Statistical identification of major genes in pigs

Promotor:	dr. ir. E.W. Brascamp
	hoogleraar in de veefokkerij
Co-promotor:	dr. ir. J.A.M. van Arendonk
	universitair hoofddocent veefokkerij

Statistical identification of major genes in pigs

L.L.G. Janss

Proefschrift

ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van de Landbouwuniversiteit Wageningen,
dr. C.M. Karssen,
in het openbaar te verdedigen
op vrijdag 10 januari 1997
des namiddags te één uur dertig in de aula
van de Landbouwuniversiteit te Wageningen.

Abstract

Janss, L.L.G., 1996. Statistical identification of major genes in pigs. Doctoral thesis, Department of Animal Breeding, Wageningen Agricultural University, P.O. Box 338, 6700 AH Wageningen, The Netherlands.

This thesis considers use of segregation analysis for detection of major genes in livestock populations. Segregation analysis has not found widespread use in livestock because of the general impossibility to perform the required computations in the large and complicated population structures encountered. In this thesis, a Bayesian approach to segregation analysis is developed, which makes use of Markov Chain Monte Carlo (MCMC) methodology to perform the otherwise intractable computations. The Bayesian approach combined with the MCMC computing methodology, proved very flexible in the construction of realistic models for the analysis of livestock data. Several analyses are reported from data on crossbred pigs, demonstrating the likely existence of several major genes affecting traits of biological importance.

Cover photographs: Alien Jalvingh
ISBN 90-5485-597-5

Stellingen

1. Wanneer verschillende allelen van een hoofdgen in ouderlijnen gefixeerd zijn, kan met fenotypische waarnemingen aan een F_2-kruising zulk een hoofdgen slecht opgespoord worden.
   Dit proefschrift

2. Bij het gebruik van gecombineerde gegevens van een F_1- en een F_2-kruising zal in een model waarin slechts een hoofdgen een variantieverhoging kan verklaren, elke variantieverhoging verklaard worden door een hoofdgen.
   Dit proefschrift

3. Met behulp van hoofdgenen, zoals de aangetoonde genen voor intramusculair vet en rugspek, kan een fokkerij-organisatie efficient en flexibel een divers produktenpakket leveren.
   Dit proefschrift

4. In de toepassing van statistiek zijn Bayesiaanse methodes een logische voortzetting van reeds aanwezige trends om absoluut gestelde zekerheden te vervangen door gemodelleerde onzekerheden.
   Dit proefschrift

5. Door het grote aantal afstammingslussen zijn simpele recursieve pel-algorithmes ongeschikt voor toepassing in uitgebreide afstammingen van landbouwhuisdieren, dit in tegenstelling tot bijvoorbeeld de suggestie van Fernando et al. (1993, Theor. Appl. Genet., 87: 89-93).
   Dit proefschrift

6. Het herhaald trekken van realisaties uit een set van conditionele kansverdelingen construeert niet noodzakelijk een valide Gibbs keten.
   Naar: Hobert en Casella, 1994, Techn. rapport BU-1221-M, Cornell University.

7. Kwantitatieve genetici maken al snel de fout te veronderstellen dat de aanwezigheid van additieve variantie en additieve fokwaardes de aanwezigheid van onderliggende additieve genen zou impliceren.

8. Door de complexiteit van genetische regulatie zal het enthousiasme voor het vinden van genen die kwantitatieve kenmerken beïnvloeden slechts stand houden tot daadwerkelijk zulke genen zijn gevonden.

9. Wetenschappelijke kennis of implicaties van wetenschappelijke technieken dienen publiek te zijn opdat de maatschappij grenzen kan stellen aan de toepassing van deze kennis of technieken.

10. Het theoretische genetische onderzoek van de verschillende genetica-vakgroepen van de LUW zou samengebracht moeten worden in een "theoretische genetica" groep.

11. Gezien de gebruikelijke betekenis van "fokken" in het Nederlands (Van Dale: doen voorttelen, aankweken van vee) is "veefokkerij" een onjuiste benaming voor het vakgebied dat zich bezig houdt met de genetische verbetering van vee.

12. Computersystemen evolueren als levende organismes.

13. De toegenomen emancipatie van de vrouw blijkt onder andere uit het verhoogde aandeel vrouwen onder de hardrijders en bumperdrukkers.

14. Rode koeien zonder billen zijn eigenlijk zwart.

15. De geest is onlosmakelijk verbonden met het lichaam.
naar: Edelman, "Bright air, brilliant fire. On the matter of the mind", Basic Books, 1991.

Stellingen bij het proefschrift van L.L.G. Janss "Statistical identification of major genes in pigs", Landbouwuniversiteit Wageningen, te verdedigen op 10 januari 1997.

Acknowledgements

The writing of a thesis requires various skills. First, one should develop a curiosity to investigate the surrounding world. To develop such a curiosity, genes as well as the environment (likely the environment in early life) will be important. Hence, my parents must have played an important role in this respect. Then, some education will be required. In retrospect, the teachers at secondary school did succeed to teach me some language skills, although at the time their efforts seemed futile. In the fields of statistics and genetics, outside the standard university curriculum, I have learned a lot on my stays abroad from people like Jean-Louis Foulley, Daniel Gianola and Robin Thompson. The first two are responsible for teaching me some Bayesian statistics and I am very proud to present in this thesis some Bayesian analyses. The latter, Robin Thompson, has been an important catalyst in developing the Monte Carlo Markov Chain approaches for segregation analysis presented in this thesis. Writing a thesis also requires a good share of practicality, common sense and just hard work. This, of course, is an outstanding quality of the Dutch and I am indebted to Julius van der Werf, Johan van Arendonk and Pim Brascamp for their coaching. And, writing a thesis requires a pleasant atmosphere, which was supplied by my room mates Imke and Sijne, and, in the crucial 'terminal phase', by Ivonne. To all these people: this booklet is a bit yours as well.

Contents

[a] Department of Animal Breeding
Wageningen Institute of Animal Sciences
Wageningen Agricultural University
P.O. Box 338
6700 AH Wageningen, The Netherlands

[b] *Present Address:*
Department of Animal Science
University of New England
Armidale, NSW 2351, Australia

[c] Department of Biometrical Genetics
AFRC Roslin Institute
Roslin, Midlothian, EH25 9PS, UK
Present Address:
Statistics Department
IACR-Rothamsted
Harpenden, Hertfordshire AL5 2JQ, UK

General introduction: the Dutch Meishan crossing experiment and aim of this thesis

Chapter 1

Development of a synthetic line with Meishan could be an interesting approach to improve fertility of Western pig lines. To investigate the potential of such approach, Dutch pig breeding companies have produced F_1 and F_2 Meishan x Western crossbreds. Aim of this thesis is development of statistical methodology to model major gene inheritance, and analysis of data collected on the produced Meishan crossbreds for presence of major genes.

Meishan crossing experiment

One of the activities of commercial pig breeding companies is the marketing of young, generally hybrid, sows, to be used for commercial weaner-production. The ideal hybrid sow should produce many piglets with high quality for fattening. The breeding goal for selection in the so-called dam-lines used to breed such hybrids, therefore, includes reproduction traits, mainly litter size, and production traits, mainly growth and backfat (Smith, 1964). Study of the economic values of genetic improvement for these traits in dam lines, shows high marginal profit for improvement of litter size under usual Western marketing conditions (De Vries, 1989) and is argued to increase further, due to decreasing marginal profits for improvement of production traits (e.g., Haley, 1988; Bidanel, 1990). Improvement of litter size, therefore, is, and will remain, a main objective in breeding dam-lines.

A first choice to improve litter size is selection within available lines. Avalos and Smith (1987) computed that, in theory, considerable annual genetic gain for litter size should be reachable, but in practice it is generally considered that large resources and consistently continued selection for several generation will be required to improve litter size (Bichard and David, 1985). Large resources can be found by using hyper-prolific schemes (Legault and Gruand, 1976), but such schemes know long generation intervals, which is not beneficial (Avalos and Smith, 1987), and seem typically designed for large herdbook-type organisations. Progress by selection within lines, in whatever manner performed, therefore, will be slow. An alternative to improve litter

size is use of new genetic material from highly fertile breeds, where the Chinese Meishan breed may be of interest. Bidanel et al. (1990) summarised comparisons of Meishan with Large White, indicating that Meishans had an advantage in litter size of about 3 piglets, but had a disadvantage in growth (200 gr/day from $2\frac{1}{2}$ to 5 months of age) and a disadvantage in carcass lean meat content (20% at 5 months of age). For commercial application, Bidanel et al. (1991) indicated that development of an improved synthetic line with 50% Meishan, to be used as one of the parents of commercial hybrid sows, could be an interesting approach. Such approach is expected to be quickest to produce a dam-line with a commercially interesting advantage in litter size and with acceptable levels for fattening traits. Actual details, however, on how well and how fast such a synthetic line could be developed are relatively vague, because genetic aspects of important traits in such a line are unknown. Genetic aspects could be very relevant, for instance, presence of a major gene affecting one of the important traits could be an important aid, while a strong unfavourable genetic correlation between litter size and lean meat contect could be a large impediment for development of such a synthetic line.

To investigate relevant genetic aspects for development of synthetic lines with Meishan, five Dutch pig breeding companies and Wageningen Agricultural University have collaborated in a crossbreeding project. This project consisted in the production of F_1 and F_2 Meishan-crossbreds, which was set-up in such a way that one large, genetically linked, population was formed. In this project, several phenotypic measurements were collected on traits like growth, backfat and litter size, and also part of the F_2 crossbreds was slaughtered to take measurements on several meat quality traits. The initially foreseen project did not consider molecular genetic analyses, but blood samples of all animals were stored, in case such analyses would appear interesting after analysis of the collected phenotypic measurements. Based on results from genetic analyses of the produced data, each company could decide whether to pursue development of a synthetic line with Meishan by further breeding with the jointly produced crossbreds. The Meishans used in this crossbreeding project are from a pure-bred Meishan population of Euribrid BV (Boxmeer, The Netherlands), housed at Wageningen Agricultural University, and which descends from the French Meishan population.

Major genes

One of the relevant genetic aspects for development of a synthetic line, is the genetic mechanism behind the inheritance of traits, where one interesting aspect is the number of genes affecting traits. Full monogenic control of the complexely regulated quantitative traits considered is not expected, but a possible interesting variant is control by a major gene. Control by a major gene implies a partly monogenic determination, with additional effects of polygenic background genes. For genetic improvement, also oligogenic control of traits would be of interest, but by statistical methods analysing phenotypic measurements, oligogenic control is expected not distinguishable from polygenic control (when genes would have similar effects), or control by a major gene (when one gene has markedly larger effect than the other genes). As a relevant hypothesis to be investigated for the Meishan crossbreds, control of traits by a major gene, vs. polygenic control, therefore is considered.

In formation of a synthetic line, influence of a major gene could be discovered by multimodality in the trait distribution in the F_2 generation. However, by plotting the mixture distributions expected due to segregation of a major gene, one can find that appearence of multimodality requires effects of genes (difference between homozygotes) of at least 4 residual standard deviations for dominant genes or 6 residual standard deviation for additive genes. For the traits considered in the Meishan crossing experiment, line differences are too small to expect genes with such large effects causing multimodality in trait distributions. Further approaches to detect presence of major genes are based on statistical modelling, which can be based on phenotypic data only (segregation analysis) or based on phenotypic as well as molecular genetic data (linkage analysis). Throughout this thesis, use of only phenotypic data is considered, therefore remaining in the field of segregation analysis. Segregation analysis can be considered as a first screening for presence of major genes, indicating traits for which further molecular genetic analyses will be promising.

Segregation analysis

Segregation analysis (Elston and Stewart, 1971; Morton and MacLean, 1974) is a generally known term for a method for major gene detection, based on statistical modelling of a monogenic and a polygenic component to explain observed phenotypes,

and use of a (close to) exact mathematical treatment of such model. The (close to) exact mathematical treatment poses large problems, for instance in requiring to consider all possible (relevant) combinations of genotypes in a population. In animal breeding populations, this is generally impossible when considering three or more generations: in animal populations large numbers of so-called pedigree loops arise due to the typical application of multiple matings (see also Chapter 3). Further mathematical complications arise due to the additional modelling of polygenic effects, which require to be integrated out from the already intractable mixture distributions resulting from the monogenic effects. As a result, in animal breeding segregation analysis is, so far, mainly considered in theoretical studies for application to simple population structures of (assumed) independent families of one father, possibly several mothers, and offspring (e.g. Le Roy et al., 1989) and with little possibilities to also model non-genetic effects. Approaches for general application of segregation analysis are lacking.

Aim and outline of this thesis

The ultimate aim of this thesis is to investigate whether traits in the Meishan crosses are influenced by a major gene. Due to the lack of insight in detectability of major genes in crosses, and due to the lack of flexible and efficient statistical methodology to model a major gene inheritance, a large part of this thesis also is dedicated to more general theoretical aspects related to major gene detection and major gene modelling. In Chapter 2, by use of simulation studies, power of statistical tests is investigated to detect a major gene using the first two generation of a synthetic line, i.e., the data available from the Meishan crossing experiment. In Chapters 3 and 4, two approaches are developed for practical application of segregation analysis in animal populations. The first approach (Chapter 3) is an analytical approach, tackling the problem of the generally highly looped pedigrees in animal populations by development of an approach for computing approximate likelihoods in looped pedigrees. The second approach (Chapter 4) considers use of Markov chain Monte Carlo methodology, which allows for a Bayesian approach to segregation analysis. This second approach was fully developed through for general modelling of major gene inheritance. Chapters 5 and 6 then describe analyses of data from the Meishan crossing experiment for presence of major genes, using the developed Bayesian approach to segregation analysis. Chapter

5 presents results of analyses of a number of meat-quality traits, while Chapter 6 considers analyses of commercially important traits litter size, growth and backfat. In a general discussion (Chapter 7) relevance of the developed statistical methodology and relevance of the findings in Chapters 5 and 6 are discussed in general animal breeding context and in the context of development of synthetic lines with Meishan. In a second discussion chapter (Chapter 8) change of genetic variance in a synthetic line is described, which could further aid in development of synthetic lines.

References

Avalos E, Smith C (1987) Genetic improvement of litter size in pigs. Anim Prod 44: 154-164

Bichard M, David PJ (1985) Effectiveness for genetic selection for prolificacy in pigs. J Repr Fert, Suppl 33: 127-138

Bidanel (1990) Potential use of prolific Chinese breeds in maternal lines of pigs. Proc 4th World Congr Genet Appl Livest Prod, Edinburgh UK, Vol 15: 481-484

Bidanel JP, Caritez JC, Legault C (1990) Ten years of experiments with Chinese pigs in France. 1. Breed evaluation. Pig News Inform 11: 345-348

Bidanel JP, Caritez JC, Legault C (1991) Ten years of experiments with Chinese pigs in France. 2. Utilisation in crossbreeding. Pig News Inform 12: 239-243

De Vries AG (1989) Selection for production and reproduction traits in pigs. Doctoral thesis, Wageningen Agricultural University, Wageningen, The Netherlands.

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21: 523-542

Haley CS (1988) Selection for litter size in the pig. Anim Breed Abstr 56: 319-332

Legault C, Gruand J (1976) Amélioration de la prolificité des truies par la creation d'une lignée 'hyper-prolifique' et l'usage de l'insemination artificielle: principes et résultats experimentaux préliminaires. Journées Rech Porcine en France 8: 383-388

Le Roy P, Elsen JM, Knott SA (1989) Comparison of four statistical methods for detection of a major gene in a progeny test design. Genet Sel Evol 21: 341-357

Morton NE, MacLean CJ (1974) Analysis of family resemblance III. Complex segregation of quantitative traits. Am J Hum Genet 26: 489-503

Smith (1964) The use of specialised sire and dam lines in selection for meat production. Anim Prod 6: 337-344

Note:

The material in this thesis is composed of articles previously published in various journals. Notation, terminology and spelling, therefore, not always is consistent between chapters.

Identification of a major gene in F_1 and F_2 data when alleles are assumed fixed in the parental lines

Chapter 2

A maximum likelihood method is described to identify a major gene using F_2, and optionally F_1, data of an experimental cross. A model which assumed fixation at the major locus in parental lines was investigated by simulation. For large data sets (1000 observations) the likelihood ratio test was conservative and yielded a type I error of 3%, at a nominal level of 5%. The power of the test reached more than 95% for additive and completely dominant effects of 4 and 2 residual standard deviations, respectively. For smaller data sets, power decreased. In this model assuming fixation, polygenic effects may be ignored, but on various other points the model is poorly robust. When F_1 data was included any increase in variance from F_1 to F_2 biases parameter estimates and leads to putative detection of a major gene. When alleles segregate in parental lines, parameter estimates were also biased, unless the average allele frequency was exactly 0.5. The model uses only the non-normality of the distribution and corrections for non-normality due to other sources can not be made. Use of data and model in which alleles segregate in parents, e.g. F_3 data, will give better robustness and power.

Introduction

In animal breeding, crosses are used to combine favourable characteristics into one synthetic line. It is useful to detect a major gene as soon as possible in such a line, because selection could be carried out more efficiently, or repeated backcrosses be made. Once a major gene has been identified it can also be used for introgression in other lines.

Major genes can be identified using maximum likelihood methods, such as segregation analysis (Elston and Stewart, 1971; Morton and MacLean, 1974). Segregation analysis is a universal method and can be applied in populations where alleles segregate in parents. However, when applied to F_1, F_2 or backcross data

assuming fixation of alleles in parental lines, genotypes of parents are assumed known and all equal and this analysis leads to the fitting of a mixture distribution without accounting for family structure.

Fitting of mixture distributions has been proposed when pure line and backcross data as well as F_1 and F_2 data are available, and when parental lines are homozygous for all loci (Elston and Stewart, 1973; Elston, 1984). Statistical properties of this method, however, were not described, and several assumptions may not hold. For example, not much is known concerning the power of this method when only F_2 data are available, which is often the case when developing a synthetic line. Furthermore, homozygosity at all loci in parental lines is not tenable in practical animal breeding. Here it is assumed that many alleles of small effect, so called polygenes, are segregating in the parental lines. Alleles at the major locus are assumed fixed. F_1 data could possibly be included, but this is not necessarily more informative because F_1 and F_2 generations may have different means and variances due to segregating polygenes.

The aim of this paper is to investigate by simulation some of the statistical properties of fitting mixture distributions, such as Type I error, power of the likelihood ratio test and bias of parameter estimates, when using only F_2 data. To study the properties of the major gene model, polygenic variance is not estimated. The robustness of this model will be checked when polygenic variance is present in the data, and when the major gene is not fixed in the parental lines. The question whether F_1 data can and should be included will be addressed.

Models used for simulation

A base-population of F_1 individuals was simulated, although the F_1 generation may not have had observed records. Consider a single locus A with alleles A_1 and A_2, where A_1 has frequencies f_p and f_m in the paternal and maternal line. Genotype frequencies, values and numeration are given for F_1 individuals as :

Genotype	A_1A_1	A_1A_2	A_2A_2
Number	1	2	3
Frequency	f_pf_m	$f_p(1-f_m)+f_m(1-f_p)$	$(1-f_p)(1-f_m)$
Value	μ_1	μ_2	μ_3

Genotypes of F_1 animals were allocated according to the frequencies given above using uniform random numbers. For the F_2 generation, genotype probabilities were calculated given the parents' genotypes using Mendelian transmission probabilities and assuming random mating and no selection. A random environmental component e_i was simulated and added to the genotype. The observation on individual i (F_1 or F_2) with genotype r (y_i^r) is:

$$y_i^r = \mu_r + e_i, \tag{1}$$

with e_i distributed $N(0, \sigma^2)$. Polygenic effects are assumed to be normally distributed. For base individuals polygenic values were sampled from $N(0, \sigma_g^2)$, where σ_g^2 is the polygenic variance. No records were simulated for F_1 individuals when polygenic effects were included. For F_2 offspring, phenotypic observations y_{ij}^r were simulated as:

$$y_{ij}^r = \mu_r + \tfrac{1}{2} a_p + \tfrac{1}{2} a_m + \phi_i + e_{ij}, \tag{2}$$

where ϕ_i is the Mendelian sampling term, sampled from $N(0, \frac{1}{2}\sigma_g^2)$, a_p and a_m are paternal and maternal polygenic values and e_{ij} is distributed $N(0, \sigma^2)$. Additionally, data were simulated with no major gene or polygenic effect :

$$y_i = e_i, \tag{3}$$

where e_i is distributed $N(0, \sigma^2)$. A balanced family structure was simulated, with an equal number of dams, nested within sire, and an equal number of offspring for each dam. Random variables were generated by the IMSL routines GGUBFS for uniform variables and GGNQF for normal variables (Imsl, 1984).

Models used for analysis

The test for the presence of a major gene is based on comparing the likelihood of a model with and without a major gene. Polygenic effects are not included in the model, and the model without a major gene therefore contains random environment only. Apart from major gene or no major gene, models can account for only F_2 data, or for

both F_1 and F_2 data. This results in a total of 4 models to be described.

Model for F_2 data with environment only

For F_2 data, with n observations, the model can be written :

$$y_i = \beta + e_i \qquad (4)$$
$$\text{with} \quad E(y_i) = \beta$$
$$\text{var}(y_i) = \text{var}(e_i) = \sigma^2$$

The logarithm of the joint likelihood for all observations, assuming normality and uncorrelated errors, is :

$$L_1 = -\tfrac{n}{2} \ln(2\pi\sigma^2) + \Sigma_{i=1}^{n} -(y_i-\beta)^2/2\sigma^2 \qquad (5)$$

Maximising (5) with respect to β and σ^2 yields as the maximum likelihood (ML) estimate for the mean, $\hat{\beta} = \Sigma_i y_i/n$, and the ML estimate for the variance is $\hat{\sigma}^2 = \Sigma_i (y_i-\hat{\beta})^2/n$.

Model for F_1 and F_2 data with environment only

Data on F_1 and F_2 are combined, with $n_1 + n_2 = N$ observations. The observation on animal j from generation i (i=1, 2) is:

$$y_{ij} = \beta_i + e_{ij} \qquad (6)$$
$$\text{with} \quad E(y_{ij}) = \beta_i$$
$$\text{var}(y_{ij}) = \text{var}(e_{ij}) = \sigma^2$$

where β_i is the mean for generation i. Observations for F_1 and F_2 are assumed to have equal environmental variance. The joint log-likelihood is given as:

$$L_1^* = -\tfrac{N}{2} \ln(2\pi\sigma^2) - \Sigma_{i=1}^{2}\Sigma_{j=1}^{ni} (y_{ij}-\beta_i)^2/2\sigma^2 \qquad (7)$$

The ML estimates for β_i are simply the observed means for each generation, i.e. $\hat{\beta}_1 = \Sigma_j y_{1j}/n_1$, and $\hat{\beta}_2 = \Sigma_j y_{2j}/n_2$. The ML estimate for the variance is $\hat{\sigma}^2 = \Sigma_i \Sigma_j (y_{ij}-\hat{\beta}_i)^2/N$.

Model with major gene and environment for F_2 data

When alleles are assumed fixed in parental lines, all F_1 individuals are known to be heterozygous. If no polygenic effects are considered, this means that all F_2 individuals have the same expectation, and conditioning on parents is redundant. In the likelihood for such data, summations over the parents' possible genotypes can be omitted and families can be pooled. The model is given as :

$$y_i^r = \mu_r + e_i \qquad (8)$$
$$\text{with} \quad e_i \sim N(0, \sigma^2)$$

and the log-likelihood equals :

$$L_2 = \Sigma_{i=1}^{n} \ln\left\{ \Sigma_{r=1}^{3} P_r f(y_i \mid G_i=r) \right\} \qquad (9)$$

In (9) G_i is the genotype of individual i, P_r denotes the prior probability that $G_i=r$, which equals $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ for r=1, 2 and 3 (or A_1A_1, A_1A_2 and A_2A_2). The total number of F_2 individuals is given as n, and the function f is given as :

$$f(y_i \mid G_i=r) = (2\pi\sigma^2)^{-0.5} \exp\left\{ -(y_i-\mu_r)^2/2\sigma^2 \right\} \qquad (10)$$

Model with major gene and environment for F_1 and F_2 data

In the F_1 generation only one genotype occurs; hence F_1 data are distributed around a single mean, with a variance equal to the residual variance in the F_2 generation. Due to possible heterosis shown by the polygenes, a separate mean is modelled, but the possible heterogeneity in variance caused by polygenes is not accounted for. The model for individual j from generation i for genotype r is:

$$y_{ij}^r = \mu_r + \beta_i + e_{ij} \qquad (11)$$
$$\text{with} \quad e_{ij} \sim N(0, \sigma^2)$$

where β_i is a fixed effect for generation i. Model (11) is overparameterised because genotype means and 2 general means are modelled. We chose to put β_2=0. In that case

the mean of F_1 individuals, which all have known genotype r=2, can be written as $\mu_{F1} = \mu_2 + \beta_1$. The joint log-likelihood for F_1 and F_2 data, using μ_{F1} is:

$$L_2^* = \sum_{j=1}^{n1} \left\{ -\tfrac{1}{2} \ln(2\pi\sigma^2) - (y_{1j}-\mu_{F1})^2/2\sigma^2 \right\} + \sum_{j=1}^{n2} \ln\left\{ \sum_{r=1}^{3} P_r f(y_{2j} | G_j=r) \right\} \quad (12)$$

where n_1 and n_2 are number of observations in the F_1 and F_2 generation. The ML estimate for μ_{F1} is equal to $\hat{\beta}_1$ in (6).

ML estimates for μ_r (r=1,2,3) and σ^2 in models (8) and (11) cannot be given explicitly. These parameters were estimated by minimising minus log-likelihoods L_2 in (9) and L_2^* in (12), using a quasi-Newton minimisation routine. A reparameterisation was made using the difference between homozygotes $t=\mu_3-\mu_1$, and a relative dominance coefficient $d=(\mu_2-\mu_1)/t$, as in Morton and MacLean (1974). By experience, this parameterisation was found more appropriate than the parameterisation using three means μ_1, μ_2 and μ_3, because convergence is generally reached faster due to smaller sampling covariances between the estimates. The mean was chosen as the midhomozygote value: $\mu = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_3$.

Parameters t and d are easier to interpret than 3 means, and therefore results are also presented using these parameters. Parameter t indicates the magnitude of the major gene effect and can be expressed either absolutely or in units of the residual standard deviation. Parameter t was constrained to be positive, which is arbitrary because the likelihood for the parameters μ, t and d is equal to the likelihood for the parameters μ, $-t$ and $(1-d)$. Parameter d was estimated in the interval [0,1]. Problems were detected when this constraint was not used, because t could become zero, leading to infinitely large estimates for d. This occurred frequently when the effects where small and dominant. Minimisation by IMSL routine ZXMIN (Imsl, 1984) specified 3 significant digits in the estimated parameters as the convergence criterion.

Hypothesis testing

The null hypothesis (H_0) is "no major gene effect", whereas the alternative hypothesis (H_1) is "a major gene effect is present". The log-likelihoods L_1 in (5) and L_2 in (9) are

the likelihoods for each hypothesis when only F_2 data are present. When F_1 data are included the likelihoods L_1* in (7) and L_2* in (12) apply. A likelihood ratio test is used to accept or reject H_0. Twice the logarithm of the likelihood ratio is given as:

$\tau = 2(L_2 - L_1)$, for F_2 data only

or $\tau = 2(L_2^* - L_1^*)$, for F_1 and F_2 data.

Two important aspects of any test are the type I and type II errors. The type I error is the percentage of cases in which H_0 is rejected, although it is true. The H_0 model is simulated by (3). The type II error is the percentage of cases in which H_1 is rejected, although it is true. Here, the type II error is not used, but its complement, the power, which is the percentage of cases in which H_1 is accepted, when H_1 is true. The H_1 model is simulated by model (1). Fixation of alleles in parental lines is simulated by taking f_p=1 and f_m=0.

Type I error

The distribution of τ when H_0 is true is expected asymptotically to be χ^2 with 2 degrees of freedom, because the H_1 model has 2 parameters more than the H_0 model (Wilks, 1938). Since in practice data sets are always of finite size, it is interesting to know whether and when the distribution of τ is close enough to the expected asymptotic distribution, so that quantiles from a χ^2 distribution can be used as critical values. Type I errors were estimated for data sets of 100 up to 2 000 observations, simulating 1 000 replicates for each size of data set. Three critical values were used, corresponding to nominal levels of 10, 5 and 1%. The nominal level is defined as the expected error rate, based on the asymptotic distribution. Exact binomial probabilities were used to test whether the estimates differed significantly from the nominal level. When the observed number of significant replicates does not differ significantly, a χ^2 distribution is considered suitable to provide critical values. Also, when the observed number is lower than expected the asymptotic distribution might remain useful. The nominal type I error is in that case an upper bound for the real type I error.

Power of the test and estimated parameters

The power is investigated for additive (d=0.5) and completely dominant (d=1) effects,

with a residual variance of 100, and t varying from 10 to 40, i.e. from 1 to 4 residual standard deviations. The additive genetic variance caused by this locus equals $t^2/8$, when t is absolute. Heritability in the narrow sense therefore varies from 0.11-0.67. Each data set contained 1 000 observations, and each situation was repeated 100 times. The power of the test for smaller data sets was investigated for one relatively small effect and one relatively large effect.

Robustness

Investigation of the type I error and the power considered situations where either H_0 or H_1 was true, satisfying all assumptions in the models. The robustness of this test and usefulness of the assumption of fixation in parents for parameter estimation was investigated for situations which violate two assumptions:

- when there is a covariance between error terms. This was induced by simulation of polygenic variance by model (2). The total variance was held constant at 100, so that the power of the test could not change due to a change in total variance.
- when fixation of alleles is not the case. The data were simulated by model (1), in which f_p and f_m were not equal to 0 and 1, resulting in segregation of alleles in the F_1 parents. Firstly, 3 situations were simulated where the average allele frequency remains 0.5. In that case only the assumption that all F_1 parents are heterozygous was violated. Secondly, 3 situations were simulated where the average allele frequency was not 0.5. In that case, the assumption that genotype frequencies in F_2 are $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ was also violated.

Inclusion of F_1 data

A major gene, which starts segregating in the F_2 not only renders the distribution non-normal, but also increases the phenotypic variance in the F_2 relative to the F_1. When F_1 data are included, this increase in variance may be taken as supplementary evidence, apart from any non-normality, for the existence of a major gene. Assessing the relative importance of the 2 sources of information is useful so as to judge the robustness of the model including F_1 data. The effects on non-normality and increased F_2 variance due to the major gene should therefore be distinguished. This was accomplished by simulating different residual variances in F_1 and F_2. Four situations were investigated, combining all combinations of non-normality in F_2 and increased

variance in F_2 (Table 1). In general, 500 F_1 and 1 000 F_2 observations were simulated. For situation 3 data sets with 1 000 F_1 and 1 000 F_2 observations were also investigated. Data for situations 1 and 3 were simulated by model (3), whereas data for situations 2 and 4 were simulated by model (1).

Table 1 The effect on variance and non-normality in the F_2, when F_1 and F_2 data are combined, for various situations investigated.

Situation	Description	F_2 distribution normal	Larger variance in F_2
1	H_0 (no major gene)	Yes	No
2	H_1 (major gene)	No	Yes
3	H_0 with increased F_2 variance	Yes	Yes
4	H_1 with decreased F_2 variance	No	No

Results

Type I error and parameter estimates under the null hypothesis

Estimated type I errors, based on 1 000 replicates, have been given in Table 2 for different sizes of the data set. Estimates decreased, and more or less stabilised when the size of the data set exceeded 1 000 observations, especially for a nominal level of 10%, which were most accurate. For these large data sets, however, the type I errors were too low ($P < 0.01$), which means that critical values obtained from a $\chi^2(2)$ distribution would provide a too conservative test. For example, application of the $\chi^2(2)$ 95-percentile to data sets with 1 000 observations will not result in the expected type I error of 5%, but rather in a type I error of ≈3%.

When no major gene effect was present, still on average a considerable effect could be found. Parameter estimates for the major gene model have been given in Table 3, simulating just a normally distributed error effect with variance 100. The empirical standard deviation for estimated t-values ranged between 7 (N=100) and 5 (N=2 000) (not in Table). The average estimate for t is therefore biased, and many of the individual estimates were significantly different from zero if a t-test was applied.

The average estimated d is 0.5, which is expected because the simulated distribution was symmetrical.

Table 2 Estimated Type I errors (%) at 3 nominal levels for different size of the data set

	Nominal level					
	10%		5%		1%	
N	Estimate	P	Estimate	P	Estimate	P
100	9.5	0.3216	5.0	0.5375	0.8	0.3317
250	7.8	0.0099	3.3	0.0059	0.9	0.4573
500	6.9	0.0004	2.9	0.0007	0.4	0.0287
1000	6.1	0.0000	3.1	0.0022	0.5	0.0661
2000	6.0	0.0000	2.5	0.0001	0.6	0.1289

N: Number of observations in the data set

P: critical level for test whether estimate is equal to the nominal level, based on exact binomial probabilities

Table 3 Average major gene parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) under the null-hypothesis for varying size of the data set

N	t	d	σ^2
100	15.90	0.50	57.1
250	13.72	0.50	67.0
500	12.54	0.49	73.2
1000	11.35	0.51	77.2
2000	10.51	0.50	81.3

Simulated: $\sigma^2 = 100$; N: Number of observations in the data set.

Parameter estimates and power of the test

Results for the different situations studied under a major gene model have been given in Table 4. The $\chi^2(2)$ 95-percentile was used as critical value for the test. The power

reached over 95% for additive effects (d=0.5) with a t value of 40, which is 4σ (residual standard deviations). For complete dominant effects (d=1), 100 % power was reached for an effect of t=20 (2σ). Phenotypic distributions for these 2 cases are unimodal, although not normal (Figure 1).

Table 4 Power of the test and average parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) in different situations (data sets with 1 000 observations, 100 replicates)

Simulated parameters			Power	Estimated parameters		
σ^2	d	t		t	d	σ^2
100	-	0	3.1	11.4	0.51	77.2
100	0.50	10	3	12.6	0.44	84.7
		15	7	14.0	0.47	95.4
		20	12	18.2	0.47	100.2
		25	29	23.4	0.48	104.4
		30	38	28.1	0.50	108.6
		35	82	34.9	0.50	99.2
		40	96	39.8	0.50	103.3
100	1.00	10	1	14.1	0.93	61.6
		15	70	18.2	0.83	90.9
		20	100	22.4	0.87	94.0
		25	100	27.2	0.89	95.6
		30	100	32.7	0.90	94.0
		35	100	37.6	0.90	94.8
		40	100	40.9	0.96	97.4

Power: Number significant at nominal 0.05 level (total=100)

First line: based on 1 000 simulations under H_0 (Tables 1 and 2).

