marker alleles the respective line origins are indicated by numbers (1 and 2). Conditional QTL genotype probabilities depend on flanking marker genotypes and the recombination rates between the markers and QTL and can be derived as described by, e. g., Carbonell *et al.* (1992, Table 1). We allow for double recombinations and assume Haldane's mapping function (Haldane

1919).

Probabilities for the genotypes G_{QQ}, G_{Qq}, and G_{qq} of an individual at the ℓth QTL conditional on flanking marker information M_i can be collected in a row vector $\boldsymbol{l}_i^{\ell}$ with

$$\boldsymbol{l}_i^{\ell} = \begin{pmatrix} \Pr(G_{QQ}|M_i) & \Pr(G_{Qq}|M_i) & \Pr(G_{qq}|M_i) \end{pmatrix} = \begin{pmatrix} p_i^{QQ} & p_i^{Qq} & p_i^{qq} \end{pmatrix},$$

where M_i denotes the observed flanking marker genotype $i \in \{1, \ldots, 9\}$ of an individual. We assume that in each marker interval either no or only a single QTL exists. The joint conditional probability for two linked QTL is just the product of both single probabilities if at least one completely informative marker is in between (Rönnegård *et al.* 2008). Thus, the probability of a two-locus QTL genotype, e. g. G_{QQHh}, given the particular marker genotypes M_i and N_j ($i, j \in \{1, \ldots, 9\}$) at QTL ℓ and k, respectively, is defined as $\Pr(G_{QQHh}|M_i, N_j) = \Pr(G_{QQ}|M_i)\Pr(G_{Hh}|N_j)$. We define $\boldsymbol{l}_{ij}^{\ell k}$ as the row vector with all joint conditional QTL genotype probabilities for a pairwise epistatic effect at QTL ℓ and k.

Elementary covariance matrices: As a second ingredient we need elementary covariance matrices between all possible QTL genotypes G_{QQ}, G_{Qq}, and G_{qq} in the F_2 populations. The elementary matrices for additive genetic QTL effects $\boldsymbol{A}$ (Xie *et al.* 1998) and dominance QTL effects $\boldsymbol{D}$ (Smith 1984; Xie *et al.* 1998) are

$$\boldsymbol{A} = \begin{array}{c} \\ G_{QQ} \\ G_{Qq} \\ G_{qq} \end{array} \begin{array}{c} \begin{array}{ccc} G_{QQ} & G_{Qq} & G_{qq} \end{array} \\ \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \end{array} \quad \text{and} \quad \boldsymbol{D} = \begin{array}{c} \\ G_{QQ} \\ G_{Qq} \\ G_{qq} \end{array} \begin{array}{c} \begin{array}{ccc} G_{QQ} & G_{Qq} & G_{qq} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}.$$

We use the Kronecker product (symbol $\otimes$) of $\boldsymbol{A}$ and $\boldsymbol{D}$ to compute the four different 9×9 elementary matrices, $\boldsymbol{A} \otimes \boldsymbol{A}, \boldsymbol{A} \otimes \boldsymbol{D}, \boldsymbol{D} \otimes \boldsymbol{A}, \boldsymbol{D} \otimes \boldsymbol{D}$, which include covariances between pairwise epistatic effects and correspond to nine genotypes (G_{QQHH}, G_{QQHh}, G_{QQhh}, G_{QqHH}, G_{QqHh}, G_{Qqhh}, G_{qqHH}, G_{qqHh}, and G_{qqhh}) for pairwise QTL combinations.

QTL relationship matrices: The $n \times n$ additive genetic, dominance, and pairwise epistatic relationship matrices for all F_2 individuals can be set up for a putative QTL position or combinations thereof with conditional QTL genotype probabilities ($\boldsymbol{l}_i^{\ell}$ and $\boldsymbol{l}_{ij}^{\ell k}$ vectors) and elementary matrices (Xie *et al.* 1998). Relationship coefficients are averages of possible QTL genotype combinations. For the additive genetic relationship

matrix $\boldsymbol{V}_{a_\ell} = \left\{a^\ell_{st}\right\}^n_{s,t=1}$ we get diagonal elements

$$a^\ell_{ss} \;=\; \boldsymbol{l}^\ell_i \mathrm{diag}\,(\boldsymbol{A}) \;=\; 2p_i^{QQ} + p_i^{Qq} + 2p_i^{qq} \tag{2.3}$$

and off-diagonals

$$\begin{aligned} a^\ell_{st} &= \boldsymbol{l}^\ell_i \boldsymbol{A} \left(\boldsymbol{l}^\ell_j\right)' \\ &= 2p_i^{QQ}\,p_j^{QQ} + p_i^{Qq}\,p_j^{QQ} + p_i^{QQ}\,p_j^{Qq} + p_i^{Qq}\,p_j^{Qq} + p_i^{qq}\,p_j^{Qq} + p_i^{Qq}\,p_j^{qq} + 2p_i^{qq}\,p_j^{qq} \end{aligned} \tag{2.4}$$

at the ℓth QTL. If both individuals s and t belong to the same marker class i, then a^ℓ_{st} can be simplified to

$$a^\ell_{st} \;=\; 2(p_i^{QQ})^2 + 2p_i^{QQ}\,p_i^{Qq} + (p_i^{Qq})^2 + 2p_i^{Qq}\,p_i^{qq} + 2(p_i^{qq})^2\,, \tag{2.5}$$

because the conditional probabilities are equal. The dominance relationship matrix $\boldsymbol{V}_{d_\ell} = \left\{d^\ell_{st}\right\}^n_{s,t=1}$ is set up equivalently, but instead of $\boldsymbol{A}$ the elementary matrix $\boldsymbol{D}$ is used, i. e. $d^\ell_{ss} = \boldsymbol{l}^\ell_i\,\mathrm{diag}(\boldsymbol{D})$ and $d^\ell_{st} = \boldsymbol{l}^\ell_i \boldsymbol{D} \left(\boldsymbol{l}^\ell_j\right)'$.

We suggest that the pairwise epistatic relationship matrices $\boldsymbol{V}_{aa_{\ell k}}, \boldsymbol{V}_{ad_{\ell k}}, \boldsymbol{V}_{da_{\ell k}}, \boldsymbol{V}_{dd_{\ell k}}$ at the ℓth and kth QTL are computed analogously to $\boldsymbol{V}_{a_\ell}$ using the appropriate Kronecker product of elementary matrices (e. g. $\boldsymbol{A} \otimes \boldsymbol{A}$). Computation of matrix elements is done as in Equations (2.3) and (2.4), employing corresponding row vectors $\boldsymbol{l}^{\ell k}_{ij}$. Note that this is equivalent to using Hadamard products of QTL relationship matrices $\boldsymbol{V}_{a_\ell}$ and $\boldsymbol{V}_{d_\ell}$ given that there is at least one completely informative marker between both QTL or no linkage between them (Rönnegård *et al.* 2008), which is always fulfilled by our assumptions. To ensure positive definiteness of covariance matrices, we assume that locations of putative QTL and markers do not coincide.

Equivalent model with average genetic effects: What we have outlined so far is termed "individual model", because each individual receives its own genetic effects for the different kinds of genetic components. For a particular QTL ℓ the LMM of (2.1) with only additive genetic effects becomes

$$\boldsymbol{Y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\ell \boldsymbol{u}_{a_\ell} + \boldsymbol{e}\,, \tag{2.6}$$

with covariance matrix of the phenotypes conditional on the observed marker genotypes

$$\boldsymbol{V} \;=\; \boldsymbol{Z}_\ell \boldsymbol{V}_{a_\ell} \boldsymbol{Z}'_\ell \sigma^2_{a_\ell} + \boldsymbol{I}\sigma^2_e\,. \tag{2.7}$$

A model equivalent to (2.6) is

$$\boldsymbol{Y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \widetilde{\boldsymbol{Z}}_\ell \widetilde{\boldsymbol{u}}_{a_\ell} + \boldsymbol{m}_{a_\ell} + \boldsymbol{e}\,, \tag{2.8}$$

where a vector $\widetilde{\boldsymbol{u}}_{a_\ell}$ with length $n_\ell = 9$ (number of different marker classes) of average additive genetic effects for all possible marker genotype classes is considered. An additional random effect $\boldsymbol{m}_{a_\ell}$ of length n appears, termed "additive genetic sampling effect", and it describes the deviations of the individual additive genetic effects from the average additive genetic effects of marker classes. The dimension of $\widetilde{\boldsymbol{Z}}_\ell$ is $n \times n_\ell$. Accordingly, the covariance matrix of the phenotypes can be expressed as

$$\boldsymbol{V} \;=\; \widetilde{\boldsymbol{Z}}_\ell \widetilde{\boldsymbol{V}}_{a_\ell} \widetilde{\boldsymbol{Z}}'_\ell \sigma^2_{a_\ell} + \boldsymbol{V}_{m_{a_\ell}} \sigma^2_{a_\ell} + \boldsymbol{I}\sigma^2_e\,, \tag{2.9}$$

where $\widetilde{\boldsymbol{V}}_{a_\ell} = \left\{\widetilde{a}^\ell_{ij}\right\}_{i,j=1}^{n_\ell}$ denotes the reduced $n_\ell \times n_\ell$ relationship matrix of the average additive genetic effects at the QTL. The additive genetic variance of the individual model (2.7) is $\sigma^2_{a_\ell}$, which is identical to $\sigma^2_{a_\ell}$ in (2.9). The variance of the additive genetic sampling effect is $\mathrm{Var}(\boldsymbol{m}_{a_\ell}) = \boldsymbol{V}_{m_{a_\ell}} \sigma^2_{a_\ell}$, where $\boldsymbol{V}_{m_{a_\ell}}$ denotes the relationship matrix of the additive genetic sampling effect. There are n^ℓ_i individuals with the same marker genotype i at the QTL. The variance of the average additive genetic effect of a certain marker class i, averaged over n^ℓ_i individuals, is given in the reduced model as

$$\widetilde{a}^\ell_{ii} \;=\; \begin{cases} \boldsymbol{l}^\ell_i \mathrm{diag}(\boldsymbol{A}) & \text{if} \quad n^\ell_i = 0\,, \\ \dfrac{1}{n^\ell_i}\boldsymbol{l}^\ell_i \mathrm{diag}\left(\boldsymbol{A}\right) + \left(1 - \dfrac{1}{n^\ell_i}\right) \boldsymbol{l}^\ell_i \boldsymbol{A} \left(\boldsymbol{l}^\ell_i\right)' & \text{else}\,. \end{cases} \tag{2.10}$$

Equation (2.10) is valid, because there are n^ℓ_i diagonal elements and $\left(n^\ell_i\right)^2 - n^\ell_i$ off-diagonal elements in the relationship matrix of the individual additive genetic effects.

The three possible cases appearing in the additive genetic relationship matrix of the individual model are further investigated (see Equations (2.3) to (2.5)). First, the variance of an individual additive genetic effect with marker class i is $2p_i^{QQ} + p_i^{Qq} + 2p_i^{qq} := \widetilde{v}^\ell_{ii}$ and second, the covariance between two additive genetic effects with the same marker class i is $2(p_i^{QQ})^2 + 2p_i^{QQ}\,p_i^{Qq} + (p_i^{Qq})^2 + 2p_i^{Qq}\,p_i^{qq} + 2(p_i^{qq})^2 := v^\ell_{ii}$. Then the element $\widetilde{a}^\ell_{ii}$ for $n^\ell_i > 0$ can be written as

$$\widetilde{a}^\ell_{ii} \;=\; \frac{1}{n^\ell_i}\widetilde{v}^\ell_{ii} + \left(1 - \frac{1}{n^\ell_i}\right) v^\ell_{ii}\,. \tag{2.11}$$

The variance of the average additive genetic effect is asymptotically equal to the co-

variance between individual additive genetic effects of the same marker class i, i. e. $\lim_{n_i^\ell \to \infty} \widetilde{a}_{ii}^\ell = v_{ii}^\ell$. Third, the covariance of additive genetic effects with marker classes i and j is $2p_i^{QQ}\,p_j^{QQ} + p_i^{Qq}\,p_j^{QQ} + p_i^{QQ}\,p_j^{Qq} + p_i^{Qq}\,p_j^{Qq} + p_i^{qq}\,p_j^{Qq} + p_i^{Qq}\,p_j^{qq} + 2p_i^{qq}\,p_j^{qq} := v_{ij}^\ell$. Now, the covariance of the average additive genetic effects of marker genotypes i and j $(i \neq j)$ can be expressed as

$$\widetilde{a}_{ij}^\ell \;=\; \boldsymbol{l}_i^\ell \boldsymbol{A} \left(\boldsymbol{l}_j^\ell\right)' \;=\; v_{ij}^\ell \,. \tag{2.12}$$

This is equal to the covariance among the two individual additive genetic effects of marker classes i and j.

The relationship matrix of the additive genetic sampling effects $\boldsymbol{V}_{m_{a_\ell}}$ can be determined as the difference between the relationship matrices of additive genetic effects from the individual model (individual genetic effects) and the reduced model (average genetic effects), which are inferred from Equations (2.7) and (2.9), i. e. $\boldsymbol{V}_{m_{a_\ell}} = \boldsymbol{Z}_\ell \boldsymbol{V}_{a\ell} \boldsymbol{Z}_\ell' - \widetilde{\boldsymbol{Z}}_\ell \widetilde{\boldsymbol{V}}_{a\ell} \widetilde{\boldsymbol{Z}}_\ell'$. Generally, $\boldsymbol{V}_{m_{a_\ell}}$ (order n) can be written as

$$\boldsymbol{V}_{m_{a_\ell}} \;=\; \begin{pmatrix} \boldsymbol{M}_{11}^{a_\ell} & \boldsymbol{M}_{12}^{a_\ell} & \dots & \boldsymbol{M}_{19}^{a_\ell} \\ \boldsymbol{M}_{21}^{a_\ell} & \boldsymbol{M}_{22}^{a_\ell} & \dots & \boldsymbol{M}_{29}^{a_\ell} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{M}_{91}^{a_\ell} & \boldsymbol{M}_{92}^{a_\ell} & \dots & \boldsymbol{M}_{99}^{a_\ell} \end{pmatrix} \tag{2.13}$$

if the individuals are arranged by marker class. To study the matrices $\boldsymbol{M}_{ij}^{a_\ell}$ we assume that each marker genotype appears at least once, i. e. $n_i^\ell \geq 1$.

Concerning the third case, the additive genetic covariance between a pair of individuals s and t with different marker genotypes i and j equals the difference of (2.4) and (2.12): $m_{ij}^{a_\ell} = v_{ij}^\ell - v_{ij}^\ell = 0$. Therefore, for $i \neq j$ $\boldsymbol{M}_{ij}^{a_\ell} = \boldsymbol{0}$ in (2.13) and $\boldsymbol{V}_{m_{a_\ell}}$ is a block diagonal matrix if the observations are ordered by marker genotypes. The diagonal block $\boldsymbol{M}_{ii}^{a_\ell}$ corresponding to marker class i has the order n_i^ℓ and can be expressed as

$$\boldsymbol{M}_{ii}^{a_\ell} \;=\; \begin{pmatrix} \widetilde{m}_{ii}^{a_\ell} + m_{ii}^{a_\ell} & m_{ii}^{a_\ell} & \dots & m_{ii}^{a_\ell} \\ m_{ii}^{a_\ell} & \widetilde{m}_{ii}^{a_\ell} + m_{ii}^{a_\ell} & \dots & m_{ii}^{a_\ell} \\ \vdots & \vdots & \ddots & \vdots \\ m_{ii}^{a_\ell} & m_{ii}^{a_\ell} & \dots & \widetilde{m}_{ii}^{a_\ell} + m_{ii}^{a_\ell} \end{pmatrix} .$$

The covariance $m_{ii}^{a_\ell}$ of the additive genetic sampling effects of two individuals s and t

given the same marker genotype i (second case) is the difference of (2.5) and (2.11),

$$m_{ii}^{a_\ell} = v_{ii}^{\ell} - \left(\frac{1}{n_i^{\ell}}\widetilde{v}_{ii}^{\ell} + \left(1 - \frac{1}{n_i^{\ell}}\right) v_{ii}^{\ell}\right) = -\frac{1}{n_i^{\ell}}\left(\widetilde{v}_{ii}^{\ell} - v_{ii}^{\ell}\right) . \tag{2.14}$$

For $n_i^{\ell} \geq 1$ $m_{ii}^{a_\ell} \in [-0.5, 0.0]$. The variance $\widetilde{m}_{ii}^{a_\ell} + m_{ii}^{a_\ell}$ of the additive genetic sampling effect given the marker genotype i (first case) is the difference of (2.3) and (2.11),

$$\widetilde{m}_{ii}^{a_\ell} + m_{ii}^{a_\ell} = \widetilde{v}_{ii}^{\ell} - \left(\frac{1}{n_i^{\ell}}\widetilde{v}_{ii}^{\ell} + \left(1 - \frac{1}{n_i^{\ell}}\right) v_{ii}^{\ell}\right) = \left(1 - \frac{1}{n_i^{\ell}}\right)\left(\widetilde{v}_{ii}^{\ell} - v_{ii}^{\ell}\right) \tag{2.15}$$

with $\widetilde{m}_{ii}^{a_\ell} \in [0.0, 0.5]$. Note that the elements $\widetilde{m}_{ii}^{a_\ell} = \widetilde{v}_{ii}^{\ell} - v_{ii}^{\ell}$ are independent of n_i^{ℓ}. However, $\widetilde{m}_{ii}^{a_\ell}$ and $m_{ii}^{a_\ell}$ depend on conditional genotype probabilities. From (2.14) and (2.15) it is obvious that $\widetilde{m}_{ii}^{a_\ell}$ is a function of the covariance of the additive genetic sampling effects from the same marker class i and the corresponding number n_i^{ℓ} of observations, $\widetilde{m}_{ii}^{a_\ell} = -n_i^{\ell}\, m_{ii}^{a_\ell}$.

The calculation of the relationship matrix of the additive genetic effect of the individual model (2.6) and the reduced model (2.8) as well as the additive genetic sampling relationship matrix $\boldsymbol{V}_{m_{a_\ell}}$ is summarized in Table 2.1.

TABLE 2.1: Correspondence of elements of additive genetic relationship matrices in the individual model and the equivalent model with additive genetic sampling effects. Each variable in the second column (individual model) is the sum from the two expressions of the third and fourth columns (equivalent model). Case 1: diagonal elements for marker class $i \in \{1, \ldots, 9\}$; case 2: two individuals with equal marker class i; case 3: two individuals with different marker classes i and j.

	Individual model	Equivalent model	
Case	$(\boldsymbol{Z}_\ell \boldsymbol{V}_{a_\ell} \boldsymbol{Z}_\ell')_{st}$	$\left(\widetilde{\boldsymbol{Z}}_\ell \widetilde{\boldsymbol{V}}_{a_\ell} \widetilde{\boldsymbol{Z}}_\ell'\right)_{st}$	$\left(\boldsymbol{V}_{m_{a_\ell}}\right)_{st}$
1	$\widetilde{v}_{ii}^{\ell}$	$\frac{1}{n_i^{\ell}}\widetilde{v}_{ii}^{\ell} + \left(1 - \frac{1}{n_i^{\ell}}\right) v_{ii}^{\ell}$	$\left(1 - \frac{1}{n_i^{\ell}}\right)\left(\widetilde{v}_{ii}^{\ell} - v_{ii}^{\ell}\right)$
2	v_{ii}^{ℓ}	$\frac{1}{n_i^{\ell}}\widetilde{v}_{ii}^{\ell} + \left(1 - \frac{1}{n_i^{\ell}}\right) v_{ii}^{\ell}$	$-\frac{1}{n_i^{\ell}}\left(\widetilde{v}_{ii}^{\ell} - v_{ii}^{\ell}\right)$
3	v_{ij}^{ℓ}	v_{ij}^{ℓ}	0

If model (2.6) includes not only additive genetic but also dominance effects, the genetic parameters for average dominance effects and dominance sampling terms can be obtained analogously. The genetic sampling relationship matrices of the pairwise epistatic effects can also be calculated similarly to the additive genetic and dominance effects, but the row vectors $\boldsymbol{l}_{ij}^{\ell k}$ that considered the joint conditional QTL genotypes

probabilities of the ℓ th and k th QTL have to be used. Then $n_{\ell k}$ different marker classes have to be considered, where $n_{\ell k} = 27$ if the QTL are in two adjacent marker intervals and $n_{\ell k} = 81$ otherwise.

If we assume that the number of F_2 individuals approaches infinity ($n \to \infty$), then the number of individuals per marker class i also increases ($n_i^\ell \to \infty$). The diagonal elements $\widetilde{a}_{ii}^\ell$ as well as $m_{ii}^{a_\ell}$ depend on n_i^ℓ, where $\frac{1}{n_i^\ell}$ tends to zero for $n \to \infty$. Hence $\lim_{n_i^\ell \to \infty} \boldsymbol{V}_{m_{a_\ell}} = \boldsymbol{\Delta}_{a_\ell}$, where $\boldsymbol{\Delta}_{a_\ell}$ is a diagonal matrix of order n of elements $\widetilde{m}_{ii}^{a_\ell}$. Therefore, the covariance matrix of the additive genetic sampling effects is asymptotically diagonal.

Reduced model: Instead of an individual model we developed a reduced model approach, which is an approximation of model (2.8), with decreased dimension of the relationship matrices. The LMM is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \widetilde{\boldsymbol{Z}}_\ell \widetilde{\boldsymbol{u}}_{a_\ell} + \boldsymbol{\varepsilon}$, where the residuals are assumed to be independently and identically normally distributed with $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\varepsilon^2)$. Here the F_2 individuals are grouped according to their marker genotypes and average genetic effects are estimated for marker classes instead of individual genetic effects, as described in (2.8). The dimension of the relationship matrices depends on the number of marker classes (n_ℓ and $n_{\ell k}$), but not on the experiment size n. We call this procedure the reduced model (*vs.* the individual model).

In general, the reduced model with respect to additive genetic, dominance, and pairwise epistatic effects is

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \sum_{\ell=1}^{\nu} \widetilde{\boldsymbol{Z}}_\ell (\widetilde{\boldsymbol{u}}_{a_\ell} + \widetilde{\boldsymbol{u}}_{d_\ell}) \\ &\quad + \sum_{\ell=1}^{\nu-1} \sum_{k=\ell+1}^{\nu} \widetilde{\boldsymbol{Z}}_{\ell k} (\widetilde{\boldsymbol{u}}_{aa_{\ell k}} + \widetilde{\boldsymbol{u}}_{ad_{\ell k}} + \widetilde{\boldsymbol{u}}_{da_{\ell k}} + \widetilde{\boldsymbol{u}}_{dd_{\ell k}}) + \boldsymbol{\varepsilon}\,, \end{aligned} \tag{2.16}$$

where the residuals are again assumed to be independently and identically normally distributed with $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\varepsilon^2)$. The vectors $\widetilde{\boldsymbol{u}}_\tau$ with $\tau \in \{a_\ell, d_\ell, aa_{\ell k}, ad_{\ell k}, da_{\ell k}, dd_{\ell k}\}$ consider the average additive genetic, dominance and pairwise epistatic effects of length n_ℓ and $n_{\ell k}$.

The calculation of the reduced dominance relationship matrix $\widetilde{\boldsymbol{V}}_{d_\ell}$ at the ℓ th QTL is done similarly to the notes above, but $\boldsymbol{A}$ has to be replaced by $\boldsymbol{D}$. Both $\widetilde{\boldsymbol{V}}_{a_\ell}$ and $\widetilde{\boldsymbol{V}}_{d_\ell}$ are matrices of order n_ℓ, where $n_\ell = 9$ if the flanking markers are fully informative. The reduced epistatic relationship matrices $\widetilde{\boldsymbol{V}}_{aa_{\ell k}}, \widetilde{\boldsymbol{V}}_{ad_{\ell k}}, \widetilde{\boldsymbol{V}}_{da_{\ell k}}, \widetilde{\boldsymbol{V}}_{dd_{\ell k}}$ of the ℓ th and k th QTL are computed analogously to $\widetilde{\boldsymbol{V}}_{a_\ell}$ from (2.10) and (2.12), but the corresponding Kronecker product is used instead of $\boldsymbol{A}$ and the row vector $\boldsymbol{l}_{ij}^{\ell k}$ for the i th and j th marker class is applied.

The difference $\widetilde{v}_{ii}^{\ell} - v_{ii}^{\ell}$ (asymptotic variance $\widetilde{m}_{ii}^{a_\ell}$) between the variance of an individual additive genetic effect and the covariance between two additive genetic effects of the same marker class decreases as the distance between flanking markers becomes smaller. Decreasing QTL effects and genetic variances lead to the same effect. In the extreme case, when the marker location and the position of the QTL coincide, the difference $\widetilde{v}_{ii}^{\ell} - v_{ii}^{\ell}$ is zero and therefore $\boldsymbol{V}_{m_{a_\ell}} = \boldsymbol{0}$. In this case the covariance of the phenotypes in the reduced and the individual model are identical. Therefore, approximating $\boldsymbol{V}_{m_{a_\ell}}\sigma^2_{a_\ell} + \boldsymbol{I}\sigma^2_e$ (or its multilocus equivalent) by $\boldsymbol{I}\sigma^2_\varepsilon$ seems to be a reasonable choice. Note that Xu (1995) and Xu (1998) investigated the inflation of the residual variance through the within marker genotype QTL variance in the RIM, which is similar to our genetic sampling effects.