For small genetic effects ($t \leq 10$, i.e. 1σ) t was overestimated, in particular when t=0, as was already mentioned. For larger genetic effects, t was overestimated for d=1 and was underestimated for d=0.5. For d=0.5, average estimates for t and d differed from the simulated values by less than 1%, when the power reached near 100 %. For d=1,

however, the bias in t was still 10% when the power had reached 100%. This bias reduced gradually, and was less than 1% for a genetic effect of t=40.

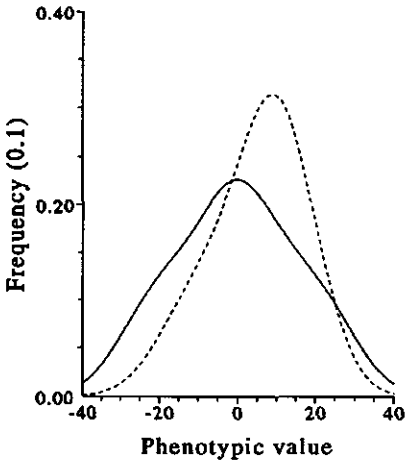


Figure 1 Phenotypic distributions on which over 95% power was reached for the identification of a major gene: t=40, d=0.5 (solid line) and t=20, d=1 (dashed line); σ=10.

In Figure 2 power of the test is depicted for varying sizes of the data set. Two additive effects were chosen, with t=25 and t=35. Each point in the figure is on average of 100 replicates. The power increased with increasing number of observations. Increasing the number of observations above 1 000 gave relatively less improvement in power, especially for the smaller effect (t=25). For a small number of observations this graph is expected to level off at the type I error (nominally 5%), but sampling makes results somewhat erratic.

Robustness when ignoring polygenic variance

Data following model (2) were simulated with d=0.5 and t=35 and different proportions of polygenic and residual variance. The data set contained 20 sires with 5 dams each and 10 offspring per dam; each situation was repeated 100 times. Estimated parameters and resulting power are in Table 5. Parameter estimates for t and d, and the power of the test were not affected when a part of the variance was polygenic. The total estimated variance was equal to the sum of simulated variances.

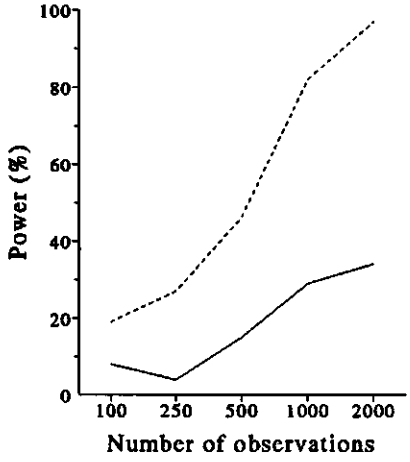


Figure 2 The power for detection of a major gene in relation to the size of the data set shown for 2 situations: t=25 (solid line) and t=35 (dashed line); d=0.5 and σ=10.

Table 5 Power of the test and average parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) when polygenic variance is present (data sets with 1000 observations, 100 replicates)

Simulated parameters		Power	Estimated parameters		
σ_g^2	σ_e^2		t	d	σ^2
0	100	82	34.9	0.50	99.2
20	80	87	35.0	0.50	99.6
40	60	80	34.4	0.51	102.5
60	40	78	34.5	0.50	101.4
80	20	90	35.3	0.50	96.7
100	0	80	34.5	0.50	100.0

σ_g^2, σ_e^2: simulated polygenic and residual variance

Other parameters simulated: t=35, d=0.5

Power: number significant at nominal 0.05 level (total=100)

Robustness when ignoring segregation in the parental lines

Data following model (1) were simulated with d=0.5, t=35, σ^2=100 and various values

for f_p and f_m. The genotype probabilities in parents (F_1) and offspring (F_2) are in Table 6. For the first three situations, genotype probabilities in the F_2 were $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ as assumed under the fixation assumption. For the last three situations, however, genotype probabilities were different, because the allele frequency was not 0.5 on average. High average allele frequencies were simulated, but because only additive effects are considered, results are equally valid for low allele frequencies. The power remained equal, as long as genotype probabilities in F_2 remained $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ and parameter estimates are unbiased (Table 7). In case the allele frequency did not average 0.5, however, parameter estimates were biased. The power of the test increased, because in this situation the distribution became skewed. The situation with d=0.5 and t=35 for data where the gene is fixed in parental lines (Table 4), with a power of 82 %, may serve as a reference.

Table 6 Genotype probabilities in F_1 and F_2 for different allele frequencies in the parental lines

		F_1 probabilities			F_2 probabilities		
f_p	f_m	A_1A_1	A_1A_2	A_2A_2	A_1A_1	A_1A_2	A_2A_2
0.9	0.1	0.09	0.82	0.09	0.25	0.50	0.25
0.8	0.2	0.16	0.68	0.16	0.25	0.50	0.25
0.6	0.4	0.24	0.52	0.24	0.25	0.50	0.25
0.9	0.3	0.07	0.66	0.27	0.16	0.48	0.36
0.9	0.5	0.05	0.50	0.45	0.09	0.42	0.49
0.9	0.7	0.03	0.34	0.63	0.04	0.32	0.64

f_p, f_m: frequency of A_1 allele in paternal and maternal line

Inclusion of F_1 data

Five hundred, or 1000, F_1 observations were also simulated, with additive major gene effects (Table 8). With no major gene effect (t=0 and hence σ^2_{mg}=0), and with equal variances in F_1 and F_2 (situation 1) the average estimated t was much smaller than in the model using only F_2 data (Table 3). In the second situation (Table 8) a major gene effect of t=20 was simulated, which corresponds to the given major gene variance of 50.

Table 7 Power of the test and parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) when alleles are segregating by various frequencies in the parental lines (data sets with 1000 observations, 100 replicates)

f_p	f_m	Power	t	d	σ^2
0.9	0.1	76	34.37	0.50	103.9
0.8	0.2	83	34.66	0.51	101.5
0.6	0.4	76	34.14	0.50	105.6
0.9	0.3	81	31.99	0.58	113.4
0.9	0.5	92	26.02	0.77	127.2
0.9	0.7	99	21.17	0.96	115.9

Simulated : t=35, d=0.5, σ^2=100; f_p, f_m: allele frequency in paternal and maternal line; Power : number significant at nominal 0.05 level (total=100)

Table 8 Power of the test and parameter estimates for genetic effects (t) and variance (σ^2) in different situations when 500 F_1 and 1 000 F_2 observations are combined

	F_1	F_2			Estimated parameters	
Situation	σ_e^2	σ_e^2	σ_{mg}^2	Power	t	σ^2
1	100	100	0	1	3.03	97.9
2	100	100	50	100	19.43	100.8
3	100	150	0	100	19.62	99.3
3	100	110	0	15	7.72	99.1
3*	100	110	0	25	8.11	99.3
4	150	100	50	2	5.05	145.3

Situation: refers to Table 2

3*: alternative with 1 000 F_1 observations instead of 500

σ_e^2, σ_{mg}^2: simulated residual and major gene variance

Power: number significant at nominal 0.05 level (total=100)

When using only F_2 data, the test had a power of only 12 % for detection of an additive effect of t=20 (Table 4). When including F_1 data, however, the power was 100

% (Table 8). From the situations 3 and 4 considered in Table 8, however, it becomes apparent that when F_1 data were included, the major gene was only detected by its effect on variance, considering a power near the type I error rate as non relevant. When the variance in F_2 increased by 50%, but when in fact no major gene was present, a major gene was found in 100 % of the cases. For smaller increases of the variance (10%) major genes were still detected, and the probability of detection increased with the size of the data set (alternative 3* with more F_1 observations). A major gene was totally not detectable, on the other hand, when the total variance in F_1 was equal to the total variance in F_2 (situation 4). This shows that the ability to detect a major gene can even be worsened when F_1 data are included. If only F_2 data was used, a major gene with similar effect was detected in 12 % of the cases (Table 4).

Discussion and conclusions

Type I error

Nominal levels for type I errors were based on Wilks (1938) who proved asymptotic convergence of the likelihood ratio test statistic to a χ^2 distribution. Type I errors decreased and stabilised for larger data sets, as expected. The estimated type I errors, however, were significantly too low. It is unlikely that the type I error, after having first decreased, would increase for even larger data sets as studied here. It can be concluded therefore, that type I errors are significantly lower than expected in the asymptotic case, and that for large data sets the likelihood ratio test is conservative. It has been investigated whether the constraint used on the dominance coefficient could have caused the too low type I errors. However, this was not the case, because even with no constraint, too low type I errors were found of 7.5% and 3.9% at nominal levels of 10 and 5%.

For the investigation of power we have chosen to use the theoretical asymptotic quantiles, although they were shown to give a conservative test. The nominal level for the type I error is then an upper bound, and the experimenter still has a reasonable good idea of the risk of making a type I error. When the actual type I error would be above the expected level, however, the test would become of less use.

A second reason for still using theoretical asymptotic quantiles is that adapting the test is difficult and of little practical use. A difficulty is, for instance, that estimated

quantiles would be subject to sampling and the obtained point estimate is therefore only expected to give the correct test. Therefore, 2 experimenters investigating the same test, will find different critical values and the test applied will depend on the experimenter. Also in practice such a procedure would be difficult to apply since the calculated quantile would only hold for the same model and data sets of similar size and structure.

Power of the test

Using only F_2 data, the power of this test was poor for additive effects (dominance coefficient = 0.5). This can be explained by the resulting symmetrical distribution which is similar to the distribution under H_0. In this case, the genetic effect has to be about 4σ to be detectable, which corresponds to an heritability of 0.67 in the F_2 generation. When the dominance coefficient is 1, an effect of 2σ was detectable. These results are based on data sets with 1 000 observations, but it was shown that the power decreased dramatically for smaller data sets.

Power increased when F_1 data was included in the analysis, and additive effects of 2σ could be detected. In that case the increase in variance in F_2, caused by the major gene, was taken as an important indication for the presence of a major gene. The power to detect a major gene in F_2 data may also increase if alleles were not fixed in the parental lines, or alternatively F_3, instead of F_2, data were used. This corresponds more to the situation in a usual population, where between-family variation will arise. For F_3 data, for example, when pure lines were homozygous, the allele frequency will be 0.5, and parents will be in Hardy-Weinberg equilibrium. For such a situation, Le Roy (1989) found a power of 25% for an additive effect of 2σ in a data set of 400 observations (20 sires with 20 half-sib offspring each). In Figure 2, the power for a data set of similar size can be seen to be only $\approx$10 % for an even larger effect of 2.5σ (t=25). This indicates that an increase in power may be expected when the F_3 generation is observed, despite that more parameters have to be estimated, and that parents' genotypes are no longer known.

The power for detection of a major gene is related to the unexplained variance in the model of analysis. The inclusion of fixed and polygenic effects will therefore make the major gene easier to detect, provided that all these effects can be accurately estimated.

Parameter estimates

For additive effects simulated (d=0.5), bias for the average estimated genetic effect t and dominance coefficient d was less than 1% when the power approached 100%. For dominant effects (d=1), however, t was overestimated by 10% when the power for detection of a major gene reached 100%. This overestimate is probably related to the underestimate for d, which resulted from the applied constraint. As mentioned, this constraint was applied to prevent t from going to zero, at which point d tended to go to infinity. When such a constraint was not applied with, for instance, an effect of t=10 and d=1, gave in 100 replicates an average estimated d of 2.93. This is an average overestimate of ≈200%. The average estimate using the constraint was 0.93, showing that indeed better estimates were obtained under the restriction, even when the true value was on the border of the allowed parameter space. In practice, of course, overdominance can not be excluded and parameter estimates could be compared with and without this constraint. A small, near zero, estimate for t and a large estimate for d would suggest a possible overestimation of d.

For very small or absent effects, the ML estimates were considerably biased. In this situation, the asymptotic properties of ML estimates, i.e. consistency, are far from being attained. In the absence of a major gene, average estimates were presented for increasing size of the data set. This showed that the average estimate decreased, and will probably reach the true value when the number of observations is very much larger. Bias of ML estimates in finite samples also resulted in significant t-values when no effect was present. This indicates that the presence of a major gene should not be judged by the estimates and their standard errors. The standard errors discussed here were empirical standard errors. In practice such standard errors will have to be obtained using the inverse of an estimated Hessian matrix, or some other quadratic approximation of the likelihood curve in the optimum. Using the estimated Hessian matrices, we found roughly the same standard errors, although they were not very accurate. In our study, the quasi-Newton algorithm was started close to the optimum and not enough iterations are then carried out to estimate the Hessian matrix accurately.

Robustness of model and test

Inclusion of F_1 data results in a poorly robust test when differences in variances would

arise between the F_1 and F_2 due to other causes than a major gene. An increase in variance from F_1 to F_2, can result in a putative major gene being detected. An increase in variance of 10% for instance gave 25% false detections when 1 000 F_1 and 1 000 F_2 observations were combined. Such increases are not unlikely, due to, for instance, polygenes. The major gene test is then merely a test for homogeneous variance in F_1 and F_2. The inclusion of F_1 data could also worsen the detection of a major gene, when the environmental variance in F_2 was less. Therefore any differences in variance, due to other causes than the major gene effect, will bias the parameter estimates. Also in a model that allows for segregation, such biases will remain.

It was shown that the model is robust when polygenic effects were ignored. This can be explained by the fact that the test uses only the non-normality of the distribution as a criterion. It must be noted however that, when polygenic effects can be accurately estimated, including a polygenic effect in the model will increase power because it reduces the residual variance.

Another aspect of robustness concerns the assumption of fixed alleles in parental lines. It was shown that parameter estimates were not biased when alleles segregated, as long as the average frequency in the 2 lines was 0.5. In that case the assumed fitting proportions $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ are still correct. If the average frequency in parental lines differed from 0.5, t was underestimated and, because skewness was introduced, estimates for d deviated from 0.5. This second situation is more likely to occurr than the situation where the average frequency is exactly 0.5. Because it could be difficult to justify the fixation assumption a-priori, application of a more general model that allows for segregation in parental lines, might have to be considered.

A final aspect of robustness concerns non-normality of the distribution not due to a major gene. As stated earlier a mixture distribution is fitted and the detection of a major gene in F_2 data, assuming fixation, relies solely on the non-normality caused by the major gene. This means that in fact only a significant non-normality is proven. The method would therefore be poorly robust against any non-normality due to another cause. The robustness might be improved using data in which alleles segregate in parents. This is guaranteed in F_3 data, but may also arise in F_2 data, when alleles were not fixed in parental lines. If segregation in parents is the case, evidence for a major gene is no longer only in the non-normality of the overall distribution, but also for instance in heterogeneous within family variances. Therefore a model that allows for

segregation is not only preferred to increase power, but also is preferred to improve robustness.

Acknowledgements

Profs Brascamp and Grossman and Drs Van Arendonk and Van Putten are acknowledged for helpful comments and editorial suggestions. Comments of 2 referees have helped to shorten the manuscript and to improve the discussion section. This research was supported financially by the Dutch Product Board for Livestock, Meat and Eggs, and the Dutch pig breeding companies Bovar, Euribrid, Fomeva, Nieuw-Dalland and NVS.

References

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21: 523-542

Elston RC, Stewart J (1973) The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data. Genetics 73: 695-711

Elston RC (1984) The genetic analysis of quantitative trait differences between two homozygous lines. Genetics 108: 733-744

IMSL (1984) Library reference manual Edition 9.2, International and statistical libraries, Houston, Texas

Le Roy P (1989). Méthodes de détection de gènes majeurs; application aux animaux domestiques. Doctoral Thesis, Université de Paris-Sud, Centre D'Orsay

Morton NE, MacLean CJ (1974). Analysis of family resemblance III. Complex segregation of quantitative traits. Am J Hum Genet 26: 489-503

Wilks SS (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9: 60-62

Computing approximate monogenic model likelihoods in large pedigrees with loops

Chapter 3

In this chapter 'iterative peeling' is introduced, a method equivalent to the traditional recursive peeling method for computing exact likelihoods in non-looped pedigrees, but which also can be used to obtain approximate likelihoods in looped pedigrees. Iterative peeling is an interesting tool for animal breeding, where exact recursive peeling is generally infeasible due to the abundant number of loops in animal pedigrees. In simulations, hypothesis testing and parameter estimation were compared based on approximated *likelihoods in looped pedigrees* and exact *likelihoods in non-looped pedigrees*, showing no biases being introduced by the approximation in looped pedigrees.

Introduction

Research into the use of major gene models in animal breeding has been aimed mainly at approximations to a mixed inheritance model, including polygenes, in one generation half-sib structures (Hoeschele, 1988; Le Roy et al., 1989; Knott et al., 1992). Because of the pedigree loops that arise in animal breeding situations, extension to multigeneration pedigrees is difficult. A pedigree loop arises when two individuals are connected by more than one path of descendance or marriage relationships. Lange and Elston (1975) described various types of loops, among which inbreeding loops, marriage rings and marriage loops. In animal breeding pedigrees these kinds of loops are very common. In particular, multiple matings which are generally applied to males and often to females, result in many marriage loops and marriage rings.

For genotype probability and likelihood computation, loops can be dealt with in an exact manner only in pedigrees with a few simple non-overlapping loops using the traditional recursive peeling method (Elston and Stewart, 1971; Cannings et al., 1976; Cannings et al., 1978). However, in highly looped pedigrees, common in animal breeding, exact recursive peeling is too demanding computationally and recursive peeling also is not flexible to allow for approximate computations.

In this study we introduce 'iterative peeling'. Iterative peeling is developed as an

exact method for application in non-looped pedigrees, equivalent to recursive peeling, but which, unlike the original recursive variant, can be used without modifications in looped pedigrees to obtain approximate likelihoods. The main objective of this paper is to introduce iterative peeling for such approximations in looped pedigrees, allowing for a more general application of major gene models in animal breeding. Using simulations, the usefulness of the approximation for likelihood-based hypothesis testing and parameter estimation in looped pedigrees is investigated. A monogenic model will be considered, which can be extended to a mixed inheritance model, as will be discussed.

Recursive and iterative peeling

In the first section, recursive peeling is described for obtaining monogenic model likelihoods in non-looped pedigrees. In the second section, 'iterative peeling' is introduced as an equivalent method for exact computations in non-looped pedigrees. The equivalent exact method in non-looped pedigrees can be used as an approximate method in looped pedigrees.

Recursive peeling

Probability and likelihood computations in non-looped pedigrees can be done by recursive peeling (Elston and Stewart, 1971; Cannings et al., 1976; Cannings et al., 1978) using two basic peeling operations of 'peeling up' and 'peeling down'. Roughly, considering a single family, a peel-up operation represents the information in a family in probabilities for the genotype G_i of a parent i, and a peel-down operation represents this information in probabilities for the genotype G_k for an offspring k. Here, notation based on Van Arendonk et al. (1989) is used, where the result of the peel-up operation is denoted by $prog(G_i)$ and the result of the peel-down operation is denoted by $prior(G_k)$. The corresponding notation in Cannings et al. (1976, 1978) is the $R^*(..;G_i)$ function for peeling up and the $R^+(..;G_k)$ function for peeling down.

Peeling operations are used recursively, e.g. computation of a *prog* term for a parent based on progeny data, may include previously computed *prog* terms of those progeny, representing information from grand-progeny. The aim of peeling is to condense all information from a pedigree into a *prior* and *prog* term for a single

individual l, obtaining the likelihood L for all data in the pedigree as :

$$L = \Sigma_{Gl}\, prior(G_l)\, f(y_l \mid G_l)\, prog(G_l) \qquad (1)$$

where $f(y_l \mid G_l)$ is the penetrance function, which is the probability for the observed data y_l on individual l, given it has genotype G_l. The individual l may be an individual from the base population, in which case the base-population genotype frequency $P(G_l)$ is used in place of $prior(G_l)$. Individual l also may have no own data or no progeny, in which case the corresponding penetrance term or *prog* term is removed. Computationally this is implemented using a penetrance or *prog* term containing 1's.

Peeling equations

A peeling equation for an individual is obtained by considering the collection of possible base-population genotype frequencies, genotype transmission probabilities, penetrance probabilities and other peeling terms pertaining to the individuals in its family and summing over all possible genotypes of the family members. The terms thus entering in a peeling equation are difficult to give in general. Here, equations will be given to use peeling in a pedigree structure with dams nested within sires. In this structure a family is a half-sib family of one sire with several mates, containing groups of full sibs which are, across groups, paternal half-sibs. Three different peeling equations are considered, two for peeling up, dependent on whether this is done for a sire or a dam, and one for peeling down. In the peeling equations, *prior*, *prog* and penetrance functions on family members are specified in all places where they can enter. When these are not relevant, e.g. when a progeny does not have progeny of its own, these are removed or, computationally, terms containing 1's are used. *Prior* terms for individuals in the base population are substituted with base-population genotype frequencies.

To condense all information in a *prog* term for a sire i the following is used :

$$prog(G_i) = \Pi_j \Sigma_{Gj}\, prior(G_j)\, f(y_j|G_j)\, \Pi_k \Sigma_{Gk}\, P(G_k|G_i,G_j)\, f(y_k|G_k) prog(G_k) \qquad (2)$$

where j=1 to n_i are mates of i, each mate having k=1 to n_{ij} progeny, and $P(G_k \mid G_i,G_j)$ is the genotype transmission probability of sire i and a dam j to offspring k. To

condense all information from a half-sib family into a *prog* term for one particular dam j^* of the family, the following is used :

$$prog(G_{j^*})=\Sigma_{Gi}\ prior(G_i)\ f(y_i|G_i)\ prog_{-j^*}(G_i)$$
$$\Pi_k\Sigma_{Gk}\ P(G_k|G_i,G_{j^*})\ f(y_k|G_k)prog(G_k) \qquad (3)$$

where i is the sire of the family, $prog_{-j^*}(G_i)$ is like in equation (2), but excluding dam j^* and $k=1,n_{ij^*}$ are progeny of dam j^*. To condense all information in a *prior* term for one particular progeny k^* with dam j^*, the following is used :

$$prior(G_{k^*}) = \Sigma_{Gi}\ prior(G_i)\ f(y_i|G_i)\ phs(G_i)$$
$$\Sigma_{Gj^*}\ prior(G_{j^*})\ f(y_{j^*}|G_{j^*})\ fs(G_i,G_{j^*})\ P(G_{k^*}|G_i,G_{j^*}) \qquad (4)$$

where i is the sire of the family, $phs(G_i)$ is a term that includes information on the paternal half-sibs of k^*, which is a function of the genotype of its sire i and is computed as :

$$phs(G_i)=\Pi_{j,j\neq j^*}\ \Sigma_{Gj}\ prior(G_j)\ f(y_j|G_j)\ \Pi_k\ \Sigma_{Gk}\ P(G_k|G_i,G_j)\ f(y_k|G_k)\ prog(G_k)$$

and where in (4) $fs(G_i,G_{j^*})$ is a term that includes information on the full-sibs of k^*, which is a function of the genotypes of its sire i and dam j^*, and is computed as :

$$fs(G_i,G_{j^*})=\Pi_{k,k\neq k^*}\ \Sigma_{Gk}\ P(G_k|G_{i^*},G_{j^*})\ f(y_k|G_k)\ prog(G_k)$$

Iterative peeling

Iterative peeling is equivalent to recursive peeling used in non-looped pedigrees. Iterative peeling is based on an algebraic partitioning of the likelihood and on repeated computation of peeling equations, based on the idea of iterative computation of genotype probabilities (Van Arendonk et al., 1989).

Partitioning of likelihood

The aim of obtaining the likelihood of all data using equation (1) requires families to be handled in a certain order and requires peeling, within each family, to be in a

certain direction. Peeling operations can be used to partition the likelihood pertaining to parts of the pedigree. This partitioning is continued until parts are obtained pertaining to single families. This allows a family-wise evaluation of the likelihood, and the requirement of peeling to have a direction within each family becomes obsolete.

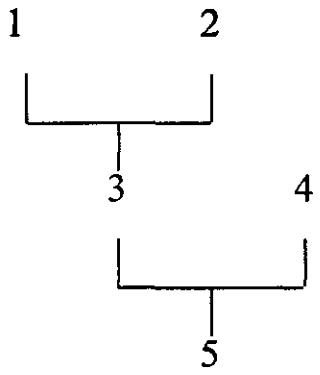


Figure 1 Example pedigree to demonstrate partitioned computation of the likelihood

Consider the pedigree with 5 individuals in Figure 1. In this pedigree two families are present, a first family with individuals 1, 2 and 3, and a second family with individuals 3, 4, and 5. Here, one partitioning above and below individual 3 divides the pedigree in two families, with individual 3 being in both families. Individual 3 is called a linking individual. The likelihood for a monogenic model, assuming data is available on all 5 individuals, is computed as :

$$L = \Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\Sigma_{G4}\Sigma_{G5}\, P(G_1)\, P(G_2)\, P(G_3 \mid G_1,G_2)\, P(G_4)\, P(G_5 \mid G_3,G_4)$$
$$f(y_1 \mid G_1)\, f(y_2 \mid G_2)\, f(y_3 \mid G_3)\, f(y_4 \mid G_4)\, f(y_5 \mid G_5)$$

Now, L is multiplied and divided by $L_1 = \Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\, P(G_1)\, P(G_2)\, P(G_3 \mid G_1,G_2)$ $f(y_1 \mid G_1)\, f(y_2 \mid G_2)$, which is the likelihood of family 1, ignoring data on progeny 3. Some reordering yields :

$$L = L_1 * \Sigma_{G3}\Sigma_{G4}\Sigma_{G5}\, \{\, \Sigma_{G1}\Sigma_{G2}\, P(G_1)P(G_2)P(G_3 \mid G_1,G_2) f(y_1 \mid G_1) f(y_2 \mid G_2) / L_1 \,\}$$
$$* P(G_4)P(G_5 \mid G_3,G_4) f(y_3 \mid G_3) f(y_4 \mid G_4) f(y_5 \mid G_5)$$

where the part $\Sigma_{G1}\Sigma_{G2}\, P(G_1)\, P(G_2)\, P(G_3 \mid G_1,G_2) f(y_1 \mid G_1) f(y_2 \mid G_2)$ has been isolated. This part is *prior*(G_3). The term defined as L_1 can be rewritten as $\Sigma_{G3}\Sigma_{G1}\Sigma_{G2}\, P(G_1)\, P(G_2)\, P(G_3 \mid G_1,G_2) f(y_1 \mid G_1) f(y_2 \mid G_2)$, which is Σ_{G3}*prior*(G_3). This simplifies L to :

$$L = L_1 \{ \Sigma_{G3}\Sigma_{G4}\Sigma_{G5}\, prior^{sc}(G_3) P(G_4) P(G_5 \mid G_3,G_4) f(y_3 \mid G_3) f(y_4 \mid G_4) f(y_5 \mid G_5) \}$$

where $prior^{sc}(G_3)$ stands for a scaled, or normalised, prior term. Now the likelihood can be written as $L = L_1 L_2$, or $\ln(L) = \ln(L_1) + \ln(L_2)$, with one likelihood term per family. This is a partitioning using a *prior* term for the linking individual. It shows that for this type of partitioning (i) in the family where the linking individual is a progeny, after the partitioning, information on the linking individual, i.e. own data and progeny data, is ignored and (ii) in the family where the linking individual is a parent, a scaled *prior* term is used for the linking individual. This term is used in a manner like a base-population genotype frequency for base individuals. The scaled *prior* term for a linking individual l, is computed in general as :

$$prior^{sc}(G_l) = prior(G_l) \,/\, \Sigma_{Gl}\ prior(G_l).$$

Although the partitioning is only shown for one example, the partitioning is very general. The term L_1 above is in general the sum of the *prior* term for a linking individual l, which is the collection of all probability terms pertaining to anterior individuals of l and the transmission probability to l, summed over all possible genotypes of l and of its anterior individuals. At the same time this term represents the likelihood of the entire anterior part of the pedigree and l, excluding data on l. The remaining part after the partitioning, L_2 in the example, is the likelihood of the posterior part of the pedigree of l, including l with a scaled *prior* term. In larger pedigrees this partitioning is repeated to yield parts corresponding to single families. When repeating the partitionings, results of earlier partitionings must be taken into account, e.g. the result that, after a partitioning, information on a linking individual is ignored in the family where the linking individual was a progeny.

The likelihood of a pedigree can be partitioned entirely using *prior* terms. However, the iterative computation, as will be introduced hereafter, can be speeded up by using also a partitioning of the likelihood using a *prog* term. Showing this based on

the example, the likelihood L is multiplied and divided by a term representing the likelihood of family 2, ignoring data on individual 3, $L_2^* = \Sigma_{G3}\Sigma_{G4}\Sigma_{G5}\, P(G_4)\, P(G_5 \mid G_3,G_4)\, f(y_3 \mid G_3\, f(y_4 \mid G_4)$, which leads to :

$$L = \Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\, P(G_1)P(G_2)P(G_3 \mid G_1,G_2)f(y_1 \mid G_1)f(y_2 \mid G_2)f(y_3 \mid G_3)$$
$$* \{ \Sigma_{G4}\Sigma_{G5}\, P(G_4)P(G_5 \mid G_3,G_4)f(y_4 \mid G_4)f(y_5 \mid G_5)/L_2^* \}\, L_2^*$$

Here a term $\Sigma_{G4}\Sigma_{G5}\, P(G_4)P(G_5 \mid G_3,G_4)f(y_4 \mid G_4)f(y_5 \mid G_5)$ has been isolated, which is $prog(G_3)$. The division by L_2^* scales this term, L_2^* being $\Sigma_{G3}\, prog(G_3)$. Hence, L is written as :

$$L = \{\Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\, P(G_1)P(G_2)P(G_3 \mid G_1,G_2)f(y_1 \mid G_1)f(y_2 \mid G_2)f(y_3 \mid G_3)prog^{sc}(G_3)\}L_2^*$$

where $prog^{sc}(G_3)$ denotes the scaled or normalised *prog* term. For a partitioning using a *prog* term it is seen that (i) in the family where the linking individual is a progeny, a $prog^{sc}$ term is added as information for the individual and (ii) in the family where the linking individual is a parent, all information from observations and from prior terms is ignored. The scaled *prog* term for a linking individual l, is computed in general as:

$$prog^{sc}(G_l) = prog(G_l) / \Sigma_{Gl}\ prog(G_l).$$

Partitioning in a nested design

In a nested design, partitionings are carried through until parts are obtained corresponding to sire families. In such families, several female parents can be present. The linking individuals are all the sires and dams of the families, except when they are in the base population. In this design we consider a partitioning using a *prog* term for each male and a *prior* term for each female that is a linking individual. When all parents of a family are in the base population, the part of the likelihood pertaining to such a family is computed as :

$$
\begin{aligned}
L_s = \{ & \Sigma_{Gi}\, P(G_i)\, f(y_i \mid G_i) \\
& \Pi_j \Sigma_{Gj}\, P(G_j)\, f(y_j \mid G_j) \\
& \Pi_k \Sigma_{Gk}\, P(G_k \mid G_i, G_j)\, f(y_k \mid G_k)\, prog^{sc}(G_k) \\
& \Pi_l \Sigma_{Gl}\, P(G_l \mid G_i, G_j) \\
& \Pi_m \Sigma_{Gm}\, P(G_m \mid G_i, G_j)\, f(y_m \mid G_m) \; \}
\end{aligned} \tag{5}
$$

where i indicates the sire of family s, j sums over the dams of the family, k indicates male progeny that are linking individuals, l indicates female progeny that are linking individuals and m indicates all other progeny. When the sire of the family is not in the base population, the term $P(G_i)f(y_i \mid G_i)$ on the first line of (5) is removed and for each dam that is not in the base population the term $P(G_j)$ on the second line of (5) is replaced with $prior^{sc}(G_j)$. The considered partitionings using *prog* terms for all male linking individuals lead to this removal of information from sires on the first line of (5) when sires are not in the base population and lead to the inclusion of the $prog^{sc}$ for males on the third line of equation 5. The considered partitionings using *prior* term for all female linking individuals, lead to the inclusion of a $prior^{sc}$ term on the second line of (5) when dams are not in the base population and the removal of all information of females on the fourth line of equation 5. Based on the results from the previous paragraph, after the partitionings the likelihood of the entire pedigree is :

$$\ln(L) = \Sigma_s \ln(L_s) \tag{6}$$

Repeated computation of peeling equations

Iterative peeling uses repeated computation of peeling equations. The repeated computation is a method to establish the order in which equations should be handled. Therefore, iterative peeling does not require to know such an order beforehand, as is required for recursive peeling.

For each individual a *prior* and a *prog* term is computed and remains stored because results of peeling terms can be required as input for the computation of other peeling terms. Iterative peeling computes a series of solutions $prior^{[0]}$, $prior^{[1]}$, etc. for these terms. Starting values are taken for individual i as $prior^{[0]}(G_i)=P(G_i)$, the genotype frequencies in the base population and $prog^{[0]}(G_i)$ equals 1 for all G_i. Iterative computation starts by computing $prior^{[1]}(G_i)$ for each individual i, in order of

descending age. Evaluation of these *prior* terms is based on $prior^{[1]}$ terms of parents, which are available because older individuals are updated before younger individuals, and on $prog^{[0]}$ terms of sibs. Subsequently, $prog^{[1]}(G_i)$ is computed for each individual i, in order of ascending age. Evaluation of these *prog* terms is based on $prior^{[1]}$ terms of mates, on $prog^{[1]}$ terms of progeny, which are available because now younger individuals are updated before older individuals, and for female parents on a $prog^{[0]}$ or $prog^{[1]}$ term of their male mate. Whether this last term is already updated as $prog^{[1]}$ depends on the order in which *prog* terms are computed. After computation of all $prior^{[1]}$ and $prog^{[1]}$ terms is completed, a new iteration starts computing $prior^{[2]}$ and $prog^{[2]}$, etc.

Starting values are such that $prior^{[0]}$ terms are correct for all individuals in the base population, and $prog^{[0]}$ terms are correct for all individuals without progeny. Terms that can be correct after the first cycle of computations are for instance $prior^{[1]}$ terms of individuals descending from two base individuals and $prog^{[1]}$ terms of parents without grandprogeny. Correct computation of a term shows, when in the next cycle recomputed terms are equal to old terms. Once it is found that a term is correctly computed, recomputation can be omitted in following iterations of the algorithm. The order in which terms are found correct gives information on the order in which recursive peeling could be used. Generally, in each iteration, reasonably large groups of terms appear correct, keeping the number of cycles required to compute all terms correctly reasonably small, typically about the number of generations in the data set. When all terms are found correctly computed, likelihood of the data can be obtained using (5) and (6).

Application in looped pedigrees

The series of solutions $prior^{[0]}$, $prior^{[1]}$, etc., obtained with iterative peeling can be considered as temporary solutions for the required terms, corresponding to solutions based on a not yet fully determined peeling order. Also 'temporary' likelihoods can be computed using (5) and (6) based on a not yet fully determined order. In non-looped pedigrees, a peeling order can eventually be found and temporary solutions become exact. In looped pedigrees, a peeling order for recursive peeling can not be determined. In the iterative peeling algorithm the impossibility to find a peeling order in looped pedigrees shows from continuing changes in peeling terms. In looped pedigrees, these

changes were found to decrease in size quickly and temporary likelihoods were found to stabilise, supplying an approximation. Because in iterative peeling every following update of terms includes information from 50% less related individuals, a geometric rate of convergence is plausible. As a stopping rule to use the approximation in looped pedigrees, we used the average absolute difference between subsequent normalised heterozygote probabilities, based on computed peeling terms. For convenience, only the heterozygote probability, which changed the most, was monitored.

Simulation study

Application of iterative peeling to obtain approximate likelihoods in looped pedigrees was the aim of this study. Simulations were therefore performed to investigate the usefulness of this approximation. Because exact computations are infeasible in large looped pedigrees, approximate likelihoods could not be compared with exact ones. Hence, an indirect way to study the approximation was found by studying the distribution of test statistics and of parameter estimates over a number of replicated analyses in looped as well as non-looped pedigrees. In non-looped pedigrees exact likelihoods could be computed, serving as a reference. Simulations and analysis are based on a biallelic autosomal locus and a normal penetrance function.

Simulated data

Data sets had a nested structure each generation, with full sibs nested within paternal half-sibs. Three different data structures were used (Table 1), one structure without loops and two structures with loops. The data structures were designed to contain approximately the same number of observations, the same number of base individuals (structure 1 vs. 2) and the same family sizes (1 vs. 3). In structures 2 and 3, the third generation was produced by taking one son from each sire and one daughter from each dam, maintaining the same breeding structure across generations. No directional selection was practised, and breeding females for a male were taken each from a different sire-family. Half and full-sib matings were avoided, so that inbreeding was absent within the 3 generations considered. The additional third generation in structures 2 and 3 caused many pedigree loops in the form of marriage loops. All individuals used for breeding the last generation, i.e. 120 for structure 2 and 60 individuals for

structure 3, were involved in one or more of such loops, often overlapping.

Table 1 Possible structures of simulated data sets

Structure	Generations (including parents)	Sires, Dams and Progeny per dam (per generation)	Total Observations
1	2	20, 5, 10	1120
2	3	20, 5, 5	1120
3	3	10, 5, 10	1060

Genotype G_i of an individual equals 1, 2, or 3 corresponding to genotypes A_1A_1, A_1A_2 and A_2A_2 on an autosomal locus. Genotypes for individuals in the base population were randomly sampled using genotype frequencies according to Hardy-Weinberg proportions, after which genotypes of other individuals were randomly sampled based on realised parental genotypes assuming Mendelian transmission probabilities. For each individual a random normally distributed environmental component was sampled and added to a pre-determined effect of each genotype to obtain a phenotypic observation. Random numbers were generated using GGUBFS and GGNQF (IMSL, 1984). Details on the parameters used for these simulations are given in the following sections.

Model and model fitting

The statistical model can be specified by the probability terms in (2), (3) and (4) which are $P(G_i)$, genotype frequency in the base population for individual i, $P(G_i \mid G_s, G_d)$, transmission probability for individual i given genotype of its sire s and dam d, and the penetrance function $f(y_i \mid G_i)$, probability for the data y_i on individual i given the genotype G_i of individual i. From these three, transmission probabilities are assumed known to be Mendelian. Genotype frequencies in the base population depend on the unknown frequency f of the A_1 allele, assuming Hardy-Weinberg proportions of genotypes. The penetrance function for an individual i is taken as :

$$f(y_i \mid G_i)=(2\pi\sigma)^{-½}\exp\{-\tfrac{1}{2}(y_i-\mu_{Gi})^2/\sigma^2\}$$

This penetrance function is a normal probability density function with variance σ^2 around the mean μ_{Gi} for genotype G_i. No dominance is assumed. For analysis, means attributed to the genotypes are expressed as $\mu_1 = \mu-½t$, $\mu_2 = \mu$ and $\mu_3 = \mu+½t$, where t is the difference between homozygotes, referred to as the gene effect. The unknown parameters in the model are then f, μ, t, and σ^2.

Likelihoods were computed using iterative peeling. For structure 1, without loops, computations were done exactly by repeating the computations until no further changes occurred, having found the order for recursive computation. For the looped pedigrees of structures 2 and 3, iterative peeling was used to obtain approximate likelihoods. The stopping rule was a change less than 10^{-8} for the average absolute heterozygote probabilities of all individuals. The maximum of the likelihood was searched using the downhill simplex algorithm (Nelder and Mead, 1965), using as convergence criterion the variance of likelihood values of points in the simplex to be less than 10^{-12}.

Comparisons

Looped and non-looped pedigrees were compared in hypothesis tests and parameter estimation. In hypothesis testing, a null hypothesis postulating the absence of a major gene is used, described by a model with parameters μ and σ^2, and an alternative hypothesis postulating the presence of a major gene is used, described by a model with parameters f, μ, t, σ^2. Tests are based on the likelihood ratio (LR) test statistic, which is twice the natural logarithm of the ratio of maximum likelihoods under each hypothesis. Type I error and power, the complement of Type II error, were investigated at their nominal level, i.e. assuming the expected classicial asymptotic χ^2 distribution for the LR test statistic under the null hypothesis (Wilks, 1938). Using the classical rules, rejection thresholds were obtained from a χ^2 distribution with 2 degrees of freedom, being the difference in number of parameters between the null- and alternative hypothesis. It should be noted that for testing mixtures, these classical rules do not lead exactly to the nominal Type I errors (Titterington et al., 1985), but this is not of importance for the comparisons between looped and nonlooped pedigrees to be made here. The likelihood L_0 for the null-hypothesis is computed as:

$$L_0 = \prod_i (2\pi\sigma^2)^{-½} \exp\{-\tfrac{1}{2}(y_i-\mu)^2/\sigma^2\}$$

where y_i are observations with i=1, ... , N, the total number of observations, assumed normally and independently distributed. Under the null-hypothesis, the maximum likelihood estimate for the mean is $\hat{\mu}=\Sigma y_i/N$ and for the variance is $\hat{\sigma}^2 = \Sigma(y_i-\hat{\mu}_0)^2/N$.

Type I error of the test for a major gene was investigated by simulating 1000 data sets of each structure (Table 1), generating for each individual only a randomly distributed error term with σ^2=100 as phenotype. Likelihoods for the null hypothesis and the alternative hypothesis were computed in each of these replicated data sets, and the likelihood ratio test statistic was obtained. The number of significant tests in these 1000 data sets was counted using rejection thresholds of 4.605 and 5.991, corresponding to nominal Type I errors of 10% and 5%. Power to detect a major gene was investigated by simulating 100 data sets of each structure (Table 1) for three different gene effects t=5, t=7.5 and t=10 and using allele frequency f=0.5 and residual variance σ^2=100. Hence, relative gene effects t/σ were 0.5 0.75 and 1. Power was based on a nominal Type I error of 5%, using a rejection threshold of 5.991. Parameter estimates were compared using the 100 data sets of each structure (Table 1) used to investigate power with t=10.

Results

Type I errors were significantly lower than their nominal, i.e. asymptotically expected, level, but comparison of Type I errors between looped and non-looped structures does not show significant differences (Table 2). This indicates that absolute values of approximate likelihoods obtained are at average close to expected and that the distribution of the test statistic over a number of replicates is not significantly altered when loops are present. Similar conclusions can be drawn by comparing power of the test under the alternative hypothesis (Table 3). Parameters estimates for gene effect under the alternative hypothesis are biased in general, but estimates for gene effect as well as allele frequency do not differ between looped and non-looped structures (Table 4). This indicates that location of the maximum is, at average over replicates, not altered for approximate likelihoods.

Table 2 Estimated Type I errors (%) under the null hypothesis of no major gene, given for non-looped structures (1) and for looped structures (2,3) based on 1000 simulated data sets for each structure

Structure	Nominal level	
	10%	5%
1	2.8	1.4
2	3.2	1.4
3	2.5	1.7

Table 3 Estimated power (%) for a major gene test under the alternative hypothesis of presence of a major gene, given for non-looped structures (1) and for looped structures (2,3) based on 100 simulated data sets for each structure and for each of three different genetic effects

Structure	Genetic Effect t/σ		
	0.5	0.75	1
1	20	66	96
2	13	58	94
3	15	72	92

Table 4 Average parameter estimates for genetic effect (t) and allele frequency (f) with empirical standard errors of the mean (±SEM) under the alternative hypothesis of presence of a major gene, given for non-looped structures (1) and for looped structures (2,3), based on 100 simulated data sets for each structure

Structure	$\hat{t}$ ± SEM	$\hat{f}$ ± SEM
1	10.95 ± 0.30	0.479 ± 0.021
2	11.33 ± 0.23	0.499 ± 0.021
3	10.87 ± 0.25	0.501 ± 0.021

Simulated parameters : t=10 and f=0.5

Discussion and conclusions

An alternative peeling algorithm, called iterative peeling, was presented. The iterative peeling algorithm includes an algorithm to find an order for evaluating peeling equations. When an order can not be found, as in looped pedigrees, an approximate likelihood is supplied. Hereto, use of a partitioned computation of the likelihood also is crucial. Traditional recursive peeling does not know such approximations, because this method considers only to the compute the, exact, likelihood once a peeling order is found and computes the likelihood by representing all pedigree information in terms for a single individual. Usefulness of iterative peeling as an approximate method in looped pedigrees was investigated by simulations. At an aggregate level, i.e. compared at average over a number of replicated data sets, no difference were found between looped and non-looped pedigrees. Exact computations were infeasible due to the large number of loops in the typical animal breeding pedigrees we considered, and properties of iterative peeling could not be studied comparing exact and approximated likelihoods in individual data sets.

The iterative peeling method may be of interest for application in animal breeding. In human populations, pedigrees are generally small and loops are not abundant so that exact computations can be considered using more complicated forms of peeling (see Cannings et al., 1978). These more complicated forms of peeling consider genotypes on sets of individuals jointly. Larger pedigrees and more abundant looping in animal breeding, however, makes the sets of genotypes considered jointly too large to make exact computations feasible. Therefore, approximate methods are required for application in animal breeding. Iterative peeling seems very suited, being exact without loops, and automatically supplying approximate likelihoods when loops are present. Note that, due to the partitioned computation of likelihood, iterative peeling also automatically handles pedigrees consisting of independent families, i.e. data traditionally handled with sire- or sire and dam models. Equations and partitionings given here could be extended to allow for more general pedigrees. In particular, allowance could be made for females being mated with several males. Hereto, partitionings should accommodate for 'linking individuals' being parents in several families, rather than just one. The monogenic model used, could also be extended to a mixed inheritance model, the model usually required for analysis of animal breeding

data. In iterative peeling only uni- and bivariate functions of genotypes are considered on single families. This can be combined with for instance a hermitian integration (Le Roy et al., 1989; Knott et al., 1992) to include a polygenic component.

Acknowledgements

This research was supported financially by the Dutch Product Board for Livestock and Meat, the Dutch Pig Herdbook Society, Bovar BV, VOC Nieuw-Dalland BV, Euribrid BV and Fomeva BV.

References

Cannings C, Thompson EA, Skolnick MH (1976) The recursive derivation of likelihoods on complex pedigrees. Advan Appl Prob 8: 622-625

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Advan Appl Prob 10: 26-61

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21: 523-542

Hoeshele I (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. Theor Appl Genet 76: 81-92

IMSL (1984) Library reference manual Edition 9.2, International and statistical libraries, Houston, Texas

Knott SA, Haley CS, Thompson R (1992) Methods of segregation analysis for animal breeding data : a comparison of power. Heredity 68: 299-311

Lange K, Elston RC (1975) Extensions to pedigree analysis I. Likelihood calculations for simple and complex pedigrees. Hum Hered 25: 95-105

Le Roy P, Elsen JM, Knott SA (1989) Comparison of four statistical methods for detection of a major gene in a progeny test design. Genet Sel Evol 21: 341-357

Nelder JA, Mead R (1965) A simplex method for function minimization. Comp J 7: 147-151

Titterington DM, Smith AFM, Makov EU (1985) Statistical analysis of finite mixture distributions. Wiley and Sons, New York.

Van Arendonk JAM, Smith C, Kennedy BW (1989) Method to estimate genotype probabilities at individual loci in farm livestock. Theor Appl Genet 78: 735-740

Wilks, SS (1938) The large sample distribution of the likelihood ratio for testing composite hypothesis. Ann Math Stat 9: 60-62

Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations

Chapter 4

Application of Gibbs sampling is considered for inference in a mixed inheritance model in animal populations. Implementation of the Gibbs sampler on scalar components, as used for human populations, appeared not to be efficient and an approach with blockwise sampling of genotypes was proposed for use in animal populations. The blockwise sampling by which genotypes of a sire ad its final progeny were sampled jointly, was effective in improving mixing, although further improvements could be looked for. From Gibbs samples posterior densities of parameters were visualised, from which highly marginalised Bayesian point- and interval estimates can be obtained.

Introduction

Gibbs sampling has been proposed for making inferences in a mixed inheritance model in human populations (Guo and Thompson, 1992). The Gibbs sampler is a sampling-based computational tool to perform marginalisations without analytical approximation (Geman and Geman, 1984; Gelfand and Smith, 1990). As such, it can marginalise the joint density of unknowns from a mixed inheritance model with respect to polygenic effects as well as genotypes. Using analytical approaches (e.g. Le Roy et al., 1989; Knott et al., 1992; Kinghorn et al., 1993) this is an impossible task in general pedigrees. Due to its potential, Gibbs sampling, or related techniques, may soon dominate other computational methods for making genetic inferences, in particular when modelling single loci, such as in major gene detection and in QTL- and marker mapping. For a review on recent applications of Gibbs sampling in animal breeding see Sorensen et al. (1994).

Use of the Gibbs sampler implementation for human populations (Guo and Thompson, 1992) in animal breeding, may show very slow mixing of genotype states, resulting in difficulty in achieving convergence. Large progeny groups in animal breeding are responsible for this effect. The aim of this study was to describe the

construction of a markov chain using a modified sampling scheme, more suited for inference in animal populations. Because this study is the first report of using Gibbs sampling in a mixed inheritance model in animal breeding, we will describe in detail the construction of the required markov chain. The effect of the modified sampling scheme on mixing will be demonstrated. A small simulation study will be presented showing the types of marginal posterior densities that can be obtained with discussion of possible methods of inference based on these marginal densities.

Mixed inheritance model

In a mixed inheritance model a trait is influenced by the genotype at a single locus and by a polygenic effect, which is the aggregate effect of a large number of loci unrelated to the single locus. The single locus is assumed to be an additive, biallelic, autosomal locus with Mendelian transmission probabilities. Alleles at the single locus are A_1 and A_2 with genotypes A_1A_1, A_1A_2, A_2A_1 and A_2A_2. The heterozygotes A_1A_2 and A_2A_1 are distinguished to provide a simple and yet flexible notation for their covariance structure. In an alternative notation, the genotype of individual i is denoted $\mathbf{w}_i$ with four possible realisations ω_{ef}, a row vector, corresponding to genotype A_eA_f : ω_{11}=(1 0 0 0), ω_{12}=(0 1 0 0), ω_{21}=(0 0 1 0) and ω_{22}=(0 0 0 1). We assume a homogeneous population of base individuals with genotypes in Hardy-Weinberg proportions. Relaxation of these assumptions is feasible by increasing the number of parameters to be estimated, which poses no particular difficulty. We also assume that each individual has one observation for the trait. Inbreeding will be accounted for in the computations.

The statistical model for the observations is :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e} \qquad (1)$$

where β is a vector of fixed nongenetic effects, $\mathbf{X}$ is a design matrix relating nongenetic effects to observations, $\mathbf{u}$ is a vector of random polygenic effects for all individuals in the pedigree, $\mathbf{Z}$ is a design matrix relating polygenic effects to observations, $\mathbf{Wm}$ is a vector of random effects at the single locus for all individuals and $\mathbf{e}$ is a vector with errors. The effects at the single locus are expressed using $\mathbf{W}=\{\mathbf{w}_i\}$, a matrix containing information on the genotype of each individual, and $\mathbf{m}$, a vector with genotype means,

where $\mathbf{m}'=(-a\ 0\ 0\ a)$. Hence, the A_2 allele is assumed to increase the trait value, hereafter called the favourable allele, no dominance is assumed and no distinction is made between the effects of the two heterozygotes A_1A_2 and A_2A_1.

The distribution of $\mathbf{e}$ is $N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where N denotes the normal distribution. The covariance structures for polygenic effects can be expressed as $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where $\mathbf{A}$ is the numerator relationship matrix and σ_u^2 the polygenic variance. The covariance structure for genotypes, however, cannot be expressed in matrix notation. To show the parallels between polygenic and monogenic effects, therefore, we will specify both in scalar notation. For individual i, the polygenic effect u_i is :

$$u_i \sim N(0, \sigma_u^2), \text{ when } i \text{ is an individual in the base population} \quad (2a)$$

$$u_i \sim N(\tfrac{1}{2}u_{S,i}+\tfrac{1}{2}u_{D,i}, \phi_i\sigma_u^2), \text{ when } i \text{ is not in the base population} \quad (2b)$$

where $u_{S,i}$ and $u_{D,i}$ in (2b) are polygenic effects of the sire and dam of i, and $\phi_i=(\frac{1}{2} - \frac{1}{4}F_{S,i} - \frac{1}{4}F_{D,i})$ is the Mendelian sampling term for individual i, where $F_{S,i}$ and $F_{D,i}$ are inbreeding coefficients of the sire and dam of i. An analogous scalar notation for the covariance structure for the genotype $\mathbf{w}_i$ of individual i is :

$$P(\mathbf{w}_i = \omega_{ef}) = p_e p_f \quad (3a)$$

$$P(\mathbf{w}_i=\omega_{ef} \mid \mathbf{w}_{S,i}=\omega_{gh}, \mathbf{w}_{D,i}=\omega_{g'h'}) = \tau_{e,gh}\tau_{f,g'h'} \quad (3b)$$

where P denotes probability, p_1 and p_2 $(=1-p_1)$ are frequencies of alleles A_1 and A_2 in the base population, $\tau_{1,gh}$ is the probability of transmission to an offspring of an A_1 alelle from a parent with genotype A_gA_h and $\tau_{2,gh}=1-\tau_{1,gh}$. In (3b) $\mathbf{w}_{S,i}$ and $\mathbf{w}_{D,i}$ are genotypes of sire and dam of i. Assuming Mendelian probabilities of transmission, $\tau_{1,11}=1$, $\tau_{1,12}=\tau_{1,21}=\frac{1}{2}$ and $\tau_{1,22}=0$.

Flat priors are assigned for nongenetic effects β, for variance components σ_e^2 and σ_u^2, and for allele effect a and allele frequency p_1, i.e. $f(\beta, \sigma_e^2, \sigma_u^2, a, p_1) \propto$ constant. Variance components are apriori positive, i.e. excluding zero, and the allele frequency is bounded between zero and one, including the bounds. The joint density of all unknowns, given data $\mathbf{y}$, is symbolically denoted :

$$f(\beta, \mathbf{u}, \mathbf{W}, \sigma_e^2, \sigma_u^2, a, p_1 \mid \mathbf{y}) \quad (4)$$

Gibbs sampling

The Gibbs sampler is based on a markov chain which is primarily used to generate samples from a joint density (Geman and Geman, 1984; Gelfand and Smith, 1990). These samples allow the study of all marginal densities from that joint density. For statistical problems, the joint density is for the set of unknowns from a statistical model, given observed data, e.g. as in (4). In statistical problems, the Gibbs sampler is generally used to study marginal posterior densities of parameters, i.e., considering other parameters as nuisances. Using the primary joint structure in the samples, also sampling correlations between parameter estimates and, e.g., two-parameter countour plots can be obtained. In analytical approaches, the study of marginal densities would require integrations or summations, often not feasible to compute, but which are circumvented when using Gibbs sampling.

Validity

Using Gibbs sampling is valid when the joint density considered has a non-zero probability over its entire domain (Tanner, 1993), which is similar to the requirement of irreducibility of the Gibbs markov chain. An irreducible chain can be characterised as a chain which, from any state, has a positive probability of transition to each other state. Irreducibility is not always straightforward, e.g., a model with a single locus with more than two alleles or a discrete penetrance function leads to a reducible chain (Sheehan and Thomas, 1993; Lin et al., 1993). Also, chains may be 'practically' reduced, i.e., transition probabilities to certain states are so low that, in practice, these states are never reached. For the model described here, the Gibbs markov chain is theoretically irreducible; possible practical reducibility will be discussed later.

The use of an improper joint density, i.e., a non-integrable function, is also invalid for application of Gibbs sampling. Hobert and Casella (1994) showed that priors $(\sigma^2)^{-(b+1)}$ for variance component estimation in linear models lead to a proper posterior density when $b<0$. Hence, a flat prior, corresponding to $b=-1$ yields a proper posterior (see also Besag et al., 1991; Wang et al., 1993; Wang et al., 1994). We assume that the result of Hobert and Casella (1994) obtained for linear models is also valid for the mixed inheritance model.

General construction

The Gibbs markov chain is a continuing series of realisations for the unknowns β, **u**, **W**, σ_e^2, σ_u^2, a and p_1. Let $\theta^{[t]}=(\beta^{[t]}, \mathbf{u}^{[t]}, \mathbf{W}^{[t]}, \sigma_e^{2[t]}, \sigma_u^{2[t]}, a^{[t]}, p_1^{[t]})$ denote the set of realisations for the unknowns at state or cycle t in the Gibbs chain. Construction of the Gibbs chain requires a set of realisations $\theta^{[t+1]}$, given the current set of realisations $\theta^{[t]}$. To initiate the chain, a set of starting realisations $\theta^{[0]}$ is required, for which we used zeros for β and **u** and initial guesses for σ_e^2, σ_u^2, a and p_1. Genotypes **W** were initiated as all heterozygotes A_1A_2.

In the most straightforward implementation of the Gibbs sampler, $\theta^{[t+1]}$ is obtained by sampling for each θ_i (i=1,r) a new realisation from the conditional distribution of θ_i, given the available realisations $\theta_1^{[t+1]},..., \theta_{i-1}^{[t+1]}, \theta_{i+1}^{[t]},..., \theta_r^{[t]}$ and given the data **y** (e.g., Gelfand and Smith, 1990). The form of the conditional densities required, often appear to be simple. For model (1), conditional densities are normal for β_i's, u_i's and a, discrete for $\mathbf{w}_i$'s, inverted chi-square for σ_e^2 and σ_u^2 and beta for p_1. The simple form of the conditional densities allows implementation of a Gibbs chain for a mixed inheritance model based on sampling from the exact small sample distributions in each step. Further, for computations on pedigrees, the 'neighbourhood set' of an individual (e.g., Sheehan and Thomas, 1993) plays an important role. This neighbourhood set consists of the polygenic values or genotypes of the parents, progeny and mates of an individual, together with the data on itself. To compute conditional densities for sampling the polygenic effect or genotype of an individual, only the elements in this neighbourhood set are required, because of redundancies arising in the conditional densities. One side-effect in particular is that exact computations of conditional genotype probabilities are automatically made in looped pedigrees, whereas analytical approaches become intractable in pedigrees containing loops. Animal pedigrees generally contain many loops due to common occurence of multiple matings and inbreeding.

Mixing and blocking

The most straightforward implementation of the Gibbs chain, sampling single parameters, may not be an efficient way to obtain Gibbs samples because strongly dependent parameters may show slow mixing (Smith and Roberts, 1993; Tanner, 1993). By mixing we generally refer to the speed of movement of the chain in the parameter

space. In the initial phase of a markov chain, mixing is important for convergence to the equilibrium distribution and for burn-in time. In the later phase, mixing determines the serial correlations in the chain which affect efficiency by which accurate inferences can be made from the chain. In a mixed inheritance model, genotypes often show slow mixing due to the dependence between genotypes of parents and progeny. This dependence is stronger and mixing is poorer when progeny groups are larger.

As described by e.g., Smith and Roberts (1993) and Tanner (1993), mixing can be improved by applying Gibbs sampling to subvectors, treating components as a 'block', rather than using a complete breakdown of the parameter vector in its scalar components. In practice, blocking can be implemented using one or more reduced conditional densities. Use of reduced conditionals densities, as in substitution sampling, was also considered by Gelfand and Smith (1990) for improving convergence. Efficiency of the Gibbs sampler was improved here by blocking genotypes of each sire and its final progeny. Final progeny are progeny that are not parents themselves. By considering only final progeny, the number of individuals involved in computing conditional densities is not increased and remains based on parents, progeny and mates. But, for final progeny, phenotypes are used instead of genotypes. Efficiency was improved further by updating genotypes starting with the youngest families. In this manner, changes appearing in younger families can cause changes in older families within the same update cycle of the Gibbs chain. The blockwise treatment of genotypes of each sire and its final progeny was also applied to polygenic effects. In the results section the effect of blocking on the changes of genotypes in a Gibbs chain will be demonstrated.

Random number generator

Construction of the Gibbs markov chain requires sampling of many random deviates, which are based on pseudo-random number generators. Because parameters in the markov chain are updated repeatedly in the same order, the absence of serial correlations in the deviates is important. We used the RAN1, GASDEV and GAMDEV routines (Press et al., 1986), which seemed to meet that requirement. The GAMDEV routine was used to generate chi-square deviates with even-numbered degrees of freedom. Deviates with odd-numbered degrees of freedom were generated by adding one squared random normal deviate.

Sampling of realisations in the Gibbs chain

In this section obtaining $\theta^{[t+1]}$ given $\theta^{[t]}$ will be described. This represents computation of one 'Gibbs cycle'. Repeating this procedure constructs the Gibbs markov chain. The described blockwise treatment for genotypes and polygenic effects of sires and final progeny is incorporated, as well as the order of sampling starting with the youngest individuals. In the Gibbs chain, sampling is applied to all unknowns, including genetic parameters to allow for Bayesian inferences. Parameters are updated in the order given in the joint density (4).

Nongenetic effects

Assume first that nongenetic effects β are levels of one factor. Then step (S1) in the construction of state t+1 from t is :

(S1) sample $\beta_i^{[t+1]}$ from $N(\tilde{y}_{i.}/n_i,\ \sigma_e^{2[t]}/n_i)$.

(S1) is based on conditional solutions to the linear model and on conditional standard errors for nongenetic effects. Conditioning on polygenic effects, genotypes and allele effect results in the use of corrected data $\tilde{\mathbf{y}}=(\mathbf{y}-\mathbf{Z}\mathbf{u}^{[t]}-\mathbf{Z}\mathbf{W}^{[t]}\mathbf{m}^{[t]})$, with $\mathbf{m}^{[t]}=(-a^{[t]},\ 0,\ 0,\ a^{[t]})$, where $\tilde{y}_{i.}$ is the total of observations from $\tilde{\mathbf{y}}$ pertaining to level i and n_i is the number of observations in level i. More effects would be handled one at a time, correcting $\tilde{\mathbf{y}}$ also for other nongenetic effects. For two effects, β is partitioned as $[\beta_1\ \beta_2]$ and $\mathbf{X}$ as $[\mathbf{X}_1\ \mathbf{X}_2]$, and β_1 is updated to state t+1 as above using $\tilde{\mathbf{y}}=(\mathbf{y}-\mathbf{X}_2\beta_2^{[t]}-\mathbf{Z}\mathbf{u}^{[t]}-\mathbf{Z}\mathbf{W}^{[t]}\mathbf{m}^{[t]})$, after which β_2 is updated in the same manner using $\tilde{\mathbf{y}}=(\mathbf{y}-\mathbf{X}_1\beta_1^{[t+1]}-\mathbf{Z}\mathbf{u}^{[t]}-\mathbf{Z}\mathbf{W}^{[t]}\mathbf{m}^{[t]})$. Note the direct use of $\beta_1^{[t+1]}$.

Polygenic effects

Steps to update polygenic effects are based on BLUP equations for the linear 'animal model' (Henderson, 1988) and on conditional standard errors for polygenic effects. The neighbourhood set of polygenic effects (e.g., Sheehan and Thomas, 1993) to be considered is represented exactly in BLUP equations. Updating polygenic effects is based on using step (S2.1) for dam j, and step (S2.2) for sire i with its final progeny l:

(S2.1) sample $u_j^{[t+1]}$ from $N(c_j/d_j,\ \sigma_e^{2[t]}/d_j)$,

(S2.2) sample $u_i^{[t+1]}$ from $N(c_i/d_i,\ \sigma_e^{2[t]}/d_i)$ and
sample $u_l^{[t+1]}$ from $N(c_l/d_l,\ \sigma_e^{2[t]}/d_l)$ for each final progeny l of sire i,

where the BLUP equations are $d_j u_j = c_j$ to solve for the polygenic effect u_j of dam j; $d_i u_i = c_i$ to solve for the polygenic effect u_i of sire i after absorption of all final progeny of i; and $d_l u_l = c_l$ to solve for the polygenic effect u_l of final progeny l. Step (S2.2) is sampling of new realisations for a sire and its final progeny jointly as a block, done in two steps. The first step draws a new realisation for the sire effect from the reduced conditional density, after absorption of final progeny. The second step finalises the joint sampling by obtaining new realisations for final progeny, conditional on the new value for the sire. Based on BLUP equations, elements in (S2.1) and (S2.2) are :

$$
\begin{aligned}
c_j &= \tilde{y}_j + \tfrac{1}{2}\alpha\delta_j(u_{S,j}^{[t]} + u_{D,j}^{[t]}) - \alpha\textstyle\sum_k (\tfrac{1}{4}\delta_k u_{S,k}^{[.]} - \tfrac{1}{2}\delta_k u_k^{[t+1]}) \\
d_j &= 1+\alpha(\delta_j + \tfrac{1}{4}\textstyle\sum_k \delta_k) \\
c_i &= \tilde{y}_i + \tfrac{1}{2}\alpha\delta_i(u_{S,i}^{[t]} + u_{D,i}^{[t]}) - \alpha\textstyle\sum_m(\tfrac{1}{4}\delta_m u_{D,m}^{[.]} - \tfrac{1}{2}\delta_m u_m^{[t+1]}) \\
&\quad -\textstyle\sum_l\{\ \tfrac{1}{4}\alpha\delta_l u_{D,l}^{[.]} - (\tilde{y}_l + \tfrac{1}{2}\alpha\delta_l u_{D,l}^{[.]})/(1+\alpha\delta_l)\ \} \\
d_i &= 1 + \alpha(\delta_i+\tfrac{1}{4}\textstyle\sum_k \delta_k) - \tfrac{1}{4}\textstyle\sum_l(\alpha\delta_l)^2/(1+\alpha\delta_l) \\
c_l &= \tilde{y}_l + \tfrac{1}{2}\alpha\delta_l(u_i^{[t+1]} + u_{D,l}^{[.]}) \\
d_l &= 1+\alpha\delta_l
\end{aligned}
$$

where for individual i, $u_{S,i}$ and $u_{D,i}$ denote polygenic effects of the sire and dam of i, $\tilde{y}_i$ is the element pertaining to i from the corrected data $\tilde{\mathbf{y}}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[t+1]}-\mathbf{Z}\mathbf{W}^{[t]}\mathbf{m}^{[t]})$, δ_i is the reciprocal of the Mendelian sampling term ϕ_i from (2b); the premises are similar for other individuals j, k, l or m. In the equation for c_j, $\sum_k$ is evaluated for each progeny k of j; in the equation for c_i, $\sum_l$ is evaluated for each final progeny l of i and $\sum_m$ is evaluated for each nonfinal progeny m of i. In the equation for c_l, i is the sire of l. Finally, α is the variance ratio $\sigma_e^{2[t]}/\sigma_u^{2[t]}$. When it is unclear whether a polygenic value used is from state t or $t+1$, the state is not specifically indicated.

Genotypes

Obtaining new realisations for genotypes is done similarly as for polygenic effects, except that discrete distributions are sampled. Conditional probabilities for genotypes

are obtained by peeling (e.g., Cannings et al., 1978), but taking genotypes of the individuals in the neighbourhood set, i.e., parents, progeny and mates, as known. Analogous to polygenic effects, updating all genotypes is based on step (S3.1) for dam j, and step (S3.2) for sire i and its final progeny l :

(S3.1) sample $\mathbf{w}_j$ according to the probabilities :

$$P(\mathbf{w}_j=\omega_{ef}) \propto f(\tilde{y}_j \mid \mathbf{w}_j=\omega_{ef}) P(\mathbf{w}_j=\omega_{ef} \mid \mathbf{w}_{S,j}^{[t]}, \mathbf{w}_{D,j}^{[t]}) \cdot \prod_k P(\mathbf{w}_k^{[t+1]} \mid \mathbf{w}_{S,k}^{[.]}, \mathbf{w}_j=\omega_{ef})$$

(S3.2) sample $\mathbf{w}_i$ according to the probabilities

$$P(\mathbf{w}_i=\omega_{ef}) \propto f(\tilde{y}_i \mid \mathbf{w}_i=\omega_{ef})\ P(\mathbf{w}_i=\omega_{ef} \mid \mathbf{w}_{S,i}^{[t]}, \mathbf{w}_{D,i}^{[t]}) \cdot \prod_m P(\mathbf{w}_m^{[t+1]} \mid \mathbf{w}_i=\omega_{ef}, \mathbf{w}_{D,m}^{[.]}) \cdot \prod_l \Sigma_{g,h}\ P(\mathbf{w}_l=\omega_{gh} \mid \mathbf{w}_i=\omega_{ef}, \mathbf{w}_{D,l}^{[.]})\ f(\tilde{y}_l \mid \mathbf{w}_l=\omega_{gh})$$

and for each final progeny l of sire i, sample $\mathbf{w}_l$ according to:

$$P(\mathbf{w}_l=\omega_{gh}) \propto f(\tilde{y}_l \mid \mathbf{w}_l=\omega_{gh})\ P(\mathbf{w}_l=\omega_{gh} \mid \mathbf{w}_i^{[t+1]}, \mathbf{w}_{D,l}^{[.]})$$

where notation is analogous to that for polygenic effects, and P denotes probability. Here $\tilde{\mathbf{y}}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[t+1]}-\mathbf{Z}\mathbf{u}^{[t+1]})$, $f(\tilde{y}_i \mid \mathbf{w}_i=\omega_{ef}) \propto \exp\{-\frac{1}{2}(\tilde{y}_i-\omega_{ef}\mathbf{m}^{[t]})^2/\sigma_e^{2[t]}\}$ is the normal penetrance function for i, and $P(\mathbf{w}_i=\omega_{ef} \mid \mathbf{w}_{S,i}^{[t]}, \mathbf{w}_{D,i}^{[t]})$ is a transmission probability for i, available from (3b). When parents of i are unkown, the transmission probability is replaced by $p_e^{[t]}p_f^{[t]}$. The products over k, l and m are evaluated for the same individuals as the sums over k, l and m for polygenic effects and the sum within $\prod_l$ is evaluated over the possible genotypes of progeny l, for g=1,2 and h=1,2. Step (S3.2) is the sampling of the genotypes of a sire and its final progeny, where in the first part a new genotype for the sire is sampled from a reduced conditional density. In the reduced conditional density for a sire, phenotypes of final progeny are used. The actual sampling of genotypes is done by evaluating the above probabilities for all possible realisations ω_{11}, ω_{12}, ω_{21}, and ω_{22}, and sampling of a new genotype according to these probabilities. Probabilities are given to proportionality, and so need to be normalised.