The approximation of the individual model by the reduced model relies on two different aspects. First, the covariances $m_{ii}^{a_\ell}$ between genetic sampling effects (deviation of individual genetic effects from average genetic effects of marker classes) are assumed to be zero. Second, the asymptotic variances $\widetilde{m}_{ii}^{a_\ell}$ of the additive genetic sampling effects are treated as equal for all marker classes. Covariances $m_{ii}^{a_\ell}$ between additive genetic sampling effects of individuals sharing the same marker class i are shown in Table 2.2 for an additive QTL in the middle of a 10 cM marker interval in dependence on sample size. The elements $m_{ii}^{a_\ell}$ were calculated using the number of expected proportions for each marker genotype according to Equation (2.14). To make sure that $n_i^\ell \geq 1$, we used $n \geq 500$. For 500 F_2 individuals this covariance $m_{ii}^{a_\ell}$ is $\leq 1\,\%$ of the QTL variance and shows a further decline when the sample size increases. Only for marker classes $G_{11/22}$ and $G_{22/11}$ is there a very high (negative) covariance (48.7 % of the additive genetic variance) and an experiment with > 2000 F_2 individuals would be required to reach a value $< 10\,\%$. These marker genotypes are rare, we expect these marker genotypes to occur twice in total among 500 F_2 genotypes. Therefore, omitting these covariances $m_{ii}^{a_\ell}$ has little effect on the likelihood.

Table 2.2: Covariances $m_{ii}^{a_\ell}$ of additive genetic sampling effects within marker class i for different numbers (n) of F_2 individuals: QTL in the middle of a 10 cM marker interval. Flanking marker genotypes $G_{./.}$ are indexed by their alleles for each i.

	$m_{ii}^{a_\ell}$								
	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
n	$G_{11/11}$	$G_{11/12}$	$G_{11/22}$	$G_{12/11}$	$G_{12/12}$	$G_{12/22}$	$G_{22/11}$	$G_{22/12}$	$G_{22/22}$
500	0.000	−0.012	−0.487	−0.012	0.000	−0.012	−0.487	−0.012	0.000
1000	0.000	−0.006	−0.243	−0.006	0.000	−0.006	−0.243	−0.006	0.000
2000	0.000	−0.003	−0.122	−0.003	0.000	−0.003	−0.122	−0.003	0.000
3000	0.000	−0.002	−0.081	−0.002	0.000	−0.002	−0.081	−0.002	0.000

The asymptotic variances $\widetilde{m}_{ii}^{a_\ell}$ of the additive genetic sampling effects for different marker classes are, however, larger than their corresponding covariances and, more importantly, they show considerable variation between more frequent marker classes. The sixth line of Table 2.3 shows the genetic sampling variances for all marker classes, again for an additive QTL in the middle of a 10 cM marker interval. For the three most frequent marker classes, the genetic sampling variance is at $\leq 1\,\%$ of the additive genetic variance (classes 1, 5 and 9) and for another four marker classes it equals 25 % (classes 2, 4, 6 and 8), while a 50 % value occurs only in the very rare classes (3 and 7). Note that the genetic sampling variances become smaller when the QTL is located closer to the boundary of the marker interval. The genetic sampling effects completely vanish if marker locations and positions of the QTL coincide (Table 2.3, first line). In such cases, the covariances of the genetic sampling effects are zero and the assumption $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{I}\sigma_\varepsilon^2$ of the reduced model (2.16) is exact.

TABLE 2.3: Asymptotic variances $\widetilde{m}_{ii}^{a_\ell}$ of additive genetic sampling effects within marker class i for differently sized marker intervals and different QTL positions within marker intervals (cM).

Marker	Position	$\widetilde{m}_{ii}^{a_\ell}$								
interval	of QTL	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
0	0	0.00				0.00				0.00
10	1	0.00	0.09	0.18	0.09	0.00	0.09	0.18	0.09	0.00
10	2	0.00	0.16	0.32	0.16	0.01	0.16	0.32	0.16	0.00
10	3	0.00	0.21	0.42	0.21	0.01	0.21	0.42	0.21	0.00
10	4	0.00	0.24	0.48	0.24	0.01	0.24	0.48	0.24	0.00
10	5	0.00	0.25	0.50	0.25	0.01	0.25	0.50	0.25	0.00
20	10	0.02	0.26	0.50	0.26	0.04	0.26	0.50	0.26	0.02
30	15	0.04	0.27	0.50	0.27	0.08	0.27	0.50	0.27	0.04
40	20	0.07	0.29	0.50	0.29	0.13	0.29	0.50	0.29	0.07

The latter considerations suggest, as a further alternative, a weighted approach, where the second part of the approximation inherent in the reduced model, i. e. equal genetic sampling variances for all marker classes, is skipped, while the assumption (first part) of zero covariances for genetic sampling effects within marker class is maintained. For a single additive QTL this results in the following mixed model equations (MME):

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\widetilde{\boldsymbol{Z}}_\ell \\ \widetilde{\boldsymbol{Z}}_\ell'\boldsymbol{W}\boldsymbol{X} & \widetilde{\boldsymbol{Z}}_\ell'\boldsymbol{W}\widetilde{\boldsymbol{Z}}_\ell + \widetilde{\boldsymbol{V}}_{a_\ell}^{-1}\lambda \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \widetilde{\boldsymbol{u}}_{a_\ell} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \\ \widetilde{\boldsymbol{Z}}_\ell'\boldsymbol{W}\boldsymbol{y} \end{pmatrix},$$

where $\lambda = \dfrac{\sigma_e^2}{\sigma_{a_\ell}^2}$. The variance of the residuals is $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathrm{Var}(\boldsymbol{m}_{a_\ell} + \boldsymbol{e}) = \boldsymbol{W}^{-1}\sigma_e^2$, where σ_e^2 is defined as in (2.8) and all other symbols as in (2.1) and (2.16). The diagonal

matrix $\boldsymbol{W}$ of order n has the entries $w_{ss} = \sigma_e^2 \left(\widetilde{m}_{ii}^{a_\ell} \sigma_{a_\ell}^2 + \sigma_e^2\right)^{-1}$, which differ between observations from different marker classes and are equal for observations from the same marker class i. If more QTL and nonadditive genetic gene actions are considered in the model, then the genetic sampling variances for different QTL and different kinds τ of genetic effects have to be summed to get the entire genetic sampling variance of an observation and w_{ss} (s th individual given the marker class i) becomes

$$w_{ss} \;=\; \frac{\sigma_e^2}{\sum \widetilde{m}_{ii}^{\tau} \sigma_\tau^2 + \sigma_e^2}\,, \tag{2.17}$$

where $\tau \in \{a_\ell, d_\ell, aa_{\ell k}, ad_{\ell k}, da_{\ell k}, dd_{\ell k}\}$. This weighted version of the reduced model retains the advantage of a reduced dimension of the QTL relationship matrices as in the reduced model, but may provide a better approximation of the exact residual log-likelihood ratio test ($RLRT$) statistics. If marker location and position of QTL coincide, the weights of (2.17) are one and $\boldsymbol{W}$ is an identity matrix. The weights of (2.17) are similar to the weights in the weighted least-squares method of QTL mapping as shown by Xu (1995) and Xu (1998).

Coincidence of markers and QTL results in singularity of $\widetilde{\boldsymbol{V}}_{a_\ell}$ (identical to $\boldsymbol{A}$ in this case) and was not further considered here. However, this situation can be treated, e. g. by regularization [adding a small quantity to the diagonal elements of $\widetilde{\boldsymbol{V}}_{a_\ell}$; Neumaier (1998)], which has little effect on the test statistics and is easy to implement, by including allelic effects in the model instead of genotypic effects, or by replacing $\widetilde{\boldsymbol{V}}_{a_\ell}$ by a reduced rank approximation (Rönnegård *et al.* 2007) obtained by spectral decomposition.

SIMULATIONS

First, a single F_2 family as the simplest case of a combination of multiple line crosses was considered to demonstrate the properties of the reduced model in comparison to the individual model (Xie *et al.* 1998) and the fixed effects methods MIM (Kao and Zeng 1997; Kao *et al.* 1999) and RIM (Haley and Knott 1992; Martínez and Curnow 1992). Experiments from four different scenarios were simulated with 1000 replications per scenario and $n = 200$ F_2 individuals per experiment. Scenarios 1 and 2 consisted of a single additive genetic QTL at 35 cM on a single chromosome of 50 cM length, whereas in the other scenarios (3 and 4) there were two linked QTL with equally sized QTL effects in repulsion. In the fourth scenario chromosome length was extended to 80 cM and an interaction effect was included. For further characteristics of all scenarios see Table 2.4. The observations were simulated using Cockerham's

F_2-metric model (Cockerham 1954; Kao and Zeng 2002, Table 3). The relative QTL variance R^2 is the proportion of the phenotypic variance σ_p^2 explained by the QTL and is $R^2 = \frac{\sigma_{QTL}^2}{\sigma_p^2}$.

Table 2.4: Brief summary of simulated scenarios: the number of QTL ν, length of the chromosome l_c (cM), QTL positions P_1 and P_2 (cM), marker positions (cM), residual variance σ_e^2, additive genetic effects (a_1, a_2), and additive-by-additive genetic effects aa_{12} as well as the relative QTL variance R^2 (%).

Scenario	ν	l_c	P_1	P_2	Marker positions	σ_e^2	a_1	a_2	aa_{12}	R^2
1	1	50	35	–	$0, 10, 20, 30, 40, 50$	9.529	1.0	–	–	5.0
2	1	50	35	–	$0, 10, 20, 30, 40, 50$	1.000	1.0	–	–	33.3
3	2	50	25	35	$0, 10, 20, 30, 40, 50$	0.181	1.0	−1.0	–	50.0
4	2	80	35	45	$0, 40, 80$	1.000	1.0	−1.0	1.0	30.0

In the second part our small simulation study focused on the performance of the reduced *vs.* the individual model in a situation with multiple families. Four independent F_2 families, each with 50 progeny ($n = 200$), were derived from a population consisting of four different inbred lines, representing all pairwise combinations of QTL genotypes (G_{QQHH}, G_{QQhh}, G_{qqHH}, G_{qqhh}). For each family F_1 individuals were generated from a random pair of inbred lines. Markers were always assumed to be fully informative. In the LMM family means were treated as fixed. Remaining parameters were chosen as previously described for the third scenario (Table 2.4). For each genetic effect a single (population-specific) variance was assumed.

Significance thresholds for the null hypothesis of no linked QTL were determined by simulating 1000 experiments of the same size for each scenario, where QTL with the same kind and size of effects were present, but unlinked to the markers. After analyzing these experiments, the 95 % quantile of the maximum values of the test statistic from all replications was taken as a significance threshold, specific for each scenario and method, which allowed the determination of experimental power. We performed the residual log-likelihood ratio test for the reduced and the individual model, the log-likelihood ratio test for MIM and the F-test for RIM. Mean QTL positions, root mean squared error (RMSE) of the QTL positions as well as their 5 % and 95 % quantiles were evaluated to characterize the precision of location estimates. For each replication we analyzed positions or combinations thereof, where marker locations and QTL positions did not coincide (step width 1 cM, both QTL in different marker intervals). Therefore, we applied RIM and MIM with the same restrictions as the VCM. Our analyses used the true genetic model for testing for segregating QTL, i. e. the model included only the simulated effects of QTL and no model selection was performed. All calculations were

done with self-written Fortran 95 programs in combination with ASReml (Gilmour *et al.* 2008) for estimation of variance components and evaluation of the restricted maximum likelihood function (Patterson and Thompson 1971).

DISCUSSION

Results for all simulated single-QTL scenarios are summarized in Table 2.5. The experimental power was 100 % (scenarios 2 and 3) or nearly so (scenario 4), with the exception of scenario 1, where the experimental power was uniformly at 82 % for all methods. There was almost no variation between methods in the mean estimated position in the single-QTL scenarios (1 and 2); even the distributions of the estimates showed identical 5 % and 95 % quantiles. Differences between methods became, however, apparent in the two-QTL scenarios. For scenario 3 (two QTL in repulsion, no interactions), MIM resulted in average estimated QTL positions at 24.7 and 34.9 cM, nearly identical to the simulated values at 25 and 35 cM. The RMSEs for positions of the QTL were < 1.4 cM for both QTL for MIM and ~ 2.0 cM for the individual model, while the reduced model and RIM performed very similarly with RMSEs of ~ 4.2 cM. In scenario 3 the reduced model, the individual model and RIM on average placed the QTL somewhat more towards the ends of the chromosome compared to MIM and the true values, resulting in an overestimation of the distance (true distance: 10 cM) between both QTL, ranging from 2.7 cM (individual model) to 6.7 cM (reduced model). For scenario 4 (two QTL in repulsion with interactions) this overestimation of the distance between the QTL was, however, very similar for all methods at $\sim 2.0 - 3.1$ cM. The RMSEs for estimated positions of the QTL were between 5.3 and 5.6 cM with little differences between the first and second QTL for RIM as well as the reduced and the individual model. However, the RMSE of MIM at the same time showed the highest deviation of 6.5 cM for the first and the smallest deviation of 3.7 cM for the second QTL.

Note that MIM was applied according to the original approach of Kao and Zeng (1997) and Kao *et al.* (1999), which ignores double recombination events (complete interference) within the marker interval. However, double recombinations were taken into account for RIM and the VCM.

As theory indicated, estimated residual variance components from methods coping better with genetic deviations from the mean of a marker class (MIM, individual model) were smaller in the two-QTL scenarios compared to RIM and the reduced model, where the genetic sampling variance (QTL genotype variability within marker genotype) is part of the residual variance.

TABLE 2.5: Average estimates (mean) for QTL positions (P_1, P_2) with associated root mean squared error (RMSE) and quantiles (quant.) together with mean estimates of the residual variance $\hat{\sigma}_r^2$ and the observed power (%) for different scenarios: 200 F_2 individuals per simulated experiment and 1000 replications per scenario.

	Reduced model		Individual model		RIM		MIM	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
			Scenario 1					
Mean	32.58		32.53		32.73		32.40	
RMSE	10.71		10.79		10.78		10.89	
5% quant.	9.00		9.00		9.00		9.00	
95% quant.	48.00		48.00		48.00		48.00	
$\hat{\sigma}_r^2$	9.49		9.38		9.50		9.37	
Power	81.70		81.80		81.70		81.90	
			Scenario 2					
Mean	34.82		34.87		34.88		34.68	
RMSE	2.80		2.66		2.68		2.57	
5% quant.	31.00		31.00		31.00		31.00	
95% quant.	39.00		39.00		39.00		39.00	
$\hat{\sigma}_r^2$	1.04		0.99		1.04		0.99	
Power	100.00		100.00		100.00		100.00	
			Scenario 3					
Mean	21.67	38.38	23.70	36.37	22.22	37.82	24.70	34.89
RMSE	4.20	4.17	2.03	2.02	4.29	4.25	1.35	1.38
5% quant.	17.00	35.00	22.00	34.00	17.00	32.00	23.00	33.00
95% quant.	25.00	42.00	26.00	38.00	28.00	42.00	27.00	37.00
$\hat{\sigma}_r^2$	0.27		0.18		0.28		0.19	
Power	100.00		100.00		100.00		100.00	
			Scenario 4					
Mean	34.03	46.22	33.56	46.70	34.13	46.09	32.29	44.44
RMSE	5.36	5.50	5.26	5.47	5.51	5.55	6.47	3.72
5% quant.	24.50	41.00	24.00	41.00	24.00	41.00	21.00	41.00
95% quant.	39.00	56.50	39.00	57.00	39.00	57.00	39.00	52.00
$\hat{\sigma}_r^2$	1.21		0.93		1.23		1.04	
Power	99.20		99.40		99.60		100.00	

The results of the analysis of the multiple families are shown in Table 2.6. The accuracy of the estimated QTL positions of the individual model under consideration of four families was slightly better than that of the reduced model. However, when multiple families were considered, the difference between both models (reduced and individual model) was less than that of the single family (scenario 3). The RMSEs for positions of the QTL as shown in Table 2.6 were increased compared to the RMSEs of the third scenario of Table 2.5, because not all families are fully informative. The observed power

of the individual and the reduced model again almost reached 100 %. As expected, the estimated residual variance was inflated by the within marker genotype QTL variance.

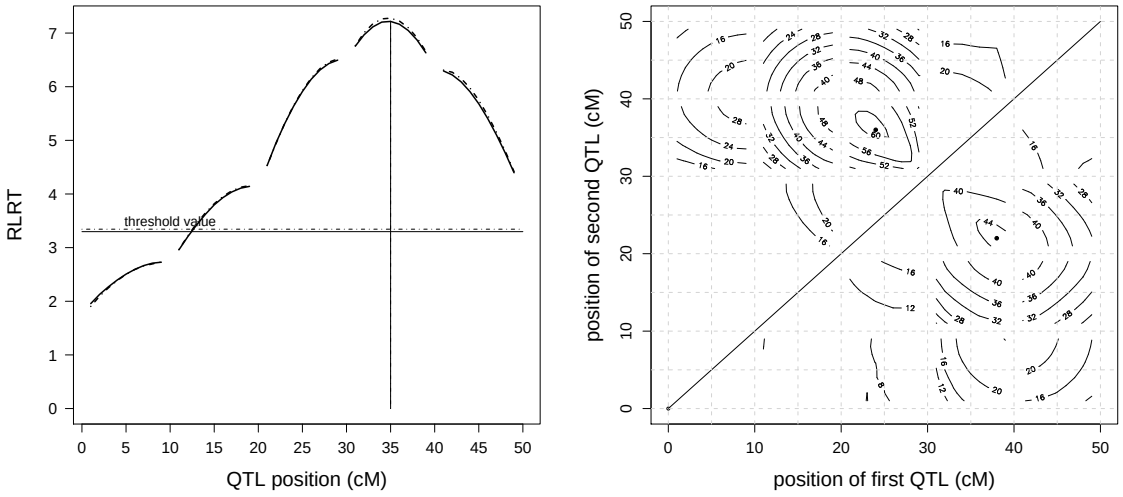
TABLE 2.6: Average estimates (mean) for QTL positions (P_1, P_2) with associated root mean squared error (RMSE) and quantiles together with mean estimates of the residual variance $\hat{\sigma}_r^2$ and the observed power (%) for the third scenario with 50 F_2 individuals for each of the four families per simulated experiment (1000 replications per scenario).

	Reduced model		Individual model	
	P_1	P_2	P_1	P_2
Mean	22.38	35.99	23.00	35.59
RMSE	6.77	4.78	6.26	4.27
5% quantile	8.00	28.00	9.00	28.00
95% quantile	28.00	45.00	28.00	44.00
$\hat{\sigma}_r^2$	0.21		0.18	
Power	99.50		99.60	

The required CPU time for ASReml (GILMOUR *et al.* 2008) of the reduced and the individual model was 26.7 and 80.1 sec for each repetition recorded on an HP DL380 G6 (72 GB RAM, 2× XEON X5570, 2.93 GHz, multiuser environment) in a two-QTL scenario with only additive genetic effects (four families), i. e. the individual model required threefold more computing time. The run time required for the evaluation of a single QTL (scenario 1 or 2) was sevenfold for the individual model compared with the reduced model for each repetition. If the number of individuals and the number of variance components increase, the speed gain of the reduced model relative to the individual model is expected to increase.

Average *RLRT* profiles from the reduced and the individual model were almost identical for the first scenario with a single QTL (Figure 2.1(a)). For two QTL in scenario 3 (Figure 2.1(b)), the shapes of the *RLRT* surfaces from both methods were again very similar, but the average size of the maximum was higher for the individual model (60.62 compared to 44.52). The *RLRT* surfaces of scenario 4 of the reduced and the individual model as well as the weighted reduced model are nearly identical (results not shown). The likelihood profile of the weighted approach was smaller than that of the reduced model, but QTL positions seemed to be estimated more accurately.

The considerable advantage of the reduced model with respect to computing time is achieved by a smaller number of genetic effects accompanied by a smaller dimension of their associated covariance matrices. Moreover, this dimension does not depend on the size of the experiment, in contrast to the individual model. The amount of saveable computing time can be expected to vary somewhat between different REML algorithms. Average information (AI) REML (GILMOUR *et al.* 1995; JOHNSON and



(a) (b)

FIGURE 2.1: For a single QTL (scenario 1) average *RLRT* profiles 2.1(a) of the individual model (dashed line) and the reduced model (solid line) nearly coincide, so do their significance thresholds. When two QTL were present (scenario 3), contour plots 2.1(b) of the *RLRT* surfaces from the reduced model (below diagonal) and the individual model (above diagonal) showed a similar shape, but different absolute heights (respective *RLRT* maxima 44.52 and 60.62). Averaging was over 1000 replications.

THOMPSON 1995) may be implemented either in an MME-based version or as a variant requiring the inversion of the covariance matrix $\boldsymbol{V}$ of phenotypes, termed the "direct method" by LEE and VAN DER WERF (2006). These authors recommend the direct method if genetic covariance matrices are dense because of both speed and numerical stability. Application of the Sherman-Morrison-Woodbury matrix identity (e. g. HENDERSON and SEARLE 1981; XU 1998) to determine the inverse of $\boldsymbol{V}$ results in

$$\boldsymbol{V}^{-1} \;=\; (\boldsymbol{HGH}' + \boldsymbol{R})^{-1} \;=\; \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{H}\left(\boldsymbol{G}^{-1} + \boldsymbol{H}'\boldsymbol{R}^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}'\boldsymbol{R}^{-1}\,,$$

where $\boldsymbol{R}$ denotes the covariance matrix of residuals, $\boldsymbol{G}$ is the covariance matrix of all genetic effects (block diagonal), and $\boldsymbol{H}$ is the corresponding incidence matrix. To obtain $\boldsymbol{V}^{-1}$ the inversion of a dense matrix of the same order as $\boldsymbol{G}$ is required, which usually is considerably smaller than the number of observations for the reduced model (e. g. $\dim(\boldsymbol{G}) = 9 \times 9$ for a single QTL with additive genetic effects and $\dim(\boldsymbol{G}) = 36 \times 36$ for two QTL with additive genetic and dominant effects). In conclusion, the increase in computing speed obtained by the reduced model may differ between algorithms, but is substantial when compared with the individual model, thus

broadening the general applicability of the VCM for mapping purposes.

The amount of possible improvement of the reduced model obtained by accounting for genetic sampling variation within marker classes remains to be investigated. A more comprehensive comparison of methods than presented here is underway to obtain a more complete picture. Despite the limited number of scenarios in our simulations, it can already be concluded that the proposed reduced model may be competitive with other standard methods for mapping of (multiple) QTL not only in terms of computing time, but also in terms of detection power and precision of estimated positions of the QTL.

The authors thank the reviewers for their helpful comments and suggestions. This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, MA 1553/3-1).

LITERATURE

Abdel-Azim, G. and A. E. Freeman, 2001 A rapid method for computing the inverse of the gametic covariance matrix between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **33**: 153–173.

Carbonell, E. A., T. M. Gerig, E. Balansard, and M. J. Asins, 1992 Interval mapping in the analysis of nonadditive quantitative trait loci. Biometrics **48**: 305–315.

Carlborg, Ö. and C. S. Haley, 2004 Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. **5**: 618–625.

Cockerham, C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics **39**: 859–882.

Crepieux, S., C. Lebreton, B. Servin, and G. Charmet, 2004 Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. Genetics **168**: 1737–1749.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson, 2008 *ASReml User Guide Release 3.0*. VSN International, Hemel Hempstead, UK.

Gilmour, A. R., R. Thompson, and B. R. Cullis, 1995 Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics **51**: 1440–1450.

Grignola, F. E., I. Hoeschele, and B. Tier, 1996 Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. Methodology. Genet. Sel. Evol. **28**: 479–490.

Haldane, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. **8**: 299–309.

Haley, C. S. and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–324.

Haseman, J. K. and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2**: 3–19.

Henderson, H. V. and S. R. Searle, 1981 On deriving the inverse of a sum of matrices. SIAM Rev. Soc. Ind. Appl. Math. **23**: 53–60.

Jannink, J.-L. and R. Jansen, 2001 Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **157**: 445–454.

Johnson, D. L. and R. Thompson, 1995 Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. J. Dairy Sci. **78**: 449–456.

Kao, C.-H., 2000 On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics **156**: 855–865.

Kao, C.-H. and Z.-B. Zeng, 1997 General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53**: 653–665.

Kao, C.-H. and Z.-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160**: 1243–1261.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

Lander, E. S. and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.

Lee, S. H. and J. H. J. Van der Werf, 2006 An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genet. Sel. Evol. **38**: 25–43.

Li, G. and Y. Cui, 2009 A statistical variance components framework for mapping imprinted quantitative trait locus in experimental crosses. J. Probab. Stat. **2009**: 1–27.