Residual and polygenic variance

Variance components follow inverted chi-square distributions, with new realisations for σ_e^2 and σ_u^2 obtained as:

(S4) sample $\sigma_e^{2[t+1]}$ as $\mathbf{e'e}/\chi^2(n\text{-}2)$

(S5) sample $\sigma_u^{2[t+1]}$ as $\mathbf{u}^{[t+1]\prime}\mathbf{A}^{-1}\mathbf{u}^{[t+1]}/\chi^2(q\text{-}2)$

where $\mathbf{e}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[t+1]}-\mathbf{Z}\mathbf{u}^{[t+1]} - \mathbf{Z}\mathbf{W}^{[t+1]}\mathbf{m}^{[t]})$, $\mathbf{A}$ is the numerator relationship matrix, n is the number of observations, q is the number of individuals, and $\chi^2(n\text{-}2)$ and $\chi^2(q\text{-}2)$ are random deviates from chi-squared distributions with n-2 and q-2 degrees of freedom. Using degrees of freedom n-2 and q-2, a flat prior for variance components is used (Wang et al., 1994). The quadratic $\mathbf{u}^{[t+1]\prime}\mathbf{A}^{-1}\mathbf{u}^{[t+1]}$ is computed as $\sum_i u_i^2 + \sum_j d_j(u_j - \frac{1}{2}u_{S_j} - \frac{1}{2}u_{D_j})^2$, a scalar computation due to the factorisation of $\mathbf{A}$ (Quaas, 1976). The first summation is over all base animals and the second summation is over all non-base animals. Further notation is as in sampling steps (S2.1) and (S2.2). To prevent accidental rounding-off of variance components to zero, variances were not allowed to be smaller than 10^{-12}. Whenever a realised value fell below 10^{-12}, the sampling, i.e. (S4) or (S5), was repeated.

Allele effect

Using genotypes as a known classification factor, effect of an allele is estimated as the deviation of homozygotes from an assumed mean of zero, yielding a linear model equation $(n_1+n_4)\hat{a}=(\tilde{y}_{4.} - \tilde{y}_{1.})$. This leads to :

(S6) sample $a^{[t+1]}$ from $N((\tilde{y}_{4.} - \tilde{y}_{1.})/(n_1+n_4),\ \sigma_e^{2[t+1]}/(n_1+n_4))$

where n_i is diagonal element i of $\mathbf{W}^{[t+1]\prime}\mathbf{Z}'\mathbf{Z}\mathbf{W}^{[t+1]}$, giving the number of genotypes of each type; $\tilde{y}_{i.}$ is element i of $\mathbf{W}^{[t+1]\prime}\mathbf{Z}'\tilde{\mathbf{y}}$, containing sums of corrected data per genotype with $\tilde{\mathbf{y}}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[t+1]}-\mathbf{Z}\mathbf{u}^{[t+1]})$. When all genotypes are A_1A_1 or A_2A_2, i.e., n_1 or n_4 is equal to the total number of animals, the effect of the allele is nonestimable and the new realisation for a is taken as zero.

Allele frequency

Given genotypes of base individuals, allele frequency in the base generation has a beta distribution. This leads to :

(S7) sample $p_1^{[t+1]}$ from $f(p_1) \propto p_1^{B1}(1-p_1)^{B2}$

where B_1 is number of A_1 alleles and B_2 number of A_2 alleles in genotypes of base individuals. An acception-rejection technique is used to sample a new allele frequency. A 'suggested' sample p_1* is generated from a uniform density. This p_1* is accepted as the new sample for p_1 with probability $f(p_1^*)/f_{max}(p_1)$, where $f_{max}(p_1)$ is the maximum value of $f(p_1)$, attained for $p_1=B_1/(B_1+B_2)$. When p_1* is rejected, the procedure is repeated.

Statistical inference

In the following we will describe a straightforward use of a Gibbs chain for making statistical inferences. In the discussion section, we will elaborate on alternative approaches. For statistical inference, a long markov chain is produced, repeating the update scheme described in the previous section to obtain vectors with subsequent realisations for parameters. The subsequent realisations, or states in the markov chain, will show serial correlations, so that not every state is used to obtain Gibbs samples. Instead, virtually independent samples are obtained by "thinning the chain". From the original chain, every K^{th} sample is taken, which is referred to as thinning 'by K'. Determining a suitable K-value or thinning parameter will be described first.

Thinning parameter

An initial run of the Gibbs sampler is required to determine a suitable K value. Following Raftery and Lewis (1992), thinning is based on a transformation of the original output into a binary process, for which transition probabilities are studied. Let $\theta^{[t]}$ be the value for a certain parameter at state t in the test run. The binary process is defined as $Z^{[t]} = \delta(\theta^{[t]} \leq c)$, where δ is the indicator function and c, in our application, is the mean of $\theta^{[t]}$s. Thus, $Z^{[t]}$ indicates whether the realisation at state t was below or above the mean. The mean was taken because we are primarily interested in a central location parameter for the posterior densities. A suitable thinning parameter is obtained as follows, using for computations the binary process $Z^{[t]}$:

(i) a thinning parameter k_1 is determined such that $Z^{[t]}$, thinned by k_1, is approximately first order markov (Raftery and Lewis, 1992);

(ii) $Z^{[t]}$ thinned by k_1, being first order markov, can be described by a simple transition mechanism with transition probabilities α and β, which are estimated

from $Z^{[t]}$ thinned by k_1;

(iii) an additional thinning parameter k_2 is determined, such that the transition probabilities in $Z^{[t]}$, thinned by $K=k_1k_2$, differ only ε from the transition probabilities for $k_2\rightarrow\infty$, i.e., for $K\rightarrow\infty$. Based on estimated transition probabilities from (ii) and powers of the corresponding transition probability matrix, $k_2=\ln(\varepsilon)/\ln(1-\alpha-\beta)$. In our application, we took $\varepsilon=0.001$.

Step (iii) differs from the approach suggested by Raftery and Lewis (1992), who thinned only by k_1, yielding serially correlated realisations. Raftery and Lewis (1992) also determined the number of 'burn in' cycles to be $\ln(\varepsilon(\alpha+\beta)/\max(\alpha,\beta))/\ln(1-\alpha-\beta)$, which, for small ε and α and β approximately equal, is close to K. Therefore, taking the first Gibbs sample at state K, therefore, generally allows for a sufficient burn in as well. In practice this was indeed observed.

Determining K can be repeated for various parameters, or for functions of parameters in the Gibbs chain, e.g., a heritability as a ratio of variance components. Different parameters or different functions of parameters may yield different K's. The approach we used is to determine K for various parameters and functions and choose the largest K to be applied to all. Hence, the Gibbs chain can be constructed and, at every K^{th} cycle, realisations for parameters at that cycle are saved as being a "Gibbs sample" for the set of model parameters. Using the same thinning for all parameters, the primary joint structure in the samples is retained, allowing, e.g., computation of sampling correlations between parameter estimates.

Inference from marginal densities

So far, we have considered sets of realisations $\theta^{[t]}$ arising in the Gibbs chain as coherent units, being joint samples. Marginal densities of parameters are studied by observing realisations of a single parameter in these samples, irrespective of realisations for other parameters. We will focus on the genetic hyper parameters, i.e., variance components and effect and frequency of the major gene. However, non-genetic effects, polygenic effects and genotypes could be studied as well from the Gibbs chains. A very general inference is made by visualising the marginal posterior densities in a density estimate. In this study, we supply nonparametric density estimates in the form of average shifted histograms (Scott, 1992). At boundaries of parameter spaces, a reflection boundary technique (e.g., Scott, 1992, pg 149) was used to smooth the

histogram up to the boundary. The posterior density can be summarised by one or more statistics. Straightforward cases are, approximately, symmetric densities where mean and standard deviation are appropriate for describing the density. To describe more complicated densities, the mode often is a valuable third statistic. For symmetric densities, the mean will correspond to a maximum likelihood or maximum a-posteriori point estimate and the standard deviation will correspond to the small sample standard error of this parameter estimate. Parameter estimates based on Gibbs samples are subject to Monte Carlo (MC) error. Because our analysis is based on nearly independent Gibbs samples, empirical MC error on the posterior mean simply can be assessed from the estimated standard deviation of the posterior density and the number of Gibbs samples generated.

In the Gibbs chain, allele effect a may appear positive as well as negative. The sign of a, however, is not relevant, being based on the arbitrary assignment of A_2A_2 as the genotype with value $+a$. From the Gibbs samples, therefore, we studied the absolute values of a. For consistency, we also studied the frequency of the favourable allele, denoted p_h. The favourable allele is A_2 when a is positive and A_1 when a is negative.

Simulated data

A population was simulated in which 10 males were mated with 4 dams each, producing 5 progeny per female, yielding 200 offspring per generation. A sex was assigned to each progeny at random on a 1:1 ratio, but requiring at least one male and one female in each full-sibship. For each subsequent generation, each sire was replaced by a son and each dam was replaced by a daughter. Generations were non overlapping and no intentional selection was practiced. Mating was at random; unintentional inbreeding could be present from the second generation onwards because of finite population size. The theoretical rate of inbreeding was ≈0.8% per generation. The population was simulated for 5 generations, which resulted in a population of 1050 individuals, including the 50 base generation individuals.

For all individuals observations were simulated according to the model of analysis. This simulation included polygenic effects from the normal densities (2a) for base animals and from (2b) for non-base animals, genotypes according to probabilities

from (3a) and (3b) and sampling of normally distributed random errors. Two data sets were simulated for which genetic parameters are in Table 1. In data set 0 no effect at the single locus was simulated. This set was used to demonstrate results found when no single gene effect is present. In data set 1 a single gene effect was simulated with a difference (25) between extreme genotypes of ≈2 standard deviations of the variation within single genes. The effect of the single gene in set 1 was expected to be clearly detectable. Average inbreeding coefficients in generation 5 were 4.1% in data set 1 and 4.2% in data set 2, matching theoretical predicted rate. Numbers of individuals with nonzero inbreeding coefficients were 450 in data set 0 and 430 in data set 1. This indicates a large number of pedigree loops in these data sets already due to inbreeding alone. Multiple matings applied in this simulated breeding structure resulted in an additional large number of loops. Inbreeding was taken into account in the simulation of polygenic effects and in the analysis at steps (S2) and (S4). An effect of sex was simulated which favoured males by +2 units and sex was used in the analysis as an explanatory nongenetic effect.

Table 1 Parameter values used in simulation

Parameter	Data set 0	Data set 1
σ_e^2	100	100
σ_u^2	50	50
a	0	12.5
p_1	-	0.3
$\sigma_m^2=2p_1p_2a^2$	0	65.6

Results

Mixing and the effect of blocked Gibbs sampling

For data set 1, with a simulated effect of a major gene, changes of genotypes were studied for three classes of individuals : final progeny, dams and sires. In Table 2 the average number of genotype changes per cycle is given for each class of individuals. Without blocking, virtually non of the sire-genotypes changed in the majority of the

Gibbs cycles. Results were such that about once in 100 cycles, one sire-genotype was changed. Hence, the genotype configuration for sires remains practically the same over many hundreds of cycles and movement of the markov chain is restricted to a small subspace. In this case, changes appearing for final progeny and dams are relatively meaningless, because these changes are limited due to the near fixation of all sire genotypes. With the blocking technique, mixing is improved, changing about 5% of the sire-genotypes each cycle. The increased changes in sire genotypes resulted in a general increase of changes in the entire pedigree, which can be seen in particular for dam genotypes.

Table 2 Average number of genotype changes per Gibbs cycle for three groups of individuals with a scalar updating of genotypes ('scalar') and with a block updating of genotypes of sires and final progeny ('block') in data set 1 (average of 10000 Gibbs cycles)

Group (total number)	Average number of changes per Gibbs cycle	
	scalar	block
Finals (800)	234 (29%)	258 (32%)
Dams (200)	12.7 (6.4%)	39.8 (20%)
Sires (50)	0.008 (0.02%)	2.62 (5.2%)

Data without a major gene

Determination of a thinning parameter K for data set 0 was based on an initial run of the Gibbs chain of 10000 cycles. Starting realisations for genetic parameters were taken as the simulated values (Table 1) which represented a pure polygenic mechanism. Allele frequency p_1 was initiated as 0.5. Thinning parameters were determined for the variance components for errors, σ_e^2, for polygenic effects, σ_u^2, and for major gene effects, $\sigma_m^2=2p_1p_2a^2$, for the absolute value of the effect of the allele $|a|$, and for the frequency of the favourable allele, p_h. Allele frequency p_h showed the strongest dependencies, requiring $K\approx890$ to yield independent samples.

For data set 0, two Gibbs chains were run, each initiated with different seeds for the random number generator. From each chain, 250 Gibbs samples were obtained

using K=890. Results from each set of Gibbs samples are in Table 3, presenting the estimated contributions of the two genetic components and error in terms of variances. Runs were consistent in the estimate for major gene variance. In each case, a unimodal density for major gene variance was found with a mode at zero (Figure 1). From such densities, we infer the variance component to be zero, which means an absence of the major gene effect.

Table 3 Estimated means and standard deviations of posterior densities for genetic parameters in data set 0 (no major gene) in two runs of the Gibbs sampler, based on 250 samples per run.

	Mean (Standard deviation)	
Parameter	Run 1	Run 2
σ_e^2	98.0 (7.4)	95.7 (6.8)
σ_u^2	38.1 (14.3)	48.5 (11.4)
$\sigma_m^2=2p_1p_2a^2$	5.3 (4.9)[a]	4.8 (5.3)[a]

[a] Mode is zero

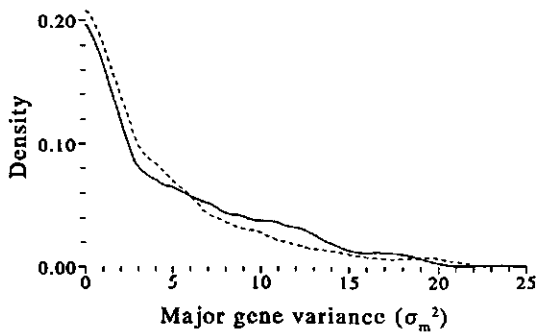


Figure 1 Estimated posterior densities (averaged histogram frequencies) for major gene variance in data set 0 for run 1 (solid line) and run 2 (dashed line) of the Gibbs sampler, based on 250 samples per run.

Estimates for variance components were different in the two runs, especially for polygenic variance. Posterior means for σ_u^2 differed about 10 units, which cannot be explained by Monte Carlo (MC) error. The empirical MC error on the means for σ_u^2 was estimated as 0.9 for run 1 and 0.7 for run 2. Differences in these estimates must be caused by a near reducibility of the chain, with allele frequency moving in a few subspaces between which mixing is relatively bad. When the probability of moving to a different subspace is low this type of behaviour is unlikely to be spotted in the tuning phase, which we based on 10000 cycles only.

Table 4 Estimated means and standard deviations of posterior densities for genetic parameters in data set 1, based on 500 Gibbs samples.

Parameter	Mean (Stand. dev.)
σ_e^2	104.6 (10.4)
σ_u^2	35.6 (16.6)
$\lvert a \rvert$	12.5 (1.82)
p_h	0.56 (0.12)
$\sigma_m^2=2p_1p_2a^2$	73.5 (19.5)

Data with a major gene

Posterior means and standard errors for parameter estimates in data set 1 are in Table 4. These estimates are based on 500 Gibbs samples from a single Gibbs chain, using K=400. Starting values for data set 1 represented a pure polygenic model, i.e., $\sigma_u^2 \approx 116$ and a=0, which does not correspond to the simulated parameters. From this polygenic starting point, the Gibbs chain was observed to move to a mixed inheritance model in a few hundred cycles. Density estimates for σ_u^2, $\lvert a \rvert$, p_h and $\sigma_m^2=2p_1p_2a^2$ are in Figure 2. The density estimate for σ_u^2 shows a unimodal density with a mode for $\sigma_u^2>0$, indicating the significance of the polygenic component in the model. The density estimate for σ_m^2 shows a local mode for $\sigma_m^2=0$, and a global mode for $\sigma_m^2>0$. The odds are 1:26 between the estimated density for $\sigma_m^2=0$ and for $\sigma_m^2>0$, which is taken as evidence for a significant single gene component. Neither posterior means nor

modes agree perfectly with the simulated values, but in each case the simulated value was well within a 90% highest posterior density region of the estimate.

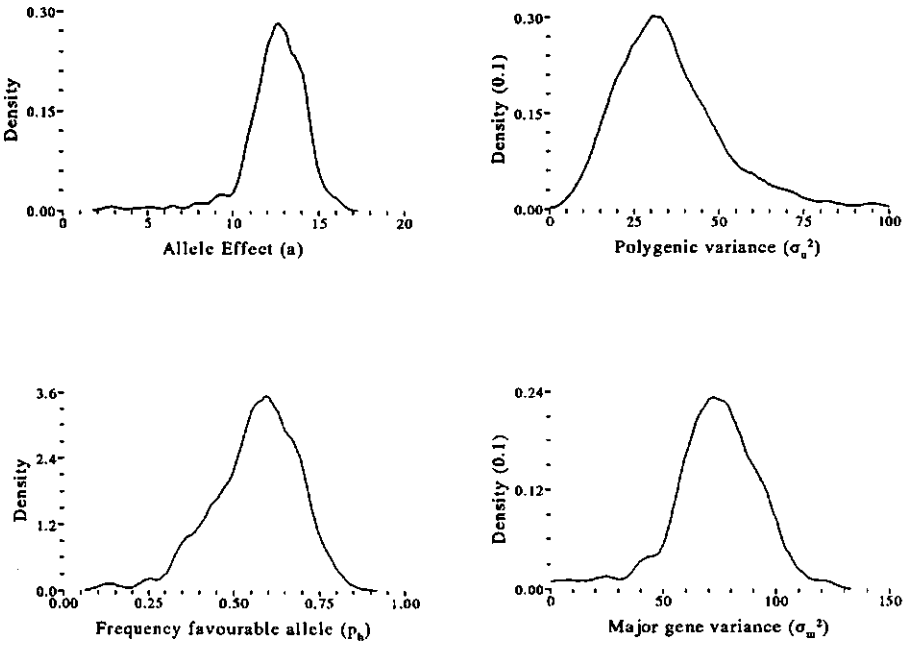


Figure 2 Estimated posterior densities (averaged histogram frequencies) for genetic parameters in data set 1, based on 500 Gibbs samples

Allele frequency in data set 1 was poorly estimable, showing values in a range between 0.15 and 0.9 (Figure 2). The influence of allele frequency on estimated polygenic variance and major gene variance is large. Two more analyses of data set 1 were performed, fixing the allele frequency of the favourable allele at 0.74 or at 0.60 (Table 5). The value of 0.74 was the true realised value in the simulation of data set 1 and the value of 0.60 was around the mode of the marginal posterior of allele frequency. Each value, therefore, can be taken as a plausible estimate which, based on the posterior from Figure 2, are not dramatically different from each other. Use of an estimated value, treated as a true value without error in a further estimation step, is a procedure common for classical inference from a joint likelihood function. Fixing allele

frequency, the portion of polygenic variance in the total genetic variance ranged from 38% for p_h=0.74 to 28% for p_h=0.60. The MC error on the posterior means for polygenic variance is about 0.5% of the estimated total genetic variance and, therefore, is too small to account for these differences. Hence, fixing the unknown allele frequency at some value substantially affects estimates for the two genetic variances. In contrast, the 'marginal' estimates (Table 4), which are averaged over all possible allele frequencies, are not affected by the arbitrary choice of a point estimate.

Table 5 Estimated means and standard deviations of posterior densities for genetic parameters in data set 1 fixing allele frequency at two different values, based on 500 Gibbs samples per case

	Mean (Standard deviation) using	
Parameter	p_h=0.74	p_h=0.60
σ_e^2	102.0 (10.5)	104.1 (10.0)
σ_u^2	40.1 (13.9)	31.0 (12.4)
$\lvert a \rvert$	13.1 (1.39)	12.6 (1.25)
$\sigma_m^2=2p_1p_2a^2$	66.7 (13.5)	81.3 (15.1)

Discussion

Mixing in the Gibbs chain

In this study we described the construction of a Gibbs markov chain for inference in a mixed inheritance model. Efficiency of the Gibbs sampler depends on the parameterisation used and on the sampling scheme applied. A Gibbs sampling approach for a mixed inheritance model applied to human populations (Guo and Thompson, 1992) is inefficient when applied to animal populations. We suggested a blockwise treatment for genotypes, yielding faster changes in the Gibbs chain without considerable complications in computing. The blocking is typically applied to parents with large progeny groups. We applied this for sires, but the technique can also be applied to dams. Without blocking, markov chains remain stuck in a subspace of the

parameter space, making a proper inference impossible. With blocking, mixing was improved, although inference in data set 0 remained difficult. Here, two Gibbs chains did not yield exactly similar results for all parameters, possibly the result of a more subtle type of bad mixing. Multiple runs of the Gibbs sampler, preferably with various starting values, can be used to spot, but not to solve, such problems of mixing. The blocking technique, therefore, is possibly only a first step to improve mixing and more methods could be developed and added. Note further that the efficiency of blocking will depend on the data structure, in particular, on the progeny group sizes and on the allele effect at the major locus. In animal breeding practice, progeny groups are generally sufficiently large to recommend the use of blocking.

Alternative uses of Gibbs chains

Efficiency in using realisations from a markov chain for statistical inference can possibly be improved. For instance, use of independent samples is not required. Posterior means and other density features, including the density itself, can be estimated direclty using serially correlated states in the chain (Geyer, 1992; Wang et al, 1994). Advantages of our approach of using independent samples is that accuracy of output from a Gibbs chain can be appreciated directly, simply by the number of samples. Independent samples also allow comparison of output from multiple chains by standard analysis-of-variance methods. A further measure to increase accuracy of the estimate of a mean is the use of Rao-Blackwell estimates (Gelfand and Smith, 1990). This procedure uses from every state the expected value for a certain parameter, rather than the realised value in the chain. Expected values are often directly available from the intermediate computations in the Gibbs chain, and vary less because the disturbance from the conditional variance is eliminated.

Statistical inference

In the mixed major gene-polygenic inheritance model, maximum likelihood (ML) inference is classically employed (e.g., Elston and Stewart, 1971; Morton and MacLean, 1974). Gibbs sampling can also be used to obtain such ML estimates (e.g., Guo and Thompson, 1992). Specification of prior densities is then circumvented by updating a parameter, e.g., a variance component, not with samples from the specified densities but with the expectation for that parameter given realisations of other parameters. This

technique is known as Monte Carlo EM (Tanner, 1993). In our model, a REML inference could be made by omitting the sampling steps for σ_e^2, σ_u^2, a and p_1, and by updating these parameters as their expectation. A ML inference could be implemented by also updating elements of β with their expectation. In this manner, based on the Gibbs sampler, a hierarchy of inferential methods can be obtained by suppressing certain sampling steps in the construction of the chain. Note that when using this Monte Carlo EM technique, fluctuations in the chain will not correspond to standard errors of parameter estimates, and density estimates of posteriors cannot be made.

ML inference and associated hypothesis testing in major gene models, however, have several shortcomings. For instance, REML (Patterson and Thompson, 1971) was developed to overcome biases in ML point estimates for variance components, and ML standard errors and likelihood ratio tests are based on asymptotic normal approximations. For application of the likelihood ratio test, moreover, assumed asymptotic distribution of the test statistic is questionable when dealing with mixture distributions (Titterington et al., 1985). Using Gibbs sampling, alternatives for ML inference are available.

Inference could possibly be improved by using the Gibbs chain as implemented in this study, including the sampling steps for all genetic hyper parameters, and making use of the marginal posterior densities of parameters obtained. This approach is generally Bayesian, and for our implementation with flat priors for all hyper parameters could be classified as 'empirical' Bayesian. With this approach, standard errors of parameter estimates, or, in general, interval estimates in any form, are directly available. Interval estimates will be based on small sample distributions and respect the natural bounds on parameter spaces. As point estimates, mean or mode of the posterior density could be used, which would be respectively marginal APE (a-posteriori expectation) or marginal MAP (maximum a-posteriori) estimates. APE is simple to compute from Gibbs chains, and APE estimates are considered more optimal than estimators locating a marginal or joint mode (Henderson, 1953; Harville, 1977). However, absence of a variance component shows a density with a gobal mode at zero, as we showed, in which case the posterior mode is an appealing point estimate. This favours, from a more practical point of view, use of MAP estimates. Further, making use of the Gibbs chains as we presented, highly marginalised densities are used, considering for each parameter all other parameters as nuisances. This provides a richer

summary and may improve estimation of the two genetic components in the mixed inheritance model. We showed for instance when fixing allele frequency, that estimates for genetic variances in the mixed inheritance model depend on the value used for allele frequency. Marginal estimates, however, take into account the error in estimating allele frequency, or any other parameter. This gives a more realistic inference, representing better uncertainty in the estimates and providing a better disentanglement between, e.g., polygenic and major gene variance.

Hypothesis testing

We did not thoroughly consider power to detect single genes or test of significance of the single gene component. It was shown for major gene variance, σ_m^2, that absence of a single gene effect leads to a global mode for $\sigma_m^2=0$. As discussed, the MAP estimate would be zero in this case, correctly indicating absence of a single gene effect. Presence of a single gene effect showed a density with a global mode for $\sigma_m^2>0$, and a local mode for $\sigma_m^2=0$. We used the odds ratio of the densities at both modes as a criterion, assuming significance at a 5% level when the odds ratio is above 1:20. This criterion, however, may be very severe. An alternative would be to assume a mixed mode of inheritance as soon as the mode for $\sigma_m^2>0$ dominates the mode for $\sigma_m^2=0$. When experimenting with smaller effects of the major gene, a gradual increase of the density at $\sigma_m^2=0$ was indeed observed, indicating less likely action of a major gene. It would be of interest to further develop hypothesis testing because a test based on small sample distributions obtained from the Gibbs sampler has the potential to improve the likelihood ratio test for presence of a major gene. Gibbs sampling approaches also can handle very large data sets, e.g., as shown using a polygenic model by Van der Lugt et al. (1994), because Gibbs sampling implementations require little memory and do not accumulate round-off errors. This facilitates use of the generally abundant amount of information in animal populations, which is a simple measure to increase power. Additional simulations showed, for instance, that a major gene with a=6 and other parameters as in data set 1, i.e. explaining about onequarter of all genetic variance, was detected easily in a data set with 5000 individuals.

Acknowledgements

This research was supported financially by the Dutch Product Board for Livestock and Meat, the Dutch Pig Herdbook Society, Bovar, Euribrid, Fomeva and Nieuw Dalland. Daniel Gianola and Daniel Sorensen are acknowledged for discussions on the use of improper priors for variance components.

References

Besag J, York J, Mollie A (1991) Bayesian image restoration with two applications in spacial statistics (with discussion). Ann Inst Statist Math 43: 1-59

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Advan Appl Prob 10: 26-61

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523-542

Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. J Am Stat Assoc 85: 398-409

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans Pattn Anal Mach Intell 6: 721-741

Geyer CJ (1992) A practical guide to Markov chain Monte Carlo. Statist Sci 7 : 467-511

Guo SW, Thompson EA (1992) A monte carlo method for combined segregation and linkage analysis. Am J Hum Genet 51: 1111-1126

Harville DA (1977) Maximum likelihood approaches to variance component estimation and related problems. J Am Stat Assoc 72: 320-339

Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9: 226-252

Henderson CR (1988) Theoretical basis and computational methods for a number of different animal models. J Dairy Sci 71 : supplement 2 : 1-16

Hobert JP, Casella G (1994) Gibbs sampling with improper prior distributions. Technical report BU-1221-M, Biometrics Unit, Cornell University

Kinghorn BP, Kennedy BW, Smith C (1993) A method of screening for genes of major effect. Genetics 134 : 351-360

Knott SA, Haley CS, Thompson R (1992) Methods of segregation analysis for animal breeding data : a comparison of power. Heredity 68 : 299-311

Le Roy P, Elsen JM, Knott SA (1989) Comparison of four statistical methods for detection of a major gene in a progeny test design. Genet Sel Evol 21 : 341-357

Lin S, Thompson E, Wijsman E (1993) Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. IMA J Math Med & Biol 10: 1-17

Morton NE, MacLean CJ (1974). Analysis of family resemblance III. Complex segregation of quantitative traits. Am J Hum Genet 26: 489-503

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545-554

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1986) Numerical recipes; The art of scientific computing. Cambridge University Press

Quaas RL (1976) Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32: 949-953

Raftery AE, Lewis SM (1992) How many iterates in the Gibbs sampler? In: Bernardo JM, Bergen JO, David AP, Smith AFM (eds) Bayesian statistics, Oxford University Press.

Scott DW (1992) Mutivariate density estimation. Wiley and Sons, New York

Sheehan N, Thomas A (1993) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics 49: 163 - 175

Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related markov chain monte carlo methods. J Roy Stat Soc B 55: 3-24

Sorensen D, Anderson S, Jensen J, Wang CS, Gianola D (1994) Inferences about genetic parameters using the Gibbs sampler. Proc 5th World Congr Genet Appl Livest Prod, Guelph Canada, 18 : 321-328

Tanner MA (1993) Tools for statistical inference. Springer-verlag, New York

Titterington DM, Smith AFM, Makov EU (1985). Statistical analysis of finite mixture distributions. Wiley and Sons, New York

Van der Lugt AW, Janss LLG, Van Arendonk JAM (1994) Estimation of variance components in large animal models using Gibbs sampling. Proc 5th World Congr Genet Appl Livest Prod, Guelph Canada, 18 : 329-332

Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet Sel Evol 25: 41-62

Wang CS, Rutledge JJ, Gianola D (1994) Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet Sel Evol 26: 91-115

Bayesian statistical analyses for presence of single genes affecting meat quality traits in a crossed pig population

Chapter 5

Presence of single genes affecting meat quality traits was investigated in F_2 individuals of a cross between Chinese Meishan and Western pig lines using phenotypic measurements on 11 traits. A Bayesian approach was used for inference about a mixed model of inheritance, postulating effects of polygenic background genes, action of a bi-allelic autosomal single gene and various non-genetic effects. Cooking loss, drip loss, two pH measurements, intramuscular fat, shearforce and back-fat thickness were traits found to be likely influenced by a single gene. In all cases, a recessive allele was found, which likely originates from the Meishan breed and is absent in the Western founder lines. By studying associations between genotypes assigned to individuals based on phenotypic measurements for various traits, it was concluded that cooking loss, two pH measurements and possibly backfat thickness are influenced by one gene, and that a second gene influences intramuscular fat and possibly shearforce and drip loss. Statistical findings were supported by demonstrating marked differences in variances of families of fathers inferred as carriers and those inferred as non-carriers. It is concluded that further molecular genetic research effort to map single genes affecting these traits based on the same experimental data has a high probability of success.

Introduction

Since the advent of modern DNA techniques, identification of single genes is receiving increased attention in fundamental and applied sciences. Study of effects of single genes can aid in unravelling physiological processes which has relevance for many life sciences, and often has relevance across species. For instance, the finding of an obesity gene in mice (Zhang et al., 1994) may have relevance for several other mammals such as humans or pigs. Use of animal populations for identification of single genes can have several advantages unseen in human populations, such as large amounts of data,

designed experiments and controlled breeding. This makes the use of animal populations, in this respect, worthy of further attention.

In commercial livestock populations, phenotypic observations are often abundantly available at low costs, making preliminary statistical analysis a worthwhile first step in the identification of single genes. Segregation analysis (Elston and Stewart 1971; Morton and MacLean, 1974) is the most powerful statistical method for identification of single genes (Hill and Knott, 1990) but, so far, has not found widespread use in animal genetics. Exact computations involved in application of this method are impossible in common situations arising in animal populations, and analytical approximations (e.g., Le Roy et al., 1989; Knott et al., 1992) limit application to simple models and simple pedigree structures. Recently, however, Gibbs sampling and related Markov chain Monte Carlo methods (Geman and Geman 1984; Gelfand and Smith 1990; Smith and Roberts, 1993) have been introduced which can facilitate computations in many statistical applications. Gibbs sampling has been used in human genetics for both likelihood based as well as Bayesian based inferences in variance component estimation (Guo and Thompson, 1991), segregation and linkage analysis (Guo and Thompson, 1992; Thomas and Cortessis, 1992), computation of genotype probabilities (e.g., Sheehan and Thomas, 1993) and gene mapping (as an example in Smith and Roberts, 1993). In animal genetics, Gibbs sampling has been introduced in Bayesian approaches for variance component estimation in linear models (e.g. Wang et al., 1993), and in non-linear models (Sorensen et al., 1995), for analysis of selection response (Sorensen, et al., 1994) and for segregation analysis (Janss et al., 1995).

The aim of this paper is to investigate whether single genes may exist which affect meat quality traits, measured in the F_2 of a cross of Chinese Meishan and Western pig lines. Detection of single genes is based on a statistical modelling approach, using a Bayesian approach to segregation analysis described by Janss et al. (1995). The secondary aim of this paper is to demonstrate the flexibility, by virtue of applying Gibbs sampling, of this Bayesian approach, which so far has not been used for the analysis of field data. This highlights details in application of this new methodology and shows the types of inferences that are produced, which are different from inferences made by classical likelihood-based segregation analysis.

Material

Data

F_2 crossbreds between the Chinese Meishan pig breed and Western pig lines were available from an experiment involving five Dutch pig breeding companies (Figure 1). Crossbreds were produced in two batches at the same time in all companies. For each batch, purebred Western females at the companies were artificially inseminated by a group of 19 purebred Meishan males from a centrally housed population of Meishan animals, producing crossbred F_1 litters. Purebred females at the breeding companies were of Dutch Landrace and Large White types. In total, 126 F_1 crossbred litters were produced. From F_1 litters, a random selection of young males and females were taken as parents to produce F_2 crossbred litters, each female producing one F_2 litter. In total, 264 F_2 litters were produced, descending from 39 F_1 fathers. The 39 F_1 fathers were used across breeding companies through the formation of a central pool and use of artificial insemination; F_1 females remained at the breeding companies where born. This yielded a 75% similar genetic background for all F_2 crossbreds. From performance tested F_2 animals, about 1200 in total, approximately 350 animals were retained for further breeding. The majority of these animals were gilts (>300), chosen at random within the framework of the experiment. The additional animals (<50) were retained by the breeding companies, likely with selection on a combination of production and reproduction traits. Due to the low number involved and emphasis on different traits, the effect of this selection on the traits used in the current analyses is expected to be negligible. Performance tested F_2 animals not retained for breeding were slaughtered at approximately 90 kg in a central slaughter-house. On these slaughtered animals, several meat quality traits were measured. For genetic analyses, a pedigree file was constructed including F_1 parents and pure line (Meishan and Western) grand-parents of the observed F_2 individuals. Presence of Halothane susceptibility (Eikelenboom and Minkema, 1974), a common known genetic defect affecting meat quality traits in western lines and which is known as malignant hyperthermia in man, was excluded, by molecular typing of pure Meishan founders and F_1 fathers, which were all found free of this Halothane susceptibility mutation. Molecular typing was done by Van Haeringen Laboratorium BV (Wageningen, The Netherlands), using methods as described by Otsu et al. (1992).

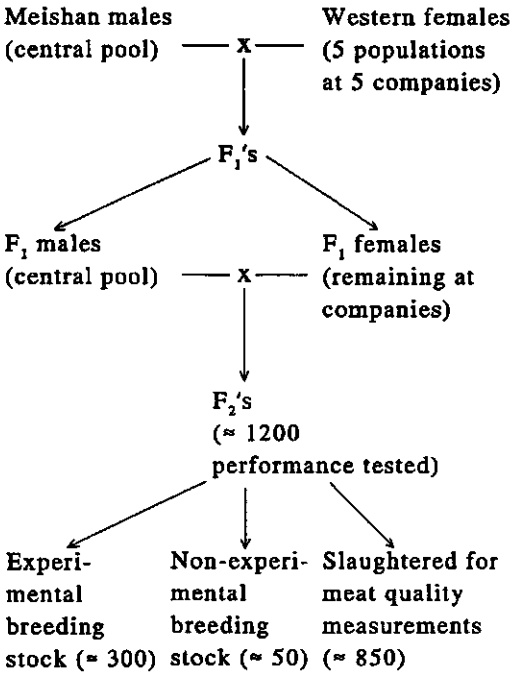


Figure 1 Design of the crossing experiment to produce F_2 crossbreds between Chinese Meishan and 5 Western pig lines. Step 1: 126 F_1 litters were produced from 19 Meishan males and 126 females of 5 Western lines in 5 companies. Step 2: 264 F_2 litters were produced from 39 centrally housed F_1 males taken equally from all companies, and from 265 F_1 females having remained in the companies. Step 3: from produced F_2 crossbreds, animals not used for breeding were centrally slaughtered to measure meat quality traits. All selection steps were random, except selection of non-experimental F_2 breeding stock.

Measurements

In Table 1 numbers of observations, raw means and standard deviations for the traits measured are given. In samples of *M. Longissimus* (loin muscle), pH, drip loss, cooking loss, shearforce, intramuscular fat and color were measured; additionally, a pH

measurement was taken in a sample of *M. Semimembranosus* (a ham muscle) and back fat- and lean thickness were measured. Color was measured as three coordinates according to the CIELAB L*a*b* system, where L* is a general indication of lightness, a* represents the degree of green-redness and b* represent the degree of blue-yellowness (MacDougall, 1986). Fat- and lean thickness are based on a single measurement with the Hennessy Grading Probe between the 3rd and 4th rib, 6 cm from the spine, as routinely done in The Netherlands to predict carcass meat percentage. Predicted meat percentages, however, were not analyzed in this study, because the employed prediction equation to predict meat percentage from fat- and lean thickness might not hold for the relatively fat Meishan crossbreds. All traits were measured 24 hours after slaughter, except fat- and lean thickness which were measured directly after slaughter. Exact details on measurement procedures for these traits can be found in Hovenier et al. (1992). In the following, abbreviations for trait names will be used as given in Table 1.

Table 1 Overview of meat quality traits measured[a], the numbers of observations (*N*) and raw means and standard deviations.

Trait	Full name, measurement unit	*N*	Mean	Std
Drip	Drip Loss, %	844	2.70	1.54
Cook	Cooking loss, %	845	26.4	3.46
Shear	Shear force, N	845	39.6	10.5
Imfat	Intramuscular fat, %	831	1.84	0.87
pH	pH	845	5.66	0.26
pH-s	pH in *M. Semimenbranosus*	846	5.82	0.30
Light	CIELAB L* color coordinate	844	53.9	4.83
Red	CIELAB a* color coordinate	846	17.3	1.90
Yellow	CIELAB b* color coordinate	845	9.59	1.92
Fat	HGP Back-fat thickness, mm	846	22.0	5.69
Lean	HGP Back-lean thickness, mm	844	40.6	6.69

[a]Measurements are in *M. Longissimus*, except for pH-s, Fat and Lean

Non-genetic influences

Meat quality traits can be largely influenced by genetic background and by environment. Well known environmental effects are transport conditions, leading often to large effects of slaughter day- or week (Cameron, 1990; Hovenier et al., 1992). Data analyzed here were collected on 26 different slaughter-days. In the described data, also an effect of breeding company where the crossbred was produced (5 levels) could be expected. This effect could have a partial genetic background, as the maternal grand-dam of the crossbreds was company specific, and may additionally have a non-genetic basis, e.g. in housing or feeding conditions at the different companies. Because semen of F_1 fathers was exchanged between companies, the genetic basis of a company effect could be separated from non-genetic sources. The design of the experiment was such that also possible effects of breeding company could be separated from effects of slaughter days by slaughtering animals from at least two companies on most of the days. Besides slaughter day and breeding company, sex of the animal (measurements were made on females and on intact males) and its carcass weight were recorded as they may have non-genetic influences on the recorded traits. Significance of these non-genetic effects was investigated by use of a fixed linear model (SAS-GLM, SAS Institute INC, 1988) fitting slaughter day, breeding company, sex and carcass weight simultaneously. Each variable was found to have a significant effect ($P<0.01$) on at least several traits; slaughter day was significant for all traits. In further genetic analyses, all non-genetic effects considered were used, thus maintaining for simplicity the same non-genetic effects for analysis of each trait.

Methods

Statistical model

A model was used with non-genetic effects of slaughter-day, breeding company, sex and carcass weight, and genetic effects of polygenic background genes and a single gene. Polygenic effects were modelled to be strictly additive. The model for the single locus assumed an autosomal biallelic locus with Mendelian transmission probabilities. A possible dominance effect at the single locus was allowed for. Estimation of polygenic variance in the described data will be based on variation between F_2 families, and this will not include possible segregation variance at polygenic loci.

Segregation variance refers to the increase of genetic variance that can arise in the F_2 due to allele frequency differences in founder lines (e.g., Lande, 1981). However, assuming the polygenic loci to be large in number and assuming no gene with large effect to be present among the polygenes, segregation variance at polygenes will be negligible. At the single locus, segregation variance was accounted for by modelling of different allele frequencies for the founder groups. The statistical model to describe phenotypic observations on F_2 crossbreds for each trait **y** is :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e} \qquad (1)$$

In (1), β is a vector of fixed non-genetic effects and **X** is a design-covariate matrix containing 0/1 dummy variables relating effects of slaughter day, breeding company and sex to observations and containing a column with carcass weights of those individuals with observations in **y**. Vectors **u** and **Wm** contain genetic effects of all individuals in the pedigree considered, which here included genetic effects of F_2 crossbreds, their parents and grandparents. Genetic effects are seperated in polygenic effects in **u** and single-gene effects in **Wm**. Matrix **Z** is an incidence matrix relating the genetic effects to observation in **y**; **Z** contains empty columns for individuals without an observation. Vector **e** contains random errors. Single-gene effects are expressed using **W**, a four-column matrix with 0/1 variables to indicate genotypes of individuals, and the vector $\mathbf{m}=(-a, d, d, a)'$ which contains the genotypic values. Four genotypes are considered here for notational convenience only; in computations, three genotypes are considered, not distinguishing between the two heterozygotes. In **W**, the four columns correspond to the possible genotypes denoted as A_LA_L, A_LA_H, A_HA_L and A_HA_H. Allele A_L, with 'L' of 'Low', is defined as the allele which decreases the values of phenotypic measurements in **y**; A_H, with 'H' of 'High', is defined as the allele which increases the values of phenotypic measurements in **y**. Alleles are defined in this manner, because 'Low' and 'High' are unique attributes that can be assigned to alleles. In **m**, a and d are referred to as the additive and dominance effect at the single locus, where a is positive, so that definition of the 'Low' and 'High' attributes is consistent. Actual computations were based on non-uniquely defined alleles A_1 and A_2 with unrestricted a; uniquely defined alleles A_L and A_H with $a \geq 0$ were obtained as a transformation (see Appendix).

Above, distinction between the two heterozygotes is made to allow for a flexible notation of pedigree genotype probabilities as follows: $\Pr(A_eA_f) = p_e\, p_f$ for Meishan founder animals, $\Pr(A_eA_f) = r_e\, r_f$ for Dutch founder animals, with $e, f \in \{L, H\}$ and where p_L and p_H ($p_L+p_H=1$) are the frequency of alleles A_L and A_H in Meishan founders and r_L and r_H ($r_L+r_H=1$) are the frequency of alleles A_L and A_H in Dutch founders; $\Pr(A_eA_f) = \tau_{e,gh}\tau_{f,g^*h^*}$, for all non founder individuals, with $e, f, g, h, g^*, h^* \in \{L, H\}$, and where A_gA_h and $A_{g^*}A_{h^*}$ are the genotypes of the sire and dam of the individual considered, $\tau_{L,gh}$ is the transmission probability for genotype A_gA_h to transmit an A_L allele and $\tau_{H,gh}=1-\tau_{L,gh}$ is the corresponding probability to transmit the A_H allele. With Mendelian inheritance $\tau_{L,LL}=1$, $\tau_{L,LH}=\tau_{L,HL}=½$ and $\tau_{L,HH}=0$. Distributional assumptions for **e** are specified as $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ and for **u** are specified as $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where **A** is the numerator of the relationship matrix. Statistical inference was based on a Bayesian approach. Specification of the statistical model for the Bayesian approach is completed by specifying use of uniform prior distributions for non-genetic effects, variance components, effects at the single locus and allele frequencies. These prior distributions were defined on $\langle -\infty, \infty \rangle$ for the non-genetic effects and effects at the single locus, on $\langle 0, \infty \rangle$ for the variance components and on $[0,1]$ for the allele frequencies. Variances were assumed a-priori positive, which was computationally implemented by defining the prior on $[10^{-12}, \infty\rangle$. The prior defined on $\langle -\infty, \infty \rangle$ for the additive effect at the single locus corresponds to the situation in the actual computations as shortly described above and as exemplified in the Appendix. With defined distributional assumptions, the complete set of parameters for model (1) then was $\theta_1=(\beta, \mathbf{u}, \mathbf{W}, \sigma_e^2, \sigma_u^2, a, d, p_L, r_L)$. In this set of parameters, the variance components, effects at the single locus and the allele frequencies are referred to as the (genetic) hyper-parameters.

As well as a model postulating mixed inheritance, a model postulating pure polygenic inheritance was used by suppressing the term for single gene effects in (1), leading to the model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \qquad (2)$$

with all specifications equal to those of model (1) and with parameters $\theta_2=(\beta, \mathbf{u}, \sigma_e^2, \sigma_u^2)$. The pure polygenic model was used to supply an overall quantification of genetic

variance for the traits analysed. As a third model, a mixed model of inheritance with a restriction on the degree of dominance was used. This model is the same as model (1) except that genotypic values are defined by $\mathbf{m}'=(-a, ca, ca, a)$, where c defines the imposed degree of dominance. This model was used to impose complete dominance of the A_L allele ($c=-1$) or complete dominance of the A_H allele ($c=1$). In this model also one of the allele frequencies was assumed known, arbitrarily taken to be p_L, such that the set of parameters for this model was $\theta_3=(\beta, \mathbf{u}, \mathbf{W}, \sigma_e^2, \sigma_u^2, a, r_L)$.

Gibbs sampling

Bayesian marginal posterior distributions of model parameters were obtained using Gibbs sampling. In such an approach, a Markov chain is constructed which is known to have a stationary distribution equal to the joint posterior distribution of all model parameters, here all parameters in θ_1 for the main model (1), θ_2 for the polygenic model, or θ_3 for the model with restricted degree of dominance. From such a Markov chain samples of marginal posterior distributions of model parameters and of functions of model parameters were obtained. The construction of such a Markov chain was described by Janss et al. (1995) and implemented in a software package (see Appendix). This implementation includes blocked sampling of genotypes of each sire with those of its final progeny and similar blocked sampling of polygenic effects of each sire with those of its final progeny (Janss et al., 1995). This blocked sampling facilitates convergence of the Gibbs sampler when analyzing typical animal breeding data sets with relatively large progeny groups. Additionally, a model-relaxation technique (Sheehan and Thomas, 1993) was applied to further improve convergence of the Gibbs sampler for the single gene component, by relaxation of transmission probabilities. Such a relaxation uses $\tau_{L,LL}=1-p_{nmt}$ and $\tau_{L,HH}=p_{nmt}$, where p_{nmt} is a small probability for 'non-Mendelian transmission'. Inference about the strict Mendelian model of interest is made by using from the constructed markov chains only those samples where the genotype configuration was Mendelian. Sheehan and Thomas (1993) showed that the rate at which Mendelian samples randomly appear in a relaxed chain, equals the likelihood ratio between the strict Mendelian model and the relaxed model, dependent therefore not only on the parameter p_{nmt}, but also on the data. To achieve a certain rate of Mendelian samples, some trial runs are required to determine, for each data set analyzed, a suitable value of p_{nmt}. In order for the relaxation technique to have

a reasonable impact on convergence, relaxation may be relatively strong, leading to a low rate of Mendelian samples in the relaxed chains: in the analyses performed here, we aimed at a rate of Mendelian samples of 1 to 10%. Gibbs chains computed for inferences in the mixed inheritance model were started as a 'hot' chain (in the terminology of Lin et al., 1993), using initially p_{nmt}=0.5, which defines a non-genetic transmission model by allowing random transmission of alleles. Subsequently, such a hot chain was annealed by slowly reducing p_{nmt} to near zero, which restricts movement of the chain to the Mendelian and near-Mendelian space. In the construction of the Gibbs sampler, sampling of random realizations for various types of distributions was based, directly or indirectly, on the uniform random number generator RAN2 (Press et al., 1992). For construction and sampling details see Janss et al. (1995) and the Appendix.

Convergence of the Gibbs sampler was judged for the hyper-parameters by comparison of samples from replicated chains by analysis-of-variance (ANOVA), testing for a significant chain effect. In this approach, Gibbs chains are run that are sufficiently long to obtain a number of independent samples from each chain. This, then, allows to test for equality of the within- and between chain variances with a standard ANOVA F-test. Significant differences between chains are considered an indication of (practical) reducibility, in which case Gibbs sampling theory (Geman and Geman 1984; Gelfand and Smith 1990) does not hold and the samples generated are not from the correct marginal distributions. In this case, the Gibbs sampler is said not to have converged. Significance of differences was assumed when the F-statistic exceeded the 1% significance level. The ANOVA requires independence of the samples, hence only a number of states from each chain, sufficiently spaced, are used. Determination of a suitable spacing yielding virtually serially independent samples was done according to the procedure exemplified by Janss et al. (1995), mainly based on Raftery and Lewis (1992). The same spacing was used for all parameters of interest. The ANOVA also acts as a post-check on the presumed independence of the Gibbs samples: when the spacing between samples is not sufficient and the assumption of independence does not hold, computed F-statistics will be inflated and chain-effects could be found significant. Insufficient spacing can be verified by increasing the spacing between samples used in the ANOVA by running longer Gibbs chains that keep the same number of samples in the ANOVA in order not to affect power to detect differences between the chains.

Statistical inference

Generated independent samples used in the ANOVA as described above for assessing convergence, are subsequently used for statistical inferences. From samples of marginal posterior distreibutions, non-parametric density estimates of posteriors were made in the form of average shifted histograms (Scott, 1992). Such a graph provides a more general and broad inference, than a specific point- and/or interval estimator. For parameters with natural boundaries on their parameter space, density estimates were smoothed up to the bound(s) of the parameter space by a reflection boundary technique (Scott, 1992). Secondly, samples from marginal posterior distributions were used to compute estimates of mean and standard deviation of the posterior distributions, which were estimated by mean and standard deviation of the Gibbs samples. These estimates converge stochastically, with increasing number of Gibbs samples generated, to the true mean and standard deviation of the marginal posterior distributions of the respective parameters (Smith and Roberts, 1993). The posterior mean was chosen used as a point estimator, falling in the class of APE (A-Posteriori Expectation) estimators. Such APE estimators have the general property of minimizing quadratic posterior loss. The higher marginalized Bayesian estimators, compared to classical maximum likelihood estimators, are expected to have the same asymptotic properties and superior non-asymptotic properties from a Bayesian viewpoint (Gianola and Foulley, 1990). In analogy to frequentist approaches for statistical inference, the posterior standard deviation can be interpreted as a standard error of the parameter estimate, but is not in general equal to a frequentist standard error.

Of primary interest for statistical inferences were the variance components σ_u^2 for polygenic variance and σ_w^2 for the variance explained by the single gene. These two genetic variances were used to judge significance of the genetic model and, in particular, to judge significance of the single gene component in the model. The variance of single gene effects σ_w^2 was computed as a function of effects at the single locus and of allele frequencies in each Gibbs sample as $2pq(a+d(q-p))^2+(2pqd)^2$ (Falconer, 1989), where a and d are as defined previously, p is the frequency of the favorable A_H allele, and $q=1-p$. The variance of single gene effects was computed to represent the variance in the F_2 generation, by using $p=(p_H+r_H)/2$ in this formula. Non-significance of a variance component (shortly σ^2) was empirically shown to lead to a posterior distribution with global mode at $\sigma^2=0$ (Janss et al., 1995). Significance of a

variance component shows a global mode for $\sigma^2>0$, which may still be accompanied by a local mode or a non-zero density at $\sigma^2=0$. For variance of single gene effects, σ_w^2, a penalty is used in order to reduce the error of falsely accepting presence of a single gene, by considering σ_w^2 to be significant only when a global mode for $\sigma_w^2>0$ has a density 20-fold larger than the density at $\sigma_w^2=0$, corresponding to a 5% significance level. In this manner the usual conservatism is applied, accepting presence of a single gene only when abundant evidence is available, or else not rejecting the null hypothesis of polygenic inheritance. The mode(s) and density ratios were determined from the non-parametric density estimates. Once a mixed inheritance model is found likely, further inferences focussed on the effects at the single locus and on allele frequencies. Allele frequencies are not uniquely identifiable with the available data on F_2 crossbreds only. For instance, it is not possible to distinguish between a case with $p_L=1$ and $r_L=0$ (origin of A_L from Meishan founders) and a case with $p_L=0$ and $r_L=1$ (origin of A_L from Dutch founders). The available data only allows unique determination of the genotype frequencies in F_1 parents, yielding estimable functions of allele frequencies being $p_L r_L$, $p_L r_H+p_H r_L$ and $p_H r_H$, representing the frequencies of $A_L A_L$, $A_L A_H$ & $A_H A_L$, and $A_H A_H$ genotypes in the F_1 generation.

From the Gibbs chains, also marginal posterior distributions of individual genotypes, referred to as genotype probabilities, were estimated from the frequency counts of the genotypes sampled in the Markov chain. Genotype probabilities could be estimated from Gibbs chains sampling (β, **u**, **W**) conditional on some point estimates for the hyper-parameters. Here, however, genotype probabilities were estimated from Gibbs chains sampling all model parameters, which will supply estimates of genotype probabilities not conditional on any point estimates for hyper-parameters and where uncertainty from estimation of hyper-parameters will be included. Estimates of genotype probabilities were used to study whether various traits found to be influenced by a single gene, actually could be influenced by the same gene, as follows: the interval [0,1] in which genotype probabilities fall, was discretized into k=1, ..., K smaller intervals; for a group of n individuals, the number of individuals was counted with a genotype probability for a first trait falling in interval k and with a genotype probability for a second trait falling in interval k^*, the count being denoted c_{kk^*}, for k=1, ..., K and k^*=1, ..., K; the quantities c_{kk^*} were collected in a K-by-K table in which association between the genotypes was tested by a chi-square test for association

with $(K\text{-}1)(K\text{-}1)$ degrees of freedom; a significant association was considered to be an indication that two traits, found to be influenced by single genes, could actually well be influenced by the same gene affecting both traits. This procedure works best with large K, but the choice for K is bounded by the number of individuals n because counts c_{kk*} must be reasonably large in order for the approximate chi-square test to be valid. Choice of the interval cut-points is arbitrarily and also can be chosen such that all counts c_{kk*} are reasonably large. Estimated genotype probabilities preferably should have high and similar accuracy.

Results

Polygenic model

For inference in the polygenic model, for each trait a trial Gibbs chain of 10000 cycles was run. From these chains, it was determined that virtually independent samples could be obtained using a spacing of 800 cycles. Subsequently, for inferences, for each trait 5 Gibbs chains of 40000 cycles were run, obtaining 50 independent Gibbs samples per chain and 250 samples in total per trait. Using an estimate of phenotypic variance from a model fitting non-genetic effects only, starting values for polygenic variances in the five replicated chains were chosen as 10, 20, 30, 40 and 50% of this estimate of phenotypic variance, and with error variance equal to the remainder. Starting values for non-genetic effects and polygenic effects were zero. A burn-in period of 1600 cycles was used to allow the Gibbs chains to reach equilibrium. Estimated posterior means and posterior standard deviations for variance components and heritability are in Table 2. Features of the posterior distribution for heritability (h^2) were obtained by computing from each Gibbs sample of variance components the corresponding value of h^2 and subsequently using these values to summarize the posterior distribution of h^2. Tests for convergence of the Gibbs sampler by comparison of multiple chain output using ANOVA on the independent samples, showed no significant differences between replicated chains for all parameters in Table 2, demonstrating convergence of the Gibbs sampler. Results indicate general existence of genetic variation for the various traits. Absolute values of heritability are generally low: Hovenier et al. (1993) indicate in a review average heritabilities for water-holding capacity traits (Cook, Drip) and pH of 0.20, for Shear and color traits of 0.30, and for Imfat of 0.50. Heritabilities in the data

analyzed are 0.10 to 0.20 below these average literature values.

Table 2 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for error variance (σ_e^2), polygenic variance (σ_u^2) and heritability (h^2) in the polygenic model, based on a total of 250 independent Gibbs samples from 5 replicated chains.

	σ_e^2		σ_u^2		h^2	
Trait	mpm	mpsd	mpm	mpsd	mpm	mpsd
Cook	7.31	0.503	1.07	0.578	0.126	0.064
Drip	1.71	0.134	0.272	0.127	0.136	0.061
Shear	51.2	2.93	2.45	2.07	0.045	0.038
Imfat	0.429	0.0544	0.258	0.0757	0.372	0.094
pH[a]	4.39	0.348	0.567	0.313	0.114	0.061
pH-s[a]	5.06	0.385	0.551	0.354	0.098	0.060
Light	15.8	1.35	3.61	1.47	0.184	0.071
Red	2.63	0.200	0.431	0.198	0.140	0.061
Yellow	2.60	0.190	0.294	0.163	0.101	0.054
Fat	19.4	1.79	6.26	2.36	0.241	0.081
Lean	33.4	2.76	8.08	3.25	0.193	0.073

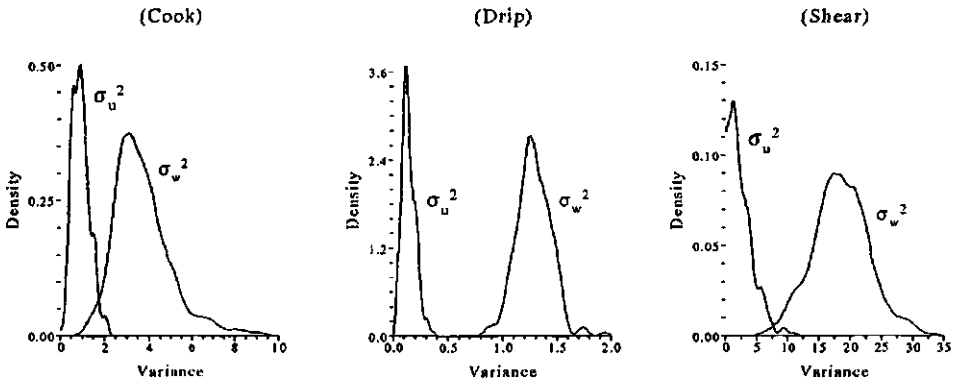
[a]For pH and pH-s variance components are in hundreds

Mixed inheritance model

For the mixed inheritance model, trial Gibbs chains were run to determine suitable values of the relaxation parameter and the spacing to be used between Gibbs cycles to yield independent Gibbs samples. Such trial runs showed that for some traits, too strong relaxation could lead to Gibbs chains settling at non-Mendelian states, without further realisations of Mendelian samples. Therefore, a variety of approaches was used to obtain Mendelian samples for the mixed inheritance model: for 'well behaved' traits, which were Cook, Imfat, Fat, Lean, White and Yellow, a relatively large relaxation was used leading to a low rate of Mendelian samples of 1 to 2% and virtually independent Mendelian samples were obtained by taking every 20th Mendelian sample occurring;

for the remaining 'less behaved' traits, a smaller relaxation was used leading to a larger rate of Mendelian samples of 5 to 10% and virtually independent Mendelian samples were obtained by taking, varying between traits, every 50th to 180th Mendelian sample occurring. In this manner, one Mendelian sample was obtained about every 1000 to 2000 cycles for all traits analyzed. Values for the relaxation parameter p_{nmt} to obtain the mentioned rates of non-Mendelian samples ranged from 2.0×10^{-3} to 6.5×10^{-3}. For inferences, 5 chains per trait were run, obtaining 50 independent Mendelian samples per chain and 250 in total per trait. All chains started with the described annealing of a hot chain, which was taken over 1000 cycles, and was followed by a burn-in of another 1000 cycles. Gibbs chains were started using the following parameter values: zeros for non-genetic effects, polygenic effects and effects at the single locus, heterozygotes for all genotypes, posterior mean estimates from the polygenic model (Table 2) for variance components and 0.5 for allele frequencies. It was observed that the approach of annealing a hot chain was quite effective in letting the Gibbs sampler converge to its equilibrium, even from such a crude starting point.

Estimated marginal posterior distributions for the two genetic variance components in the mixed inheritance model are in Figure 2 and estimated marginal posterior means and standard deviations of all three variance components are in Table 3. Analysis of differences between replicated chains indicated two traits where lack of convergence of the Gibbs sampler was diagnosed: pH-s and Fat, for parameters residual variance and polygenic variance. Convergence was found for estimation of single-gene variances for all traits.



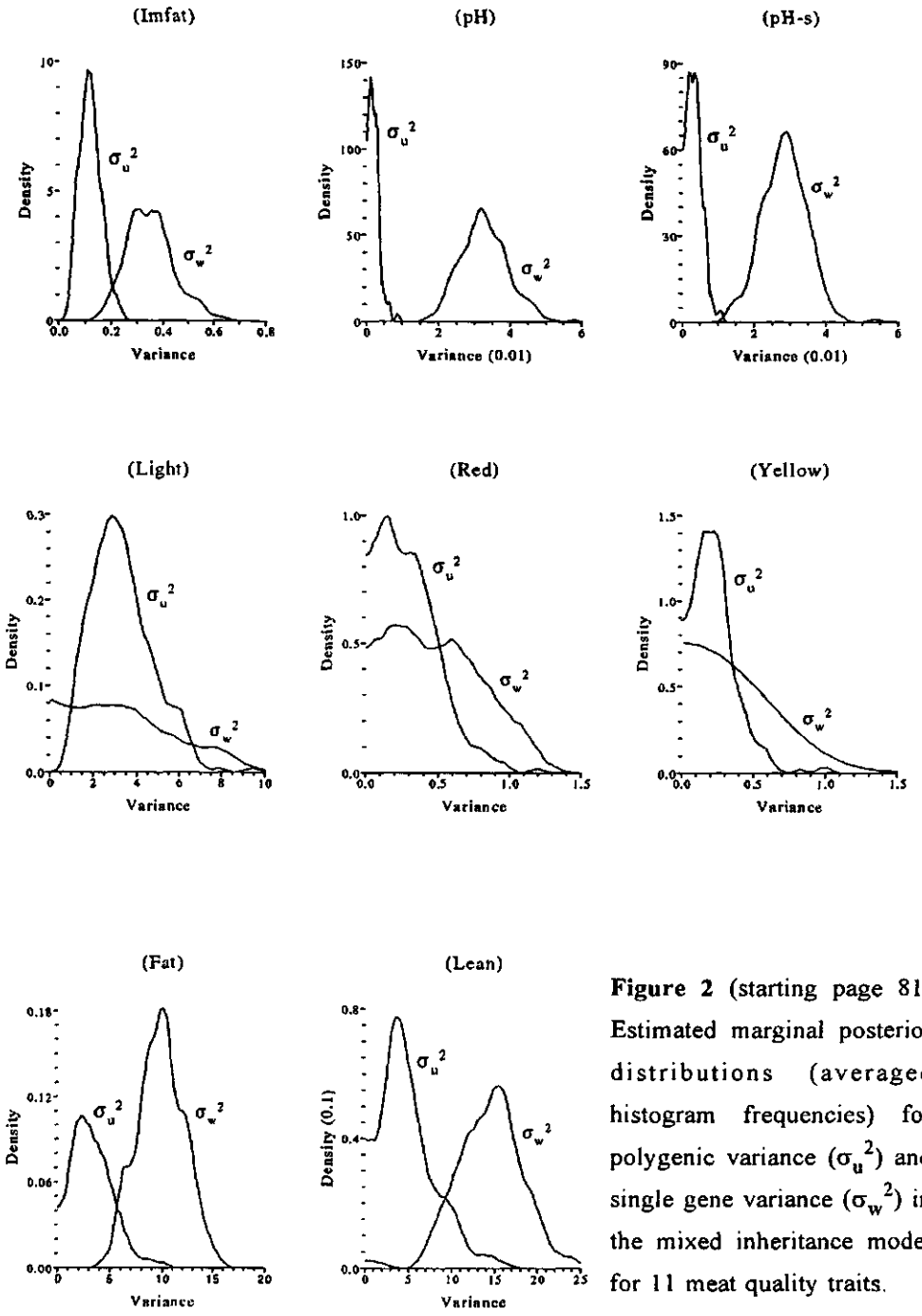


Figure 2 (starting page 81) Estimated marginal posterior distributions (averaged histogram frequencies) for polygenic variance (σ_u^2) and single gene variance (σ_w^2) in the mixed inheritance model for 11 meat quality traits.

Table 3 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for error variance (σ_e^2), polygenic variance (σ_u^2) and single gene variance (σ_w^2) in the mixed inheritance model, based on a total of 250 independent Gibbs samples from 5 replicated chains.

	σ_e^2		σ_u^2		σ_w^2	
Trait	mpm	mpsd	mpm	mpsd	mpm	mpsd
Cook	5.03	0.417	0.931	0.399	3.73	1.30
Drip	0.582	0.0697	0.142	0.0627	1.29	0.158
Shear	35.2	3.32	2.55	1.98	18.7	4.79
Imfat	0.260	0.0310	0.120	0.0416	0.351	0.0914
pH[a]	2.03	0.165	0.220	0.156	3.28	0.663
pH-s[a]	2.72**	0.377	0.344**	0.216	2.85	0.623
Light	12.9	2.61	3.34	1.42	3.59	2.46
Red	2.29	0.304	0.302	0.209	0.501	0.314
Yellow	2.29	0.274	0.233	0.156	0.436	0.797
Fat	12.6**	1.82	3.34**	1.94	9.92	2.24
Lean	21.7	3.97	5.01	3.19	14.4	3.81

[a]For pH and pH-s variance components are in hundreds

**Significant differences between replicated chains ($P<0.01$)

Based on the marginal posterior distributions depicted in Figure 2, traits were grouped according to significance of the two genetic variances as follows: (1) traits Cook, Drip and Imfat showing significant influence of a single gene in presence of additional significant polygenic variance; (2) traits Shear, pH, pH-s, Fat and Lean, showing significant influence of a single gene, but with low polygenic variances showing non-negligible densities at $\sigma_u^2=0$; (3) color traits Light, Red and Yellow, showing non-significant single gene variance. Group (2) includes the two traits with no convergence of the Gibbs sampler for some parameters. The total of genetic variance inferred in the mixed inheritance model appears larger than genetic variance inferred in the pure polygenic model. This can be explained by the segregation variance at the single locus, which will be attributed to error variance in the polygenic model, but can be included

in genetic variance in the mixed inheritance model.

Table 4 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for additive effect (a) and dominance effect (d) at the single locus, and estimated 95% highest posterior density (HPD) regions for their difference in the mixed inheritance model, based on a total of 250 independent Gibbs samples from 5 replicated chains, shown for traits with significant contributions of single gene variance.

	a		d		$a-\|d\|$ 95% HPD region	
Trait	mpm	mpsd	mpm	mpsd	from	to
Cook	4.67	0.436	4.64	0.562	-1.08	1.29
Drip	1.40	0.0886	-1.53	0.175	-0.631	0.323
Shear	5.72	0.844	-8.54	1.74	-5.99	1.21
Imfat	1.14	0.0767	-1.09	0.128	-0.239	0.359
pH	0.319	0.0170	-0.313	0.0241	-0.0475	0.0651
pH-s	0.233	0.0202	-0.269**	0.0493	-0.139	0.0831
Fat	4.39	0.584	-3.85**	1.07	-2.32	3.38
Lean	4.39	1.05	-4.16	1.67	-4.83	4.67

** Significant differences between replicated chains ($P<0.01$)

The eight traits in groups (1) and (2) described above were considered for further investigations on estimates of effects at the single locus (Table 4) and F_1 genotype frequencies (Table 5). Estimates for additive effect a and dominance effect d at the single locus indicated that d was likely to be of the same absolute value than a: Table 4 shows the estimated 95% highest posterior density (HPD) regions for the difference $a-|d|$, which in all cases included the value zero. HPD regions were obtained by computing, from each Gibbs sample of a and d, the difference $a-|d|$, subsequently making a non-parametric density estimate for this difference and obtaining from this density estimate the left- and right 2.5% quantiles. Hence, a single gene with complete dominance for one of its alleles was inferred for all traits listed in Table 4.