Liu, Y., G. B. Jansen, and C. Y. Lin, 2002 The covariance between relatives conditional on genetic markers. Genet. Sel. Evol. **34**: 657–678.

Martínez, O. and R. N. Curnow, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85**: 480–488.

Mayer, M., 2005 A comparison of regression interval mapping and multiple interval mapping for linked QTL. Heredity **94**: 599–605.

Mayer, M., Y. Liu, and G. Freyer, 2004 A simulation study on the accuracy of position and effect estimates of linked QTL and their asymptotic standard deviations using multiple interval mapping in an F_2 scheme. Genet. Sel. Evol. **36**: 455–479.

Neumaier, A., 1998 Solving ill-conditioned and singular linear systems: A tutorial on regularization. SIAM Rev. Soc. Ind. Appl. Math. **40**: 636–666.

Patterson, H. D. and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika **58**: 545–554.

Rönnegård, L., K. Mischenko, S. Holmgren, and Ö. Carlborg, 2007 Increasing the efficiency of variance component quantitative trait loci analysis by using reduced-rank identity-by-descent matrices. Genetics **176**: 1935–1938.

Rönnegård, L., R. Pong-Wong, and Ö. Carlborg, 2008 Defining the assumptions underlying modeling of epistatic QTL using variance component methods. J. Hered. **99**: 421–425.

Smith, S. P., 1984 *Dominance Relationship Matrix and Inverse for an Inbred Population*. Mimeo, Department of Dairy Science, Ohio State University, Columbus, OH.

Wang, T., R. L. Fernando, S. van der Beek, M. Grossman, and J. A. M. van Arendonk, 1995 Covariance between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **27**: 251–274.

Xie, C., D. D. Gessler, and S. Xu, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics **149**: 1139–1146.

Xu, S., 1995 A comment on the simple regression method for interval mapping. Genetics **141**: 1657–1659.

Xu, S., 1996 Mapping quantitative trait loci using four-way crosses. Genet. Res. **68**: 175–181.

Xu, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics **148**: 517–524.

Xu, S. and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141**: 1189–1197.

CHAPTER TWO

COMPETITIVENESS OF A REDUCED RANDOM MODEL VERSUS FIXED AND RANDOM ALTERNATIVES FOR MAPPING MULTIPLE QTL IN F_2 POPULATIONS

Daisy Zimmer, Manfred Mayer, Norbert Reinsch

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

ABSTRACT

Recently in the literature a computationally tractable random model (RRM) has been proposed for mapping multiple QTL in F_2 populations in the presence of additive and nonadditive genetic effects. The RRM approximates genetic covariances by replacing individual genetic effects by average genetic effects for each marker class. By simulating a series of line cross mapping experiments a comprehensive comparison between the RRM and other standard methods was obtained. Additional methods were a random model with individual genetic effects and two fixed model approaches: regression interval mapping and multiple interval mapping. The underlying genetic model considered two linked QTL with additive and additive-by-additive genetic effects. Our simulations show that the RRM is a competitive method to map multiple linked QTL with additive and nonadditive genetic effects. Criteria of evaluation were the observed power of one- and two-QTL models, the root mean squared errors of estimated QTL positions and effects as well as the deviation of the estimated residual variance from its simulated value. The RRM exhibits no major difference to the other methods with respect to QTL detection power and accuracy of estimated effects. The RRM however, clearly outperforms the individual random model with respect to computational speed and is therefore recommended for application.

INTRODUCTION

Estimation of genetic parameters and mapping of quantitative trait loci (QTL) are important for detection of individual loci responsible for quantitative genetic variation. Interactions of QTL alleles within (dominance) or between (epistasis) loci have to be taken into account. Multiple QTL models contribute to understand the genetic architecture of quantitative traits and their variation in a population (Zeng *et al.* 1999). Considering multiple QTL simultaneously improves the power to detect linked QTL (e. g. Kao and Zeng 2002; Mayer *et al.* 2004; Mayer 2005) and avoids the so-called "ghost QTL" phenomenon (e. g. Haley and Knott 1992). The precision of the estimated QTL positions and effects are possibly improved (e. g. Jannink and Jansen 2001; Carlborg and Haley 2004). Numerous methods for mapping QTL have been proposed in the literature as non-parametric methods (e. g. Kruglyak and Lander 1995), least-squares based methods (e. g. Haley and Knott 1992; Martínez and Curnow 1992), maximum likelihood (ML) based methods (e. g. Lander and Botstein 1989) and Bayesian methods (e. g. Thomas and Cortessis 1992; Hoeschele and VanRaden 1993a,b).

Lander and Botstein (1989) suggested a ML based method of interval mapping (IM) for QTL mapping using flanking marker information in F_2 and backcross (BC) populations derived from inbred lines, hereafter called F_2 or BC. An extension of IM considers fixed additive genetic, dominance and epistatic QTL effects as parts of the likelihood function for a mixture model, called multiple interval mapping (MIM, Kao and Zeng 1997; Kao *et al.* 1999). An approximation of IM, called regression interval mapping (RIM), was proposed by Haley and Knott (1992) and Martínez and Curnow (1992) in F_2 and BC designs. The advantages of RIM are its simplicity and its computational speed, whereby many parameters can be fitted simultaneously.

For a single-QTL model there are several studies which show that RIM and ML based methods produced nearly the same results (e. g. Haley and Knott 1992; Martínez and Curnow 1992; Xu 1995, 1998a,b; Kao 2000) in experimental populations. Considering multiple linked QTL, these fixed models (e. g. RIM and MIM) are again well suited to detect multiple, possibly interacting QTL in experimental populations. However, the accuracy of estimated QTL positions and effects from RIM is less compared to MIM in some situations (e. g. QTL in repulsion; Kao 2000; Mayer *et al.* 2004; Mayer 2005). By simulations and analytical treatments Kao (2000) investigated RIM and MIM. It was shown that the similarity of both methods depends, among others, on interval size, difference between QTL effects, intensity of epistasis and distance between QTL in BC populations. In the work of Kao (2000) QTL parameters were estimated and likelihood test statistics were determined at true QTL positions. The estimation of QTL positions is also an important issue in QTL mapping with known differences between methods. In a comparison of MIM and RIM, Mayer *et al.* (2004) and Mayer (2005) pointed out that in some situations, especially with (closely) linked QTL or QTL in repulsion, RIM may produced unsuitable and inaccurate parameter estimates, particularly QTL effects. There is, to the authors knowledge, no report on the precision of estimated QTL positions from RIM and MIM in presence of epistatic interactions. The residual variance was clearly overestimated (e. g. Mayer 2005) by RIM and therefore the relative QTL variance was underestimated. A theoretical derivation of the bias of the residual variance in RIM was given by Xu (1995) and Xu (1998c), emphasizing that the residual variance is inflated by the within marker genotype QTL variance.

Linear mixed models with random additive genetic effects are applied for mapping QTL in a broad range of species (Xu and Atchley 1995; Grignola *et al.* 1996a,b; Xie *et al.* 1998; Crepieux *et al.* 2004; Li and Cui 2009; Zimmer *et al.* 2011). Such variance component methods (VCM) are especially advantageous in scenarios with a large number of small families. Multiple line cross experiments were often evaluated

using the VCM (e. g. Xie *et al.* 1998; Xu 1998c; Crepieux *et al.* 2004; Li and Cui 2009; Zimmer *et al.* 2011), because different lines are considered simultaneously for QTL detection to avoid the so-called "genetic drift error" (Xu 1996). Xu (1998c) compared fixed and random models in multiple families of line crosses for QTL mapping by extensive Monte Carlo simulations. Fixed models provided good estimates of genetic parameters and QTL positions. Random models were recommended in combining data from a large number of families. In the literature alternative procedures were proposed to set up the additive genetic relationship matrices required for covariance structures from marker data and pedigree information in inbred populations (e. g. Wang *et al.* 1995; Abdel-Azim and Freeman 2001) as well as the marker-based relationship matrices for QTL with additive and nonadditive genetic (dominance, epistasis) effects in noninbred populations (e. g. Liu *et al.* 2002). Xie *et al.* (1998) suggested a VCM with individual genetic effects, called individual random model (IRM) in Zimmer *et al.* (2011), to map QTL with respect to additive genetic and dominance effects. Xie *et al.* (1998) set up the required additive genetic and dominance relationship matrices from conditional QTL genotype probabilities (given the flanking marker genotypes) to map QTL in BC and F_2 designs. Their efficient approach was extended to relationship matrices of pairwise epistatic effects (Zimmer *et al.* 2011). A reduced version of the random model (RRM) replaces individual genetic effects by average genetic effects for different marker classes in experimental populations, offering considerable savings in computing time (Zimmer *et al.* 2011).

In this article we compare the recently developed RRM with the VCM which employs individual genetic effects (IRM) and the fixed models RIM and MIM in terms of their ability to map QTL in an F_2 design. Multiple linked QTL with additive and additive-by-additive genetic effects were studied, where the QTL are in coupling, in repulsion or without main effects. The investigated methods were compared, particularly in terms of their observed power, accuracy of estimated QTL positions and effects.

MATERIAL AND METHODS

We used an F_2 design derived from two inbred lines (single line cross), where alternative alleles in the parental lines were fixed for markers and QTL. Using flanking markers of the F_2 individuals the conditional QTL genotypes were inferred from Haldane's mapping function (Haldane 1919). Double recombinations were taken into account. Considering pairwise epistatic effects, the joint conditional QTL genotype probabilities are the product of the marginal conditional QTL genotype probabilities of individual QTL, because we assumed that there is at most one QTL within a marker interval

(e. g. Xu and Atchley 1995; Rönnegård *et al.* 2008). Note that if both QTL are in the same marker interval, they are not independent and the joint conditional QTL genotype probabilities can not be obtained this way (e. g. Mayer 2007). Furthermore, we assumed that location of QTL and position of the markers do not coincide. The conditional probabilities of QTL genotypes given the nine flanking marker genotypes of a single QTL have been provided by several authors, see e. g. Haley and Knott (1992). If two QTL are in adjacent marker intervals, the number of joint flanking markers is reduced. Instead of 81 possible flanking marker genotypes there are only 27 marker genotypes, because both QTL share a flanking marker.

Regression interval mapping (RIM): RIM is a least-squares method and was developed for BC experiments by Martínez and Curnow (1992) and for F_2 designs by Haley and Knott (1992). In RIM phenotypic values of offspring are regressed on the coefficients of the genetic effects (additive and nonadditive genetic) for putative QTL at a fixed position or a combination thereof. The coefficients of the genetic effects model are the conditional expectations of QTL genotypes given the flanking marker genotypes derived for the presumed positions of the QTL.

Multiple interval mapping (MIM): MIM was suggested by Kao and Zeng (1997) to map multiple QTL with respect to additive and nonadditive genetic effects in experimental populations. Further descriptions of MIM are given by Kao *et al.* (1999), Zeng *et al.* (1999) and Kao (2000). The statistical model including additive genetic, dominance and pairwise epistatic effects for F_2 populations is shown by, e. g., Mayer *et al.* (2004) with detailed explanation of the different components. MIM is a ML based method, which uses multiple marker intervals simultaneously for QTL mapping. The likelihood function of MIM is a finite mixture of 3^ν (2^ν) normal distributions for the F_2 (BC) from inbred populations, where ν is the number of QTL. Kao and Zeng (1997) proposed general formulas to obtain the maximum likelihood estimates using an expectation maximization algorithm (Dempster *et al.* 1977). Note that the original approach of MIM (Kao and Zeng 1997; Kao *et al.* 1999) ignores double recombination. While the distribution of the phenotypic trait in each marker genotype class is a mixture of normal distributions according to the QTL genotype, RIM approximates this mixture distribution by a single normal one.

Variance component method (VCM) with an individual model (IRM): In contrast to RIM and MIM, the VCM deals with a random model. For F_2 populations, the required additive genetic and dominance relationship matrices can be derived from

elementary covariance matrices and conditional QTL genotype probabilities (Xie *et al.* 1998). An extension to interacting QTL with pairwise epistatic effects is given in Zimmer *et al.* (2011). The dimension of all relationship matrices is $n \times n$, where n is the number of F_2 individuals and individual genetic effects are estimated (IRM). If n is large or multiple QTL are considered simultaneously with additive and nonadditive genetic effects, the IRM is computationally demanding or infeasible.

Reduced model VCM (RRM): Here F_2 individuals are grouped according to their marker genotypes. In the reduced model (RRM) an average conditional genotypic effect for individuals within the same marker class is estimated instead of individual genetic effects. The RRM and the IRM are asymptotically equivalent (Zimmer *et al.* 2011). To set up the required relationship matrices, again elementary covariance matrices and conditional QTL genotype probabilities can be used. The main advantage of the RRM is the decreased dimension of all required relationship matrices, being independent of the number of F_2 individuals and only determined by the number of different marker classes. The dimension of the additive genetic and dominance relationship matrices is 9×9 in our case. If two QTL are located in adjacent marker intervals, the dimension of the relationship matrices for pairwise epistatic effects is 27×27 and 81×81 otherwise.

Simulations: Comprehensive simulations of mapping experiments with a single F_2 family were done to explore the characteristics of the different methods. Two linked QTL ($\nu = 2$) at positions $P_1 = 35$ and $P_2 = 45\,\text{cM}$ on a single chromosome of length 50 cM with six markers at 10 cM spacing ($0, 10, 20, 30, 40, 50\,\text{cM}$) were postulated. Four different scenarios were studied. Additive (a_1, a_2) and additive-by-additive genetic effects (aa) as well as the residual variance σ_e^2 were simulated as shown in Table 3.1. All other types of genetic effects were assumed zero. Additive genetic effects of both QTL and additive-by-additive effects contribute to the genetic variance σ_{QTL}^2. According to Kao and Zeng (2002, Equation 34) it is $\sigma_{QTL}^2 = \frac{1}{2}\,a_1^2 + \frac{1}{2}\,a_2^2 + \frac{1}{4}\,aa^2 + \lambda a_1 a_2$, where $\lambda = 1 - 2\theta$ is the linkage parameter and θ is the recombination rate between QTL. Each simulated scenario varied the number of F_2 individuals ($n \in \{200, 500\}$) and the relative QTL variance R^2 ($R^2 \in \{0.05, 0.10, 0.25\}$). Here R^2 is the proportion of the phenotypic variance which is explained by QTL, i. e. the broad sense heritability.

RRM was compared to IRM, RIM and MIM in terms of the observed power and the accuracy of the estimated QTL positions and effects. Means ($\bar{P}_1$, $\bar{P}_2$) and root mean squared errors (RMSEs) of the estimated QTL positions ($\hat{P}_1$, $\hat{P}_2$) were evaluated to characterize the precision of location estimates. The average estimated residual vari-

TABLE 3.1: Summary of simulated scenarios. Additive genetic effects (a_1, a_2) and additive-by-additive genetic effects (aa) as well as QTL variances (σ^2_{QTL}) are listed. Residual variances σ^2_e are shown for each relative QTL variance R^2 (%).

					σ^2_e		
scenario	a_1	a_2	aa	σ^2_{QTL}	$R^2 = 5$	$R^2 = 10$	$R^2 = 25$
1	1.0	1.0	1.0	2.07	39.30	18.61	6.20
2	1.0	−1.0	1.0	0.43	8.17	3.87	1.29
3	0.5	−1.0	1.0	0.47	8.81	4.17	1.39
4	0.0	0.0	1.0	0.25	4.75	2.25	0.75

ance ($\hat{\sigma}^2_r$) from all runs also was documented. To quantify the number of identified QTL (one or two QTL) the observed power was determined as proportion of replications that exceeded an empirical threshold. The RMSE was calculated to assess the precision of parameter estimates and is defined as the square root of the expected value of the squared difference between the estimator and the true parameter. RMSEs were determined for QTL effects and positions.

In contrast to fixed models, the RRM is a random model that estimates conditional genetic values for each marker class or a combination thereof, dependent on the putative QTL positions. For the purpose of comparison, the estimated genetic effects of the QTL obtained from RIM and MIM were transformed into conditional genotypic values (sum of effects) of marker classes. The estimated genotypic values of QTL genotypes were stored in a vector $\boldsymbol{g}$ with $\dim(\boldsymbol{g}) = 9 \times 1$. Let $\boldsymbol{L}$ be the matrix with conditional QTL genotype probabilities with one row for each marker class and one column for each QTL genotype. The vector of conditional genotypic effects $\boldsymbol{c}$ was obtained as $\boldsymbol{c} = \boldsymbol{L}\boldsymbol{g}$. The simulated genotypic value for a certain marker class was calculated as mean of the genotypic values of all individuals given the flanking marker genotypes.

The simulated data were evaluated with own Fortran 95 programs for RIM, RRM and IRM. ASReml (GILMOUR *et al.* 2008) was used for the estimation of variance components in the VCM and for the evaluation of the restricted maximum likelihood function (PATTERSON and THOMPSON 1971). The implementation of MIM as described in more detail by MAYER *et al.* (2004) was applied.

Simulations were repeated $N = 500$ times with one exception: The IRM with a family size of $n = 500$ was computationally demanding and therefore only $N = 50$ runs were performed and all related estimates were marked by asterisk in the following. All scenarios were analyzed using a one- and a two-QTL model as explained in the next paragraph.

Hypothesis testing: The simulated data were analyzed by a one-QTL model (only an additive genetic QTL effect) and a two-QTL model (additive genetic effects for each QTL and an additive-by-additive effect of both QTL) to determine the number of identified QTL. The one-QTL model was "nested" in the two-QTL model. More complicated models were not considered.

Testing for the presence of a single QTL with an additive genetic effect in the fixed model is equivalent to testing the corresponding variance component being larger than zero in the random model. The testing problem for a single QTL is

$$H_0^{(1)} : \sigma^2_{a_1} \;=\; 0 \qquad vs. \qquad H_A^{(1)} : \sigma^2_{a_1} \;>\; 0\,. \tag{3.1}$$

Considering two linked QTL with additive genetic effects at both QTL and an additive-by-additive effect ($\tau \in \{a_1, a_2, aa\}$) the testing problem of (3.1) can be extended to

$$H_0^{(2)} : \forall\tau,\ \sigma^2_\tau \;=\; 0 \qquad vs. \qquad H_A^{(2)} : \exists\tau,\ \sigma^2_\tau \;>\; 0\,. \tag{3.2}$$

Both QTL positions were estimated simultaneously in a two-dimensional search procedure. The rejection of $H_0^{(1)}$ or $H_0^{(2)}$ leads to the general statement of a marked QTL. But the rejection of the null hypothesis $H_0^{(2)}$ substantiate only that there is at least one variance component larger than zero. Aiming for inferences on the number of QTL, it makes sense to test two QTL independently of the requirement that a single QTL was identified previously, e. g. for the detection of QTL in repulsion or two QTL with only epistasis. Determining the number of identified QTL (two QTL *vs.* one QTL) leads to the testing problem

$$\begin{aligned} H_0^{(12)} :&\ \left((\sigma^2_{a_1} > 0 \ \wedge\ \sigma^2_{a_2} = 0) \ \vee\ (\sigma^2_{a_1} = 0 \ \wedge\ \sigma^2_{a_2} > 0)\right) \ \wedge\ (\sigma^2_{aa} = 0) \\ &\qquad vs. \\ H_A^{(12)} :&\ (\sigma^2_{a_1} > 0 \ \wedge\ \sigma^2_{a_2} > 0) \ \vee\ (\sigma^2_{aa} > 0)\,. \end{aligned} \tag{3.3}$$

The testing problem (3.3) considers two different states of both QTL. Two QTL were identified if both QTL had main QTL effects different from zero ($\sigma^2_{a_1} > 0$, $\sigma^2_{a_2} > 0$) or if the epistatic effect between both loci was positive ($\sigma^2_{aa} > 0$). The model under $H_0^{(12)}$ is equivalent to a single QTL model.

For simplification of the testing problem (3.3) we used the maximum values of the test statistics of (3.1) and (3.2) to test for a second QTL which contributes significantly to the variation of the phenotypes, regardless of the decision of the testing problem of a

single QTL (3.1). In this way the maximum test statistic of the one-QTL model and the two-QTL model of a certain run were used to create the corresponding test statistic of hypothesis test (3.3). There was only one QTL identified if the null hypothesis $H_0^{(1)}$ was rejected but not $H_0^{(12)}$. Two QTL were identified if either both null hypotheses $H_0^{(1)}$ and $H_0^{(12)}$ were rejected or $H_0^{(1)}$ was accepted and $H_0^{(12)}$ was rejected. There was no evidence for any QTL if both null hypotheses $H_0^{(1)}$ and $H_0^{(12)}$ were not rejected.

For each replication we successively analyzed (step width 1 cM) putative QTL positions along the chromosome or combinations thereof, and determined the maximum value of the test statistic. We assumed that both QTL were in different marker intervals and also that marker locations and QTL positions did not coincide, because the relationship matrices of the VCM are singular there. Therefore, we modified RIM and MIM to work with the same restrictions as the VCM. The test statistics were the residual likelihood ratio test statistic $RLRT$ of VCM, the likelihood ratio test statistic LRT of MIM and the F-test statistic F of RIM. In this way a test statistic profile along the chromosome was produced with gaps at the marker positions. Note that under the null hypotheses the genetic variance components lie on the boundaries of the parameter space. The null distribution of the test statistic was determined via simulation.

Chromosome-wide thresholds: To take into account multiple testing in QTL detection, empirical thresholds should be determined. For that purpose simulated data under the null hypothesis of no marked QTL were used to derive the null distribution of the test statistic and to obtain an empirical threshold. The empirical chromosome-wide threshold (type I error $\alpha = 0.05$) was obtained by choosing the 95 th percentile of the empirical distribution function using all determined maximum values of the test statistic from N ($N = 500$, with exception as mentioned above) simulations under the null hypothesis of no segregating QTL. The residual variance under the null hypothesis of no linked QTL was chosen in such a way that the phenotypic variance remained unchanged. The observed power to detect QTL was determined as the percentage of runs where the test statistic exceeded the corresponding empirical threshold.

RESULTS

Observed power: The observed power and the number of identified QTL are shown in Table 3.2. In scenario 1 (QTL in coupling) the VCM (RRM, IRM) had a higher power to detect both QTL than fixed models (RIM, MIM) which produced nearly the same observed power of QTL detection. Observed power values for one and two QTL showed a general tendency to identify a single QTL instead of two, when there was

a lack of information (small R^2 and n) in the experiment. This tendency was less pronounced for the VCM. Only if the number of individuals was $n = 500$ and the relative QTL variance was $R^2 = 25\,\%$, two QTL were identified correctly with a power ≥ 94.2. In the second scenario with QTL effects of the same size, but in repulsion ($a_1 = -a_2 = aa = 1.0$), the two-QTL model always was identified with a higher power compared to the single-QTL model. Observed power for two QTL always was $> 42\,\%$, while the single QTL power never exceeded $5\,\%$. Comparison of observed power values of the four methods showed very similar results and no remarkable differences. The observed power to detect both QTL in scenario 3 with unequal QTL effects in repulsion ($a_1 = 0.5$, $a_2 = -aa = -1.0$) was again very similar for all methods, except for small experiments ($n = 200$) and little genetic variance ($R^2 = 5\,\%$ and $10\,\%$), where the VCM showed a gain in power of $\sim 5\,\%$. Accordingly their power to detect a single QTL was smaller than for fixed models. In the fourth scenario with only an epistatic additive-by-additive genetic effect, significant identifications of a single QTL were very rare ($\leq 2.4\,\%$) in all methods. Ability to identify both QTL showed almost no variation between methods.

Over all scenarios, the RRM essentially showed the same power to detect two QTL compared to the IRM. Differences in scenario 1 in favor of the RRM should be considered with caution, because of the limited number of runs for the IRM. Therefore, the RRM provides equal capabilities for QTL detection, but with a considerably lower computational workload.

Estimated QTL positions from single-QTL models: The average estimated QTL positions from significant runs of single-QTL models and their standard deviations (SDs) are shown in Table 3.3 for selected cases of the first three scenarios ($R^2 = 5\,\%$), where the power to detect a single QTL was $\geq 4.0\,\%$ and $N = 500$.

In scenario 1 the mean estimated position of a single QTL ranged from 36.43 (IRM) to 36.79 cM (RRM). The SDs were ≈ 11 cM. Hence, the position of the QTL covered almost the hole chromosome. If the number of individuals increased from $n = 200$ to 500, the average estimated QTL position moved in the middle between both simulated QTL and a so-called ghost QTL (e. g. Haley and Knott 1992) was identified. In the second scenario the position of the QTL was in the middle of the simulated chromosome, but again SDs were quite high (≈ 16 cM). The estimated QTL positions in scenario 3 with $n = 200$ were very similar between methods with SDs ranging from 13 (MIM) to 14 cM (RRM). If the number of individuals was $n = 500$, the average estimated QTL position moved towards the simulated position of the second QTL

TABLE 3.2: Observed power (%) to identify one or two QTL of RRM, IRM, RIM and MIM for the four scenarios. The number of individuals (n) per experiment were 200 and 500. The relative QTL variances (R^2) were 5, 10 and 25 %. The observed power based on $N = 500$ runs, except for the values indicated by an asterisk ($N = 50$).

R^2	RRM		IRM		RIM		MIM	
(%)	1 QTL	2 QTL	1 QTL	2 QTL	1 QTL	2 QTL	1 QTL	2 QTL
				Scenario 1				
				$n = 200$				
5	66.0	11.4	66.4	11.0	68.4	9.4	71.0	5.8
10	73.2	24.2	74.6	22.0	81.4	15.8	86.2	11.0
25	33.6	66.4	34.6	65.4	50.2	49.8	54.6	45.4
				$n = 500$				
5	67.2	32.0	82.0*	18.0*	81.8	17.4	83.2	16.0
10	41.6	58.4	56.0*	44.0*	57.8	42.2	58.0	42.0
25	2.8	97.2	2.0*	98.0*	5.8	94.2	5.8	94.2
				Scenario 2				
				$n = 200$				
5	4.0	44.6	4.2	42.2	4.4	47.4	4.8	44.2
10	1.8	80.0	2.0	78.8	1.8	83.0	2.0	80.2
25	0.0	99.8	0.0	100.0	0.0	99.8	0.0	100.0
				$n = 500$				
5	1.2	92.0	0.0*	84.0*	1.4	91.4	1.6	91.2
10	0.2	99.8	0.0*	100.0*	0.0	100.0	0.2	99.8
25	0.0	100.0	0.0*	100.0*	0.0	100.0	0.0	100.0
				Scenario 3				
				$n = 200$				
5	16.2	42.8	17.0	41.0	18.6	37.8	20.8	35.6
10	14.0	74.0	14.2	73.2	17.6	68.4	17.8	69.2
25	0.4	99.6	0.0	100.0	0.2	99.8	0.4	99.6
				$n = 500$				
5	7.8	88.8	8.0*	86.0*	12.4	83.8	11.8	84.2
10	0.2	99.8	0.0*	100.0*	0.4	99.6	0.6	99.4
25	0.0	100.0	0.0*	100.0*	0.0	100.0	0.0	100.0
				Scenario 4				
				$n = 200$				
5	1.6	60.4	1.8	57.8	2.4	60.2	2.2	58.0
10	0.4	91.4	0.4	91.6	0.4	91.8	0.4	92.4
25	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
				$n = 500$				
5	0.2	98.4	0.0*	100.0*	0.4	98.0	0.6	97.8
10	0.0	100.0	0.0*	100.0*	0.0	100.0	0.0	100.0
25	0.0	100.0	0.0*	100.0*	0.0	100.0	0.0	100.0

TABLE 3.3: Mean estimated QTL positions based on significant repetitions with one-QTL models with associated standard deviations (SDs) from RRM, IRM, RIM and MIM for scenario 1, 2 and 3 with a relative QTL variance of $R^2 = 5\,\%$, n is the number of individuals per simulated experiment.

	$n = 200$				$n = 500$			
	RRM	IRM	RIM	MIM	RRM	IRM	RIM	MIM
				Scenario 1				
Mean	36.79	36.43	36.59	36.71	39.25	–	39.07	39.06
SD	10.72	11.14	10.90	10.84	6.56	–	6.75	6.73
				Scenario 2				
Mean	22.35	23.71	22.95	25.46				
SD	16.07	16.20	15.45	16.02				
				Scenario 3				
Mean	40.37	40.69	39.92	40.62	44.09	–	44.56	44.39
SD	13.94	13.67	13.55	13.04	10.77	–	9.12	9.31

($P_2 = 45\,\text{cM}$) for RRM, RIM and MIM. The reason is, of course, that the additive genetic effect of the second QTL ($a_2 = -1.0$) contributed more to the genetic variance than the first QTL ($a_1 = 0.5$). Although the average estimated QTL position was on the right side of the 50 cM chromosome, the SDs of the third scenario were surprisingly high also in comparison to scenario 1.

Parameter estimated from two-QTL models: The average estimated QTL positions based on significant runs of two-QTL models, the associated root mean squared errors (RMSEs) and the estimated residual variances $\hat{\sigma}_r^2$ can be found in Tables 3.4-3.7 for each scenario. The number of runs with significant detection of two QTL which identified the correct marker interval combination of the first QTL $\hat{P}_1 \in (30, 40]\,\text{cM}$ and the second QTL $\hat{P}_2 \in (40, 50]\,\text{cM}$ is denoted as s_{45}. The parameter s_{45} comprises both the observed power and the ability to locate QTL positions in the correct marker intervals.

Position estimates and root mean squared errors: A common observation for all scenarios was that mean estimates of QTL positions were biased to the left side of the chromosome and RMSEs increased by a factor of $\sim 1.5 - 3.5$ when the relative QTL variance R^2 was small. Lower numbers of individuals n lead to a similar inflation of RMSEs for constant R^2. RMSEs of the mean estimated first position $\bar{P}_1$ in all scenarios were higher than the RMSEs of the mean estimated second QTL $\bar{P}_2$, because both QTL were simulated on the right side of the chromosome. Therefore, $\bar{P}_1$ can vary

more on the left side of the chromosome.

The observed power of scenario 1 (Table 3.4) to identify two QTL was highest in RRM and IRM. Surprisingly the RMSEs for $n = 200$ at both estimated QTL positions were smaller than in RRM and IRM. MIM and RIM based on less significant replications and it was expected that their RMSEs will be smaller than those of the VCM. With $n = 500$, RMSE differences between the four methods nearly vanished and fluctuated around small values of $\sim 1\,\text{cM}$, with exceptions of some less reliable results for the IRM. The distance between both QTL (true distance: 10 cM) was overestimated and ranged from 18.59 (RRM) to 20.39 cM (IRM) for $n = 200$ and $R^2 = 5\,\%$. Mean estimates of QTL positions were, however, nearly identical to the simulated values with $n = 500$ and $R^2 = 25\,\%$. The corresponding estimated distance between both QTL accordingly decreased and ranged from 10.77 (RRM) to 11.98 cM (RIM). RMSEs of the first QTL positions were approximately twice as large as RMSEs of the second QTL positions in all methods and variations. Remarkably, the number of runs which identified the QTL positions in the correct marker interval s_{45} was highest for RRM and much higher than in RIM and MIM in this particular scenario.

Table 3.5 shows the results of scenario 2 from significant runs of two-QTL models. The average estimates of QTL positions were very similar among the investigated methods. The mean estimated QTL positions of the second QTL $\bar{P}_2$ were close to the simulated position. In contrast, the mean estimated first positions $\bar{P}_1$ were biased to the left side of the chromosome. The mean estimates of $\bar{P}_1$ became closer to the simulated value if R^2 and n were increased. Differences between RMSEs of QTL positions between methods did not exceed $\sim 1\,\text{cM}$. Identification of correct marker intervals s_{45} showed a different picture compared to scenario 1 with some tendency to better s_{45} values for RIM and MIM with $n = 200$ and smaller relative QTL variance. RMSEs as well as the related parameter s_{45} generally indicated a higher precision of mapping in the second scenario (QTL in repulsion) compared to the first scenario (QTL in coupling).

Analogous results of scenario 3 are shown in Table 3.6. Again, $\bar{P}_1$ was biased and closer to the simulated positions if R^2 and n increased. The position of the second QTL was mapped quite accurately, particularly when compared to scenario 1. In all methods, s_{45} and RMSEs showed little variation and no pattern in favor of any method.

Finally Table 3.7 gives an overview over estimates of QTL positions and their precision in the last scenario. Mean estimated QTL positions were close to simulated ones for all methods, with the largest deviations from true positions of $\sim 5\,\text{cM}$ for $n = 200$ and $R^2 = 5\,\%$. Distances between mean estimated QTL positions as well as RMSEs for QTL positions were very uniform over methods. Variations of s_{45} values between

methods was considerably smaller than in the previous scenarios.

TABLE 3.4: Scenario 1: Mean estimates of QTL positions based on significant runs with two-QTL models (simulated $P_1 = 35\,\text{cM}$, $P_2 = 45\,\text{cM}$) with associated root mean squared errors (RMSEs) and average estimated residual variances $\hat{\sigma}_r^2$ of RRM, IRM, RIM and MIM. The number of significant runs where the positions of both QTL were in the correct marker intervals is s_{45}. The number of individuals per simulated experiment is n and R^2 denotes the relative QTL variances. All values are based on $N = 500$ runs, except of values indicated by an asterisk ($N = 50$).

	RRM		IRM		RIM		MIM	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
				$n = 200$, $R^2 = 5\%$				
Mean	20.12	38.71	18.18	38.57	18.30	37.72	16.28	35.90
RMSE	19.96	10.46	21.17	10.03	21.14	11.55	22.56	13.47
$\hat{\sigma}_r^2$	37.70		36.07		37.80		36.20	
s_{45}	15		11		8		4	
				$n = 200$, $R^2 = 10\%$				
Mean	24.56	41.76	24.15	41.45	22.89	40.35	21.20	39.38
RMSE	16.73	7.33	17.02	7.45	17.46	9.19	18.69	9.42
$\hat{\sigma}_r^2$	18.04		17.41		18.03		17.09	
s_{45}	53		47		26		15	
				$n = 200$, $R^2 = 25\%$				
Mean	30.26	43.67	30.26	43.80	28.71	43.61	29.26	43.97
RMSE	10.77	4.31	10.82	4.25	11.65	5.26	11.40	4.81
$\hat{\sigma}_r^2$	6.13		5.86		6.16		5.86	
s_{45}	217		214		138		135	
				$n = 500$, $R^2 = 5\%$				
Mean	28.19	43.03	31.94*	43.50*	27.72	42.15	27.54	42.83
RMSE	13.46	5.47	6.74*	4.98*	12.88	6.50	13.35	6.05
$\hat{\sigma}_r^2$	39.08		38.17*		38.77		38.09	
s_{45}	94		6*		44		42	
				$n = 500$, $R^2 = 10\%$				
Mean	30.70	43.88	30.68*	44.95*	29.99	43.84	29.73	43.82
RMSE	10.36	3.96	10.29*	2.81*	10.32	4.92	10.65	4.72
$\hat{\sigma}_r^2$	18.58		18.05*		18.58		18.21	
s_{45}	199		14*		128		131	
				$n = 500$, $R^2 = 25\%$				
Mean	33.80	44.57	33.92*	45.20*	32.80	44.78	33.03	44.74
RMSE	5.47	2.73	4.46*	2.16*	6.24	3.19	5.79	2.90
$\hat{\sigma}_r^2$	6.28		6.05*		6.30		6.12	
s_{45}	418		41*		374		390	

TABLE 3.5: Scenario 2: Mean estimates of QTL positions based on significant runs with two-QTL models (simulated $P_1 = 35\,\text{cM}$, $P_2 = 45\,\text{cM}$) with associated root mean squared errors (RMSEs) and average estimated residual variances $\hat{\sigma}_r^2$ of RRM, IRM, RIM and MIM. The number of significant runs where the positions of both QTL were in the correct marker intervals is s_{45}. The number of individuals per simulated experiment is n and R^2 denotes the relative QTL variances. All values are based on $N = 500$ runs, except of values indicated by an asterisk ($N = 50$).

	RRM		IRM		RIM		MIM	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
				$n = 200$, $R^2 = 5\%$				
Mean	27.89	43.56	27.75	43.49	29.69	43.24	28.86	42.98
RMSE	12.60	7.83	12.71	8.02	11.83	7.50	12.32	8.07
$\hat{\sigma}_r^2$	8.02		7.79		8.06		7.66	
s_{45}	116		108		150		134	
				$n = 200$, $R^2 = 10\%$				
Mean	30.74	44.52	31.18	44.50	32.05	44.21	32.02	44.32
RMSE	9.09	6.11	8.60	5.76	8.34	5.41	7.70	5.29
$\hat{\sigma}_r^2$	3.91		3.75		3.92		3.69	
s_{45}	263		268		297		290	
				$n = 200$, $R^2 = 25\%$				
Mean	33.26	45.61	33.69	45.49	34.19	45.01	34.14	45.07
RMSE	4.41	3.21	3.90	2.81	4.19	2.98	3.54	2.48
$\hat{\sigma}_r^2$	1.39		1.26		1.41		1.26	
s_{45}	415		443		435		458	
				$n = 500$, $R^2 = 5\%$				
Mean	32.02	44.88	30.81*	45.23*	33.02	44.19	32.84	44.19
RMSE	6.93	4.53	8.04*	3.94*	6.45	4.62	6.44	4.46
$\hat{\sigma}_r^2$	8.25		8.01*		8.26		8.03	
s_{45}	331		26*		360		359	
				$n = 500$, $R^2 = 10\%$				
Mean	33.44	45.31	33.22*	45.42*	34.24	44.69	33.93	44.76
RMSE	4.19	3.20	3.93*	3.04*	3.99	3.12	4.06	2.92
$\hat{\sigma}_r^2$	3.98		3.82*		3.99		3.83	
s_{45}	424		40*		444		442	
				$n = 500$, $R^2 = 25\%$				
Mean	34.41	45.43	34.40*	45.40*	34.87	45.02	34.73	44.86
RMSE	2.16	1.64	1.91*	1.48*	2.06	1.71	1.53	1.39
$\hat{\sigma}_r^2$	1.42		1.28*		1.43		1.29	
s_{45}	482		49*		487		498	

TABLE 3.6: Scenario 3: Mean estimates of QTL positions based on significant runs with two-QTL models (simulated $P_1 = 35\,\text{cM}$, $P_2 = 45\,\text{cM}$) with associated root mean squared errors (RMSEs) and average estimated residual variances $\hat{\sigma}_r^2$ of RRM, IRM, RIM and MIM. The number of significant runs where the positions of both QTL were in the correct marker intervals is s_{45}. The number of individuals per simulated experiment is n and R^2 denotes the relative QTL variances. All values are based on $N = 500$ runs, except of values indicated by an asterisk ($N = 50$).

	RRM		IRM		RIM		MIM	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
				$n = 200$, $R^2 = 5\%$				
Mean	27.50	43.66	27.23	43.17	27.89	43.30	26.99	42.71
RMSE	13.74	6.94	14.10	7.86	13.66	7.28	14.04	8.37
$\hat{\sigma}_r^2$	8.57		8.37		8.61		8.22	
s_{45}	111		108		108		93	
				$n = 200$, $R^2 = 10\%$				
Mean	30.68	45.06	30.82	44.89	30.91	44.37	30.78	44.51
RMSE	10.28	4.65	10.06	4.73	9.96	4.99	9.88	4.66
$\hat{\sigma}_r^2$	4.16		4.02		4.17		3.96	
s_{45}	242		245		227		236	
				$n = 200$, $R^2 = 25\%$				
Mean	33.57	45.59	33.90	45.47	33.88	45.10	33.73	45.13
RMSE	5.25	2.52	4.63	2.44	5.04	2.60	4.86	2.50
$\hat{\sigma}_r^2$	1.46		1.35		1.47		1.36	
s_{45}	407		425		417		426	
				$n = 500$, $R^2 = 5\%$				
Mean	32.30	44.92	31.43*	45.64*	32.41	44.27	32.00	44.25
RMSE	8.05	3.94	8.90*	2.40*	7.98	4.24	8.34	4.20
$\hat{\sigma}_r^2$	8.84		8.70*		8.85		8.64	
s_{45}	327		28*		317		308	
				$n = 500$, $R^2 = 10\%$				
Mean	33.81	45.29	33.56*	45.47*	33.85	44.74	33.78	44.71
RMSE	5.12	2.77	4.47*	2.12*	5.12	3.03	5.03	2.80
$\hat{\sigma}_r^2$	4.25		4.10*		4.25		4.12	
s_{45}	420		37*		418		425	
				$n = 500$, $R^2 = 25\%$				
Mean	34.74	45.25	34.62*	45.20*	34.85	45.01	34.63	44.84
RMSE	2.32	1.55	2.30*	1.26*	2.33	1.63	2.28	1.45
$\hat{\sigma}_r^2$	1.48		1.37*		1.49		1.39	
s_{45}	481		48*		483		484	

TABLE 3.7: Scenario 4: Mean estimates of QTL positions based on significant runs with two-QTL models (simulated $P_1 = 35\,\text{cM}$, $P_2 = 45\,\text{cM}$) with associated root mean squared errors (RMSEs) and average estimated residual variances $\hat{\sigma}_r^2$ of RRM, IRM, RIM and MIM. The number of significant runs where the positions of both QTL were in the correct marker intervals is s_{45}. The number of individuals per simulated experiment is n and R^2 denotes the relative QTL variances. All values are based on $N = 500$ runs, except of values indicated by an asterisk ($N = 50$).

	RRM		IRM		RIM		MIM	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
				$n = 200$, $R^2 = 5\%$				
Mean	30.29	42.65	30.06	42.60	30.00	42.66	29.70	42.37
RMSE	10.79	7.24	11.02	7.33	10.85	7.40	11.25	7.86
$\hat{\sigma}_r^2$	4.62		4.53		4.65		4.45	
s_{45}	193		182		182		172	
				$n = 200$, $R^2 = 10\%$				
Mean	32.56	43.84	32.55	43.79	32.03	43.88	31.98	43.76
RMSE	7.74	5.18	7.54	5.11	8.13	5.34	8.13	5.57
$\hat{\sigma}_r^2$	2.23		2.18		2.25		2.15	
s_{45}	338		338		323		329	
				$n = 200$, $R^2 = 25\%$				
Mean	34.58	44.54	34.67	44.59	34.38	44.76	34.24	44.73
RMSE	3.60	2.82	3.34	2.71	3.53	2.79	3.42	2.81
$\hat{\sigma}_r^2$	0.77		0.73		0.78		0.73	
s_{45}	445		449		447		447	
				$n = 500$, $R^2 = 5\%$				
Mean	33.55	43.91	31.77*	43.93*	32.99	44.08	32.92	44.03
RMSE	6.02	4.29	7.80*	4.88*	6.58	4.48	6.50	4.25
$\hat{\sigma}_r^2$	4.75		4.68*		4.76		4.67	
s_{45}	393		32*		378		377	
				$n = 500$, $R^2 = 10\%$				
Mean	34.81	44.45	34.42*	44.96*	34.45	44.70	34.34	44.47
RMSE	3.31	2.77	3.16*	2.31*	3.46	2.70	3.46	2.81
$\hat{\sigma}_r^2$	2.28		2.21*		2.28		2.23	
s_{45}	455		44*		449		450	
				$n = 500$, $R^2 = 25\%$				
Mean	35.16	44.72	34.71*	45.02*	35.00	44.90	34.81	44.70
RMSE	1.90	1.70	2.19*	1.54*	1.79	1.66	1.78	1.64
$\hat{\sigma}_r^2$	0.79		0.74*		0.79		0.75	
s_{45}	493		48*		495		495	

Estimated residual variance: The estimated mean residual variances $\hat{\sigma}_r^2$ of RRM and RIM were nearly identical as well as $\hat{\sigma}_r^2$ between IRM and MIM in all scenarios (see Tables 3.4-3.7). This result was expected, because it is known that $\hat{\sigma}_r^2$ of RRM and RIM contains both the residual variance σ_e^2 and the within marker genotype QTL variance (Xu 1995; Zimmer *et al.* 2011). Estimates of the residual variances $\hat{\sigma}_r^2$ were taken from significant runs of the two-QTL model only. Especially in simulated experiments with low power an underestimation of σ_r^2 could be observed, in accordance with the so-called "Beavis effect" (Beavis 1994, 1998; Xu 2003) where a strong upward bias of estimated QTL effects from mapping experiments with significant outcome was reported.

QTL effects estimates: There are 27 different marker classes if both QTL are in adjacent marker intervals and 81 otherwise. To make sure that there are only 27 marker classes, we used only those runs of scenario 3 ($n = 500$, $R^2 = 25\,\%$), where each method identified the QTL in the correct marker interval combination for comparing conditional genotypic effects and their RMSEs between methods (see Table 3.8). The total number of considered runs was $N = 471$. For the three most frequent marker classes (222, 111, 000) and other frequent marker classes (221, 211, 210, 122, 121, 112, 110, 100, 011, 001), the RMSEs of the conditional genotypic effects were very similar between the investigated methods. Differences between the RMSEs were < 0.030. MIM often produced smaller RMSEs than the RIM. In very rare marker classes (202, 201, 120, 102, 021, 020), differences between the methods occur and RMSEs of RRM were smaller than those of RIM and MIM, or equal, except the marker class 120. Moderate frequent marker classes as 220, 212, 200, 101, 022, 012, 010 and 002 showed the largest differences between the methods. RRM was often the method with smallest RMSEs, except marker classes 220 and 200. Note that the RRM estimated conditional genotypic values also for marker classes without observations due to the correlations between different marker classes.

DISCUSSION

We compared RRM with the standard mapping procedures IRM, RIM and MIM with regard to experimental power as well as precision of estimated QTL positions and effects for an F_2 population. The observed power of scenarios 2 and 4 shows that it is necessary to use a multiple QTL model including additive and nonadditive genetic effects, because some QTL can only be identified if interactions are considered, in agreement with e. g. Carlborg and Haley (2004), and a multi-dimensional search approach is performed. Forward selection by adding QTL one by one tends to miss

TABLE 3.8: Comparison of root mean squared errors (RMSEs) of estimated conditional genotypic effects and simulated genotypic values (sgv). The data of scenario 3 with $n = 500$ and $R^2 = 25\,\%$ were used, where RRM, RIM and MIM identified the correct marker interval combination ($N = 471$). Homozygous marker genotypes are indicated by 2 and 0, heterozygous by 1. The frequency indicates the number of replications in which a specific marker state was observed.

marker			RRM	RIM	MIM
states	sgv	frequency		RMSEs	
222	0.492	471	0.151	0.149	0.149
221	0.495	471	0.171	0.169	0.143
220	0.492	284	0.455	0.308	0.243
212	−0.068	394	0.471	0.519	0.522
211	0.238	471	0.142	0.143	0.137
210	0.497	388	0.362	0.364	0.341
202	0.500	3	0.876	0.835	1.025
201	0.619	61	0.585	0.623	0.604
200	0.984	277	0.464	0.450	0.408
122	−0.247	471	0.249	0.253	0.251
121	0.005	457	0.382	0.411	0.409
120	0.201	82	0.636	0.592	0.578
112	−0.495	471	0.170	0.176	0.170
111	0.000	471	0.008	0.008	0.008
110	0.498	471	0.175	0.177	0.171
102	−0.251	81	0.971	1.035	1.027
101	0.523	455	0.450	0.506	0.495
100	1.240	471	0.209	0.220	0.186
022	−0.975	273	0.870	1.012	1.049
021	−0.619	78	0.811	0.960	0.998
020	0.375	4	0.562	1.148	1.116
012	−0.992	392	0.651	0.736	0.747
011	−0.247	471	0.148	0.148	0.140
010	0.485	390	0.527	0.603	0.605
002	−0.531	280	0.881	1.160	1.221
001	0.484	471	0.259	0.280	0.279
000	1.488	471	0.143	0.139	0.136

QTL compared to applying complex models and backward selection. As shown in Table 3.2, the observed power, e. g. from scenario 2, of a one-QTL model was small, but the power of a two-QTL model was $\approx 80.0\,\%$ ($n = 200$, $R^2 = 10\,\%$). Using a stepwise selection procedure (only adding a single QTL) as suggested e. g. by KAO *et al.* (1999), no QTL was found instead of two linked QTL. A sequential search for QTL, e. g. if QTL are in repulsion or without main QTL effects, failed to identify the number of QTL correctly. Therefore, a multiple QTL-model should be used for QTL mapping.

ASReml (Gilmour *et al.* 2008) uses the average information algorithm. With additive and nonadditive genetic effects of multiple QTL, the likelihood surface is very complex and therefore the identification of a global maximum may be difficult. The choice of starting values of parameters is very important. In some cases, we got negative likelihood ratio values from RRM, which is a numerical problem. In such positions we recalculated the likelihood value of the alternative model with changed starting values. An alternative approach was suggested by Li and Cui (2009), who used estimated values of the model under the null hypothesis as starting values of the alternative model, additional variance components set to small positive numbers. On this way Li and Cui (2009) guarantee at least positive likelihood ratio values.

Kao (2000) investigated the differences between RIM and MIM using BC populations. He found that the differences between QTL effects (e. g. QTL in repulsion), consideration of epistatic effects and linkage between QTL may influence the accuracy of RIM. Furthermore, RIM may have a serious problem and be less powerful to separate closely linked QTL compared to MIM, especially if epistasis is present. In our scenarios RIM had no problem to detect QTL in coupling or repulsion with respect to additive genetic effects of both QTL and an additive-by-additive effect, although two linked QTL were simulated 10 cM apart. Nevertheless, the estimates of QTL positions of RIM were less accurate than from MIM, especially if the number of F_2 individuals or the relative QTL variance was high.