Reservations on this conclusion should be made for pH-s and Fat, where non-convergence of the Gibbs sampler was diagnosed for estimation of the dominance effect at the single locus. Estimates of genotype frequencies in the F_1 (Table 5) indicated absence of the homozygote recessive genotype in the F_1 parents. The frequency of homozygote recessives in F_1 is $p_L r_L$ for Cook (with d positive) and $p_H r_H$ for other traits (with d negative). Posterior distributions for $p_L r_L$ for Cook and $p_H r_H$ for other traits are in Figure 3 for the eight traits considered, and show global modes at zero for all these frequencies. Absence of the homozygote recessive genotype in F_1 indicates that the recessive allele must be absent in one of the founder lines, although the observations on F_2 used here, do not allow one to determine the founder line. Convergence of the Gibbs sampler for estimation of the genotype frequency of the homozygote recessives was confirmed for all traits.

Table 5 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for estimable functions of allele frequencies p_L and r_L in the mixed inheritance model, based on a total of 250 independent Gibbs samples from 5 replicated chains, shown for traits with significant contributions of single gene variance.

	$p_L r_L$ [a]		$p_L r_H + p_H r_L$ [a]		$p_H r_H$ [a]	
Trait	mpm	mpsd	mpm	mpsd	mpm	mpsd
Cook	0.0189[b]	0.0166[b]	0.379	0.0631	0.602	0.0683
Drip	0.161	0.0575	0.806	0.0657	0.0325[b]	0.0309[b]
Shear	0.381	0.105	0.568	0.106	0.0515[b]	0.0419[b]
Imfat	0.495	0.0666	0.471	0.0639	0.0334[b]	0.0265[b]
pH	0.427	0.0601	0.551	0.0606	0.0218[b]	0.0208[b]
pH-s	0.294**	0.111	0.673**	0.115	0.0337[b]	0.0305[b]
Fat	0.254	0.101	0.702	0.108	0.0431[b]	0.0409[b]
Lean	0.129	0.0905	0.778	0.113	0.0937[b]	0.0749[b]

[a] Estimable functions of allele frequencies represent genotype frequencies in F_1

[b] Global mode at zero

** Significant differences between replicated chains (P<0.01)

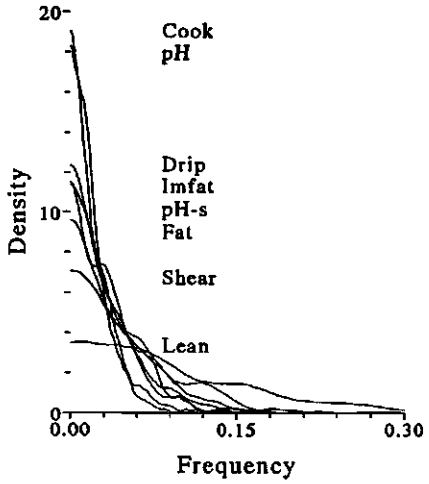


Figure 3 Estimated marginal posterior distributions (averaged histogram frequencies) for the frequency of the double recessive genotype in F_1 parents for 8 meat quality traits; vertical positions of trait names indicate starts of the graphs at the left boudary for the respective traits.

Restricted model

Results presented in Tables 4 and 5 led to the conclusion that a number of traits might be influenced by a recessive gene, absent in one of the founder lines. However, for traits pH-s and Fat, convergence of the Gibbs sampler was not observed, which did not allow one to draw definite conclusions. Also, the finding of several traits being influenced by a single gene, brings up the interesting hypothesis of whether all or some traits might actually be influenced by the same gene, acting pleiotropically. By using a restricted model, which assumed complete dominance and absence of the recessive allele in one of the founder lines, it was attempted to improve inference for pH-s and Fat, and to obtain a more accurate estimation of genotype probabilities compared to an unrestricted model. It was not considered to obtain genotype probabilities conditional on some point estimates for all hyper-parameters because we thought that such an approach could endanger attempts to elucidate pleiotropic effects of the genes.

Restrictions to impose complete dominance and absence of the recessive allele in one of the founder lines were applied as follows (cf Tables 4 and 5): for Cook, $d=a$, $p_L=0$ and $p_H=1$; for all other traits, $d=-a$, $p_L=1$ and $p_H=0$. Application of the restriction to allele frequency p_L is arbitrary from a modelling viewpoint, as only the genotype frequencies in F_1 are uniquely estimable, but was thought to be better from a

computing viewpoint: fixing p_L, frequency in the (smaller) paternal founder line, and estimating r_L, frequency in the (larger) maternal founder line was thought to supply a more stable Gibbs chain. Gibbs chains were started using the following parameter values : zeros for non-genetic effects and polygenic effects, heterozygotes for all genotypes, posterior mean estimates in the mixed inheritance model (Table 3, Table 4) for variance components and additive effect at the sintgle locus a, and 0.5 for allele frequency r_L. For inference in this restricted model, an initial phase where a hot chain was annealed was omitted: starting with plausible values for effects at the single locus, the Gibbs sampler converged equally well without such phase. Other details were the same as for the full mixed inheritance model, generating a total of 250 Gibbs samples per trait, from 5 replicated chains.

Table 6 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for additive effect at the single locus (a) and allele frequency for allele A_H in one of the founder lines (r_H) in a restricted mixed inheritance model[a], based on a total of 250 independent Gibbs samples from 5 replicated chains, shown for traits with significant contributions of single gene variance.

	a		r_H	
Trait	mpm	mpsd	mpm	mpsd
Cook	4.64	0.368	0.614	0.0652
Drip	1.44	0.0460	0.842	0.0590
Imfat	1.12	0.0665	0.504	0.0580
Shear	6.29	0.720	0.676	0.1000
pH [b]	0.316	0.0153	0.565	0.0606
pH-s [b]	0.240	0.0154	0.734	0.0905
Fat	4.18	0.365	0.782	0.0852
Lean	4.32	0.475	0.91** [c]	0.0752 [c]

[a] Restrictions: for Cook, $d=a$, and allele A_H is dominant; for other traits, $d=-a$, and allele A_H is recessive; the recessive allele was forced absent in the other founder line

[b] For pH and pH-s variance components are in hundreds

[c] global mode at $r_H=1$

** Significant differences between replicated chains ($P<0.01$)

Estimates of variance components with the restricted model confirmed estimates for the full model (not shown). Estimates for the effect at the single locus and frequency of the favorable A_H allele are in Table 6. Due to the restriction in the genetic components of the model, estimates of genetic variances were slightly lower, and estimates of error variance were slightly higher. The Gibbs sampler showed good convergence for the estimation of variance components for all traits, except for the estimation of residual variance for Lean and results showed significant contributions of single gene variance for all traits. For Lean, influence of a single gene was rejected, firstly because of poor convergence of the Gibbs sampler, and secondly, because for Lean frequency of the recessive allele, which was forced to zero in one of the founder lines, was estimated very close to one in the other founder line (Table 6). In such case, inferred significant single gene variance could be caused by a general non-normality in the data (see the discussion section). For the seven remaining traits, Cook, Drip, Shear, Imfat, pH, pH-s and Fat, influence of a single gene is considered very likely.

From the Gibbs chains used to estimate genetic parameters in the reduced model (partly in Table 6), estimates of genotype probabilities were obtained as well. Using genotype probabilities inferred for different traits, association tests were carried out as described in the Methods section to obtain indications of whether traits presumed to be affected by a single gene could actually be affected by the same gene. Associations were studied between seven traits, which were all traits in Table 6 except Lean. Tests for associations were based on estimated probabilities of female F_1 parents to be heterozygote. Use of the female parents supplied a reasonable compromise between requiring a large number of individuals and requiring individuals with precise estimates. Using female parents, 251 individuals were available and the genotype probabilities were counted in three intervals: those falling between 0 and 0.65, between 0.65 and 0.80 and between 0.80 and 1. These intervals supplied a good distribution of numbers of probabilities falling in all combinations of intervals for pairs of traits and resulted in an appropriate use of the chi-square approximation to test for association for all combinations of traits except one. Using a discretization into three intervals, tests for associations were based on a 3-by-3 table and test-statistics followed a χ^2 distribution with four degrees of freedom under the null-hypothesis. Significant associations were considered as those where the test-statistic exceeded the 1% level of significance.

Table 7 Chi-square tests statistics for test of association between inferred genotypes for traits influenced by single genes [a]

Trait	pH	pH-s	Fat	Imfat	Shear	Drip
Cook	57.7**	24.1**	6.14	3.20	5.83	6.72
pH	-	34.7**	5.44	10.1[b]	1.94	6.61
pH-s		-	13.6**	5.09	5.05	5.14
Fat			-	1.98	6.91	3.20
Imfat				-	20.7**	3.83
Shear					-	21.3**

[a] Probabilities (P_c) were estimated for 251 female F_1 parents to carry the recessive allele. Estimation was done from Gibbs chains for inference in the restricted model assuming complete dominance and absence of the recessive allele in one of the founder lines. Test for association was based on transformation of estimated probabilities to a three-class variable, indicating whether $P_c < 0.65$, $0.65 < P_c < 0.80$, or $P_c > 0.80$.

[b] Chi-square approximation not good

** Significant association ($P<0.01$)

Test statistics for associations between genotypes for the seven traits considered are in Table 7. Traits in this table are ordered corresponding to a suggested division into two groups: Cook, pH, pH-s and Fat as a first group and Imfat, Shear and Drip as a second group. Test statistics for association of genotypes between pairs of traits across these groups are non-significant, whereas associations between pairs of traits within groups often are significant. In the first group, clear associations were found between combinations of Cook, pH and pH-s, which strongly suggest that these three traits are influenced by the same gene. Also Fat may be influenced by this Cook/pH gene, but here the situation is not clear: Fat was associated with pH-s, but not with any of the other traits in this group. In the second group, the situation also is not fully clear: here Imfat is associated with Shear, Shear with Drip, but Imfat is not associated with Drip.

Discussion

Validation

To argue for presence of a single affecting the traits considered, one must demonstrate the presence of typical data characteristics, and argue that these characteristics are due to segregation of a single gene, rather than some other mechanism. The data characteristics that are typical for traits influenced by a single gene are heterogeneous within family variation for the traits measured and general- or family specific skewness and/or kurtosis in the distribution of the trait. These characteristics, when they can be observed in certain families but not in others, are fairly robust identifiers for presence of a single gene. However, non family specific data characteristics such as general skewness or kurtosis, are much less robust identifiers (Le Roy and Elsen, 1992). For the trait Lean, the recessive allele was inferred to be fixed in the founder line where it originated from. This implies that all F_1 parents were inferred to be heterozygous and, apparently, the trait did not show differences in within family variation or family specific skewnesses. This led us to reject influence of a single gene on this trait: the general skewness could be the result of a segregating single gene, but not necessarily, so that convincing evidence is not supplied.

For the remaining traits, Cook, pH, pH-s, Fat, Imfat, Shear and Drip, the recessive allele was inferred to segregate in the founder line where it originated from. This implies presence of two genotypes in F_1 parents: parents which carry the recessive allele, and parents which do not carry the recessive allele. When mating various combinations of such parents, heterogenous within-family variation and family specific skewnesses should arise. Such effects were indeed present for the traits analyzed (Table 8): families of carrier fathers showed markedly higher variance in the measured traits than families of non-carrier fathers. This heterogeneity in within family variances is not easily explained by effects other than the segregation of a single gene because design and analysis of the experiment excluded confounding with any common factors known to cause possible heterogeneity in variance. In particular, animals were raised at the same time in all locations/companies, which were geographically not widely spread, slaughtered in one slaughter house and a confounding between fathers/families and locations/companies was eliminated by formation of the described central pool of fathers, so that each resulting father-family was a mix of individuals from different

locations/companies. This all supplies sufficient evidence in favor of the heterogeneity in family variances for Cook, pH, pH-s, Fat, Imfat, Shear and Drip to be of genetic origin, caused by segregation of a single gene.

Table 8 Number of observations (N), mean and standard deviation of raw phenotypic measurements for traits inferred to be influenced by single genes, in families of non-carrier fathers, families with dubious status of the father and in families of carrier fathers[a]

	Non-carrier father families			Dubious father families			Carrier father families		
Trait	N	mean	std	N	mean	std	N	mean	std
Cook	322	26.5	2.87	205	26.4	3.12	318	26.2	4.15
Drip	0	-	-	97	2.04	0.943	747	2.78	1.58
Imfat	227	1.65	0.654	260	1.75	0.704	344	2.02	1.05
Shear	62	34.8	6.89	408	36.5	8.51	375	43.7	11.4
pH	162	5.60	0.161	107	5.65	0.229	576	5.68	0.289
pH-s	47	5.71	0.137	149	5.77	0.240	650	5.84	0.322
Fat	26	19.9	4.48	225	20.4	4.76	595	22.7	5.92

[a] Probabilities (P_c) were estimated for F_1 fathers to carry the recessive allele. Estimation was done from Gibbs chains for inference in the restricted model assuming complete dominance and absence of the recessive allele in one of the founder lines. Fathers with $P_c < 0.20$ were considered 'non-carriers', fathers with $P_c > 0.80$ were considered 'carriers', and those remaining were considered 'dubious'. For Drip, non of the fathers was found 'carrier'.

Single genes

By simple association tests between inferred genotypes, indications were obtained whether (groups of) traits actually could be influenced by the same gene. We postulated as a working hypothesis that the effects observed could be caused by two genes: one gene that influences cooking loss, pH and possibly backfat thickness, and a second gene that influences intramuscular fat, shearforce and possibly drip loss. The

presumed first gene is called Meishan Cooking loss gene (*MC*), the second gene is called Meishan Intramuscular fat gene (*MI*). A joint effect of the *MC* gene on cooking loss and pH is physiologically well understandable and estimated effects of the inferred recessive allele to decrease cooking loss and increase pH agree with expectations from a physiological viewpoint. Whether the *MC* gene also influences backfat thickness remains unclear: an association was found between backfat and one of the pH measures, but not with cooking loss and a second pH measure, and physiologically such an association is also not immediately obvious. For the presumed *MI* gene, the situation is less clear: from a physiological viewpoint, higher intramuscular fat could be associated with lower shearforce, indicative of more tender meat, but from the analyses, higher intramuscular fat appeared associated with higher shearforce. For the *MI* gene, a possible relationship with drip loss is also debatable, as such an association was only made via shearforce. Therefore, the working hypothesis of only two genes, *MC* and *MI*, influencing the traits analyzed could well be too restrictive and could require extension, postulating effects of more genes.

The recessive alleles of the inferred single genes were found to originate from one of the founder lines only. This raises the interesting question whether this was the Chinese Meishan founder line or the collection of Western founder lines. In the data, some additional evidence for one or the other hypothesis was available. In the analyses, Western founder animals were treated as a homogeneous group, but in fact these founders consisted of different lines, one from each company (Figure 1). Among F_1 fathers that were carriers of the recessive allele for various traits, descendants from all Western founder lines were present. Using this additional information, it is unlikely that such recessive allele would have been present in all these Western lines, and a more plausible explanation is that these recessive alleles originated from the common Meishan fathers of F_1 fathers.

The above described *MC* gene, affecting pH, superficially might be presumed to be actually the known Halothane gene (Eikelenboom and Minkema, 1974), or the *RN* gene (Le Roy et al., 1990), either of which also affects pH in meat. However, presence of the mutation causing Halothane susceptibility was excluded by molecular typing and presence of the RN^- allele of the *RN* gene is unlikely because this allele is thought to be specific for the Hampshire pig breed. Moreover, effects of the presumed *MC* gene are opposite to those known for the mutation of the Halothane gene and for

the RN^--allele: these two alleles are (partly) dominant and increase cooking loss and decrease pH, whereas the dominant allele of the *MC* gene decreases cooking loss and increases pH.

Bayesian segregation analysis

The secondary goal of this study was to apply a recently developed Bayesian approach for segregation analysis in extensive field data analyses. The approach uses Gibbs sampling for computing marginal posterior distributions and appeared generally feasible. The approach was effective in generating a reasonable number of independent samples from the marginal posterior distributions of parameters, and convergence was found, at least for the variance components, in practically all cases. Gibbs sampling allows use of looped pedigrees, incorporation of many relationships, and circumvents the intrinsic problems (e.g., Hasstedt, 1982) in marginalizing a joint distribution with respect to both discrete parameters (genotypes) and continuous parameters (polygenic effects and others). Because of these advantages, Gibbs sampling can, and has, also been used in maximum likelihood approaches to segregation analysis (e.g., Guo and Thompson, 1992). In combination with the Bayesian approach, Gibbs sampling provides even more flexibility than a maximum likelihood approach, for instance in the estimation of means. Models used in analysis of livestock data often comprise a large number of means; in the analysis presented here 33 means and one regression coefficient were fitted. In a Bayesian approach, these means are straightforwardly included in the Gibbs chains and treated as nuisance parameters, yielding a REML-type approach by accounting for uncertainty originating from the estimation of these fixed effects. Accounting for such uncertainty is known, from linear model applications, to remove bias in estimation of variance components. Apart from flexibility in the model, the Bayesian approach supplies posterior distributions and small-sample 'standard errors' of parameters, while the maximum likelihood approach relies on asymptotic properties.

Conclusions

The primary aim of this study was to investigate whether meat quality traits in a crossed pig population were influenced by single genes. The statistical analyses

presented showed convincingly that seven meat quality traits measured in this population are indeed influenced by single genes, which most likely originate from the Chinese Meishan breed. These genes are different from genes so-far identified to affect meat quality and further study of this population will be worthwhile. Currently the animals are being genotyped for a large number of genetic markers which will enable a linkage analysis to estimate location of the genes. Based on the results of the present study it can be concluded that the material from the F_2 cross is very suited to locate genes affecting meat quality. The results of the linkage analysis will, in particular, be helpful to determine the number of genes that are actually responsible for the observed effects on the seven traits analyzed. As a working hypothesis we postulated presence of two genes, called *MC* and *MI*, but the analysis based on phenotypic measurements only leaves considerable uncertainty about this point.

Acknowledgements

Dutch breeding companies participating in the described crossing experiment were NVS, Bovar, Euribrid, Fomeva and Nieuw-Dalland. Meishan founders used in the crossing experiment are from a pure Meishan herd at Wageningen Agricultural University, made available by Euribrid (Boxmeer, The Netherlands). Research was supported financially by the Dutch Product Board for Livestock, Meat and Eggs, and by the aforementioned breeding companies participating in the experiment.

References

Cameron ND (1990) Genetic and phenotypic parameters for carcass traits, meat and eating quality in pigs. Livest. Prod. Sci. 26: 119-135

Eikelenboom G, Minkema D (1974) Prediction of pale, soft, exudative muscle with a non-lethal test for the halothane-induced porcine malignant hyperthermia syndrome. Tijdschr. Diergeneeskunde 99: 421-426

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum. Hered. 21: 523-542

Falconer DS (1989) Introduction to quantitative genetics, 3rd ed. Longman, Harlow, London

Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. J. Am. Stat. Assoc. 85: 398-409

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattn. Anal. Mach. Intell. 6: 721-741

Gianola D, Foulley JL (1990) Variance estimation from integrated likelihoods (VEIL). Genet Sel Evol 22: 403-417

Guo SW, Thompson EA (1991) Monte carlo estimation of variance component models for large complex pedigrees. IMA J Math Appl Med Biol 8: 171-189

Guo SW, Thompson EA (1992) A monte carlo method for combined segregation and linkage analysis. Am. J. Hum. Genet. 51: 1111-1126

Hasstedt SJ (1982) A mixed-model likelihood approximation on large pedigrees. Computer and Biomedical Research 15: 295-307

Hill WG, Knott SA (1990) Identification of genes with large effects. In: Advances in statistical methods for genetic improvement of livestock, edited by D. Gianola and K. Hammond, Springer-verlag

Hovenier R, Kanis E, Van Asseldonk Th, Westerink NG (1992) Genetic parameters of pig meat quality traits in a Halothane negative population. Livest. Prod. Sci. 32: 309-321

Hovenier R, Kanis E, Van Asseldonk Th, Westerink NG (1993) Breeding for meat quality in Halothane negative populations - a review. Pig News and Information 14: 17N-25N

[1]Janss LLG, Thompson R, Van Arendonk JAM (1995) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theor. Appl. Genet. 91: 1137-1147

Knott SA, Haley CS, Thompson R (1992) Methods of segregation analysis for animal breeding data : a comparison of power. Heredity 68: 299-311

Lande R (1981) The minimum number of genes contributing to quantitative variation between and within populations. Genetics 99: 541-553

Le Roy P, Elsen JM (1992) Simple test statistics for major gene detection: a numerical comparison. Theor. Appl. Genet. 83: 635-644

Le Roy P, Elsen JM, Knott SA (1989) Comparison of four statistical methods for detection of a major gene in a progeny test design. Genet. Sel. Evol. 21: 341-357

Le Roy P, Naveau J, Elsen JM, Sellier P (1990) Evidence for a new major gene influencing meat quality in pigs. Genet. Res. 55: 33-40

[1]Chapter 4 of this thesis

Lin S, Thompson E, Wijsman E (1993) Achieving irreducibility if the Markov chain Monte Carlo method applied to pedigree data. IMA J Math Appl Med Biol 10: 1-7

MacDougall DB (1986) The chemistry of color and appearance. Food Chemistry 21: 283-299

Morton NE, MacLean CJ (1974) Analysis of family resemblance III. Complex segregation of quantitative traits. Am. J. Hum. Genet. 26: 489-503

Otsu K, Phillips MS, Khanna VK, De Leon S, MacLennan DH (1992). Refinement of diagnostic assays for a probable causal mutation for porcine and human malignant hyperthermia. Genomics 13: 835-837

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1992) Numerical recipes; The art of scientific computing, 2nd Ed. Cambridge University Press, Cambridge, Mass

Raftery AE, Lewis SM (1992) How many iterates in the Gibbs sampler? In: Bayesian statistics IV, edited by Bernardo JM, Berger JO, David AP, Smith AFM, Oxford University Press

SAS Institute Inc, 1988 SAS/STAT Users Guide, release 6.06. Cary, North Carolina

Scott DW (1992) Multivariate density estimation. Wiley and Sons, New York

Sheehan N, Thomas A (1993) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics 49: 163-175

Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related markov chain monte carlo methods. J. Roy. Stat. Soc. B 55: 3-24

Sorensen DA, Wang CS, Jensen J, Gianola D (1994) Bayesian analysis of genetic change due to selection using Gibbs sampling. Genet. Sel. Evol. 26: 333-360

Sorensen DA, Andersen S, Gianola D, Korsgaard I (1995) Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. 27: 229-249

Tanner MA (1993) Tools for statistical inference. Springer, Berlin Heidelberg New York

Thomas DC, Cortessis V (1992) A Gibbs sampling approach to linkage analysis. Hum. Hered. 42: 63-76

Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. 25: 41-62

Zhang Y, Proenca R, Maffei H, Barone M, Leopold L, et al. (1994) Positional cloning of the mouse *obese* gene and its human homologue. Nature 372: 425-432

Appendix

Details on construction of gibbs samplers for inference in mixed inheritance models

A software package was developed to construct Gibbs samplers for inference in the mixed inheritance model (1) described in the Methods section. The main theory on the construction of the required Gibbs sampler is described by Janss et al. (1995). Hereafter, some general information on the developed package is given and extensions to Janss et al. (1995) used in the present study and some computational remarks are described.

General: The set-up of the package is largely a 'help-yourself tool-kit', consisting of a set of FORTRAN-variables, corresponding to model-parameters, and a set of FORTRAN-77 routines to sample (groups of) parameters. By repeatedly calling these routines, Gibbs samplers are generated. Additional routines are supplied which read, order and code pedigree and data files and which make information from these files available to the routines for sampling model parameters. Set-up of the package also allows fitting of sub-models of the mixed inheritance model, e.g. the pure polygenic model, and allows suppressing the sampling of parameters, in which case parameters are updated by their 'current' expectation. As described by Janss et al. (1995), the latter allows application of a hierarchy of inferential approaches, for example a Monte Carlo EM likelihood-approach or Gauss-Seidel schemes for iteratively solving of linear model equations can be specified. The software package can be obtained from the authors.

Sampling details: The Gibbs sampler constructed by Janss et al. (1995) included the following: sampling of levels of one or more fixed categorical non-genetic effects; sampling of polygenic effects assuming each individual in the pedigree has one observation and applying blocking of effects of each sire and those of its final progeny; sampling of genotypes in similar blocks as polygenic effects; sampling of variance components using flat priors; sampling of an additive effect at the single locus; and sampling of allele frequency in the (one) founder population. The software package developed contains routines to sample these parameters. For inference in the mixed inheritance models used in the analyses presented here, the following features were added:

- covariates were allowed for by allowing the design matrix **X** for a non-genetic effect to be a single column vector, containing measured covariates;
- modelling of a dominance effect and modelling of additive and domiance effect with restricted relative dominance effect at the single locus were allowed for. The equation to sample the additive effect a at the single locus given by Janss et al. (1995) can be shown to be based on a linear model to regress the corrected data on a dummy vector **W'Z'**(−1, 0, 0, 1). By analogy, sampling of the dominance effect d at the single locus uses the dummy vector **W'Z'**(0, 1, 1, 0); and sampling of the additive effect at the single locus, assuming $d=ca$, uses the dummy vector **W'Z'**(−1, c, c, 1). In the model with restricted dominance effect, after sampling of a new a, d is set to ca.
- missing observations were allowed for. For polygenic effects and effects at the single locus this follows straightforwardly from linear model methodology by allowing the **Z** matrix to contain columns with all zero's. For the single gene, a missing observation for an individual is accommodated for by use of a penetrance function which equals 1 for each genotype.
- identification of several groups within founder individuals was allowed for, and the procedure to sample allele frequency in a single founder population was then extended to sample allele frequencies in each founder group separately.

For general use of the package, more extensions were made, allowing, for instance, for categorical non-genetic effects to be random and for repeated measurements. These extensions are not described in detail, not being relevant for the Gibbs sampler implementations used in the analyses presented here.

In the software package, alleles for the single gene are referred to by labels '1' and '2' and the additive effect at the single locus a is not restricted to be positive, so that these allele labels will not be unique. Unique labels to identify alleles and, for instance, allele frequencies, are the 'Low' and 'High' labels as defined in the Methods section. Correspondence between the two sets of labels is obtaines as follows: for $a \geq 0$, label '1' corresponds to 'Low' and label '2' corresponds to 'High'; for $a < 0$, label '2' corresponds to 'Low' and label '1' corresponds to 'High'. Using the so obtained 'Low' and 'High' labels, unique inferences could subsequently be made on allele frequencies, genotypes and genotype frequencies. For the additive effect at the single locus, only the absolute value was considered.

Segregation analyses for presence of major genes to affect growth, backfat and litter size in Dutch Meishan-crossbreds

Chapter 6

Presence of major genes was investigated for two growth traits, backfat thickness and two litter size traits in the F_1 and F_2 population of a cross between Meishan and Western pig lines. Segregation analyses were performed in a Bayesian setting, estimating the contribution of background polygenes and the contribution of a possible major gene to the expression of the traits considered. In a first analysis, joint analysis of F_1 and F_2 crossbred data was performed, in which different error variances were fitted for F_1 and for F_2 observations. In this first analysis, significant contributions of major-gene variance were found for the two growth traits, for backfat, and for litter size at first parity. In a second analysis, analysis of F_2 data only was performed to check whether no biases were introduced in the joint analysis of F_1 and F_2 data. In the second analysis, no major genes were found for growth traits. Major genes affecting backfat and litter size at first parity were confirmed. The gene identified to affect backfat is a dominant gene, where the homozygote recessive genotype has an increased level of backfat of about 6 mm. The gene identified to affect litter size at first parity also is a dominant gene, where the homozygote recessive genotype has a decreased litter size of about 5 to 6 piglets.

Introduction

The Chinese Meishan pig-breed has characteristics that are quite different from those found in Western breeds (e.g., Bidanel et al., 1990; Haley and Lee, 1990; Haley et al., 1992). In particular the extreme fertility of the Meishan breed has attracted the attention of physiological research (e.g., Bolet et al., 1986) and of commercial pig-breeding companies. In order to investigate the potential of the Meishan breed for commercial pig-breeding, the Dutch pig-breeding companies Bovar, Euribrid, Fomeva, Nieuw-Dalland and NVS have set up an experiment to produce F_1 and F_2 crossbreds between

Meishan and Western lines. One aim of this experiment, considered in the present study, was to investigate presence of major genes affecting traits of interest in these crossbreds. Presence or absence of major genes will be a main criterion to decide on further utilization of the crossbreds: when major genes are present, backcrossing of the crossbreds to one of the parental lines could be used to develop a lean Meishan line or to develop a fertile Western line; when major genes are absent, continued intercrossing and selection of the crossbreds could be used to develop a synthetic line. At present, a few indications for presence of major genes in Meishan crosses have been obtained: the estrogen receptor locus was found associated with litter size (Rothschild, 1996) and in a previous analysis of meat quality data from Dutch Meishan F_2-crosses, presence of major genes affecting pH, intramuscular fat and backfat were found (Janss et al., 1996).

In the Dutch Meishan crossing experiment typing of animals for genetic markers was not a priori considered. Therefore, as also in Janss et al. (1996), segregation analyses are considered to investigate the presence of major genes and to see whether typing of animals could be interesting. Application of segregation analysis has become well-feasible by use of Markov chain Monte Carlo methodology as developed by Guo and Thompson (1992) and, for animal populations in particular, by Janss et al. (1995). For analysis of animal populations, a Bayesian approach to segregation analysis appears interesting, for instance because many non-genetic 'fixed' effects can be included in the model as nuisance parameters. In contrast, a classical likelihood-based segregation analysis is based on a joint maximization for genetic parameters and fixed effects.

In this paper we report on analysis of growth, backfat and litter size, measured on F_1 and F_2 crossbreds from the Dutch Meishan crossing experiment, for presence of major genes. A Bayesian segregation analysis is considered, as also used by Janss et al. (1996).

Material and methods

Meishan crossbreds

F_1 and F_2 crossbreds between Chinese Meishan and Western pig lines were available from an experiment involving five Dutch pig breeding companies. Western females at the companies were of Dutch Landrace and Large White types. Figure 1 shows the

design of the crossbreeding experiment and the numbers of litters produced.

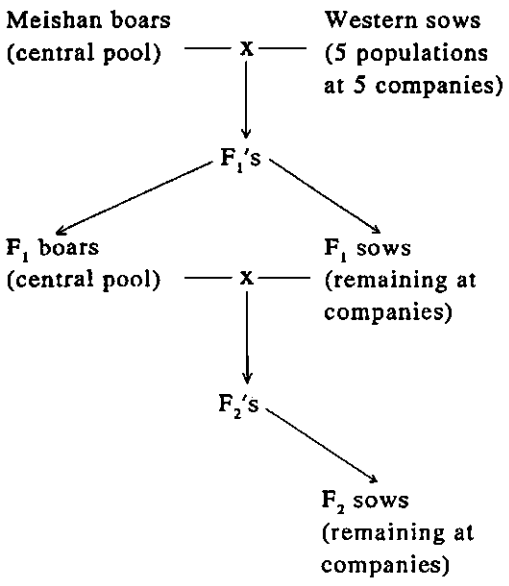


Figure 1 Design of the crossing experiment to produce F_1 and F_2 crossbreds between Chinese Meishan and Western pig lines: (1) 126 F_1 litters were produced from 19 Meishan boars and 126 Western sows of 5 lines in 5 companies; (2) from F_1 litters, a selection of boars was transferred to a central location and a selection of sows remained at the companies; (3) 265 F_2 litters were produced from 39 F_1 boars and 265 F_1 sows; (4) from F_2 litters, a selection of sows was maintained at the companies to obtain data on litter size. All selection steps were random within family.

The design created genetic links between the crossbreds produced at the five breeding companies, firstly by use of one pool of pure Meishan boars, and secondly by creation of a central pool of F_1 boars, which were taken equally from all companies and

subsequently used to inseminate sows at all companies. Due to this design, boar-families of pure Meishan boars and of F_1 boars were not confounded with companies. Mating was at random, apart from avoiding mating of full-sibs in F_1 matings. From performance tested F_1 animals, a selection of young boars and gilts were taken as parents to produce F_2 crossbred litters, where F_2 litters were born from first litters of F_1 sows. From performance tested F_2 animals, only a selection of gilts was maintained (Figure 1). Selection in both cases was done at random within families, i.e., preserving each parental lineage in the selected offspring. Sows were kept to obtain data on litter size at first and second parity; for the second parity of F_1 sows and for first and second parity of F_2 sows, these sows were inseminated with boars from a commercial sire line. Each company used its own commercial sire line to obtain these litters. This implies that, if the line-type of these commercial sires influenced the sizes of the litters they conceived, such effect will be accounted for by a general effect of company. Litter size at each parity was considered a different trait, denoted LS1 and LS2 and was the litter size at birth, i.e. the total of alive and stillborn piglets. Numbers of observations for each litter-size trait in F_1 and F_2 are given in Table 1.

Table 1 Abbreviations, computing details, units of measurement and number of measurements in F_1 and F_2 for considered traits.

Trait	Description and details	N° of measurements F_1	N° of measurements F_2
LGR	life-growth : weight/age at approx. 90 kg life-weight (gr/day)	1057	1250
TGR	test growth : (weight gain) / (days) from approx 25 to 90 kg life-weight (gr/day)	758	1022
BF	backfat thickness at approx. 90 kg life-weight, ultrasonic measurement (mm)	1056	1222
LS1	littersize (piglets born dead or alive) at 1st farrowing	262	268
LS2	littersize (piglets born dead or alive) at 2nd farrowing	246	222

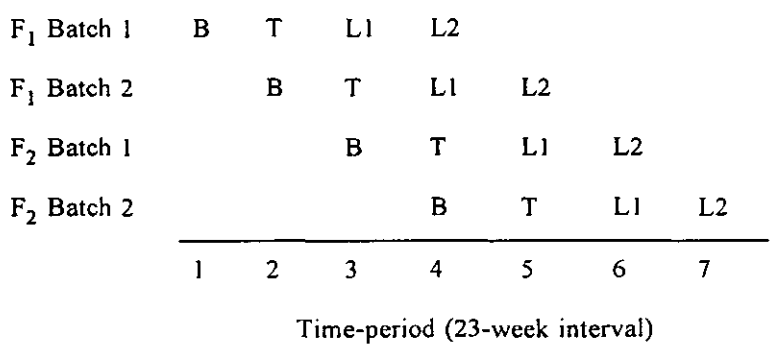


Figure 2 Lay-out of production, performance testing and farrowing of crossbreds in time, indicating the periods where crossbreds were born (B), finished performance test (T), and produced first litter (L1) and second litter (L2). F_2 crossbreds born were from first litters of F_1 crossbreds.