RIM and MIM are based on a fixed model and these methods directly estimate the QTL effects according to Cockerham's model applying the properties of orthogonal contrasts (e. g. Kao and Zeng 2002). The mean estimated genetic effects based on significant replications with a two-QTL model for the third scenario ($n = 500$, $R^2 = 25\,\%$) and their corresponding standard deviations (SDs, in parentheses) of RIM were 0.573 (0.426) for a_1, -1.064 (0.430) for a_2 and 0.999 (0.116) for aa. In comparison, the estimates of MIM were 0.517 (0.246) for a_1, -1.007 (0.252) for a_2 and 0.989 (0.114) for aa. Both additive genetic effects were estimated more accurate using MIM compared to RIM. Especially if R^2 and n were large, MIM produced more precise estimates than RIM, as expected, because MIM is a ML based method and uses a mixture model.

That the conditional QTL genotype probabilities, which used in RIM, have to be replaced by the conditional posterior probabilities as used in MIM to obtain similar results was shown theoretically by Kao (2000). This is only the case if the QTL is located at a marker position. RIM assumes that the variance of the residuals are equivalent to the variance of the phenotype given the conditional QTL genotype prob-

abilities (Xu 1995). In this way the effects of other QTL not explained by the markers are ignored in the residuals. Note that the residual variance σ_r^2 of IRM and MIM correspond to the simulated residual variance σ_e^2, but σ_r^2 of RRM and RIM is inflated by the within marker genotype QTL variance (Xu 1995, 1998c; Zimmer *et al.* 2011). Based on that fact $\hat{\sigma}_r^2$ is expected to be higher in RRM and RIM than in IRM and MIM, which should be considered in the interpretation of $\hat{\sigma}_r^2$. The additive-by-additive genetic effect was estimated very similar by both methods.

The VCM tends to be powerful and precise in detecting multiple linked QTL. Occasionally, in scenario 1, the RRM was much more powerful in QTL detection than fixed models. The estimated parameters from RRM and IRM were very similar, except of the residual variance. The RRM substantially saves computational requirements in contrast to the IRM, especially if the number of F_2 individuals and the number of variance components are increased. The non-detection of QTL because of the so-called "genetic drift error" (Xu 1996) can be reduced by using multiple line crosses. If the number of small families is large (e. g. Haseman and Elston 1972; Xu and Atchley 1995), VCM is preferred over fixed models, because of the easier implementation and the computational advantages in this context (Xu 1998c). VCM offers an advantage over RIM and MIM, because the number of segregating alleles as well as the coupling phases have not to be known to map QTL. With a small number of large families the power to detect a QTL is larger as compared to a large number of small families (Xie *et al.* 1998). Therefore, the RRM is always much more faster and computationally tractable than the IRM.

The VCM performed something like a selection procedure, because some variance components were estimated on the boundary of zero in the parameter space. Therefore, the VCM using a two-QTL model did not estimate two QTL positions necessarily. In contrast, RIM and MIM always estimated positions of the QTL if the additional QTL are considered in the genetic model. Mostly, the estimated parameters of RRM and MIM were very similar in the simulated scenarios of a single family.

Our simulation shows that the RRM is a useful method to map multiple linked QTL with additive and nonadditive genetic effects. While RRM saves a lot of computational requirements compared to IRM, RRM is competitive with other standard methods like IRM, RIM and MIM in terms of detection power and precision of estimated QTL positions and effects.

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, MA 1553/3-1).

LITERATURE

Abdel-Azim, G. and A. E. Freeman, 2001 A rapid method for computing the inverse of the gametic covariance matrix between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **33**: 153–173.

Beavis, W. D., 1994 The power and deceit of QTL experiments: Lessons from comparitive QTL studies. In *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference*, p. 250–266, Washington, DC, American Seed Trade Association.

Beavis, W. D., 1998 QTL Analyses: Power, Precision, and Accuracy. In *Molecular Dissection of Complex Traits*, edited by A. H. Paterson, p. 145–162, New York, CRC Press.

Carlborg, Ö. and C. S. Haley, 2004 Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. **5**: 618–625.

Crepieux, S., C. Lebreton, B. Servin, and G. Charmet, 2004 Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. Genetics **168**: 1737–1749.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**: 1–38.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson, 2008 *ASReml User Guide Release 3.0*. VSN International, Hemel Hempstead, UK.

Grignola, F. E., I. Hoeschele, and B. Tier, 1996a Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. Methodology. Genet. Sel. Evol. **28**: 479–490.

Grignola, F. E., I. Hoeschele, Q. Zhang, and G. Thaller, 1996b Mapping quantitative trait loci in outcross populations via residual maximum likelihood. II. A simulation study. Genet. Sel. Evol. **28**: 491–504.

Haldane, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. **8**: 299–309.

Haley, C. S. and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–324.

Haseman, J. K. and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2**: 3–19.

Hoeschele, I. and P. M. VanRaden, 1993a Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theor. Appl. Genet. **85**: 953–960.

Hoeschele, I. and P. M. VanRaden, 1993b Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor. Appl. Genet. **85**: 946–952.

Jannink, J.-L. and R. Jansen, 2001 Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **157**: 445–454.

Kao, C.-H., 2000 On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics **156**: 855–865.

Kao, C.-H. and Z.-B. Zeng, 1997 General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53**: 653–665.

Kao, C.-H. and Z.-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160**: 1243–1261.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

Kruglyak, L. and E. S. Lander, 1995 A nonparametric approach for mapping quantitative trait loci. Genetics **139**: 1421–1428.

Lander, E. S. and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.

Li, G. and Y. Cui, 2009 A statistical variance components framework for mapping imprinted quantitative trait locus in experimental crosses. J. Probab. Stat. **2009**: 1–27.

Liu, Y., G. B. Jansen, and C. Y. Lin, 2002 The covariance between relatives conditional on genetic markers. Genet. Sel. Evol. **34**: 657–678.

Martínez, O. and R. N. Curnow, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85**: 480–488.

Mayer, M., 2005 A comparison of regression interval mapping and multiple interval mapping for linked QTL. Heredity **94**: 599–605.

Mayer, M., 2007 On the mapping of 2 QTLs in the same marker interval using multiple interval mapping and a moment method. In *XI QTLMAS 2007*.

Mayer, M., Y. Liu, and G. Freyer, 2004 A simulation study on the accuracy of position and effect estimates of linked QTL and their asymptotic standard deviations using multiple interval mapping in an F_2 scheme. Genet. Sel. Evol. **36**: 455–479.

Patterson, H. D. and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika **58**: 545–554.

Rönnegård, L., R. Pong-Wong, and Ö. Carlborg, 2008 Defining the assumptions underlying modeling of epistatic QTL using variance component methods. J. Hered. **99**: 421–425.

Thomas, D. C. and V. Cortessis, 1992 A gibbs sampling approach to linkage analysis. Hum. Hered. **42**: 63–76.

Wang, T., R. L. Fernando, S. van der Beek, M. Grossman, and J. A. M. van Arendonk, 1995 Covariance between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **27**: 251–274.

Xie, C., D. D. Gessler, and S. Xu, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics **149**: 1139–1146.

Xu, S., 1995 A comment on the simple regression method for interval mapping. Genetics **141**: 1657–1659.

Xu, S., 1996 Mapping quantitative trait loci using four-way crosses. Genet. Res. **68**: 175–181.

Xu, S., 1998a Further investigation on the regression method of mapping quantitative trait loci. Heredity **80**: 364–373.

Xu, S., 1998b Iteratively reweighted least squares mapping of quantitative trait loci. Behav. Genet. **28**: 341–355.

Xu, S., 1998c Mapping quantitative trait loci using multiple families of line crosses. Genetics **148**: 517–524.

Xu, S., 2003 Theoretical basis of the Beavis Effect. Genetics **165**: 2259–2268.

Xu, S. and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141**: 1189–1197.

Zeng, Z.-B., C.-H. Kao, and C. J. Basten, 1999 Estimating the genetic architecture of quantitative traits. Genet. Res. **74**: 279–289.

Zimmer, D., M. Mayer, and N. Reinsch, 2011 Complex genetic effects in quantitative trait locus identification: A computationally tractable random model for use in F_2 populations. Genetics **187**: 261–270.

CHAPTER THREE

SPARSE COVARIANCE MATRICES IN RANDOM MODELS FOR QUANTITATIVE TRAIT LOCUS DISCOVERY IN F_2 POPULATIONS

Daisy Zimmer, Norbert Reinsch

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

ABSTRACT

In F_2 families derived from inbred lines marker-based relationship matrices for random genetic effects (additive genetic, dominance and pairwise interactions) can be derived from simple elementary matrices, describing the covariance of those effects for known QTL genotypes. A novel kind of such elementary covariances for additive and additive-by-additive genetic effects is proposed, which reflects expected perfect correlations between genetic effects of certain genotypes and, depending on the kind of effect considered, leads to a more sparse representation of genetic covariance. It is shown theoretically and by simulated examples that these covariance matrices lead to identical restricted log-likelihood values and, hence, provide the same parameter estimates as previously reported versions, while nominal standard errors of estimated genetic effects are improved and more realistic. Computational speed is considerably enhanced, when genetic effects are estimated for each genotyped F_2 individual. Moreover, when marker positions and QTL locations coincide, this kind of additive genetic covariance matrix is equivalent to models with random regression coefficients.

INTRODUCTION

Variance component methods (VCM) considering QTL effects as random in a linear mixed model (LMM) are often applied for QTL mapping in combined data, i. e. data of a large number of (small) families (e. g. Xu 1998). Combining data of multiple families apparently reduced the failure of non-detection of QTL (Xu 1998) due to fixation of a single allele in both parental lines, the so-called genetic drift error (Xu 1996). Different strategies to map multiple, possibly interacting, QTL were investigated by simulations using multiple line cross experiments (Xie *et al.* 1998; Xu 1998; Crepieux *et al.* 2004; Li and Cui 2009; Zimmer *et al.* 2011b).

An approach to set up required relationship matrices in experimental populations was suggested by Xie *et al.* (1998) and was extended through Zimmer *et al.* (2011b) using conditional QTL genotype probabilities given the flanking marker information and elementary covariance matrices describing the covariance of genetic effects for known QTL genotypes. In this way each individual gets an individual genotypic effect, which is called individual random model (IRM). Increasing numbers of F_2 individuals, genetic effects and putative QTL make the IRM computationally slow. Therefore, a reduced random model (RRM; Zimmer *et al.* 2011b) was proposed, which considers an average genotypic effect for each marker class instead of individual genotypic effects and results in considerably enhanced computing speed compared to the IRM. The competitiveness of RRM to IRM was investigated and it could be shown that RRM approximates IRM

very well for simultaneous mapping of multiple linked QTL in terms of detection power and precision of the estimated QTL positions (Zimmer *et al.* 2011a,b).

In this article we propose alternative elementary covariance matrices for additive and additive-by-additive genetic effects which are used to set up the required relationship matrices. It is shown theoretically and by simulated examples that these covariance matrices lead to the same restricted likelihood values as applying the elementary covariance matrices suggested in the literature. The impact on computing speed and the nominal standard errors of the estimated effects are investigated.

THEORY

Individual random model (IRM): We consider inbred line-derived F_2 populations, where each individual receives one observation. In a single F_2 family two segregating QTL alleles occur, say Q and q, with allele frequencies of a half. The three possible QTL genotypes are G_{QQ}, G_{Qq} and G_{qq}. In the following we restrict ourselves to additive and additive-by-additive genetic effects. The trait vector $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$ (length n, number of F_2 individuals) with respect to two QTL for IRM is modeled as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_{a_1} + \boldsymbol{Z}_2\boldsymbol{u}_{a_2} + \boldsymbol{Z}_3\boldsymbol{u}_{aa} + \boldsymbol{e}$. The vector of fixed effects is $\boldsymbol{\beta}$ (length p) and $\boldsymbol{X}$ is the related design matrix of suitable size, $\boldsymbol{u}_\tau$ with $\tau \in \{a_1, a_2, aa\}$ are the vectors of random additive and additive-by-additive genetic QTL effects and $\boldsymbol{Z}_\ell$ with $\ell \in \{1, 2, 3\}$ are the related incidence matrices. Note that $\boldsymbol{Z}_\ell$ is an identity matrix $\boldsymbol{I}$ for IRM. Expectations and variances of the genetic QTL effects are $E(\boldsymbol{u}_\tau) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{u}_\tau) = \boldsymbol{V}_\tau\sigma^2_\tau$, where $\boldsymbol{V}_\tau$ is the QTL relationship matrix conditional on the observed marker genotypes and σ^2_τ is the related QTL variance. Residuals are assumed to follow a multivariate normal distribution with $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2_e)$, whereby σ^2_e is the residual variance. Random effects, i. e. genetic effects and residuals, are assumed to be mutually uncorrelated.

Elementary additive genetic covariance matrix: To set up the required additive genetic relationship matrix $\boldsymbol{V}_{a_\ell} = \{a^\ell_{st}\}$ at the putative QTL $\ell \in \{1, 2\}$, Xie *et al.* (1998) used first, conditional QTL genotype probabilities and second, an elementary additive genetic covariance matrix.

For an individual s with $s \in \{1, \ldots, n\}$ the conditional QTL genotype probabilities (p_i^{QQ}, p_i^{Qq}, p_i^{qq}) are inferred from flanking markers of the F_2 individuals based e. g. on Haldane's mapping function (Haldane 1919). The nine possible genotypes of the flanking markers are denoted by the index i with $i \in \{1, \ldots, 9\}$. We assume at most a single QTL within a marker interval.

Following Xie *et al.* (1998), the elementary additive genetic covariance matrix of the three possible QTL genotypes (G_{QQ}, G_{Qq}, G_{qq}) is

$$\boldsymbol{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Elements of $\boldsymbol{A}$ can be interpreted as the number of identical by descent (IBD) alleles between pairs of known QTL genotypes.

The variances of $\boldsymbol{V}_{a_\ell}$ with $\ell \in \{1, 2\}$ are $a^\ell_{ss} = 2p_i^{QQ} + p_i^{Qq} + 2p_i^{qq}$ and covariances are $a^\ell_{st} = (2p_i^{QQ} + p_i^{Qq})p_j^{QQ} + p_j^{Qq} + (p_i^{Qq} + 2p_i^{qq})p_j^{qq}$ for a pair of individuals s and t with $s, t \in \{1, \ldots, n\}$ belonging to marker classes i and j with $i, j \in \{1, \ldots, 9\}$.

Modified elementary additive genetic covariance matrix: The additive genetic effect of the homozygous individuals are perfectly negatively correlated, i. e. the additive genetic effect of the QTL genotype G_{QQ} has the opposite sign as the additive genetic effect of G_{qq}. These additional information can be used for QTL mapping, where adapted elementary covariance matrices can be derived from effect specific coefficients of QTL genotypes.

Coding coefficients of a random additive effect (a_ℓ) at the putative QTL $\ell \in \{1, 2\}$ as 1 for QTL genotype G_{QQ}, 0 for G_{Qq}, -1 for G_{qq} results in a vector $\boldsymbol{X}'_{a_\ell} = (1\,, 0\,, -1)$. The variance of the additive genetic effect for the three possible QTL genotypes is $\mathrm{Var}(\boldsymbol{u}_{a_\ell}) = \mathrm{Var}(\boldsymbol{X}_{a_\ell} a_\ell) = \boldsymbol{X}_{a_\ell}\boldsymbol{X}'_{a_\ell}\sigma^2_{a_\ell} = \boldsymbol{\Lambda}\sigma^2_{a_\ell}$ with

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Matrix $\boldsymbol{\Lambda}$ represents the perfect negative correlation between both homozygous QTL genotypes and the fact that heterozygous individuals, by definition, have a breeding value equal zero, which is known in advance and is therefore treated as a constant without variance. Hence, elements for heterozygous QTL genotypes are zero in $\boldsymbol{\Lambda}$.

When replacing $\boldsymbol{A}$ by $\boldsymbol{\Lambda}$ and applying the formulas given e. g. by Zimmer *et al.* (2011b), the additive genetic relationship matrix $\boldsymbol{V}_{\lambda_\ell} = \{\lambda^\ell_{st}\}$ at the putative QTL ℓ has covariances of $\lambda^\ell_{st} = p_i^{QQ}p_j^{QQ} + p_i^{qq}p_j^{qq} - p_i^{QQ}p_j^{qq} - p_i^{qq}p_j^{QQ}$, the probability of two individuals s and t with observed marker genotypes i and j to share the same additive genetic effect minus the probability of additive genetic effects with opposite

sign. Variances obtained as $\lambda^{\ell}_{ss} = p_i^{QQ} + p_i^{qq}$, the marker-derived probability of a homozygous genotype for individual s. In this way the relationship $a^{\ell}_{st} = \lambda^{\ell}_{st} + 1$ occurs, in matrix notation $\boldsymbol{V}_{a_\ell} = \boldsymbol{V}_{\lambda_\ell} + \boldsymbol{O}_\ell$ with $\ell \in \{1, 2\}$. In a single F_2 family, matrices $\boldsymbol{O}_\ell$ have all elements equal to one. Considering multiple families, where all families are assumed to be uncorrelated, the corresponding relationship matrices are block diagonal matrices if the individuals are sorted by family. In this way matrices $\boldsymbol{O}_\ell$ are also block diagonal with blocks equal to one and zero off-diagonal blocks. Note that for IRM both $\boldsymbol{O}_\ell$ are equal.

The design matrix $\boldsymbol{Z}_\ell$ with $\ell \in \{1, 2\}$ of the IRM, relating observations to the additive genetic effects of the three possible genotypes in the case that the position of the QTL and a marker coincide, has three columns and n rows. Then the additive genetic covariance is obtained as $\boldsymbol{Z}_\ell \boldsymbol{X}_{a_\ell} \sigma^2_{a_\ell} \boldsymbol{X}'_{a_\ell} \boldsymbol{Z}'_\ell = \boldsymbol{V}_{a_\ell} \sigma^2_{a_\ell}$. In the matrix product $\boldsymbol{Z}_\ell \boldsymbol{X}_{a_\ell}$ only a single column of length n remains with elements equal to 1, 0 or -1 corresponding to the three possible QTL genotypes G_{QQ}, G_{Qq} and G_{qq}, just as in $\boldsymbol{X}_{a_\ell}$. Therefore, at marker positions the restricted log-likelihood can be evaluated and genetic effects estimated by a random regression model (RR) with design matrix $\boldsymbol{Z}_\ell \boldsymbol{X}_{a_\ell}$ for a single random additive genetic effect with variance $\sigma^2_{a_\ell}$, thereby avoiding any numerical problems due to singularity of the elementary covariance matrix $\boldsymbol{\Lambda}$. In this context it may be worthwhile to notice the eigenvalues 3, 2 and 0 of $\boldsymbol{A}$, compared to 2 and two times 0 of $\boldsymbol{\Lambda}$, as a further illustration of the fact, that at a marker location the additive genetic effect can be described by a single random variable when only two alleles are segregating within a family.

Modified additive-by-additive genetic covariances: As an elementary covariance matrix for additive-by-additive genetic effects with one row and one column for each two-QTL genotype the Kronecker product $\boldsymbol{A} \otimes \boldsymbol{A}$ has been proposed. Therefrom, in case that QTL do not coincide with markers, the epistatic relationship matrix $\boldsymbol{V}_{aa} = \{aa_{st}\}$ is derived (Zimmer *et al.* 2011b) with the help of marker-derived conditional QTL genotype probabilities. Here we examine $\boldsymbol{\Lambda} \otimes \boldsymbol{\Lambda}$ as an alternative and derive an additive-by-additive genetic relationship matrix $\boldsymbol{V}_{\lambda\lambda} = \{\lambda\lambda_{st}\}$ with diagonal elements $\lambda\lambda_{ss} = (p_i^{QQ} + p_i^{qq})(p_i^{HH} + p_i^{hh})$ as variances and covariances $\lambda\lambda_{st} = (p_i^{QQ} - p_i^{qq})(p_i^{HH} - p_i^{hh})(p_j^{QQ} - p_j^{qq})(p_j^{HH} - p_j^{hh})$ with $i, j \in \{1, \ldots, 27\}$ if both QTL are in adjacent marker intervals and $i, j \in \{1, \ldots, 81\}$ otherwise. Recall that for the first QTL the relationship $a^1_{st} = \lambda^1_{st} + 1$ occurs and for the second one we have $a^2_{st} = \lambda^2_{st} + 1$. Therefore the elements aa_{st} are $aa_{st} = a^1_{st} a^2_{st} = \lambda\lambda_{st} + \lambda^1_{st} + \lambda^2_{st} + 1$. By defining a matrix $\boldsymbol{H} = \{h_{st}\}$ with elements $h_{st} = \lambda^1_{st} + \lambda^2_{st}$, specifically with diagonal elements $h_{ss} = p_i^{QQ} + p_i^{qq} + p_i^{HH} + p_i^{hh}$ and off-diagonals

$h_{st} = (p_i^{QQ} - p_i^{qq})(p_j^{QQ} - p_j^{qq}) + (p_i^{HH} - p_i^{hh})(p_j^{HH} - p_j^{hh})$. Then $\boldsymbol{V}_{aa}$ can be decomposed as $\boldsymbol{V}_{aa} = \boldsymbol{V}_{\lambda\lambda} + \boldsymbol{H} + \boldsymbol{O}_3$, where $\boldsymbol{H} = \boldsymbol{V}_{\lambda_1} + \boldsymbol{V}_{\lambda_2}$ as shown above and $\boldsymbol{O}_3$ is a matrix with all elements equal to one in the case of a single family. Otherwise, in the case of multiple uncorrelated families and observations ordered by families, $\boldsymbol{O}_3$ is block diagonal with zero off-diagonal blocks and all elements equal one in the diagonal blocks, analogously to $\boldsymbol{O}_\ell$ with $\ell \in \{1, 2\}$.

Equivalence of the restricted log-likelihood: Applying the restricted maximum likelihood (REML) approach as developed by Patterson and Thompson (1971), a LMM is multiplied with a transformation matrix $\boldsymbol{K}$ which yields in $\boldsymbol{y}^* := \boldsymbol{K}\boldsymbol{y} = \boldsymbol{K}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{K}\boldsymbol{e}$. The matrix $\boldsymbol{K}$ with $\dim(\boldsymbol{K}) = (n - \operatorname{rank}(\boldsymbol{X})) \times n$ fulfills the properties $\boldsymbol{K}\boldsymbol{X} = \boldsymbol{0}$ and $\operatorname{rank}(\boldsymbol{K}) = n - \operatorname{rank}(\boldsymbol{X})$ (Searle *et al.* 1992, p. 250ff). The application of the linear transformation leads to $\boldsymbol{y}^* \sim N(\boldsymbol{0}, \boldsymbol{K}\boldsymbol{V}\boldsymbol{K}')$. The restricted log-likelihood function is

$$\mathcal{L}(\boldsymbol{y}; \boldsymbol{\theta}) \;=\; -\frac{1}{2}\left[(n - \operatorname{rank}(\boldsymbol{X}))\log 2\pi + \log\det(\boldsymbol{K}\boldsymbol{V}\boldsymbol{K}') + \boldsymbol{y}'\boldsymbol{K}'(\boldsymbol{K}\boldsymbol{V}\boldsymbol{K}')^{-1}\boldsymbol{K}\boldsymbol{y}\right] ,$$

where the vector $\boldsymbol{\theta}$ includes all unknown variance components. The covariance matrix of the phenotypes $\boldsymbol{V}$ given the flanking marker genotypes for the IRM considering both kinds of elementary covariance matrices can be expressed as

$$\begin{aligned}
\boldsymbol{V} &= \boldsymbol{Z}_1\boldsymbol{V}_{a_1}\boldsymbol{Z}_1'\sigma_{a_1}^2 + \boldsymbol{Z}_2\boldsymbol{V}_{a_2}\boldsymbol{Z}_2'\sigma_{a_2}^2 + \boldsymbol{Z}_3\boldsymbol{V}_{aa}\boldsymbol{Z}_3'\sigma_{aa}^2 + \boldsymbol{I}\sigma_e^2 \\
&= \boldsymbol{Z}_1\left(\boldsymbol{V}_{\lambda_1} + \boldsymbol{O}_1\right)\boldsymbol{Z}_1'\sigma_{a_1}^2 + \boldsymbol{Z}_2\left(\boldsymbol{V}_{\lambda_2} + \boldsymbol{O}_2\right)\boldsymbol{Z}_2'\sigma_{a_2}^2 \\
&\quad + \boldsymbol{Z}_3\left(\boldsymbol{V}_{\lambda\lambda} + \boldsymbol{H} + \boldsymbol{O}_3\right)\boldsymbol{Z}_3'\sigma_{aa}^2 + \boldsymbol{I}\sigma_e^2 .
\end{aligned}$$

Using the restricted log-likelihood function $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\theta})$ with transformation matrix $\boldsymbol{K}$, it turns out that the matrix products $\boldsymbol{K}\boldsymbol{Z}_\ell\boldsymbol{O}_\ell\boldsymbol{Z}_\ell'\boldsymbol{K}'$ for $\ell \in \{1, 2, 3\}$ are equal zero. Therefore $\boldsymbol{V}$ can be equivalently written as

$$\boldsymbol{V}_E \;=\; \boldsymbol{Z}_1\boldsymbol{V}_{\lambda_1}\boldsymbol{Z}_1'\sigma_{a_1}^2 + \boldsymbol{Z}_2\boldsymbol{V}_{\lambda_2}\boldsymbol{Z}_2'\sigma_{a_2}^2 + \boldsymbol{Z}_3\left(\boldsymbol{V}_{\lambda\lambda} + \boldsymbol{H}\right)\boldsymbol{Z}_3'\sigma_{aa}^2 + \boldsymbol{I}\sigma_e^2 .$$

It follows that the matrix product $\boldsymbol{K}\boldsymbol{V}\boldsymbol{K}'$ for both equivalent spellings of $\boldsymbol{V}$ for the IRM is the same. Furthermore, using the decomposition $\boldsymbol{H} = \boldsymbol{V}_{\lambda_1} + \boldsymbol{V}_{\lambda_2}$ as well as the property $\boldsymbol{Z}_\ell = \boldsymbol{I}$ with $\ell \in \{1, 2, 3\}$ for the IRM we get

$$\boldsymbol{V}_R \;=\; \boldsymbol{V}_{\lambda_1}\left(\sigma_{a_1}^2 + \sigma_{aa}^2\right) + \boldsymbol{V}_{\lambda_2}\left(\sigma_{a_2}^2 + \sigma_{aa}^2\right) + \boldsymbol{V}_{\lambda\lambda}\sigma_{aa}^2 + \boldsymbol{I}\sigma_e^2 .$$

Hence, restricted log-likelihood functions $\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta})$ using $\boldsymbol{V}$ and $\boldsymbol{V}_E$ are equivalent and identical estimates of all parameters are obtained. Using $\boldsymbol{V}_R$ to estimate variance components is discussed later.

Because of the equivalence of the restricted log-likelihoods we suggest to use the covariance matrix of phenotypes $\boldsymbol{V}_E$ to map multiple, possible linked, QTL in single and multiple families, because matrices $\boldsymbol{V}_{\lambda_1}$, $\boldsymbol{V}_{\lambda_2}$ and $\boldsymbol{V}_{\lambda\lambda} + \boldsymbol{H}$ as well as their inverses have (numerous) entries equal zero, while matrices $\boldsymbol{V}_\tau$ with $\tau \in \{a_1, a_2, aa\}$ have no zero elements if positions of QTL and marker locations do not coincide. Algorithm using mixed model equations and sparse matrix structures of the here suggested covariance matrices are expected to be computationally faster.

Reduced random model (RRM): The RRM considers an average genotypic effect for each possible marker class instead of individual genotypic effects as obtained in the IRM. The trait vector $\boldsymbol{y}$ is modeled as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \widetilde{\boldsymbol{Z}}_1\widetilde{\boldsymbol{u}}_{a_1} + \widetilde{\boldsymbol{Z}}_2\widetilde{\boldsymbol{u}}_{a_2} + \widetilde{\boldsymbol{Z}}_3\widetilde{\boldsymbol{u}}_{aa} + \boldsymbol{\epsilon}$, where $\widetilde{\boldsymbol{Z}}_\ell$ are incidence matrices of suited order which related the observations to the corresponding marker classes. Again, expectations of the average genetic effects are zero and variances are $\mathrm{Var}(\widetilde{\boldsymbol{u}}_\tau) = \widetilde{\boldsymbol{V}}_\tau\sigma^2_\tau$ with $\tau \in \{a_1, a_2, aa\}$. The relationship matrix of average genetic effects are $\widetilde{\boldsymbol{V}}_\tau$ considering one row and one column for each possible marker class in each family. All remaining variables are defined as described for IRM. Again, residuals are assumed to be independently and identically normally distributed with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2_\epsilon)$, where the residual variance σ^2_ϵ is inflated by the genetic sampling effects (Zimmer *et al.* 2011b), similar to the within-marker genotype QTL variance as described by Xu (1995) and Xu (1998).

The average additive genetic relationship matrix $\widetilde{\boldsymbol{V}}_{\lambda_\ell} = \{\widetilde{\lambda}^\ell_{ij}\}$ with $i, j = 1, \ldots, 9$ using $\boldsymbol{\Lambda}$ has variances $\widetilde{\lambda}^\ell_{ii} = \frac{1}{n_i}(p_i^{QQ} + p_i^{qq} + (n_i - 1)(p_i^{QQ} - p_i^{qq})^2)$ and covariances $\widetilde{\lambda}^\ell_{ij} = (p_i^{QQ} - p_i^{qq})(p_j^{QQ} - p_j^{qq})$. Note that for the calculation of the variance of an average additive genotypic effect $\widetilde{\lambda}^\ell_{ii}$ at least a single individual in marker class i has to be assumed, i. e. $n_i > 0$ with n_i is the number of individuals in marker class i. Again, the additive genetic relationship matrix $\widetilde{\boldsymbol{V}}_{a_\ell} = \{\widetilde{a}^\ell_{ij}\}$ considering $\boldsymbol{A}$ shows the relationship $\widetilde{a}^\ell_{ij} = \widetilde{\lambda}^\ell_{ij} + 1$.

The average additive-by-additive genetic relationship matrix $\widetilde{\boldsymbol{V}}_{\lambda\lambda} = \{\widetilde{\lambda\lambda}_{ij}\}$ using the elementary covariance matrix $\boldsymbol{\Lambda} \otimes \boldsymbol{\Lambda}$ leads to variances

$$\widetilde{\lambda\lambda}_{ii} = \frac{1}{n_i}((p_i^{QQ} + p_i^{qq})(p_i^{HH} + p_i^{hh}) + (n_i - 1)(p_i^{QQ} - p_i^{qq})^2(p_i^{HH} - p_i^{hh})^2)$$

with assumption of $n_i > 0$. Covariances are

$$\widetilde{\lambda\lambda}_{ij} = (p_i^{QQ} - p_i^{qq})(p_i^{HH} - p_i^{hh})(p_j^{QQ} - p_j^{qq})(p_j^{HH} - p_j^{hh})$$

with $i, j \in \{1, \ldots, 27\}$ if both QTL are in adjacent marker intervals and otherwise $i, j \in \{1, \ldots, 81\}$. Considering the additive-by-additive genetic relationship matrix $\widetilde{\boldsymbol{V}}_{aa} = \{\widetilde{aa}_{ij}\}$ with respect to $\boldsymbol{A} \otimes \boldsymbol{A}$ leads to $\widetilde{aa}_{ij} = \widetilde{\lambda\lambda}_{ij} + \widetilde{h}_{ij} + 1$. Differences between $\widetilde{aa}_{ij}$ and $\widetilde{\lambda\lambda}_{ij} + 1$ result in matrix $\widetilde{\boldsymbol{H}} = \{\widetilde{h}_{ij}\}$ with variances

$$\widetilde{h}_{ii} = \frac{1}{n_i}(p_i^{QQ} + p_i^{qq} + p_i^{HH} + p_i^{hh} + (n_i - 1)((p_i^{QQ} - p_i^{qq})^2 + (p_i^{HH} - p_i^{hh})^2))$$

and covariances

$$\widetilde{h}_{ij} = (p_i^{QQ} - p_i^{qq})(p_j^{QQ} - p_j^{qq}) + (p_i^{HH} - p_i^{hh})(p_j^{HH} - p_j^{hh})\,.$$

The covariance matrix of the phenotypes $\widetilde{\boldsymbol{V}}$ for the RRM applying both kinds of elementary covariance matrices can be equivalently expressed in two ways as

$$\begin{aligned}\widetilde{\boldsymbol{V}} &= \widetilde{\boldsymbol{Z}}_1\widetilde{\boldsymbol{V}}_{a_1}\widetilde{\boldsymbol{Z}}_1'\sigma_{a_1}^2 + \widetilde{\boldsymbol{Z}}_2\widetilde{\boldsymbol{V}}_{a_2}\widetilde{\boldsymbol{Z}}_2'\sigma_{a_2}^2 + \widetilde{\boldsymbol{Z}}_3\widetilde{\boldsymbol{V}}_{aa}\widetilde{\boldsymbol{Z}}_3'\sigma_{aa}^2 + \boldsymbol{I}\sigma_\epsilon^2 \\ &= \widetilde{\boldsymbol{Z}}_1(\widetilde{\boldsymbol{V}}_{\lambda_1} + \widetilde{\boldsymbol{O}}_1)\widetilde{\boldsymbol{Z}}_1'\sigma_{a_1}^2 + \widetilde{\boldsymbol{Z}}_2(\widetilde{\boldsymbol{V}}_{\lambda_2} + \widetilde{\boldsymbol{O}}_2)\widetilde{\boldsymbol{Z}}_2'\sigma_{a_2}^2 \\ &\quad + \widetilde{\boldsymbol{Z}}_3(\widetilde{\boldsymbol{V}}_{\lambda\lambda} + \widetilde{\boldsymbol{H}} + \widetilde{\boldsymbol{O}}_3)\widetilde{\boldsymbol{Z}}_3'\sigma_{aa}^2 + \boldsymbol{I}\sigma_\epsilon^2\,.\end{aligned}$$

Again, using $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\theta})$ with transformation matrix $\boldsymbol{K}$ the matrix products $\boldsymbol{K}\widetilde{\boldsymbol{Z}}_\ell\widetilde{\boldsymbol{O}}_\ell\widetilde{\boldsymbol{Z}}_\ell'\boldsymbol{K}'$ with $\ell \in \{1, 2, 3\}$ are equal zero and therefore $\boldsymbol{K}\widetilde{\boldsymbol{V}}\boldsymbol{K}'$ is the same for both equivalent spellings of the RRM. The matrices $\widetilde{\boldsymbol{O}}_\ell$ are defined like for the IRM in single and multiple families, but with suited order, referring to the different marker classes. Analogously to the IRM, the covariance matrix of phenotypes $\widetilde{\boldsymbol{V}}$ is equivalent to

$$\widetilde{\boldsymbol{V}}_E = \widetilde{\boldsymbol{Z}}_1\widetilde{\boldsymbol{V}}_{\lambda_1}\widetilde{\boldsymbol{Z}}_1'\sigma_{a_1}^2 + \widetilde{\boldsymbol{Z}}_2\widetilde{\boldsymbol{V}}_{\lambda_2}\widetilde{\boldsymbol{Z}}_2'\sigma_{a_2}^2 + \widetilde{\boldsymbol{Z}}_3(\widetilde{\boldsymbol{V}}_{\lambda\lambda} + \widetilde{\boldsymbol{H}})\widetilde{\boldsymbol{Z}}_3'\sigma_{aa}^2 + \boldsymbol{I}\sigma_\epsilon^2\,,$$

leading to the same restricted log-likelihood functions $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\theta})$. Covariance matrices $\widetilde{\boldsymbol{V}}_{\lambda_\ell}$ with $\ell \in \{1, 2\}$ and $\widetilde{\boldsymbol{V}}_{\lambda\lambda} + \widetilde{\boldsymbol{H}}$ have also entries equal zero in contrast to $\widetilde{\boldsymbol{V}}_\tau$ with $\tau \in \{a_1, a_2, aa\}$ and may help to save computing time if algorithm use such sparse types of matrix structures.

RESULTS FROM APPLICATION TO SIMULATED EXAMPLES

Three different scenarios were simulated to demonstrate that both kinds of relationship matrices in both equivalent models lead to the same restricted log-likelihood profiles, to evaluate the estimated genetic conditional genotypic effects (QTL effects associated with markers) and their nominal standard errors (SEs) as well as the required computing time.

First, there are models with covariance matrices $\boldsymbol{V}_\tau$ for IRM and $\widetilde{\boldsymbol{V}}_\tau$ for RRM with $\tau \in \{a_1, a_2, aa\}$, called hereafter IRM_1 and RRM_1. Second, there are models with the proposed new covariance matrices $\boldsymbol{V}_{\lambda_\ell}$ with $\ell \in \{1, 2\}$ and $\boldsymbol{V}_{\lambda\lambda} + \boldsymbol{H}$ for IRM as well as $\widetilde{\boldsymbol{V}}_{\lambda_\ell}$ and $\widetilde{\boldsymbol{V}}_{\lambda\lambda} + \widetilde{\boldsymbol{H}}$ for RRM, called hereafter IRM_2 and RRM_2.

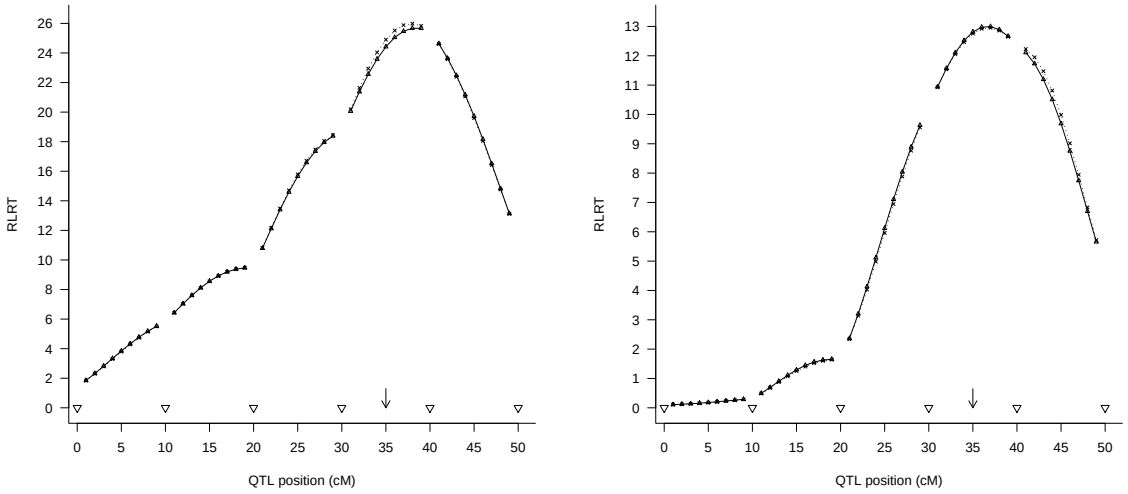
Simulations: A chromosome segment of 50 cM length, covered by markers every 10 cM (0, 10, 20, 30, 40, 50 cM) was simulated in all scenarios. A single run ($N = 1$) was investigated considering a QTL at $P = 35$ cM with only an additive genetic effect ($a = 1.0$) in a single F_2 family (scenario 1) and in four families (scenario 2). In total $n = 500$ F_2 individuals were considered and proportions of the phenotypic variance explained by the QTL were set at $R^2 = 10\,\%$ in both scenarios. The population mean μ was considered as fixed effect in a LMM for scenario 1. In scenario 2, an effect for each family was regarded, but only a single (population-specific) variance. QTL alleles within a family are not segregating necessarily, because inbred founders were randomly chosen from a population consisting of two inbred lines, the first line with QTL genotype G_{QQ} and the second line with G_{qq}. Thus each F_1 individual was produced from a randomly chosen pair of inbred parents. Always, fully informative markers were assumed. Cockerham's F_2-metric model (Cockerham 1954; Kao and Zeng 2002, Table 3) was used for simulating the observations.

In scenario 3, five independent F_2 families, each with 200 progeny ($n = 1000$), were derived from four inbred lines with QTL genotypes G_{QQHH}, G_{QQhh}, G_{qqHH} and G_{qqhh} considering two linked QTL. The experiment was repeated $N = 200$ times. Two QTL ($P_1 = 25$ and $P_2 = 35$ cM), each with an additive genetic effect ($a_1 = 1.0$, $a_2 = 0.5$) and an additive-by-additive genetic effect ($aa = 1.0$) were simulated. The relative QTL variance was set at $R^2 = 15\,\%$. Each family got a family specific mean (fixed effect). Again a single variance component was estimated for each genetic effect.

The estimation of variance components was done with ASReml (Gilmour *et al.* 2008), which uses the average information algorithm (Gilmour *et al.* 1995). ASReml exploits sparse matrix structures and is therefore efficient. For reasons of comparability between the required CPU time using both kinds of covariance matrices, these matri-

ces are completely stored in the data file which is required by ASReml. In this case, it is assured that ASReml requires the same time for reading our defined relationship matrices.

Likelihood profiles of the different methods for scenario 1 are shown in Figure 4.1(a) and for scenario 2 in Figure 4.1(b). As expected from theory, the restricted log-likelihood ratio test profiles *RLRT* of RRM_1 and RRM_2 were completely identical as well as the profiles of IRM_1 and IRM_2. For scenario 3 both kinds of covariance matrices for IRM_1 and IRM_2 (analogously for RRM_1 and RRM_2) provided the same likelihood profiles (results not shown).



(a) Scenario 1 (b) Scenario 2

FIGURE 4.1: For a single-QTL model the *RLRT* profiles of a single run of RRM_1 (solid line), RRM_2 (line with triangle), IRM_1 (dashed line) and IRM_2 (line with cross) are shown for scenarios 1 and 2.

For scenario 1 (single F_2 family) the estimates of the population mean μ, the conditional genotypic effects for RRM_1 and RRM_2, the variance components and the corresponding nominal standard errors (SEs) are depicted in Table 4.1. Also, the empirical SEs calculated from the estimated conditional genotypic effects for scenario from $N = 100$ runs (at true QTL position) are shown as well as the expected number of individuals within the different marker classes calculated for $n = 500$ individuals.

Estimates for RRM_1 and RRM_2 indicated that both methods provided the same estimates, but nominal SEs were decreased for RRM_2 compared to RRM_1. Comparison of the nominal SEs from RRM_2 with the empirical SEs showed that these SEs were in good agreement. For RRM_1 the nominal SEs had all the same magnitude and

TABLE 4.1: For scenario 1, the estimates (Est) and the corresponding nominal standard errors (SEs) of the populations mean μ, the estimated conditional genotypic effects and the estimated variance components for RRM_1 and RRM_2 are shown. Additionally, the empirical standard errors (emp. SEs) calculated from $N = 100$ runs at the true QTL position ($P = 35$cM) were presented. The expected number (no.) of individuals within a certain marker class (rounded) was given for a total number of $n = 500$ individuals. Flanking marker genotypes were denoted as ./. with entries left of the slash indicating the left marker alleles.

	RRM_1		RRM_2		emp.	
	Est	SE	Est	SE	SE	no.
μ	−0.08	0.73	−0.08	0.10	0.10	
11/11	0.74	0.74	0.74	0.14	0.14	103
11/12	0.14	0.73	0.14	0.07	0.09	21
11/22	−0.41	0.83	−0.41	0.41	0.27	1
12/11	0.60	0.73	0.60	0.12	0.09	21
12/12	0.00	0.72	0.00	0.00	0.00	209
12/22	−0.59	0.73	−0.59	0.13	0.10	21
22/11	0.43	0.78	0.43	0.29	0.48	1
22/12	−0.14	0.73	−0.14	0.07	0.07	21
22/22	−0.74	0.74	−0.74	0.14	0.14	103
$\hat{\sigma}_r^2$	4.56	0.73	4.56	0.73	0.31	
$\hat{\sigma}_a^2$	0.52	0.29	0.52	0.29	2.56	

were inflated compared to SEs from RRM_2. The conditional genotypic effect of the marker genotype 12/12 was estimated nearly as zero and therefore the SE was about zero for RRM_2. Empirical SEs showed also that nominal SEs for residual variances were increased for both RRMs, but nominal SEs of the additive genetic variance were much smaller than the empirical SE. However, the estimated variance components and corresponding SEs for RRM_1 and RRM_2 were identical. Using IRM_1 and IRM_2 for scenario 1 or both RRMs and IRMs for multiple families as in scenarios 2 and 3, it could also be shown that these methods provided the same estimated parameters, but with decreased and also more realistic nominal SEs for RRM_2 and IRM_2 (results not shown).

The required CPU time for ASReml (GILMOUR *et al.* 2008) in scenarios 1 and 2 for both RRMs was nearly identical for a repetition recorded on an HP DL380 G6 (72 GB RAM, 2× XEON X5570, 2.93 GHz, multiuser environment). In scenario 1 (2) IRM_2 needed only one quarter (half) of the time required by IRM_1, measured for a single run. In scenario 3 the required CPU time for RRM_1 was 86.71 sec in average for each run, whereby RRM_2 needed 77.24 sec. In contrast the required CPU time for IRM_1 was more than fourfold compared to IRM_2 in average for each repetition. However, the CPU time for IRM_1 (IRM_2) compared to RRM_1 (RRM_2) was more than hundredfold

(twenty-sixfold).

DISCUSSION

The savings in computing time for IRM_2 and RRM_2 base on sparse types of covariance matrices and their inverses. The application of such covariance matrices lead to less computational requirements if algorithms are implemented for the estimation of variance components which can use these sparse matrix structures like e. g. ASReml (Gilmour *et al.* 2008). As shown by simulated examples, the required CPU time for RRM_2 was slightly smaller than using RRM_1, whereby IRM_2 needed essentially less computing time compared to IRM_1. There are no zero elements for RRM_1 and IRM_1 if marker locations and QTL positions do not coincide and hence dense covariance matrices occur.

To determine the number of zero elements an operational zero of 10^{-8} was assumed in the following, i. e. elements which absolute values lower than 10^{-8} were treated as zero. At the putative QTL $\ell \in \{1, 2\}$, the corresponding additive relationship matrix $\boldsymbol{V}_{\lambda_\ell}$ for IRM_2 derived from flanking markers have at least $n_h^2 - n_h + 2n_h(n_f - n_h)$ elements equal to zero, where n_h denotes the number of individuals with double heterozygous marker genotype, i.e. the flanking marker genotype is 12/12 with entries left of the slash indicating the left marker alleles. The number of F_2 individuals from a certain family is n_f. Note that n_h do not depend on the putative position of the QTL using the operational zero, just on the width of the marker interval where the putative QTL is assumed. Only in the case where the putative QTL position is in the middle of the flanking marker interval, the number of zero elements exceeds the expected values. One explanation is, that both homozygous conditional QTL genotype probabilities for individuals belong to a certain marker class are equal in size and therefore the difference between both probabilities, as required in the calculation of elements λ_{st}^{ℓ}, are zero. This is the case for marker classes 11/22 and 22/11 besides marker genotype 12/12. If certain marker classes were not observed, this fact had no effect applying IRM_2. The value n_h can be determined from the present data set or expected values can be used. Assuming a marker interval of 10 cM, the frequency of the marker genotype 12/12 for a single QTL was about 42 %. In this way the expected number of individuals was about 84 for $n_f = 200$, then at least two third of the elements were zero.

In the additive-by-additive genetic relationship matrix $\boldsymbol{V}_{\lambda\lambda}$ for IRM_2 the number of zero elements is at least $n_{h^*}^2 - n_{h^*} + 2n_{h^*}(n_f - n_{h^*})$, where n_{h^*} is the number of individuals which have double heterozygous flanking marker genotypes, i. e. the flanking marker genotypes of the first and the second putative QTL are each 12/12. Again,

additional zero elements may appear if at least a putative QTL is in the middle of the marker interval. Note that n_{h^*} depends on the distance between both marker intervals. If the distance between both putative QTL increases, n_{h^*} decreases.

To ensure that ASReml required nearly the same time for reading our defined relationship matrices using both kinds of elementary covariance matrices, our data file considered all elements, i. e. also elements which are zero, exactly or treated as such, for IRM_2 and RRM_2. Hence, the computational speed up using the here suggested covariance matrices is attributable to the sparse matrix structure. Therefore, the CPU time from ASReml is even faster if only non-zero elements are read from file. This results in faster processing, especially for the relationship matrices which considered a genetic effect for each individual (IRM). However, both RRMs still are much more faster than the "fast" IRM_2.

Using the here suggested covariance matrices, the nominal standard errors of the estimated genetic effects were decreased compared to applying the covariance matrices as proposed for RRM_1 and IRM_1. Hence, the application of RRM_2 and IRM_2 have also an advantage if one is interested in realistic standard errors for these effects.

Whatever kind of both additive covariance matrices is used, an equivalent possibility to obtain the REML log-likelihood is a random regression (RR) on the expected number of Q-alleles, provided marker locus and QTL coincide. This may serve as a convenient alternative to applying the Sherman-Morrison-Woodbury matrix identity (e. g Henderson and Searle 1981) for calculating the inverse covariance matrix of observations and applying a REML algorithm using this matrix, as was suggested by Lee and Van der Werf (2006).

Even between markers RR results in a very similar restricted log-likelihood, as is demonstrated in Figure 4.2 for a single run of scenario 2 (multiple F_2 families, only an additive genetic effect). Exact identity is, however, only given at marker positions.

An equivalent covariance matrix of phenotypes $\boldsymbol{V}$ for IRM_2 was $\boldsymbol{V}_R$ as shown in the theory section. In principle the application of $\boldsymbol{V}_R$, or rather the individual covariance matrices, to estimate the variance components is possible by using an algorithm considering restrictions for the different variances. This variant was, however, not further investigated here.