Crossbreds were produced at the same time at the five companies in two batches. Synchronization between the companies was achieved by insemination of all sows at a similar age in three-week-periods, where batches and generations followed each other in 23-week intervals, leading to the scheme for production, testing and farrowing shown in Figure 2. Figure 2 shows that for recorded performance tests batches and generations are completely confounded with periods, such that a period-effect is sufficient to account for batch- and generation effects in the current analyses. The same holds for recorded litter sizes at first parity and for recorded litter sizes at second parity. In order to be able to compare mean levels of the different batches and generations, measurements on control lines were obtained as well, but such comparison of mean levels is outside the scope of the present study and data on control lines is not considered.

Performance tests

In performance tests, measurements were obtained on life-growth (LGR), test-growth (TGR) and on backfat thickness (BF). Performance tests were conducted for a

minimum of 2 boars and 3 gilts per litter. Table 1 shows definition of these traits and numbers of observations for each trait in F_1 and F_2. Performance test results were available on boars and gilts and on a small number of castrates; these castrates occurred in one of the F_1 batches at one of the companies and were included in the analyses regrouped with the gilts.

Table 2 Numbers of observations, housing system, feeding regime and raw means and standard deviations for production traits per company and per sex.

Company:	A	B	C	D	E
Males					
N° F_1, F_2[a]	94,178	68, 100	66, 100	126, 81	95, 100
Housing	group	group	individual	individual	group
Feeding	ad lib	ad lib	ad lib	ad lib	ad lib
LGR	557±75.5	550±93.4	605±85.8	580±89.0	552±91.6
TGR	699±123	662±139	777±145	845±154	674±140
BF	14.1±2.64	13.3±3.07	15.1±2.88	16.6±3.69	15.9±3.52
Females					
N° F_1, F_2[a]	87, 174	132, 148	86, 102	136, 126	121, 143
Housing	group	group	group	group	group
Feeding	restricted	ad lib	ad lib	restricted	ad lib
LGR	550±77.0	573±78.2	569±87.4	494±62.7	560±88.0
TGR	686±126	706±116	-	-	694±132
BF	15.0±3.08	16.9±3.36	16.6±3.55	17.6±3.54	20.4±4.53

[a] Indicated is the number of animals with one or more traits observed, with generally few missing records except for trait TGR in F_1 males at company D, where 77 observations were missing. Also see Table 1 for exact total numbers per trait in F_1 and F_2.

Testing conditions were not uniform between companies and/or sexes. To illustrate this, Table 2 shows housing system, feeding regime and unadjusted means and standard deviations for the traits per company and sex. Sex-differences appeared not to be

constant over the 5 companies, showing even different signs for growth: individually housed males grew faster than (group-housed) females, but group housed males grew slower than (group-housed) females. Also standard deviations for the traits varied between companies and sexes, but differences in standard deviations did not appear to be associated with housing, feeding, sex, mean levels for the traits or, over traits, with particular companies. For analyses, traits measured at the various companies and on the two sexes were considered as the same traits, after correction for possible differences in mean level between companies and sex-difference within company.

Non-genetic effects

A first main effect considered in analysis of the traits was time-period. As described, due to the scheme for producing, testing and farrowing of the crossbreds (Figure 2), the period-effect also accounts for differences between the generations (F_1 and F_2) and for differences between the batches within each generation. A second main effect considered was a sex by company interaction, accounting also for any differences arising due to different housing and feeding of males and/or females at the companies, as described in Table 2. Significance of effects was investigated using a fixed linear model (SAS-GLM, SAS Institute Inc, 1988), considering initially sex, company, period, and all two-way and the three-way interaction between these effects. This model was applied to the production traits life-growth, test-growth and backfat thickness. In this model, the three-way interaction appeared not significant ($P>0.01$) for all three production traits and the period by sex interaction was not significant for the two growth traits. The three-way interaction and the period by sex interaction were then dropped for all traits. Remaining significant interactions were a company by sex interaction, as expected from the data presented in Table 2, and a company by period interaction, showing that periodical fluctuations are not uniform over companies. Subsequently, it was investigated whether the company by sex interaction could be replaced by effects company, sexe, housing system and feeding regime, by considering type I sums of squares for the company by sex interaction after fitting of housing system and feeding regime effects. For the two growth traits, company by sex interaction remained significant ($P<0.01$) in such model, so that it was decided to keep the company by sex interaction in the model. The remaining model can be reformulated as consisting of one sex by company effect (10 levels for backfat and life-growth, 8

levels for test-growth) and one period by company effect (20 levels). For litter size, sex-effect is not relevant, and only period by company was considered as non-genetic effect.

Genetic models

For analyses on presence of major genes, a model was used with non-genetic effects, effects of background polygenes and effect of a single gene, called major gene. The major gene was modelled as an autosomal bi-allelic locus with Mendelian transmission probabilities. Two groups of founders, differing in allele frequency, were modelled, one group being the paternal Meishan founders of F_1 crossbreds, and one group being the maternal Dutch founders of F_1 crossbreds, similar to the model used by Janss et al. (1996). Modelling of these two founder populations with different allele frequencies allows to model a deviation of genotype frequencies from Hardy Weinberg proportions in the F_1 population caused by unequal frequencies of alleles in paternal and maternal gametes forming F_1 individuals. The use of two founder populations also allows to explain a possible difference in variance at the major locus in the F_1 and F_2 population caused by an allele frequency difference at the major locus in the paternal and maternal founder line. In the founder populations and in the F_2 population, genotypes were assumed in Hardy-Weinberg proportions with the frequency in F_2 equal to the average of the frequencies in the founder populations.

Janss and Van der Werf (1992) showed that differences in error variance between the F_1 and F_2 population led to biased estimates of major gene parameters and affected testing for significance of the major gene component. A larger error variance in F_2 was found to lead to over-estimation of the effect of the major gene and to increased probability of falsely identifying a major gene. For the present analyses, therefore, the model used by Janss et al. (1996) was extended to fit different error variances for F_1 and F_2 observations. To further safeguard against possibly erroneous interpretation of results obtained from combined analysis of F_1 and F_2 data, also analyses using F_2 data only were performed.

Main model

The main model, applied to the 5 traits described, was:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e} \quad (1)$$

where **y** are observations, β non-genetic effects, **u** polygenic effects, **W** genotypes, **m** genotype means, **e** errors and **X** and **Z** incidence matrices for non-genetic effects and polygenic effects, respectively. The vector with observations is partitioned as $\mathbf{y}'=(\mathbf{y}_1'\ \mathbf{y}_2')$, where $\mathbf{y}_1$ contains observations on F_1 individuals and $\mathbf{y}_2$ contains observations on F_2 individuals; similarly, errors are partitioned as $\mathbf{e}'=(\mathbf{e}_1'\ \mathbf{e}_2')$, where $\mathbf{e}_1$ contains the residuals of observations on F_1 individuals and $\mathbf{e}_2$ contains the residuals of observations on F_2 individuals. Non-genetic effects in β include effects of company by period (for all traits) and of company by sex (for growth traits and backfat thickness), as described in the previous paragraph.

To denote genotypes at the major locus, distinction is made between an allele A_L, with 'L' of 'Low', which is defined as the allele that decreases the value of phenotypic measurements, and A_H, with 'H' of 'High', which is defined as the allele that increases the value of phenotypic measurements. Attributes 'Low' and 'High' are unique and allow unique assignment of alleles and related parameters, such as allele frequencies. Matrix **W** is a four-column matrix indicating the genotype of each animal, where columns correspond to the possible genotypes A_LA_L, A_LA_H, A_HA_L and A_HA_H. Four genotypes were considered because this allowed a flexible notation of genotype probabilities in founder populations and of genotype transmission probabilities. In actual computations, however, only three genotypes were considered, i.e. not distinguishing between the two heterozygotes. Effects of the genotypes are represented by **m**, with $\mathbf{m}'=(-a, d, d, a)$, where a is referred to as the additive effect and d is referred to as the dominant effect at the major locus. The additive effect can only take positive values, so that **m** is consistent with the definition of the 'Low' and 'High' allele-attributes. Genotype frequencies in the founder populations are modelled defining allele frequencies for A_L and A_H to be $p_{M.L}$ and $p_{M.H}$ in Meishan founders, and $p_{D.L}$ and $p_{D.H}$ in Dutch founders, with $p_{M.L}+p_{M.H}=1$ and $p_{D.L}+p_{D.H}=1$.

Distributional assumptions for genotypes are specified by genotype probabilities for founder animals and genotype transmission probabilities for non-founders, given their parental genotypes (see Janss et al., 1995). Assuming Hardy-Weinberg proportions in the founder populations, this yields $\Pr(A_e A_f) = p_{M.e}\, p_{M.f}$ for Meishan founder animals and $\Pr(A_e A_f) = p_{D.e}\, p_{D.f}$ for Dutch founder animals, with $e, f \in \{L,H\}$. For

non-base animals, $\Pr(A_e A_f) = \tau_{e,gh}\tau_{f,g^*h^*}$, with $e, f, g, h, g^*, h^* \in \{L,H\}$ and where $A_g A_h$ and $A_{g^*}A_{h^*}$ are the genotypes of the sire and dam of the animal considered, $\tau_{L,gh}$ is the transmission probability for genotype $A_g A_h$ to transmit an A_L allele, and $\tau_{H,gh} = 1 - \tau_{L,gh}$ is the corresponding probability to transmit the A_H allele. To specify Mendelian transmission, $\tau_{L,LL}=1$, $\tau_{L,LH}= \tau_{L,HL}= \frac{1}{2}$, and $\tau_{L,HH}=0$. Distributional assumptions for polygenic effects are specified as $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where **A** is the numerator relationship matrix. Errors are assumed distributed as $\mathbf{e}_1 \sim N(\mathbf{0}, \mathbf{I}\sigma_{e1}^2)$ and $\mathbf{e}_2 \sim N(\mathbf{0}, \mathbf{I}\sigma_{e2}^2)$. Alternatively, the variance structure for errors can be denoted $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}=\text{diag}\{\mathbf{I}\sigma_{e1}^2, \mathbf{I}\sigma_{e2}^2\}$, as used in the Appendix. Specification of the statistical model for the Bayesian approach is completed by specifying use of uniform prior distributions on $<-\infty,\infty>$ for non-genetic effects and effects at the major locus, uniform prior distributions on $<0,\infty>$ for variance components, and uniform prior distributions on [0,1] for allele frequencies. In the prior distribution for variances, a-priori a value of zero is excluded, which is computationally implemented by use of priors defined on $[10^{-12}, \infty>$. The restriction for the additive effect at the major locus to be positive was not imposed through its prior distribution. Rather, a transformation was applied to obtain uniquely identified alleles and to obtain strictly positive additive effects at the major locus (see the Gibbs Sampling section).

Parameters

The complete set of unknowns used for the model (1) with specified distributional assumptions is denoted $\theta_{Gib} = (\beta, \mathbf{u}, \mathbf{W}, \sigma_{e1}^2, \sigma_{e2}^2, \sigma_u^2, a, d, p_{M,L}, p_{D,L})$. All the parameters in θ_{Gib} are used in the construction of Gibbs samplers, but non-genetic effects, polygenic effects and genotypes were not of interest in the present analyses. Further, the two allele frequencies $p_{M,L}$ and $p_{D,L}$ are not uniquely estimable because the data contained observations on crossbreds only. If, for instance, only heterozygotes were found present in F_1, it can not be distinguished whether $p_{M,L}=0$ and $p_{D,L}=1$, or whether $p_{M,L}=1$ and $p_{D,L}=0$. One estimable function of $p_{M,L}$ and $p_{D,L}$ is the allele frequency in the crossbreds. The frequency of A_L in crossbreds is denoted $p_{C,L}$, and was assumed equal to the average of the allele frequencies in founders, hence, $p_{C,L} = \frac{1}{2}(p_{M,L} + p_{D,L})$. The frequency of A_H in crossbreds is denoted $p_{C,H}$ and $p_{C,L}+p_{C,H}=1$. A second set of estimable functions of $p_{M,L}$ and $p_{D,L}$ is the set of genotype frequencies in the F_1. These genotype frequencies can deviate from Hardy-Weinberg proportions

and, therefore, can deviate from genotype frequencies in F_2, although allele frequencies in the two crossbred populations are the same. These genotype frequencies in F_1 are: $p_{F1,LL} = p_{M,L}p_{D,L}$ for the frequency of A_LA_L, $p_{F1,LH}+p_{F1,HL} = p_{M,L}p_{D,H} + p_{M,H}p_{D,L}$ for the frequency of heterozygotes A_LA_H and A_HA_L, and $p_{F1,HH} = p_{M,H}p_{D,H}$ for the frequency of A_HA_H. Based on θ_{Gib}, also the variances explained by the major gene in F_1, denoted σ_{w1}^2, and in F_2, denoted σ_{w2}^2, were computed. Major-gene variances were computed from the genotypic effects (a, d) and from genotype frequencies in F_1, or genotype frequencies in F_2, the latter computed from $p_{C,L}$ and $p_{C,H}$ assuming Hardy-Weinberg proportions. This computation of major gene variance therefore is based on assumptions of random mating and absence of directional selection. The computed variances include both additive and dominance variance at the major locus. In conclusion, the set of parameters of interest for statistical inferences was $\theta_{Inf} = (\sigma_{e1}^2, \sigma_{e2}^2, \sigma_u^2, \sigma_{w1}^2, \sigma_{w2}^2, a, d, p_{F1,LL}, p_{F1,LH}+p_{F1,HL}, p_{F1,HH}, p_{C,L})$.

Sub-models

Two sub-models of model (1) were used. The first sub-model used was a polygenic model, specified as $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$, with all specifications equal to those for model (1), including the heterogeneous error variance. The parameters of interest for statistical inferences in the polygenic model were error variances, polygenic variance and heritability in the F_2, $h_2^2 = \sigma_u^2/(\sigma_{c2}^2+\sigma_u^2)$. A second sub-model used was a model for analysis of F_2 data only, which can be specified as $\mathbf{y}_2 = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e}_2$. In this model, F_1 observations are not included and consequently σ_{e1}^2 is not estimated. In this 'F_2-only-analysis', the non-genetic effect of time-period contained two levels instead of four (see Figure 2). The parameters of interest for statistical inferences in analysis of F_2 data were those given in θ_{Inf} except σ_{e1}^2.

Gibbs sampling

Construction of Gibbs samplers

Marginal posterior distributions of model parameters were obtained using Gibbs sampling, constructing a markov chain with stationary distribution equal to the joint posterior distribution of θ_{Gib}. Construction of a markov chain using these parameters was based on Janss et al. (1995), extended to allow for the dominance effect at the

major locus, for two founder populations differing in allele frequency and for two error variance components. Inclusion of the dominance effect at the major locus and of more than one allele frequency were described in the application of Janss et al. (1996). Inclusion of two error variances is described in the Appendix.

The implementation of the Gibbs sampler generally applied single-variate sampling for all model parameters except for genotypes. For genotypes, 'blocks' were constructed containing the genotype of a sire with all its final offspring, and where genotypes in each block were sampled from their joint distribution conditional on remaining parameters and data (Janss et al., 1995). Blocked sampling of polygenic effects, also considered by Janss et al. (1995), was not applied here. Full single-variate sampling of polygenic effects was used instead, which could easier be modified to allow for two error variance components (see Appendix). To improve mixing of genotypes the relaxation technique of Sheehan and Thomas (1993) was applied. This involves relaxation of the transmission probabilities to slightly non-Mendelian probabilities by use of $\tau_{L,LL}=1-p_{rel}$, and $\tau_{L,HH}=p_{rel}$. Here, p_{rel} is referred to as the relaxation probability, which is taken small and specifies the probability of non-Mendelian transmission of alleles. From a Gibbs chain with relaxed transmission probabilities, cycles with a Mendelian genotype configuration are filtered out, providing a correct set of samples for inferences on a strict Mendelian model (Sheehan and Thomas, 1993). In order for the relaxation technique to have a reasonable impact on mixing, relaxation may be relatively strong (high p_{rel}), leading to a low rate of Mendelian samples in the relaxed chain. In the analyses performed here, we aimed at a rate of Mendelian samples of 1 to 5%. Trial runs are required to find a suitable corresponding value for p_{rel}, which may be different for each data set.

The Gibbs sampler is implemented on alleles denoted A_1 and A_2, with corresponding allele-frequencies $p_{1,M}$, $p_{2,M}$, etc., and with additive genotypic effect at the major locus a defined on $<-\infty,\infty>$. To make inferences on uniquely defined alleles A_L and A_H with allele frequencies $p_{L,M}$, $p_{H,M}$, etc., and with strictly positive a, the following was done: for $a\geq 0$, the label '1' was set to correspond to the label 'L' and the label '2' was set to correspond to the label 'H', i.e. then $p_{L,M}=p_{1,M}$ etc.; the reverse was applied for $a<0$. For inferences on a, always the absolute values of a were taken. To start computation of Gibbs chains, parameters were generally initialized as: heterozygotes for genotypes; some positive value for variances; 0.5 for allele

frequencies; and zeros for all others. Sampling of random realizations in construction of Gibbs samplers was based, directly or indirectly, on the uniform random number generator RAN2 (Press et al., 1992).

Initial trials

Trial Gibbs chains were constructed to investigate convergence behavior, burn-in periods, suitable values for p_{rel} and the degree of dependency in the chains for parameters given in θ_{Inf}. The following applies to the main model (1). Convergence behavior was investigated by 'annealing a hot chain'. Lin et al. (1993) refer to a hot chain as a chain with a high relaxation probability, showing therefore very liberal movement and virtually no Mendelian samples. An initially hot chain with p_{rel}=0.5 was annealed by slowly decreasing the relaxation probability to 10^{-3} over 1000 cycles. This gradually restricts movement to the Mendelian and near-Mendelian space and increases the proportion of Mendelian samples appearing in the chain. The same procedure was used by Janss et al. (1996) and was found to lead efficiently to convergence of the chain. From cycle 1000 onwards, the relaxation probability was kept constant at 10^{-3} and another 5000 cycles were computed to observe the parameter values for the Mendelian model to which the chain had converged. Such a procedure of an annealed hot chain was repeated to investigate whether the Mendelian parameter-space consisted of two or more separated sub-spaces. Secondly, dependency in the Gibbs chains was investigated by producing relaxed chains with $p_{rel}=10^{-3}$ and of 12 000 cycles in total, including a burn-in of 2000 cycles. Mendelian samples filtered out from such chains, were analyzed using the method of Raftery and Lewis (1992) to determine serial dependency by analyzing transition of values in the chains around the mean of the chain. From transition rates, spacing between Gibbs cycles that should yield virtual independence was predicted as exemplified by Janss et al. (1995). For the polygenic model, only dependency in the Gibbs chains was studied for the three relevant variance component (σ_{c1}^2, σ_{c2}^2, σ_u^2). For analysis of F_2 data only no specific pre-investigations were performed.

Estimation runs

Estimation of posterior distributions of parameters for each model and trait was based on five replicated Gibbs chains of such length that each chain produced 50 virtually

independent samples for all parameters in θ_{Inf}. Chain-length was determined as $51k$, where k is the largest predicted spacing for any of the parameters in θ_{Inf}. From such chain, samples of parameters in θ_{Inf} were stored from cycles $2k$, $3k$, ..., $51k$, totalling 50 samples per chain. Cycles 1 to $2k$ allowed for burn-in of the chains. Only independent samples were stored in order to largely reduce output from the Gibbs samplers and to facilitate and speed-up post-analyses. Post-analyses supplied a final check to see whether the produced samples could indeed be considered independent (see below). In estimation runs, relaxation probability was kept constant from the first cycle onwards.

Post-analyses and statistical inference

Convergence of the Gibbs sampler was judged by use of the generated 250 samples from 5 chains in an analysis-of-variance (ANOVA), testing for a significant chain-effect. Significant differences between chains are considered as an indication of (practical) reducibility, in which case Gibbs sampling theory (Geman and Geman, 1984; Gelfand and Smith, 1990) does not hold. In such case, the Gibbs sampler is said not to have converged and generated samples are not from the correct posterior distribution. Significance of chain-effects was assumed when the F-statistic exceeded the 1% significance level. The significance level of 1%, compared to a more usual level of 5%, was applied to account for the multiple tests which were performed. Wrongly assumed independence will increase the F-statistics and also can lead to significance of chain-effects. Hence, the ANOVA at the same time acts as a post-check whether the obtained samples could indeed be considered independent. When significant chain effects were found, the estimation procedure was repeated with a larger spacing between samples, to see whether this could improve convergence.

Statistical inferences were based on summarizing the generated samples in the form of estimated marginal posterior distributions or estimated features thereof. Non-parametric density estimates of posteriors were made in the form of average shifted histograms (Scott, 1992). At natural boundaries of parameter-spaces, these histograms were smoothed up to the boundaries using a reflection boundary technique (Scott, 1992). Such a histogram provides a general and broad inference, combining information on various point- and interval estimates. As features of the marginal posterior distributions, estimated means and standard deviations are presented. Posterior means

were used as point estimates for the parameters. Posterior means fall in the class of APE (A-Posteriori Expectation) estimators which have the general property of minimizing quadratic posterior loss. The higher marginalized Bayesian estimators, compared to classical ML estimators, are expected to have the same asymptotic properties and superior non-asymptotic properties from a Bayesian viewpoint (Gianola and Foulley, 1990). Statistical inferences first focussed on the genetic variance components (σ_u^2, σ_{w1}^2, σ_{w2}^2) and in particular on major gene variance in F_2 (σ_{w2}^2) to determine significance of the major gene in the model. Judgements are based on the shapes of estimated posterior distributions of variance components (Janss et al., 1995), where a non-significant variance shows a distribution with global mode at $\sigma^2=0$ and significance of a variance shows a global mode for $\sigma^2>0$. Major gene variance was concluded to be significant when the global mode had a density 20-fold larger than the density at $\sigma_{w2}^2=0$. This reflects the general conservatism for accepting presence of a major gene. Once significant major gene variance is found, further inferences focussed on the effects at the major locus and on estimable functions of allele-frequencies.

Results

Polygenic model

Inferences for a polygenic model were obtained for the full data set, estimating two error variance components, by omission of the major gene component from model (1). Required chain lengths to obtain 50 independent samples per chain were determined in initial trials to be 7 500 for traits LGR, TGR and BF, and 50 000 for traits LS1 and LS2. Analysis of the 5 Gibbs chains with 50 samples each indicated good convergence: all F-values for chain-effects were non-significant ($P>0.01$) for the variance components for each trait. Posterior means of variance components and heritabilities in the F_2 are in Table 3. Considerable differences in error variance for F_1 and F_2 were estimated, with F_2 variance being higher for all traits. Largest differences were found for TGR, BF and LS1, with error variances in F_2 more than 50% higher than in F_1. Estimated standard deviations of the marginal posterior distributions of these variance components (not shown) indicated that these differences were significant except for LS2. Hence, use of the model with two error variances appears warranted. Estimated polygenic variances indicated reasonable amounts of genetic variance to be present: indicated heritabilities,

computed relative to estimated phenotypic variance in F_2, were 0.15 and 0.20 for the litter size traits and ranged from 0.29 to 0.41 for the production traits (Table 3).

Table 3 Estimated marginal posterior means (mpm) for variance components in a polygenic model (environmental variance in F_1, σ_{e1}^2, environmental variance in F_2, σ_{e2}^2, general polygenic variance, σ_u^2, and heritability in F_2, h_2^2), based on a total of 250 independent Gibbs samples from 5 replicated chains

Trait	mpm σ_{e1}^2	mpm σ_{e2}^2	mpm σ_u^2	mpm h_2^2
LGR	3024	4224	2251	0.347
TGR	5889	9444	6502	0.407
BF	6.568	9.987	4.074	0.289
LS1	5.851	9.070	1.706	0.158
LS2	6.658	8.510	2.126	0.199

Mixed inheritance model

Initial trials

For the mixed inheritance model (1), convergence behavior of the Gibbs sampler was investigated using the described method of annealing a hot chain, which was repeated four times for each trait, using the data set with F_1 and F_2 observations. For BF, LS1 and LS2, the Gibbs sampler was found to converge to the same region of the parameter space in the different runs. For LGR and TGR, however, the Gibbs sampler converged to two different regions of the parameter-space: one region with $d<0$ and $a\approx 0$; and one region with $d>0$ and $a>0$ where d was much larger than a. No mixing was observed between these two regions. Both cases appear to describe a similar phenomenon of a low and a high group with random transmission between groups. For estimation of parameters for LGR and TGR we focussed on the case with $d>0$ by starting Gibbs chains with positive d's. The relaxation probability of 10^{-3} used in the initial trials led to 8 to 13% Mendelian samples in the chains. For the remaining of this study relaxation probabilities were slightly increased to 1.5×10^{-3} for TGR and 2×10^{-3} for other traits in order to obtain the desired rate of 1 to 5% Mendelian samples. Analysis

of dependencies in the chains indicated that chain lengths from 80 000 to 150 000 cycles were required in order to obtain 50 independent samples per chain.

Table 4 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for variance components in a mixed inheritance model (environmental variance in F_1, σ_{e1}^2, environmental variance in F_2, σ_{e2}^2, general polygenic variance, σ_u^2, major-gene variance in F_1, σ_{w1}^2 and major-gene variance in F_2, σ_{w2}^2), based on a total of 250 independent Gibbs samples from 5 replicated chains.

Trait		σ_{e1}^2	σ_{e2}^2	σ_u^2	σ_{w1}^2	σ_{w2}^2
LGR	mpm	1770	2496	1630	1264	1475
	mpsd	256	299	279	296	332
TGR	mpm	4246*	7707*	4927*	2501	3853
	mpsd	721	1443	901	980	1518
BF	mpm	4.28	6.37	2.92	2.36	3.19
	mpsd	0.471	0.614	0.562	0.683	0.919
LS1	mpm	3.54	5.42	1.51	3.19	3.99
	mpsd	0.658	1.21	0.658	1.16	1.46
LS2	mpm	5.29	6.85	1.23	2.40^{NS}	3.09^{NS}
	mpsd	1.13	1.25	0.84	1.18	1.35

* convergence not good, using ANOVA F-test for comparison of within and between chain variances ($P<0.01$).

^{NS} *Not significantly different from zero: ratio of maximum density and density at zero less than 20.*

Full-data analyses

The full data set was analyzed using the described mixed inheritance model (1), estimating two error variance components. Table 4 shows means and standard deviations of the estimated marginal posterior distributions of variance components for all traits. Analysis of samples from repeated chains showed good convergence for major gene variances for all traits, enabling to draw conclusions on the presence or

absence of a major gene to affect the traits. Other variance components also showed good convergence except for TGR. Non-convergence of the error variance and polygenic variance for TGR is likely related to non-convergence of the additive effect at the major locus for this trait, as is described later.

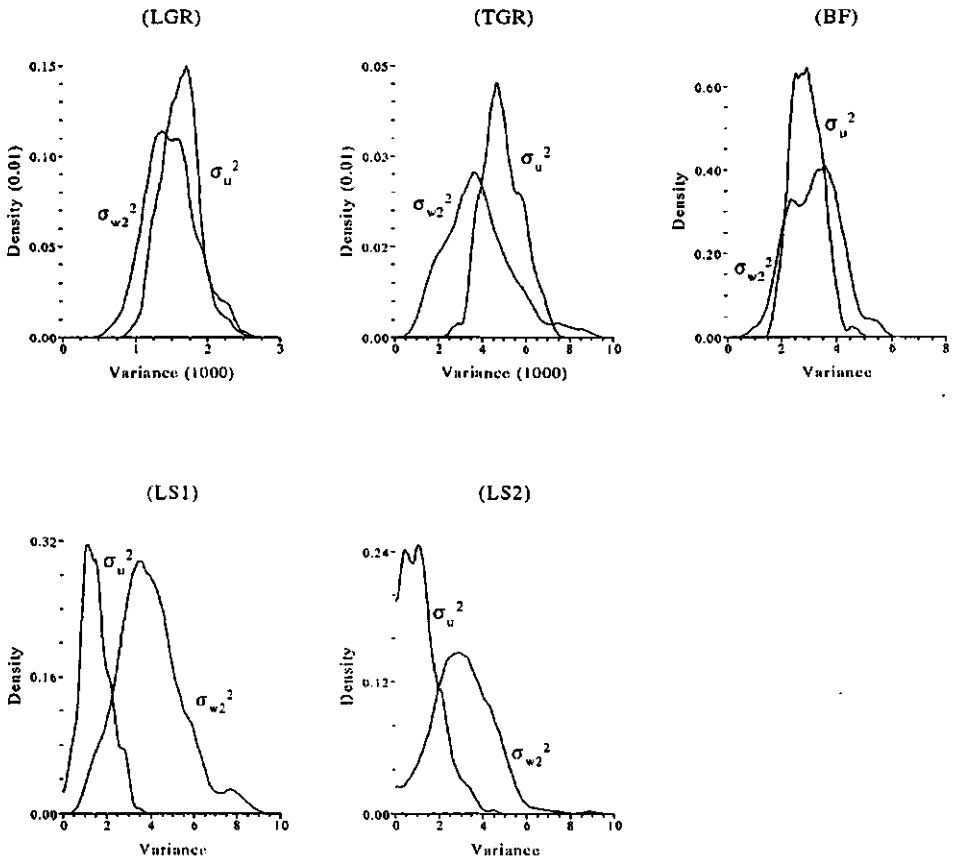


Figure 3 Estimated marginal posterior distributions (averaged histogram frequencies) of polygenic variance (σ_u^2) and of major gene variance in F_2 (σ_{w2}^2) for traits life growth (LGR), test-growth (TGR), backfat thickness (BF), litter size at first parity (LS1) and litter size at second parity (LS2).

To judge significance of the genetic variance components, density estimates for the marginal posterior distributions of polygenic variance (σ_u^2) and major gene variance in F_2 (σ_{w2}^2) are shown in Figure 3. The posterior distribution of major gene variance in F_2 for LS2 shows a non-negligible density at $\sigma_{w2}^2=0$. The density ratio of the density at $\sigma_{w2}^2=0$ relative to the maximum density was estimated as 1:6.0, so that presence of a major gene affecting LS2 was rejected. For other traits, significant contributions of major gene variance in F_2 were found (Figure 3), and the same conclusions were obtained for major gene variance in F_1 (densities not shown, conclusions in Table 4). Major gene variances in F_1 were all lower which results from differences in genotype frequencies as is described below.

For those traits with significant major gene variances (LGR, TGR, BF, LS1), Table 5 shows estimated posterior means and posterior standard deviations for the effects at the major locus, genotype frequencies of homozygotes in F_1 and the allele frequency of A_L in the crossbreds. Estimation of effects at the major locus showed good convergence, except for the additive effect a for TGR. Bad convergence of a for TGR was found caused by some chains showing estimates around 100, while other chains showed estimates around 150. These between-chain differences in estimates for the additive effect likely caused the bad convergence of error variances and polygenic variance for TGR as well. Evidence was found for dominance or over-dominance of the A_H allele for genes affecting LGR and LS1 and dominance of the A_L allele for a gene affecting BF. Estimation of genotype frequencies in F_1 showed non-convergence for a number of traits, which in these cases appeared caused by insufficient spacing between the samples. Estimates for the frequency of the A_L allele in crossbreds ($p_{C,L}$) showed good convergence. Comparison of the frequencies of homozygotes in F_1 with the allele-frequency in crossbreds, $p_{C,L}$, reveals a departure from Hardy-Weinberg proportions in F_1. For instance for BF, estimated frequency of A_L in crossbreds is 0.68, which corresponds to Hardy-Weinberg genotype frequencies of 0.46 for A_LA_L and 0.10 for A_HA_H. In the F_1, however, estimated genotype frequencies for these homozygotes were 0.44 and 0.07. The computation of major gene variance in the F_1 is based on these latter frequencies, whereas for major gene variance in the F_2 the frequencies according to Hardy-Weinberg proportions were used, which explains the differences between major gene variances in F_1 and F_2 in Table 4. Estimates of genotype frequencies in F_1 indicate low frequency of the A_LA_L genotype for TGR and LS1, and

low frequency of the A_HA_H genotype for BF, which in all cases is the recessive genotype.

Table 5 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for major-gene parameters in a mixed inheritance model (additive effect *a* and dominant effect *d* at the major locus, frequency of the 'double low' genotype A_LA_L in F_1, frequency of the 'double high' genotype A_HA_H in F_1, and frequency of the 'low' allele A_L in crossbreds $p_{C,L}$), based on a total of 250 independent Gibbs samples from 5 replicated chains.

				F_1		
Trait		*a*	*d*	freq A_LA_L	freq A_HA_H	$p_{C,L}$
LGR	mpm	37.4	75.0	0.115*	0.347*	0.384
	mpsd	7.39	7.06	0.040	0.093	0.060
TGR	mpm	122*	151	0.036*	0.548*	0.244
	mpsd	39.4	16.5	0.020	0.139	0.077
BF	mpm	2.97	-2.84	0.442	0.073	0.684
	mpsd	0.271	0.350	0.094	0.025	0.056
LS1	mpm	3.12	4.36	0.065*	0.490	0.288
	mpsd	0.423	0.534	0.029	0.104	0.062

* convergence not good, using ANOVA F-test for comparison of within and between chain variances ($P<0.01$)

F_2-only analyses

The analyses as described above were repeated for analysis of F_2 data only for traits LGR, TGR, BF and LS1. Analysis of LS2 was not considered, since the previous analysis indicated absence of a major gene for this trait. Tables 6 shows means and standard deviations of the estimated marginal posterior distributions of variance components for the four traits. Analysis of samples from repeated chains showed good convergence for all variances for all traits except again for error variance and polygenic variance for TGR. Compared to the analysis of the full data set, lower genetic variance was inferred for LGR and higher genetic variance was inferred for BF and LS1.

Opposite differences were found for the error variances for these traits. For TGR, peculiarly, all variances were lower in the analysis of F_2 data only, but here actually no further conclusions could be drawn because in both analyses non-convergence was diagnosed for some of the variances. Except for the estimate of σ_{w2}^2 for BF, all posterior standard deviations were larger in the analysis of F_2 data only, as expected due to the smaller amount of data considered. Figure 4 shows non-parametric density estimates of the posterior distributions of polygenic and major gene variance in F_2 for the four traits. The same horizontal and vertical scales were used as in Figure 3 and, consequently, the spread and height can be compared directly with the posterior distributions shown for analysis of the full data set. Major gene variances were not significantly different from zero for LGR, TGR and LS1, due to decreased means (LGR, TGR) and increased standard deviations (LGR, TGR, LS1) of the posterior distributions. The ratios of the densities at $\sigma_{w2}^2=0$ and the global mode for σ_{w2}^2 were 1:1.2 for LGR, 1:2.7 for TGR and 1:5.0 for LS1.

Table 6 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for variance components in a mixed inheritance model using F_2 data only (environmental variance in F_2, σ_{e2}^2, polygenic variance, σ_u^2, and major-gene variance in F_2, σ_{w2}^2), based on a total of 250 independent Gibbs samples from 5 replicated chains.