Applying the concept of perfect correlations also to dominance deviations for allele frequency of a half leads to an elementary covariance matrix of dominance deviations,

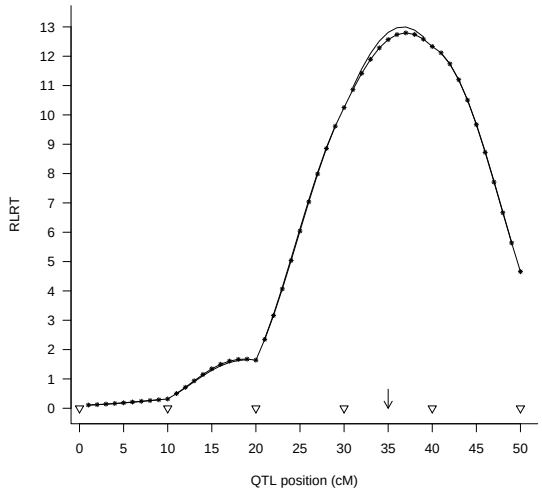


FIGURE 4.2: For scenario 2 (single-QTL model with only an additive genetic effect, multiple families) the *RLRT* profiles of a single run of RRM_2 (solid line) and RR (line with asterisk) are shown.

which is

$$\boldsymbol{\Delta} = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

Using $\boldsymbol{\Delta}$ to derive the dominance relationship matrix, the REML log-likelihood was not the same compared to applying $\boldsymbol{D}$ as suggested by XIE *et al.* (1998). For a single run in a single family the *RLRT* profiles for RRM_1 and RRM_2 shown in Figure 4.3. On marker positions the gaps were filled using RR to obtain a continuous likelihood profile for RRM_2. Note that the surface of RRM_2 using $\boldsymbol{\Delta}$ was higher compared to RRM_1. Considering $\boldsymbol{\Delta}$ the nominal SEs were also reduced compared to using $\boldsymbol{D}$ (results not shown).

Under the current state of knowledge, the application of the here suggested additive and additive-by-additive genetic covariance matrices can be suggested for QTL analyses in inbred line-derived F_2 populations, where only two segregating QTL alleles occur (allele frequency of a half). Useful areas of application can be found in plant and in animal breeding. In maize breeding the genetic variation among testcrosses using often a purely additive genetic model, because epistatic interactions seem to be of no or minor importance (e. g. JOHNSON 2004). As mentioned in several studies, the most important epistatic interaction appears due to additive-by-additive genetic effects, like shown by CHEVERUD *et al.* (2001) for adiposity of mouse inbred strains.

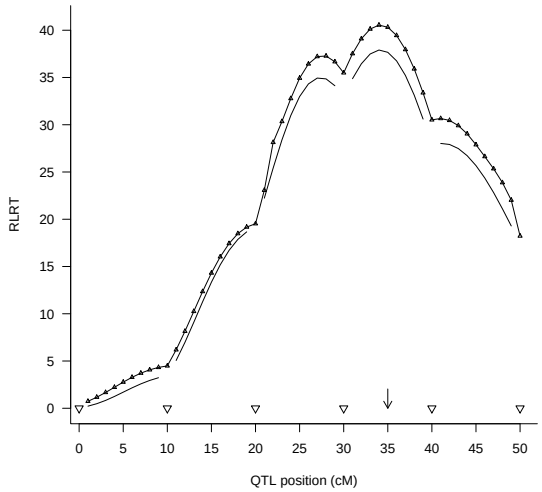


FIGURE 4.3: For a single-QTL model the *RLRT* profiles of a single run are shown for a scenario considering only a dominance effect ($d = 1.0$, $R^2 = 10\,\%$) for RRM_1 (solid line) and RRM_2 (line with triangle).

In conclusion the suggested kind of sparse covariance matrices provide the possibility to save considerable amounts of computing time, especially a random model with individual genetic effects (IRM). Moreover, nominal standard errors for genetic effects become more realistic and similar to empirical ones.

LITERATURE

CHEVERUD, J. M., T. T. VAUGHN, L. S. PLETSCHER, A. C. PERIPATO, E. S. ADAMS, C. F. ERIKSON, and K. J. KING-ELLISON, 2001 Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice. Mamm. Genome **12**: 3–12.

COCKERHAM, C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics **39**: 859–882.

CREPIEUX, S., C. LEBRETON, B. SERVIN, and G. CHARMET, 2004 Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. Genetics **168**: 1737–1749.

GILMOUR, A. R., B. J. GOGEL, B. R. CULLIS, and R. THOMPSON, 2008 *ASReml User Guide Release 3.0*. VSN International, Hemel Hempstead, UK.

GILMOUR, A. R., R. THOMPSON, and B. R. CULLIS, 1995 Average Information

REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics **51**: 1440–1450.

Haldane, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. **8**: 299–309.

Henderson, H. V. and S. R. Searle, 1981 On deriving the inverse of a sum of matrices. SIAM Rev. Soc. Ind. Appl. Math. **23**: 53–60.

Johnson, R., 2004 pp. 293–309 in *Marker-Assisted Selection*, John Wiley & Sons, Inc.

Kao, C.-H. and Z.-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160**: 1243–1261.

Lee, S. H. and J. H. J. Van der Werf, 2006 An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genet. Sel. Evol. **38**: 25–43.

Li, G. and Y. Cui, 2009 A statistical variance components framework for mapping imprinted quantitative trait locus in experimental crosses. J. Probab. Stat. **2009**: 1–27.

Patterson, H. D. and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika **58**: 545–554.

Searle, S. R., G. Casella, and C. E. McCulloch, 1992 *Variance Components*. John Wiley & Sons, New York.

Xie, C., D. D. Gessler, and S. Xu, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics **149**: 1139–1146.

Xu, S., 1995 A comment on the simple regression method for interval mapping. Genetics **141**: 1657–1659.

Xu, S., 1996 Mapping quantitative trait loci using four-way crosses. Genet. Res. **68**: 175–181.

Xu, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics **148**: 517–524.

Zimmer, D., M. Mayer, and N. Reinsch, 2011a Competitiveness of a reduced random model versus fixed and random alternatives for mapping multiple QTL in F_2 populations **unpublished**.

Zimmer, D., M. Mayer, and N. Reinsch, 2011b Complex genetic effects in quantitative trait locus identification: A computationally tractable random model for use in F_2 populations. Genetics **187**: 261–270.

GENERAL DISCUSSION

In mapping experiments it is often desirable to fit multiple QTL models with additive and nonadditive genetic effects simultaneously, because most quantitative traits are influenced by multiple, possibly linked, QTL which have a contribution to the phenotypic variation (e. g. Li and Cui 2009). Epistatic effects, which contribute to the total phenotypic variance, were found in different species (e. g. Doebley *et al.* 1995; Cheverud *et al.* 2001; Yi *et al.* 2006; Xu and Jia 2007; Radoev *et al.* 2008). Molecular genetic markers, like SNPs, are increasingly becoming available due to advanced DNA chip technology covering the whole genome. Both, more complex genetic models and dense genetic markers lead to an increasing number of parameters which have to be considered.

Variance component methods (VCM) consider additive genetic, dominance and epistatic interaction effects as random in a linear mixed model. Using VCM the estimated genetic variance components can be tested directly of significance. Hence, VCM is a flexible approach to map multiple linked QTL simultaneously regarding multiple flanking marker intervals, is easy to implement and may have computational advantages, e. g. with multiple families, compared to fixed models (Xu 1998). In this way fixed models consider a family-specific genetic effect, whereby the VCM takes into account a single population-specific variance, as e. g. described by Xu (1998). Considering a large number of families, individuals of different families are assumed to be uncorrelated and covariances between individuals belong to different families are assumed to be zero, like in this thesis and found in the literature (Xu 1998; Xie *et al.* 1998; Li and Cui 2009). Hence, only block diagonals may be different from zero. If the number of considered families is increased, the number of zero elements due to the off-diagonal blocks increases.

To set up customized relationship matrices, conditional QTL genotype probabilities and elementary covariance matrices, where QTL genotypes are assumed to be known, were used as suggested by Xie *et al.* (1998). The order of each relationship matrix given the flanking marker genotypes for the individual random model (IRM) is determined through the number of F_2 individuals. For each investigated genetic effect such a relationship matrix is necessary. Considering epistatic effects in QTL mapping, the number of F_2 individuals should be sufficiently large. Hence, the IRM is computationally expensive, because for each individual and genetic effect an individual genetic effect is taken into account. Therefore a large number of parameters is

involved, leading to high computational requirements.

In this thesis two different ways to reduce the required computing time for QTL mapping in F_2 populations derived from inbred lines were proposed.

First, a reduced random model (RRM) which considers average genetic effects instead of individual genetic effects (IRM) was suggested. It could be shown that the genetic covariance structure of the RRM is asymptotically equivalent to the genetic covariance in the IRM. By simulations it was clearly demonstrated that the RRM was competitive to the traditional approach IRM in terms of the observed power, the accuracy of the QTL positions (mean, mean squared errors) and their effects considering single and multiple families. The RRM speed up the QTL analyses in inbred line-derived F_2 families due to the reduced number of parameters.

Second, in inbred line-derived F_2 populations two QTL alleles segregate within a family and allele frequencies are a half. Using that fact, the coefficients of the additive genetic effect of the homozygous individuals are perfectly negatively correlated, which was used to set up the required covariance matrices. Application of this concept arises in sparse additive and additive-by-additive genetic relationship matrices, which have a considerable amount of zero elements and results in faster proceedings if this sparse structure is exploited in the estimation of genetic variance components, especially for the IRM. What's remarkable here is that the estimated parameters are the same as well as the restricted log-likelihood values compared to using the elementary covariance matrices as suggested in Chapter One, but the nominal standard errors of estimated effects are reduced and more realistic. Nominal standard errors are standard errors obtained by applying a certain method given the variance components, i. e. these standard errors are method specific. Furthermore, it could be shown that an additive genetic effect can be described by a single random variable at marker locations, as used in a random regression (RR) approach, when only two alleles are segregating within a family. The RR was also suitable to approximate the likelihood surface of the RRM and the IRM if the putative QTL was between the flanking markers and again, less computing time is required.

Population-specific *vs.* **family-specific variances:** Multiple families were simulated in Chapters One and Three, where for each genetic effect a single (population-specific) variance was assumed. A family was derived from mating two randomly sampled inbred lines out of the four possible inbred lines considering two QTL, which represent all pairwise combinations of QTL genotypes. In the same experimental design, Xu (1998) also used a population-specific variance, because he was interested in

the variance of the substitution effect among different families. If the number of F_2 individuals within a family is small, a population-specific variance is recommended. It is also possible to consider family-specific variance components, because there are possible differences between families due to choice a random pair of inbred lines. QTL alleles within a family are not segregating necessarily. Such differences between the families perhaps can be used for QTL mapping and therefore family-specific variances are probably useful, especially if individual families should be elected for further breeding. In this case the number of variance components which have to be estimated increases. If the family size increases, it is expected that the accuracy of QTL discovery is increased using family-specific variances.

Regression to the mean: In genome wide scans estimated QTL effects from mapping experiments are known to be (strongly) biased upwards, the so-called "Beavis effect" (Beavis 1994, 1998; Xu 2003b) if only significant outcomes are reported from experiments with low power and few information. The overestimation of the QTL effects (Beavis effect) can probably partially avoided through regression to the mean using the VCM. For a given amount of information, like sample size or marker density or relative QTL variance, the more the estimated genetic effects using the VCM will be regressed to the mean due to less information from the data, because individual QTL genotypes are uncertain when less observations per marker genotype exist or when markers are far away from each other. Further simulations in Chapter Two showed that with low power the residual variance was underestimated reporting only the significant outcomes. Hence, the actual effect of shrinkage should be investigated.

Coincidence of putative QTL and marker locations: In this thesis the required relationship matrices are set up when marker locations and QTL positions do not coincide applying the RRM and the IRM. Relationship matrices are positive definite at non-marker positions. Coincidence of markers and QTL results in singularity of covariance matrices. ASReml (Gilmour *et al.* 2008) offers the possibility of using positive semidefinite relationship matrices with an appropriate qualifier, because a covariance matrix is per definition positive (semi-) definite. But in this way convergence problems may appear. However, this situation can be treated in different ways. First, allelic effects can be included in the random models instead of genetic effects. Second, adding a small quantity to variances, called regularization (Neumaier 1998). It is expected that this treatment has a little effect on the test statistics. Third, applying a reduced rank approximation obtained by spectral decomposition as suggested by Rönnegård *et al.* (2007). Finally, for additive genetic and dominance effects in Chapter Three a

random regression approach was suggested, which is able to fill the empty space at marker positions in inbred line-derived populations considering two segregating alleles within a family and allele frequency of a half. In this way a single random variable is sufficient to explain the variation.

Marker information: The kind of genetic markers, like SNPs or microsatellites, is not a limited factor applying the IRM and the RRM, because only the marker density and the marker informativity is of interest. If sufficient dense markers are available, most of the QTL effects will be picked up by the markers and QTL can be identified.

In F_2 populations derived from inbred lines it is assumed that there is at least a completely informative marker between two QTL due to the assumption made, that there is at most a single QTL within a marker interval. Considering two QTL within a marker interval, both single conditional QTL genotype probabilities are not independent of each other. Hence, the product of both single probabilities is not equal to the joint conditional QTL genotype probability (e. g. Rönnegård *et al.* 2008). Due to increasing number of genetic markers through high-throughput genotyping technology and the availability of numerous SNPs, the assumption of only a single QTL within a marker interval is certainly understandable and apparently holds. However, Mayer (2004) showed the conditional QTL genotype probabilities for two closely linked QTL for maximum likelihood methods, like the multiple interval mapping.

Furthermore, if there are missing marker genotypes or not fully informative markers in other investigated populations, it is generally known that information from other markers in a linkage group with the putative QTL can be used. In this situations additional marker classes occur. Note that missing marker information introduces uncertainty into the analysis. A general algorithm which can handle dominant and missing markers for various populations derived from two inbred lines using a Markov chain process was suggested by Jiang and Zeng (1997) assuming no crossover interference. Any of the numerous programs available to calculate adapted conditional QTL genotype probabilities for line-cross experiments can be applied. Hence, VCM is flexible, efficient and it is comparatively easy to calculate these probabilities.

Application to different population structures: In this thesis, the RRM and the IRM were only described to map multiple linked QTL simultaneously using flanking marker information in single and multiple families, i. e. a mixture of uncorrelated families, in F_2 populations derived from parental inbred lines, i. e. lines which are fixed for alternative QTL alleles, which are common in plant and laboratory animals.

In principle, the VCM can be applied to map QTL with additive and nonadditive genetic effects in arbitrary populations (inbred or noninbred) and various experimental designs, because the knowledge of the number of segregating alleles is not presupposed (e. g. Li and Cui 2009). The VCM is even profitable in combining data from different line crosses and multiple families, assuming only progenies derived from the cross of two parental lines. A so-called "consensus mapping strategy" for combining or updating data was suggested by Xie *et al.* (1998), where advantages and disadvantages are described there. The generalization of the VCM to different population structures is possible, whereby principal modifications may be necessary. Suitable elementary covariance matrices and/or conditional QTL genotype probabilities for the underlying populations have to be constructed. Elementary covariance matrices as suggested in Chapter One can be extended to allow for multiple QTL alleles (more than two) through simple modifications.

Inbred lines are e. g. recombinant inbred lines (RILs), doubled haploid lines (DHLs) or lines which are homozygous at the QTL, i. e. two genotypically different matings. RILs are produced through crossing inbred lines and repeated siblings mating or consecutive selfing. Hence, per generation the mean homozygosity increases by 50 %. In this way new inbred lines are derived and advantages of RILs are shown by, e. g., Teuscher and Broman (2007). Crossing such RILs is profitable in high-resolution mapping and can be analyzed using the VCM.

Considering biallelic QTL in DHLs as found in plant breeding, only additive and additive-by-additive genetic effects may be important to explain genetic variation (e. g. Choo and Reinbergs 1979; Choo *et al.* 1979; Gallais 1990; Malmberg *et al.* 2005). DHLs are completely homozygous (perfect homozygous) at each locus, i. e. there are only two segregating genotypes, because haploid cells are doubled. In this way large numbers of individuals can be produced. Both alleles at each locus are equally frequent. DHLs are used instead of F_2 populations, because the production of inbred lines is time consuming, and therefore costly. For instance, DHLs enhanced the maize breeding and these lines are more effective for selection through the higher genetic variance among DHLs compared to F_2 populations (e. g. Mayor and Bernardo 2009). Crosses between DHLs and testcrosses are conceivable as suggested by, e. g., Radoev *et al.* (2008) in rapeseed. Using a testcross hybrid, i. e. the testcross population is genetically equivalent to a backcross (BC) population (Radoev *et al.* 2008), can be analyzed applying the VCM. Also, an independent tester can be used, but there are may occur additional QTL alleles (Radoev *et al.* 2008).

Using the elementary covariance matrices as suggested in Chapter One, the RRM and

the IRM can also be applied to BC populations derived from inbred lines. Only few modifications are necessary, where non-existent QTL-genotypes are eliminated from the elementary covariance matrices as well as from the conditional QTL genotype probabilities. In BC populations the number of possible marker genotypes and QTL genotypes is reduced, i. e. there are only two possible QTL genotypes at each locus, analogously to RILs and DHLs.

The VCM can be also applied for more advanced generations from inbred parental lines. Advanced intercross lines derived from two inbred lines have the same QTL genotypes as F_2 populations and are applied in high-resolution mapping (Darvasi and Soller 1995). Mapping of imprinted QTL in a combination of multiple line crosses derived from inbred lines using the VCM was suggested by Li and Cui (2009).

A general extension of the VCM to any type of multicross designs derived from inbred lines at any generation was proposed by Crepieux *et al.* (2004), where estimated coefficients of coancestries between parental lines are considered, i. e. these lines are not assumed to be unrelated and hence uncorrelated a priori, because these authors pointed out that e. g. in plant breeding often crosses between highly related elite lines are of interest. In this way an approach was suggested, where individuals from different families may be related and these information are used to improve the accuracy of the estimated parameters (Crepieux *et al.* 2004).

Furthermore, the IRM and the RRM can be applied to noninbred full-sib (FS) families, i. e. regular FS families, where the fact that such individuals are noninbred have to be taken into account in the elementary covariance matrices, see also Xie *et al.* (1998) for detailed explanations.

In general, it is important to remark that the methodology as suggested in Chapter One can be adapted to a wide range of designs as describe above. Hence, the VCM is a flexible approach for QTL mapping in different populations. Adapted conditional QTL genotype probabilities as well as elementary covariance matrices are required.

Advantages and opportunities of the RRM: Recall that the average covariance matrices of the RRM have order equal to the number of different marker classes for each considered family, i. e. they are independent of the experimental size, because F_2 individuals are grouped according to their flanking marker genotypes. In this way an average genotypic effect for each possible marker class in each family is estimated. The number of parameters which have to be estimated of the RRM is reduced as long as the number of individuals per family is high. The power to detect QTL is larger when using a small number of large families than a larger number of small families (Xie

et al. 1998). Li and Cui (2009) recommended the use of a balance of the number of families and their offspring size in multiple line crosses. However, the advantage of the RRM becomes larger with increasing experimental size, increasing marker density and increasing complexity of the genetic model (number of QTL, additive and nonadditive genetic effects). If nonadditive genetic effects are also be considered for QTL discovery, sufficient large populations must be investigated.

The vector of residuals of the RRM is the sum of deviations of individual genetic effects from average genetic effects, termed genetic sampling effects, plus the residual deviation from the IRM. Justified by arguments of asymptotics the covariances between genetic sampling effects are assumed to be zero, and in a simplified manner, the variances of genetic sampling effects are treated as equal for all marker classes in the RRM. Taken into account the genetic sampling effects, adequate weights were suggested in Chapter One, which may provide a better approximation of the exact restricted log-likelihood. In this way the statistical power may be increased and QTL positions may be estimated more precisely, but the computational burden increases due to adjusting the weights. The reduced dimension of the covariance matrices is persistent, but convergence of the weighted RRM seems to be more sensitive to starting values of the variance components. The weights are similar to that weights used in the weighted least-squares based method as proposed by Xu (1995) and Xu (1998).

In the RRM individuals with the same marker genotype are something like repeated measurements. The phenotypes of individuals with the same marker class can be handled like repeated measurements in longitudinal data analysis, whereby an adapted variance structure is necessary. Such an analysis can be equal to the weighted RRM, where the correct inference of the genetic sampling effects from the data are considered.

In this thesis the observed number of individuals with a particular marker genotype was used to calculate the average genetic variances (diagonal elements) of the RRM. Alternatively, the expected number of observations per marker class could be used to compute the corresponding variance in the average relationship matrix, e. g. there are less observed levels of certain marker genotypes. Furthermore, the RRM allows the estimation of genetic effects for marker genotypes which are not observed due to (high) correlations between different marker classes, apparently with increased nominal standard errors. In contrast, the IRM does not estimate such effects without imputing an individual without phenotype (pseudo individual).

Results from the IRM may be used in marker assisted selection, because this method estimates individual genetic values. Using the RRM individuals can be selected on the (corrected) phenotypic ranking within a certain marker class. If different genetic

effects influence the variation of a quantitative trait, conditional genotypic effects, i. e. QTL effects associated with markers, are estimated for the IRM and the RRM.

Furthermore, the RRM offers the advantage of easy implementation and this method is also numerically stable, because multicollinearity due to high correlations between individuals with the same marker genotype is eliminated. Until now, the RRM was investigated in terms of the empirical power to detect QTL, the precision of QTL positions and the estimation of their effects by simulated data, hence the application to practical data is missing yet. But it is expected that the RRM doing this also very well. Therefore, the RRM is recommended for QTL mapping, because this method is computationally tractable and much faster than the IRM due to reducing the number of parameters.

Model selection: Understanding the genetic architecture of complex genetic traits is important in quantitative genetics. Multiple QTL mapping including possibly additive and nonadditive genetic effects requires large mapping populations, because the number of statistical tests increases exponentially with the number of loci (Carlborg *et al.* 2006). A problem in multiple QTL analysis comes from model selection, because it is necessary to know how many QTL and which genetic effects should be fitted in the statistical model (e. g. Xu 2003a; Meuwissen and Goddard 2004).

Mostly, QTL mapping studies start with an one-QTL model. The so-called forward selection as proposed e. g. by Kao *et al.* (1999), i. e. adding a single QTL per time, tends to miss QTL compared to applying complex genetic models and backward selection, especially if interacting QTL are present. In this way a "conditional QTL search" is carried out and is often applied (Carlborg *et al.* 2000). In Chapter Two it was shown that a forward selection failed to identify both linked QTL if they are in repulsion or if they have no main QTL effects. In general, forward selection is expected to perform well if the QTL are independent, i. e. there are no interactions and no linked QTL (Carlborg *et al.* 2005). Considering a multiple QTL model, a multi-dimensional search approach should be performed to increase the power to detect multiple, possibly linked, QTL.

Model selection behaves similar to so-called genetic algorithm (GA) approaches, i. e. search algorithm, as proposed e. g. by Carlborg *et al.* (2000) or by Nakamichi *et al.* (2001). The GA estimates the optimum number of QTL, their positions and genetic effects, whereby this approach is computationally more tractable compared to a continuous QTL search, especially for a complex genetic model. Such algorithms should be also powerful and highly robust. Nakamichi *et al.* (2001) proposed a GA

to detect closely linked QTL without epistatic interaction in F_2 populations. A GA for simultaneous mapping of multiple QTL considering also interactions was suggested by Carlborg *et al.* (2000). Their approach can be applied for any method and is a more efficient way for discovering QTL in the genome. Also a Bayesian variable selection is an option (Xu 2003a).

Remarkably, the VCM performs somewhat like a model selection procedure, because some genetic variance components are estimated on the boundary of zero in the parameter space if there is no support for additional QTL or the respective genetic components. Applying the restricted log-likelihood function to genetic models gives rise to the best fitted genetic model which have the largest restricted log-likelihood value (Jansen 1993). This property of the VCM is particularly advantageous if more complex genetic models for simultaneous mapping of multiple QTL are suggested, i. e. a comprehensive model is used to determine the number of segregating QTL and their corresponding genetic components, because putative QTL which have no influence on a quantitative trait get variance components equal zero or nearly so and the restricted log-likelihood is the same as applying a genetic model without these genetic effects and QTL. Therefore, a repeated calculation of the restricted log-likelihood is not necessary. Hence, the VCM seems to have advantages in determining the best fitted genetic model which explained the variation due to QTL.

Conclusions: In this thesis a reduced random model, named RRM, was developed for QTL analyses in inbred line-derive F_2 families. The key point of the proposed RRM based on the consideration of average genetic effects for all possible marker classes instead of individual genetic effects, because the individuals are grouped according to their flanking marker genotypes. In this way the number of parameters is essentially reduced and hence, the RRM has a considerably lower computational workload than the IRM. The basic idea of the RRM can be applied to different population structures. Furthermore, an existing method was picked up and modified to produce a sparse type of covariance matrices considering additive genetic and additive-by-additive genetic effects in the case that only two alleles are segregating within a family, whereby the allele frequency is a half. This approach is particularly advantageous for the IRM, because the computational requirements are substantially decreased in that special case and more realistic standard errors are produced. However, the RRM generally offers considerable savings in computing time compared to the IRM. In conclusion, the RRM is a computationally tractable model to map multiple linked QTL simultaneously and is recommended for QTL analyses.

LITERATURE

Beavis, W. D., 1994 The power and deceit of QTL experiments: Lessons from comparitive QTL studies. In *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference*, p. 250–266, Washington, DC, American Seed Trade Association.

Beavis, W. D., 1998 QTL Analyses: Power, Precision, and Accuracy. In *Molecular Dissection of Complex Traits*, edited by A. H. Paterson, p. 145–162, New York, CRC Press.

Carlborg, O., L. Andersson, and B. Kinghorn, 2000 The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics **155**: 2003–2010.

Carlborg, O., G. A. Brockmann, and C. S. Haley, 2005 Simultaneous mapping of epistatic QTL in DU6i x DBA/2 mice. Mamm. Genome **16**: 481–494.

Carlborg, O., L. Jacobsson, P. Ahgren, P. Siegel, and L. Andersson, 2006 Epistasis and the release of genetic variation during long-term selection. Nat. Genet. **38**: 418–420.

Cheverud, J. M., T. T. Vaughn, L. S. Pletscher, A. C. Peripato, E. S. Adams, C. F. Erikson, and K. J. King-Ellison, 2001 Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice. Mamm. Genome **12**: 3–12.

Choo, T., B. Christie, and E. Reinbergs, 1979 Doubled haploids for estimating genetic variances and a scheme for population improvement in self-pollinating crops. Theor. Appl. Genet. **54**: 267–271.

Choo, T. and E. Reinbergs, 1979 Doubled haploids for estimating genetic variances in presence of linkage and gene association. Theor. Appl. Genet. **55**: 129–132.

Crepieux, S., C. Lebreton, B. Servin, and G. Charmet, 2004 Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. Genetics **168**: 1737–1749.

Darvasi, A. and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141**: 1199–1207.

Doebley, J., A. Stec, and C. Gustus, 1995 Teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics **141**: 333–346.

Gallais, A., 1990 Quantitative genetics of doubled haploid populations and application to the theory of line development. Genetics **124**: 199–206.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson, 2008 *ASReml User Guide Release 3.0*. VSN International, Hemel Hempstead, UK.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135**: 205–211.

Jiang, C. and Z.-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica **101**: 47–58.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

Li, G. and Y. Cui, 2009 A statistical variance components framework for mapping imprinted quantitative trait locus in experimental crosses. J. Probab. Stat. **2009**: 1–27.

Malmberg, R. L., S. Held, A. Waits, and R. Mauricio, 2005 Epistasis for fitness-related quantitative traits in Arabidopsis thaliana grown in the field and in the greenhouse. Genetics **171**: 2013–2027.

Mayer, M., 2004 Limitations of a two-step moment method for mapping linked multiple QTL. Genet. Res. **84**: 145–152.

Mayor, P. J. and R. Bernardo, 2009 Genomewide selection and market-assisted recurrent selection in doubied haploid versus F_2 populations. Crop Sci. **49**: 1719–1725.

Meuwissen, T. H. E. and M. E. Goddard, 2004 Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. **36**: 261–279.

Nakamichi, R., Y. Ukai, and H. Kishino, 2001 Detection of closely linked multiple quantitative trait loci using a genetic algorithm. Genetics **158**: 463–475.

Neumaier, A., 1998 Solving ill-conditioned and singular linear systems: A tutorial on regularization. SIAM Rev. Soc. Ind. Appl. Math. **40**: 636–666.

Radoev, M., H. C. Becker, and W. Ecke, 2008 Genetic analysis of heterosis for yield and yield components in rapeseed (Brassica napus L.) by quantitative trait locus mapping. Genetics **179**: 1547–1558.

Rönnegård, L., K. Mischenko, S. Holmgren, and Ö. Carlborg, 2007 Increasing the efficiency of variance component quantitative trait loci analysis by using reduced-rank identity-by-descent matrices. Genetics **176**: 1935–1938.

Rönnegård, L., R. Pong-Wong, and Ö. Carlborg, 2008 Defining the assumptions underlying modeling of epistatic QTL using variance component methods. J. Hered. **99**: 421–425.

Teuscher, F. and K. W. Broman, 2007 Haplotype probabilities for multiple-strain recombinant inbred lines. Genetics **175**: 1267–1274.

Xie, C., D. D. Gessler, and S. Xu, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics **149**: 1139–1146.

Xu, S., 1995 A comment on the simple regression method for interval mapping. Genetics **141**: 1657–1659.

Xu, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics **148**: 517–524.

Xu, S., 2003a Estimating polygenic effects using markers of the entire genome. Genetics **163**: 789–801.

Xu, S., 2003b Theoretical basis of the Beavis Effect. Genetics **165**: 2259–2268.

Xu, S. and Z. Jia, 2007 Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics **175**: 1955–1963.

Yi, N., D. K. Zinniel, K. Kim, E. J. Eisen, A. Bartolucci, D. B. Allison, and D. Pomp, 2006 Bayesian analyses of multiple epistatic QTL models for body weight and body composition in mice. Genet. Res. **87**: 45–60.

SUMMARY

A variance component method (VCM) considering quantitative trait loci (QTL) effects as random in a linear model may be computational expensive, especially with multiple QTL and interactions, because of a large number of genetic effects. This thesis deals with possibilities to reduce this computational burden by modifying elements of the underlying random model for inbred line-derived F_2 populations, considering additive genetic, dominance and epistatic interaction effects.

A reduced random model (RRM), as suggested in Chapter One, considers average genetic effects for all possible marker genotypes instead of genetic effects for each individual as the traditional individual random model (IRM) does. It could be shown that the genetic covariance structure of the RRM is asymptotically equivalent to the genetic covariance in the IRM. Because the number of parameters to be estimated is essentially decreased in the RRM it clearly outperforms the IRM with respect to computational speed.

Comprehensive comparisons as done in Chapter Two show that the RRM is competitive to the IRM in terms of the precision of the estimated QTL positions and the observed power. Both VCM were also compared to fixed models like regression interval mapping (RIM) and multiple interval mapping (MIM). No major differences between RRM compared to IRM, RIM and MIM in terms of the QTL detection power and the accuracy of the estimated QTL positions and their effects occurred. Hence, the RRM is a computationally tractable method and is recommended for QTL analyses instead of the IRM, especially in experiments with multiple families.

Chapter Three revisits additive and additive-by-additive genetic relationship matrices. Alternative covariance matrices are proposed, capitalizing on the prior knowledge of only two different QTL alleles in the considered type of experiments. The resulting covariance matrices and their inverses have a considerable amount of zero elements, leading to a remarkable gain in computational speed if this sparse structure is exploited in the estimation of genetic variance components. Thereby the restricted log-likelihood function remains unaltered. Moreover, more realistic standard errors for genetic effects are obtained.

In conclusion QTL analyses in inbred line-derived F_2 families can be speeded up by applying the RRM and, in case of only additive genetic effects and their interactions, by applying a sparse type of genetic covariance matrices.

ZUSAMMENFASSUNG

Varianzkomponentenmethoden mit als zufällig betrachteten QTL-Effekten (quantitative trait loci, QTL) können wegen einer großen Anzahl an genetischen Effekten sehr rechenintensiv sein, insbesondere wenn multiple QTL mit additiven, dominanten und epistatischen Effekten berücksichtigt werden. Diese Arbeit untersucht Möglichkeiten, den Rechenaufwand durch geeignete Anpassungen des zufälligen Modells zu reduzieren. Hierbei werden aus Inzuchtlinien abgeleitete F_2-Populationen betrachtet.

Im ersten Kapitel wird ein reduziertes zufälliges Modell (RRM) mit durchschnittlichen genetischen Effekten für alle auftretenden Markergenotypen vorgeschlagen, anstelle von individuellen genetischen Effekten, wie im bekannten individuellen zufälligen Modell (IRM). Die asymptotische Äquivalenz der genetischen Kovarianz von RRM und IRM wurde gezeigt. Durch die geringere Anzahl von Parametern im RRM ergibt sich eine deutliche Verringerung der Rechenzeiten im Vergleich zu IRM.

Umfangreiche Simulationen im zweiten Kapitel demonstrieren die Gleichwertigkeit von RRM und IRM hinsichtlich der Leistungsfähigkeit (Güte) der Kartierung und der Genauigkeit der geschätzten QTL-Positionen, sowie deren Effekte. Auch zu fixen Modellen, namentlich der Regressionsmethode und der multiple Intervallkartierung, traten keine nennenswerten Unterschiede auf. Wegen der rechentechnischen Vorteile kann deshalb das RRM für QTL-Analysen empfohlen werden, wobei Experimente mit multiplen Familien das Hauptanwendungsgebiet darstellen.

Im dritten Kapitel werden alternative additive und additiv-mal-additive Kovarianzmatrizen vorgeschlagen, welche die Vorkenntnis nutzen, dass nur zwei QTL-Allele im untersuchten F_2-Versuchsdesign vorliegen. Diese neuen Kovarianzmatrizen und ihre Inversen haben einen beachtlichen Anteil an Nullelementen, was zu einer höheren Rechengeschwindigkeit führt, wenn diese dünn besetzten Strukturen für die Schätzung der Varianzkomponenten genutzt werden. Dabei bleibt die restringierte log-Likelihoodfunktion unverändert, allerdings werden realistischere Standardfehler für die geschätzten genetischen Effekte erhalten.

Somit können QTL-Analysen in aus Inzuchtlinien abgeleiteten F_2-Populationen unter Verwendung von RRM beschleunigt werden, außerdem durch die Anwendung der vorgeschlagenen dünn besetzten Verwandtschaftsmatrizen für additive Effekte und deren Interaktionen.

APPENDIX

LINEAR MODEL WITH FIXED EFFECTS

RIM and MIM consider QTL effects as fixed and these methods directly estimate them. An F_2-metric model was suggested by KAO and ZENG (2002) and ZENG *et al.* (2005), which is an orthogonal model for allele frequency a half (two alleles) in an equilibrium population.

A two-QTL model with additive genetic effects for both QTL (a_1 and a_2) and an additive-by-additive genetic effect between both QTL (aa) is assumed. The orthogonal partition of genotypic effects can be done, e. g. using KAO and ZENG (2002),

$$\begin{aligned} a_1 &= \frac{\mu_{QQ} - \mu_{qq}}{2}, & a_2 &= \frac{\mu_{HH} - \mu_{hh}}{2}, \\ aa &= \frac{\mu_{QQHH} - \mu_{QQhh} - \mu_{qqHH} + \mu_{qqhh}}{4}, \end{aligned}$$

where the genotypic value of the QTL genotype G_{QQHH} is denoted by μ_{QQHH} with $\mu_{QQHH} = \mu + a_1 + a_2 + aa$. Other genetic values are defined analogously to (KAO and ZENG 2002, Table 3).

Multiple interval mapping (MIM): MIM is a maximum likelihood approach, using multiple marker intervals simultaneously for multiple QTL mapping. In F_2 populations the quantitative trait value y_t of individual t with $t = 1, \ldots, n$ (n is the number of F_2 individuals) modeling a population mean μ as well as two QTL with additive genetic effects for each QTL (a_1 and a_2) and an additive-by-additive genetic effects (aa) between both QTL is

$$y_t = \mu + x_{a_1,t} a_1 + x_{a_2,t} a_2 + x_{a_1,t} x_{a_2,t} aa + e_t \,.$$

The residual error e_t of individual t is assumed to follow a normal distribution with mean zero and variance σ_e^2. If the QTL genotype is known, which is usually not the case, the coded variables ($\ell \in 1, 2$) are

$$x_{a_\ell,t} = \begin{cases} 1 & \text{if genotype is } G_{QQ} \text{ at } \ell\,\text{th QTL} \\ 0 & \text{if genotype is } G_{Qq} \text{ at } \ell\,\text{th QTL} \\ -1 & \text{if genotype is } G_{qq} \text{ at } \ell\,\text{th QTL} \end{cases} .$$

In general, QTL genotypes are uncertain. Additionally to the vector of phenotypes $\boldsymbol{y}$ (length n), marker genotypes $\boldsymbol{M}$ are considered. The likelihood function is a normal mixture of nine possible QTL genotypes. Then, the likelihood function of parameters $\boldsymbol{\theta} = (\mu, a_1, a_2, aa, \sigma_e^2)'$, which have to be estimated, is generally expressed as

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{M}) \;=\; \prod_{t=1}^{n} \left(\sum_{i=1}^{9} p_{ti}\, N(\mu_{ti}, \sigma_e^2) \right) ,$$

where $N(\mu_{ti}, \sigma_e^2)$ denotes the normal density function and means μ_{ti} are the corresponding genotypic values of the nine different QTL genotypes in the population. Mixing proportions p_{ti}, where i denotes the possible QTL genotypes, are functions of the putative QTL positions and they are conditional QTL genotype probabilities given the flanking markers and the putative QTL positions. Kao and Zeng (1997) presented general formulas for deriving maximum likelihood estimates and the asymptotic variance-covariance matrix of the estimates using the expectation maximization algorithm. More details and extensions, e. g. to multiple QTL, are given by Kao *et al.* (1999), Zeng *et al.* (1999) and Kao (2000).

Test statistic: The detection of the QTL with MIM as well as VCM can be achieved by a (restricted) log-likelihood ratio test statistic as $(R)LRT = 2\,(\ln \mathcal{L}_A - \ln \mathcal{L}_0)$, where $\mathcal{L}_A$ is the (restricted) log-likelihood function under the alternative hypothesis H_A and $\mathcal{L}_0$ is the (restricted) log-likelihood function under the null hypothesis H_0.

In Chapter Two the alternative hypothesis $H_A^{(12)}$ (3.3) was used to determine the number of identified QTL (two QTL *vs.* one QTL). It was described that the maximum values of the test statistics of Equations (3.1) and (3.2) were used to test for the second QTL using the alternative hypothesis $H_A^{(12)}$,

$$(R)LRT \;=\; \max_{1 \le j \le c} (R)LRT_j \qquad \text{with} \quad (R)LRT_j \;=\; 2\left(\ln \mathcal{L}_A^{(2)} - \ln \mathcal{L}_A^{(1)} \right) ,$$

where $\mathcal{L}_A^{(1)}$ is the (restricted) log-likelihood under the alternative hypothesis $H_A^{(1)}$ (3.1) and $\mathcal{L}_A^{(2)}$ is the (restricted) log-likelihood under the alternative hypothesis $H_A^{(2)}$ (3.2). The number of possible QTL positions or combinations thereof is c.

Regression interval mapping (RIM): RIM is a least-squares based method, which regresses quantitative trait values onto conditional expected genotypic values. The complexity of the linear regression is given by the number of available regression parameters in the genetic model. In an F_2 population the phenotypic value y_t with $t = 1, \ldots, n$ of a quantitative trait of an individual t considering a population mean

μ as well as two QTL with additive genetic effects for each QTL (a_1 and a_2) and an additive-by-additive genetic effects (aa) between both QTL is modeled as

$$y_t = \mu + r_{a_1,t}a_1 + r_{a_2,t}a_2 + r_{aa,t}aa + e_t\,.$$

The residual error e_t of individual t is assumed to follow a normal distribution with mean zero and variance σ_e^2. Conditional QTL genotype probabilities are used as regressor variables in the design matrix in RIM. The regressor for the additive genetic effect is $r_{a_\ell,t} = p_i^{QQ} - p_i^{qq}$ at a putative QTL $\ell \in \{1,2\}$ and for additive-by-additive genetic effect is $r_{aa,t} = p_i^{QQHH} - p_i^{QQhh} - p_i^{qqHH} + p_i^{qqhh}$. RIM approximates the mixture of normal distributions of MIM by a single normal one,

$$\mathcal{L}\left(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{M}\right) = \prod_{t=1}^{n} N\left(y_t - e_t,\, \sigma_e^2\right)\,.$$

Test statistic: Analogously to MIM, the alternative hypothesis $H_A^{(12)}$ (3.3) was used to determine the number of identified QTL (two QTL *vs.* one QTL) in Chapter Two. The residual sum of squares (RSS) of the QTL model under the alternative hypothesis H_A is RSS_A (full model, γ parameters) and RSS of the model under the null hypothesis H_0 is RSS_0 (reduced model, ω parameters).

Note that simple transformations are necessary to calculate the test statistic F of $H_A^{(12)}$ (3.3) using $H_A^{(1)}$ (3.1) and $H_A^{(2)}$ (3.2) of RIM. The test statistic F using $H_A^{(1)}$ (3.1) is denoted by $F_{H_A^{(1)}}$ and using $H_A^{(2)}$ (3.2) is denoted by $F_{H_A^{(2)}}$,

$$F_{H_A^{(z)}} = \frac{n-\gamma^{(z)}}{\gamma^{(z)}-\omega}\left(\frac{RSS_0}{RSS_A^{(z)}} - 1\right)\,,$$

where ω denotes the number of parameter under the null hypothesis of no QTL. The number of parameters using $H_A^{(z)}$ with $z \in \{1,2\}$ is $\gamma^{(z)}$ with corresponding $RSS_A^{(z)}$. Using the test statistic of (3.1) and (3.2) leads to

$$F_{H_A^{(z)}}\frac{\gamma^{(z)}-\omega}{n-\gamma^{(z)}} + 1 = \frac{RSS_0}{RSS_A^{(z)}}\,,$$

where all parameters are known and the proportion is used to construct the test statistic F of $H_A^{(12)}$ (3.3),

$$F = \max_{1\le j\le c} F_j \qquad \text{with} \quad F_j = \frac{n-\gamma^{(2)}}{\gamma^{(2)}-\gamma^{(1)}}\left(\frac{RSS_A^{(1)}}{RSS_A^{(2)}} - 1\right)\,.$$

LINEAR MIXED MODEL FOR RANDOM REGRESSION

A random regression approach for multiple family analysis considers the QTL effects as random in a linear mixed model. The vector of phenotypes $\boldsymbol{y}$ (length n) is modeled with respect to an additive genetic effect as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{a}_1 + \boldsymbol{e}$, where $\boldsymbol{\beta}$ considers a fixed effect for each family and $\boldsymbol{X}$ is the related design matrix of suited order. The vector $\boldsymbol{a}_1' = (a_{11}', \ldots, a_{1f}')$ includes the random additive genetic regression coefficients, one for each family. The expectation of the additive genetic regression coefficients is $E(\boldsymbol{a}_1) = \boldsymbol{0}$ and the covariance matrix is assumed as $\mathrm{Var}(\boldsymbol{a}_1) = \boldsymbol{I}\sigma^2_{a_1}$ with $\boldsymbol{I}$ is an identity matrix (order f), because individuals from different families are assumed to be uncorrelated. Matrix $\boldsymbol{Z}_1 = \{z_{sc}\}$ has the dimension $n \times f$ and includes the expected difference of homozygous genotype coefficients, where $s = 1, \ldots, n$ and $c = 1, \ldots, f$. More precisely, for individual s with marker genotype $i \in \{1, \ldots, 9\}$ the regressor is $z_{sc} = p_i^{QQ} - p_i^{qq}$ if this individual is related to family c and zero otherwise. Residuals are assumed to follow a normal distribution with $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2_e)$, where $\boldsymbol{I}$ is an identity matrix of adequate dimension. This assumption does not hold, because the residual variance also contains the genetic sampling variance.

CALL SHELL FROM A FORTRAN PROGRAM

ASReml (Gilmour *et al.* 2008) is used to estimate variance components in combination with the restricted maximum likelihood function (Patterson and Thompson 1971). For calculating the conditional QTL genotype probabilities, for setting up the required relationship matrices or for assignment of the levels, e. g. marker classes, to the individuals, self written Fortran 95 programs are used. Calling shell and execute ASReml from an active Fortran program is done as follows.

```
shell = 'ASReml -LNS11 Sim.as '//trim(parameter)//' '  ! string: shell
io = system(shell)                                     ! integer: io
IF(io /= 0) WRITE(*,*),'ERROR'                         ! error message
```

The job file "Sim.as" is executed. The variable "io" is declared as integer and the variable "shell" is a character string of sufficient length. The system function is not necessarily thread-safe and therefore system should only be used to execute a shell command. If the shell command was successful, the variable "io" is zero.

Here, "shell" is produced through two parts, combined with slashes. The parameters, stored in character variables, are passed to the Fortran program, e. g. starting values of the variance components which differs between changing positions of the QTL.

If the CPU time of ASReml should be measured, the variable “shell” is for example

```
shell = 'time (ASReml -LNS11 Sim.as '//trim(parameter)//')
                    >& time.txt >> asreml.txt'
```

where the measured time is stored in the file “time.txt” and the ASReml output in “asreml.txt”. The file “asreml.txt” is written continuously, while the file “time.txt” is overwritten.

ASREML JOB FILE

An example job file with the filename extension “.as” for VCM is shown below.

```
Mixed Model Results             # title line
                                # labels of the data file
  run                           # repetition
  animal                        # animal ID
  family  $3                    # family ID for related animal
  a       $4                    # level for additive effects
  d                             # level for dominance effects
  y                             # phenotypes
  wg                            # weights

A.grm                           # file name of covariance matrix
Levels.txt !NOREORDER !MAXIT $1 !DOPART $2 # data file name, qualifiers

!Part 1
y ~ family !r giv(a,1) $5 !GP   # definition of LMM for VCM

!Part 2
y ~ family !r a                 # analogous definition of LMM for VCM
0 0 1                           # variance model: R=I, 1 cov. structure
a 1                             # specification of variance model
a 0 GIV1 $5 !GP

!Part 3
y !WT wg ~ family !r a          # definition of LMM for VCM with weights
0 0 1                           # variance model: R=I, 1 cov. structure
```

```
a 1                                # specification of variance model
a 0 GIV1 $5 !GP

!Part 4
y ~ family !r family.a $5 !GP   # definition of LMM for RR
```

The data structure is given in the definition of the data file, in this case called "Levels.txt", where the columns include information to "run", "animal", "family", "a", "d", "y" and "wg" in that order. The qualifier "!NOREORDER" prevents the reorganization and "!MAXIT" specifies the maximum number of iterations. The linear mixed model considers a family specific mean "family" as fixed effect and the additive genetic effect as random. The covariance structure is given in file "A.grm" and is considered through "giv(a,1)". The qualifier "!GP" restrict the updating of the variances to be in the theoretical parameter space. The covariance matrices with ending ".grm" denotes the covariance matrix itself, whereas ".giv" denotes the inverse covariance matrix. The qualifiers "$i" passed from the ASReml call, e. g. the variable "parameter" from Paragraph "Call Shell from a Fortran program".

The qualifier "!DOPART" defined the part of the job file which should be used for analysis. Part two is equal to part one, where the covariance structure is defined explicitly. The covariance structure of the residuals is equal to an identity matrix and a covariance structure of genetic effect is well-defined. In part three the linear mixed model of VCM using weights is shown. For a random regression approach an example of the job file is given in part four.

LITERATURE

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson, 2008 *ASReml User Guide Release 3.0*. VSN International, Hemel Hempstead, UK.

Kao, C.-H., 2000 On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics **156**: 855–865.

Kao, C.-H. and Z.-B. Zeng, 1997 General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53**: 653–665.

Kao, C.-H. and Z.-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160**: 1243–1261.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

Patterson, H. D. and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika **58**: 545–554.

Zeng, Z.-B., C.-H. Kao, and C. J. Basten, 1999 Estimating the genetic architecture of quantitative traits. Genet. Res. **74**: 279–289.

Zeng, Z.-B., T. Wang, and W. Zou, 2005 Modeling quantitative trait loci and interpretation of models. Genetics **169**: 1711–1725.

ACKNOWLEDGEMENT

Many thanks to my supervisor, Norbert Reinsch, for supporting my work, for reading the manuscript carefully, for helpful discussions and also for a lot of constructive comments during the whole project.

I am grateful to Manfred Mayer that I could have worked on this project which was financially supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, MA 1553/3-1). I am also thankful to Manfred Mayer and Friedrich Teuscher for valuable discussions.

My special thanks to Dörte Wittenburg for all her help accompanying my thesis, many discussions and emotional support. Thank you Nina Melzer, Sandra Andorf and Vinzent Börner for all the beautiful evenings, for technical support and for discussions. Thanks to Denis Scherpeltz for reading some parts of the manuscript, but rather for assistance in all possible ways.

CURRICULUM VITAE

Persönliche Daten	
Name	Daisy Zimmer
Geburtsdatum	12. September 1984
Geburtsort	Räckelwitz (Sachsen)
Staatsangehörigkeit	deutsch
Schulausbildung	
1991 – 1995	Grundschule Kamenz und Elstra
1995 – 1996	Mittelschule Elstra
1996 – 2003	Gotthold-Ephraim-Lessing-Gymnasium Kamenz Abschluss: Allgemeine Hochschulreife
Hochschulausbildung	
09/2003 – 11/2007	Studium an der Hochschule Zittau/Görlitz (FH) - University of Applied Sciences Abschluss: Diplom-Biomathematikerin (FH)
Promotion	
seit 12/2007	Leibniz-Institut für Nutztierbiologie, Forschungsbereich Genetik und Biometrie FBN Dummerstorf