Trait		σ_{e2}^2	σ_u^2	σ_{w2}^2
LGR	mpm	3172	1165	784^{NS}
	mpsd	465	505	473
TGR	mpm	7453*	3202*	3292^{NS}
	mpsd	1827	1268	1663
BF	mpm	5.21	2.66	4.34
	mpsd	0.767	0.887	0.850
LS1	mpm	4.03	2.66	4.51^{NS}
	mpsd	2.23	1.57	1.97

* convergence not good, using ANOVA F-test for comparison of within and between chain variances ($P<0.01$); ^{NS} Not significantly different from zero: ratio of maximum density and density at zero less than 20.

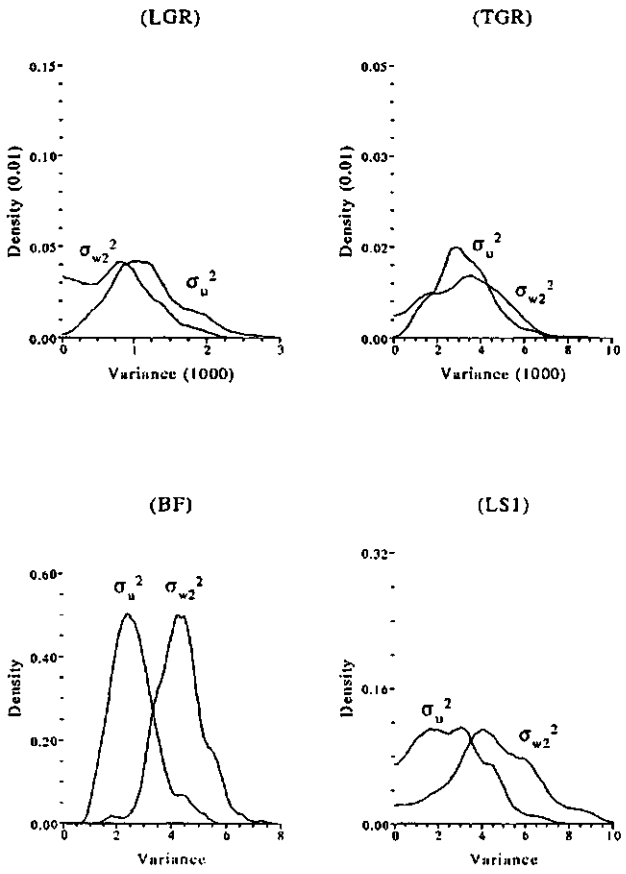


Figure 3 Estimated marginal posterior distributions (averaged histogram frequencies) of polygenic variance (σ_u^2) and of major gene variance in F_2 (σ_{w2}^2) in analysis of F_2 data only for traits life growth (LGR), test-growth (TGR), backfat thickness (BF) and litter size at first parity (LS1).

Table 7 shows estimated posterior means and posterior standard deviations for effects at the major locus, genotype frequencies of homozygotes in F_1 and the allele frequency of A_L in the crossbreds. Genotype frequencies in F_1 and possible departures from Hardy-Weinberg proportions of these genotype frequencies are also estimable from the analysis of F_2 data. Major gene parameters are shown for traits which did not show

significant major gene variance, because these estimates provide additional evidence for the presence or absence of major genes when compared to the results in Table 5.

Table 7 Estimated marginal posterior means (mpm) and marginal posterior standard deviations (mpsd) for majore-gene paramteres in a mixed inheritance model using F_2 data only (additive effect a and dominant effect d at the major locus, frequency of the 'double low' genotype A_LA_L in F_1, frequency of the 'double high' genotype A_HA_H in F_1, and frequency of the 'low' allele A_L in crossbreds $p_{C,L}$), based on a total of 250 independent Gibbs samples from 5 replicated chains.

				F_1		
Trait		a	d	freq A_LA_L	freq A_HA_H	$p_{C,L}$
LGR	mpm	21.6	12.0	0.190	0.255	0.467
	mpsd	14.3	48.2	0.133	0.157	0.133
TGR	mpm	45.3*	39.7*	0.167*	0.232*	0.468
	mpsd	24.1	83.9	0.116	0.140	0.111
BF	mpm	2.92	-2.85*	0.255	0.024^{MZ}	0.615
	mpsd	0.325	0.549	0.095	0.026	0.050
LS1	mpm	2.56	4.43	0.085	0.433	0.326
	mpsd	0.776	1.90	0.098	0.136	0.108

* convergence not good, using ANOVA F-test for comparison of within and between chain variances (P<0.01); ^{MZ} (Global) mode at zero

For LGR and TGR, estimates of effects at the major locus and genotype- and allele frequencies were very different from those found in the analysis of the full data. TGR again showed non-convergence for all parameters except for the allele frequency in crossbreds. The differences in estimates between the two analyses do not support presence of major genes influencing LGR and TGR as found in the analysis of the full data. For BF, analysis of F_2 data showed different F_1 genotype frequencies, now indicating a larger portion of heterozygotes and absence of the A_HA_H genotype in F_1.

Overall, analysis of BF using F_2 data only was considered to agree well with the analysis using the full data set. Major gene variance for LS1 was not significant in the analysis of F_2 data (Table 6, Figure 4), but similar estimates for effects at the major locus and for genotype- and allele frequencies were found as in the analysis of the full data. Therefore, we concluded that analysis of F_2 data for BF and LS1 confirmed presence of major genes affecting these traits.

Discussion and conclusions

In this study, segregation analyses were used to investigate presence of major genes affecting five commercially important traits measured on Meishan crossbreds. For combined analysis of data on F_1 and F_2 crossbreds in segregation analysis, a concern was brought up by Janss and Van der Werf (1992), showing that biases arose and that major genes could erroneously be found when error variances were different in the two generations. In the present analyses, therefore, care was taken to safeguard against such biases and false conclusions, firstly by estimating two error variance components when F_1 and F_2 data were combined, and secondly by considering also F_2 data only for analyses.

For life-growth and test-growth, large discrepancies were found between analysis of the full data and analysis of F_2 data, showing different estimates for effects at the major locus and allele frequencies and with major gene variance significant in the analysis of the full data, but not in the analysis of F_2 data. This indicates that the F_1 data had certain features which led to a significant estimate of major gene variance, and that these features were not present in the F_2 data. For example, the F_1 data may have been more skewed than the F_2 data. However, differences between the analyses of growth traits may also have been caused by analysis on the observed scale, whereas log-scale would be more appropriate, or by presence of more than one gene, or a gene with more than 2 alleles. Further investigation of the growth data therefore remains of interest. Due to the discrepancies found for the analyses of growth traits, it is concluded that presence of a single major gene affecting these traits is not likely.

Results from the analysis of the full data and the F_2 data for backfat and litter size (at first parity) agreed reasonably well, with only one marked difference in estimated genotype frequencies in F_1 for backfat. In the analysis of backfat using F_2

data, a lower frequency for the A_LA_L genotype and absence of the A_HA_H genotype was found. Due to generally well comparable estimates for major genes affecting backfat and litter size in the analysis of the full data and in the analysis of F_2 data, presence of major genes affecting these traits was found likely. Differences between homozygote genotypes were estimated as 6 mm for the gene affecting backfat and 5 to 6 piglets for the gene affecting litter size. Raw means in the F_2 were 16.8 mm backfat and 11.0 piglets at first parity, so that for backfat the 'normal' genotype corresponded to a mean level of around 16 mm vs. a level of 22 mm for the homozygous recessive genotype. For litter size, these figures would be 11.5 piglets for the 'normal' genotype and 6 piglets for the homozygous recessive genotype.

In the present study, backfat was measured ultrasonically on the life animal. Finding of a major gene for backfat is supported by the previous finding of a major gene affecting backfat measured on carcasses of F_2 crossbreds using a HGP measurement (Janss et al., 1996). In the analysis of Janss et al. (1996) a recessive allele was found that increased backfat and with absence of the homozygote recessive genotype in the F_1. Recessiveness of the allele to increase backfat agrees with the present analysis, and absence of the homozygote recessive genotype in the F_1 agrees with the present analysis of F_2 data. It is plausible therefore, that the gene identified here to affect backfat is the same as the gene found to affect backfat identified by Janss et al. (1996). Effect of the previously found major gene was larger (8.4 vs 5.8 mm), which may be explained by use of the different measurement of backfat and by use of older animals in the previous analysis. Frequencies in crossbreds of the recessive allele were very close, i.e. 0.39 in the analysis of Janss et al. (1996) and likewise 0.39 in the current analysis of F_2 data. To validate presence of the major gene affecting backfat, Janss et al. (1996) showed differences in family-variances, with larger variances in families of boars that carried the recessive allele. They also concluded that the recessive allele most likely originates from the Meishan breed.

Validation of presence of a major gene affecting litter size was found in plotting the distributions of the raw data for the F_1 and F_2 observations (Figure 5). These plots showed a slightly left-skewed distribution in the F_1, and a markedly more left-skewed distribution, even a faint bimodality, in the F_2. The difference in these distributions for F_1 and F_2 is a strong indication for an underlying genetic mechanism. The group of animals with extreme low litter sizes appearing in the F_2 were found at all five

companies and were descendants of specific boars only. This also implicitly is apparent from the statistical analyses, in which company-effects were fitted and in which two different genotypes were found present in the F_1. Due to the well-balanced design of the data, confounding with some non-genetic effect is unlikely. Estimated effects of the gene found to affect litter size showed some over-dominance and genotype frequency estimates in F_1 indicated presence of the homozygote recessive in the F_1 and, hence, presence of the recessive allele in both founder populations. However, presence of a dominant gene with the recessive allele present in one of the founder populations only, could also explain the finding. In that case, one needs to attribute the slight left-skewness seen in F_1 (Figure 5) to a general natural skewness of the observations, rather than to the effects of a major gene. This would also imply that parameter estimates must be somewhat biased due to such natural skewness, and that the difference between homozygotes could actually be larger than the difference estimated.

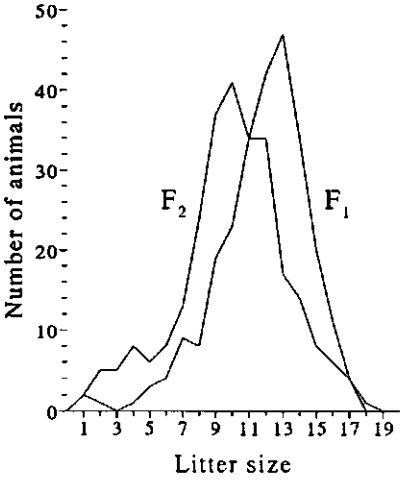


Figure 5 Distribution of raw observations for litter size at first parity for F_1 and F_2 animals.

The major gene identified to affect litter size is unlikely to be the ESR-effect identified by Rothschild et al. (1996), due to the larger magnitude of the effect found here. The major gene identified here results in a 5 to 6 piglets difference between homozygotes, whereas the effect for 50% Meishan animals associated with the ESR locus was reported to be about $2\frac{1}{2}$ at first parity. The major gene found to affect litter size affected

litter size at first parity. In the analysis of litter size at second parity, no significant effect of a major gene was found. This could imply that the currently found gene is specific for first litters, or that the effect of the same gene on second litters is smaller, and therefore could not be identified. In the experiment, mating of young sows was at fixed age, such that variation in the onset of puberty can affect the litter size at first parity. It can, therefore, not be excluded that the major gene found is (partly) related to the onset of puberty.

A possible reason for appearance of a group of F_2 sows with small litters, could also be an infection of animals by PEARS (Porcine Epidemic Abortion and Respiratory Syndrome). Such an infection prevailed in The Netherlands during the experiment. PEARS generally infects all animals at a farm at the same moment, and then may have variable effects on litter size, dependent on the pregnancy-stage of animals at that moment. When considering litter size at birth including stillborn piglets, which was the trait analyzed, the group of animals with reduced litter size should have been markedly earlier in pregnancy at the moment of infection than the other animals (P.C. Vesseur, Research institute for pig husbandry, Rosmalen, The Netherlands, personal communication). In the experiment, however, pregnancy-stage of the animals was very similar, such that PEARS should have had similar effect on all animals within each company. Also, animals with reduced litter sizes should then show increased numbers of mummified piglets, which was not found when comparing the percentages of mummified piglets in litters of size ≤ 7 with those in litters of size ≥ 8. PEARS, or any disease, can therefore not have caused appearance of the group of F_2 sows with small litters.

For breeding, the recessive alleles of the major genes identified (A_H for the gene affecting backfat, A_L for the gene affecting litter size), will be unfavorable. Selection against the recessive alleles in a constituted 50% Meishan synthetic line, would somewhat improve the line, i.e. given estimates of effects and frequencies of the genes identified, the gene affecting backfat accounts for an increased level of backfat of about 1 mm, and the gene affecting litter size accounts for a decreased level of litter size of about 0.5 piglet. For application of a Meishan synthetic line as grandparent-line in a breeding program, presence of these genes is not directly important, assuming that the recessive unfavorable allele originates from Meishan only and, hence, would not be present in white breeds used in commercial crosses. As already noted above, Janss et

al. (1996) found that the recessive allele for the gene affecting backfat appeared to originate from Meishan, but further validation of this assumption will be important for commercial use of the Meishan crossbreds.

Acknowledgements

Dutch breeding companies participating in the described crossing experiment were NVS, Bovar, Euribrid, Fomeva and Nieuw-Dalland. Meishan founders used in the crossing experiment are from a pure Meishan herd at Wageningen Agricultural University, made available by Euribrid (Boxmeer, The Netherlands). Research was supported financially by the Dutch Product Board for Livestock, Meat and Eggs, and by the aforementioned breeding companies participating in the experiment.

References

Andersson L, Haley CS, Ellegren H, Knott SA, et al (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science 263: 1771-1774

Bidanel JP, Caritez JC, Legault C (1990) Ten years of experiments with Chinese pigs in France. 1. Breed evaluation. Pig News Inform 11: 345-348

Bolet G, Martinat-Botté F, Locatelli P, Gruand J, et al (1986) Composantes de la prolificité de truies Large White hyperprolifiques en comparaison avec des truies de races Meishan et Large White. Genet Sel Evol 18: 333-342

Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. J Am Stat Assoc 85: 398-409

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans Pattn Anal Mach Intell 6: 721-741

Gianola D, Foulley JL (1990) Variance estimation from integrated likelihoods (VEIL). Genet Sel Evol 22: 403-417

Guo SW, Thompson EA (1992) A monte carlo method for combined segregation and linkage analysis. Am J Hum Genet 51: 1111-1126

Haley CS, Lee GJ (1990) Genetic components of litter size in Meishan and Large White pigs and their crosses. Proc 4th World Congr Genet Appl Livest Prod, July 1990 Edinburgh UK, 15: 458-461

Haley CS, D'Agaro E, Ellis M (1992) Genetic components of growth and ultrasonic fat depth traits in Meishan and Large White pigs and their reciprocal crosses. Anim Prod 54: 105-115

[1]Janss LLG, Van der Werf JHJ (1992) Identification of a major gene in F_1 and F_2 data when alleles are assumed fixed in the parental lines. Genet Sel Evol 24: 511-526

[2]Janss LLG, Thompson R, Van Arendonk JAM (1995) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theor Appl Genet 91: 1137-1147

[3]Janss LLG, Van Arendonk JAM, Brascamp EW (1996) Bayesian statistical analyses for presence of single genes affecting meat quality traits in a crossed pig population. Genetics (accepted)

Lin S, Thompson E, Wijsman E (1993) Achieving irreducibility of the markov chain monte carlo method applied to pedigree data. IMA J Math Appl Med Biol 10: 1-17

Sheehan N, Thomas A (1993) On the irreducibility of a markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics 49: 163-175

Rothschild M, Jacobson C, Vaske D, Tuggle C, et al (1996) The estrogen receptor locus is associated with a major gene influencing litter size in pigs. Proc Natl Acad Sci USA, 93: 201-205

Wang CS, Rutledge JJ, Gianola D (1994) Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet Sel Evol 26: 91-115

[1]Chapter 2, [2]Chapter 4 and [3]Chapter 5 of this thesis

Appendix

Sampling of parameters in the Gibbs sampler in a model with two error variance components

Construction of Gibbs samplers to make inferences in model (1) uses the set of parameters given by θ_{Gib} in the main section. Sampling of the linear model components of θ_{Gib}, which are non-genetic effects, polygenic effects and effects at the major locus, is straightforward by construction of 'conditional' linear model equations for these parameters in turn, i.e. taking other parameters as known (e.g., Wang et al., 1994). In general, solutions from linear model equations are used as means and the inverse of the left-hand-side of the linear model equations is used as variance of a (multivariate) normal distribution from which new effects are sampled. The heterogeneous variance structure for errors is accounted for by considering the most general form of linear model equations, which implicitly involves the variance structure for errors, as defined with model (1) by **R**. Conditioning on other linear model components can conveniently be described by use of 'corrected data', which is the data corrected for current values of all effects other than the effect updated (e.g., Janss et al., 1995).

Sampling of non-genetic effects is based on linear model equations $(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})\beta = (\mathbf{X}'\mathbf{R}^{-1}\tilde{\mathbf{y}})$, where $\tilde{\mathbf{y}}$ is the corrected data. To update the level i of a non-genetic effect, this leads to sampling:

$$\beta_i \sim N((\tilde{y}_{1i\cdot}/(n_{1i}\sigma^2_{e1})+\tilde{y}_{2i\cdot}/(n_{2i}\sigma^2_{e2})),\ (\sigma^2_{e1}/n_{1i}+\sigma^2_{e2}/n_{2i}))$$

where $\tilde{y}_{ji\cdot}$ is the total of observations from $\tilde{\mathbf{y}}$ from generation F_j in level i and n_{ji} is the number of observations from generation F_j in level i. Sampling of polygenic effects was based on linear model equations $(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\sigma_u^{-2}\mathbf{A}^{-1})\mathbf{u}=(\mathbf{Z}'\mathbf{R}^{-1}\tilde{\mathbf{y}})$, where $\tilde{\mathbf{y}}$ again is corrected data. A single-variate equation to solve polygenic effect u_i of individual i with one observation and of generation F_j, can be obtained from this set of linear model equations as $d_i u_i = c_i$, where

$$d_i=\sigma_{e_j}^{-2}+\sigma_u^{-2}(\delta_i+\tfrac{1}{4}\textstyle\sum_k\delta_k)$$
$$c_i=\tilde{y}_i\sigma_{e_j}^{-2}+\tfrac{1}{2}\sigma_u^{-2}\delta_i(u_{S.i}+u_{D.i})-\sigma_u^{-2}\textstyle\sum_k(\tfrac{1}{4}\delta_k u_{M.k}-\tfrac{1}{2}\delta_k u_k)$$

which is the equation given for dams by Janss et al. (1995), but with d_i and c_i divided by error variances. Based on this equation, u_i is sampled $N(c_i/d_i, d_i^{-1})$. In this equation, assuming no inbreeding, δ_i=1 when i is a founder and δ_i=2 when i is a non-founder, $u_{S.i}$ and $u_{D.i}$ are polygenic effects of the sire of i and dam of i, sums over k sum over all progeny of i (when present), where u_k is the polygenic effect of progeny k, $u_{M.k}$ is the polygenic effect of the mate of i, other parent of k. When i is a founder, $u_{S.i}$ and $u_{D.i}$ are taken as zero and for individuals without an observation, $\sigma_{e_j}^{-2}$ in d_i and $\tilde{y}_i\sigma_{e_j}^{-2}$ in c_i are omitted. Polygenic effects were sampled individually, using the above given equation to sample polygenic effect of all individuals. Sampling of the additive effect at the major locus was based on linear model equation $\mathbf{k'Z'R^{-1}Zk}a=\mathbf{k'Z'R^{-1}}\tilde{\mathbf{y}}$, where $\tilde{\mathbf{y}}$ is again corrected data and where $\mathbf{k}=\mathbf{W}(-1, 0, 0, 1)'$, i.e. a dummy-vector which indicates individuals with the A_LA_L genotype by -1's and individuals with the A_HA_H genotype by +1's. The equation can be worked out to yield

$$l=\mathbf{k'Z'R^{-1}Zk}=n_{1.HH}/\sigma_{e1}^2+n_{2.HH}/\sigma_{e2}^2+n_{1.LL}/\sigma_{e1}^2+n_{2.LL}/\sigma_{e2}^2$$
$$r=\mathbf{k'Z'R^{-1}}\tilde{\mathbf{y}}=\tilde{y}_{1.HH}./\sigma_{e1}^2+\tilde{y}_{2.HH}./\sigma_{e2}^2-\tilde{y}_{1.LL}./\sigma_{e1}^2-\tilde{y}_{2.LL}./\sigma_{e2}^2$$

which leads to sampling a from $N(r/l, l^{-1})$, and where $\tilde{y}_{j.HH}.$ is the total of corrected data on individuals with the A_HA_H genotype in generation F_j, $n_{j.HH}$ is the number of individuals with an observation in F_j and with the A_HA_H genotype, and $\tilde{y}_{j.LL}.$ and $n_{j.LL}$ is similarly for individuals with the A_LA_L genotype. For sampling of the dominant effect at the major locus, $\mathbf{k}=\mathbf{W}(0, 1, 1, 0)'$, which leads to similar sampling of d from $N(r/l, l^{-1})$, but where l and r can be worked out to be:

$$l=n_{1.LH\&HL}/\sigma_{e1}^2+n_{2.LH\&HL}/\sigma_{e2}^2$$
$$r=\tilde{y}_{1.LH\&HL}./\sigma_{e1}^2+\tilde{y}_{2.LH\&HL}./\sigma_{e2}^2$$

where now $\tilde{y}_{j.LH\&HL}.$ and $n_{j.LH\&HL}$ are totals of corrected data and numbers of individuals with an observation in F_j for individuals with the heterozygous genotype.

Sampling of genotypes and error variances are not based on linear model equations. Sampling of genotypes is affected by the heterogeneous variance structure for errors through modification of the penetrance function, which is used to compute (relative) probabilities for an observation on an individual, given alternative statuses for the genotype of the individual. The modified penetrance function used was:

$$f(\tilde{y}_{ji} \mid k) \propto \exp\{ -\tfrac{1}{2}(\tilde{y}_{ji}-\mu_k)^2/\sigma_{ej}^{\,2}\}$$

where $\tilde{y}_{ji}$ denotes corrected data on individual i of generation F_j, k indicates the various possible genotypes, and genotype k has mean μ_k. Computation of conditional genotype probabilities and sampling of new genotypes then follows as given by Janss et al. (1995). Sampling of error variances uses quadratics $\mathbf{e}_1'\mathbf{e}_1$ for error variance in F_1 and $\mathbf{e}_2'\mathbf{e}_2$ for error variance in F_2, and subsequently follows the sampling procedure as exemplified by Janss et al. (1995). Sampling of polygenic variance and of allele frequencies is not affected by the heterogeneous variance structure for errors, and therefore follows directly from the steps described by Janss et al. (1995).

General discussion 1. Application of segregation analysis and use of major genes

Chapter 7

Statistical methodology to model and detect major genes in livestock is advanced by introduction of a Bayesian approach to segregation analysis, feasible by virtue of Gibbs sampling methodology. Use of Bayesian approaches fits in a general trend to better account for uncertainty in statistical estimation procedures. In the statistical analyses of data from crossbred Meishan pigs, evidence was found for the presence of several major genes affecting traits of interest. Search for the actual genes and their gene products could generate more knowledge on the regulation of quantitative traits in general. Actual utilisation of these major genes in pig-breeding will require further genetic analyses, for instance to determine multivariate effects of the genes. For practical breeding, selection against the unfavourable recessive alleles of major genes affecting backfat and litter size will improve performance of a synthetic line. Utilisation of the favourable recessive allele of the major gene affecting intramuscular fat would require a sire-line that also contains the recessive allele. Then, litters can be produced that contain, for instance, 50% 'high intramuscular fat' piglets.

Statistical methodology

The main aim of this thesis was to investigate the presence of major genes in Meishan crosses. To do so, a large part of this thesis (Chapter 2-4) concentrated on development of statistical methodology to generally model a mixed inheritance and on tests to detect major genes in crossbreds.

Detection of major genes

Chapter 2 focused on power to detect major genes using F_1 and F_2 data using classical likelihood-ratio tests. The conclusions from Chapter 2 are expected to be equally valid for other inferential procedures, such as the Bayesian approaches applied in Chapters 5 and 6. Chapter 2 showed that identification of a major gene by segregation analysis using F_2 data is not very powerful when alleles at the major locus were fixed in the

founder populations. A separate study (Janss and Van der Werf, 1991) indeed showed that genes with smaller effects could be identified when alleles at the major locus segregated in the founder lines. In later analyses, therefore, segregation of alleles in founder lines was allowed for. The major genes identified all showed dominant gene-action. This appears plausible because dominant genes were found easier to identify and because dominant genes are more likely to segregate in the founder populations. Additive genes are more difficult to detect and would more likely have alleles fixed in the founder populations, therefore limiting the possibility to find additive genes.

Chapter 2 also showed that inclusion of F_1 data can lead to biased estimates and false conclusions regarding presence of a major gene when residual variances are not equal in F_1 and F_2. Hence, two residual variances were modelled when F_1 and F_2 data were analysed jointly in Chapter 6. Results from Chapter 6, where inferences from a combination of F_1 and F_2 data were compared with inferences from F_2 data only, nevertheless showed that robustness could remain poor when F_1 and F_2 data were combined. Likely, not only residual variances should be equal in F_1 and F_2, but also other distributional properties of the data such as skewness. Concerns raised in Chapter 2 on the robustness of segregation analysis when F_1 and F_2 data are combined, therefore, were confirmed in the practical analyses, and in general it can be concluded that care should be taken when F_1 and F_2 data are combined.

Analytical approaches to segregation analysis

Analytical approaches have been extensively investigated in human genetics, but approaches developed in human genetics can not be applied in large animal breeding pedigrees due to presence of many pedigree loops. Therefore, typically, software packages developed for analysis of human pedigrees can not be applied to analyse animal breeding pedigrees (e.g., Stricker et al., 1995). The iterative peeling approach described in Chapter 3 offers a solution to handle looped pedigrees. An alternative approach to handle looped pedigrees was proposed by Stricker et al. (1995). Both approaches recognised that exact computations were infeasible and developed an approximation by ignoring some dependencies arising due to pedigree loops. The approximations were developed using monogenic models, and extensions to also treat a mixed inheritance model with the same type of approximations are possible. Therefore, these approximations offer a significant advancement for the application of

analytical approaches to segregation analysis in animal breeding. The current value of such approaches lies mainly in the computations of genotype probabilities, for instance in genetic evaluations. Kinghorn et al. (1993) developed an iterative linear model approach for a mixed inheritance based on the same idea of iterative peeling which appears suited for such genetic evaluations.

Bayesian approaches to segregation analysis using Gibbs sampling

Gibbs sampling, or any Markov chain Monte Carlo method, offers another solution to handle pedigree loops, but then without requiring analytical approximation. At the same time, by use of Gibbs sampling, also some weaknesses in the estimation- and testing procedures of the likelihood-based analytical approaches to segregation analysis (see Chapter 4, 5) can be improved by applying this methodology in a Bayesian inference. In my view, the Bayesian approach to segregation analysis is to be preferred over the Maximum Likelihood (ML) approach from a theoretical viewpoint as well as from a practical viewpoint. A theoretical argument to reject ML approaches is that the properties of ML are only known asymptotically, and, hence, are not defined for any real situation. One practical argument to adopt the Bayesian approach is the handling of fixed effects. In the Bayesian approach, fixed effects are treated as nuisance parameters in a 'REML'-way, instead of an 'ML'-way.

Use of a Bayesian approach can be set in a wider perspective. In the application of statistics, better modelling of uncertainty is a general trend: BLUP, compared to BLP (selection index), takes into account uncertainty from the estimation of means or 'fixed effects'; REML, compared to ML, similarly takes into account uncertainty from the estimation of means in estimation of variance components. Marginal Bayesian estimators take into account uncertainty in a single parameter due to uncertainty in all other parameters in the model, and therefore seem a logic further step in this trend. Marginal Bayesian estimators have been proposed already, e.g. by Gianola and Foulley (1990) for the estimation of variance components. Here, for each variance component uncertainty was taken into account from estimation of other variance components. Harville and Carriquiry (1992) suggested the use of marginal Bayesian estimators for the estimation of breeding values accounting for uncertainty on variance components, and a similar idea was proposed by Sorensen et al. (1994) for the estimation of selection response accounting for uncertainty on variance components. In this thesis,

marginal Bayesian estimators were proposed for hyper-parameters in a mixed inheritance model, which can be viewed as a logic extension of the estimators proposed by Gianola and Foulley (1990) for linear models.

With the methodology presented in this thesis, use of segregation analysis can be expected to become a valuable aid in animal breeding for the identification of major genes affecting quantitative traits. Segregation analysis can be used complementary to linkage analysis, as each method has its strengths and weaknesses in particular situations. Segregation analysis will be valuable for analysis of field data which is primarily collected for different purposes. In such situations, genetic markers are generally not available, while phenotypic data is abundant. In contrast, linkage analysis would be a typical method for analysis of small experiments, where genetic markers are likely obtained as well and where segregation analysis would probably lack power. A combination of both approaches seems appropriate when in a large data set some animals are genotyped for genetic markers. In the search for functional genes, there are some subtle differences between the methods: segregation analysis directly identifies a functional gene and could genotype animals for such a functional gene, whereas linkage analysis is based on associations. Here, therefore, the two methods also could be used complementary to aid molecular geneticists in the identification of functional genes affecting quantitative traits.

Use of Gibbs sampling

From the experiences in using Gibbs sampling, I will shortly review here what can be considered to be the main problems in the use of Gibbs sampling:

(1) *Model building:* When using Gibbs sampling for Bayesian inferences, the statistical model should correspond to a proper (integrable) joint posterior distribution. This may not always be the case when certain non-informative priors are used. The danger is particularly apparent because 'Gibbs samplers' also can be constructed for such invalid applications (e.g. Hobert and Casella, 1993). One example of a model with an improper joint posterior distribution is a variance component model with so-called 'naive' priors for variance components (Hobert and Casella, 1993), notably a commonly used variance-component model (e.g. Box and Tiao, 1973). In our application we did not use the naive priors for variance components, but uniform priors. Hobert and Casella (1993) proved that use of uniform priors for variance components leads to a

proper joint posterior distribution in linear models. It seems plausible to assume that this conclusion is correct for a mixed inheritance model as well, as was done in Chapter 4, because the mixed inheritance model can be seen as a weighted sum of many linear models. It may be possible to detect construction of an improper joint posterior distribution by computation of the normalising constant for the (likelihood) x (prior) function (Hoeschele and Tier, 1995). When such an approach would not be feasible, integrability of the joint posterior distribution will have to be shown theoretically.

(2) *Construction of a (practically) irreducible chain:* Certain sampling schemes to construct Markov chains can lead to a reducible chain, or to a practically reducible chain, i.e. a poorly mixing chain. When modelling a single locus and using a single-variate sampling scheme to construct Gibbs samplers, irreducibility often does not hold (see e.g. Sheehan and Thomas, 1992). Moreover, also practical reducibility can arise, which can not be excluded a-priori on theoretical grounds. Therefore, a convergence diagnosing tool that compares output from multiple chains is to be preferred (see below). To alleviate (practical) reducibility, many variations on a straightforward single-variate sampling scheme can be developed and many such variations already have been proposed. In Chapter 4, for instance, so-called blocked sampling was used, as described by e.g. Smith and Roberts (1993) and Tanner (1993). Other variations can be described as using Metropolis schemes: the relaxation technique used in Chapters 5 and 6, which was suggested by Sheehan and Thomas (1992), can be seen as a Metropolis scheme within a single chain where Mendelian samples are accepted with probability 1 and others are rejected; Lin et al. (1993) also proposed Metropolis schemes using multiple chains. Further research in this area of Gibbs sampling schemes is expected to generate a large number of algorithms, a development which also has been seen for algorithms to solve linear models or to compute and maximise likelihoods.

(3) *Assessing convergence:* A good convergence diagnostic should give confidence that constructed chains moved freely through the entire parameter space. This will indicate practical irreducibility and, hence, one can be confident that the Markov chain indeed converged to the correct posterior distribution, provided that a proper joint posterior distribution was used. Comparison of between and within chain variances seems a simple and powerful method to conclude that chains moved freely through the

entire parameter space. Such comparison was suggested by Gelman and Rubin (1992), but their method lacked a definite test to conclude whether the between and within chain variance could be considered equal. The ANOVA approach used in Chapters 5 and 6 does supply such a test, which therefore appears to be a valuable extension. A practical difficulty to apply the ANOVA approach is that samples within a chain should be independent, or otherwise F-statistics will be inflated and non-convergence could be diagnosed. This procedure could be improved by computing within- and between chain variances based on dependent samples, and subsequently assume, for the computation of the F-statistic, variances having been computed on a fictitious smaller number of independent samples. This would not require to estimate before-hand a spacing to obtain independence of the samples. Further, in assessing convergence, it is important to realise that convergence will only occur for uniquely estimable parameters. When, for instance, fixed effects are over-parameterised, as was the case in our applications, only estimable contrast of fixed effects will appear to be equal in replicated chains and, hence, will appear to have converged.

Use of identified major genes

The application of the developed statistical methodology to the Meishan crosses demonstrated the presence of a number of major genes. Some of the genes identified could be of interest for pig-breeding, for instance genes influencing litter size, backfat thickness and intramuscular fat. However, showing presence of such genes only is a first step towards use of identified genes in actual breeding.

General genetic inferences of interest

Inferences on single genes presented in this thesis were based on univariate use of phenotypic data. When a major gene is found for two traits there is little information to determine whether this results from action of a pleiotropic gene or from action of two different genes. In this study, genotype probabilities were used to identify a possible pleiotropic effect, but such an approach may lack power. Although in certain cases, such as for the *MC* gene affecting cooking loss and pH measures, action of a pleiotropic gene is very plausible, in other cases reasonable uncertainty remains on this point. Further genetic inferences could therefore focus on:

(1) *Multivariate segregation analysis:* Phenotypic data could be exploited better by the use of multivariate approaches for segregation analysis. Models could be envisaged in which two linked genes affect two traits, and where recombination rate between the two loci is estimated. In such a model, also environmental and genetic variances and covariance should be estimated, in order to account for covariances between the observations. Significant non-linkage would then confirm existence of two different genes, while significant linkage confirms that genes affecting two traits are, at least, closely linked and possibly the same.

(2) *Linkage to a marker-map:* In pigs, a map with genetic markers has been made available (in Europe by Archibald et al., 1995), which can be used to link inferred major genes to specific linkage groups and to chromosomes. The aim of such analyses would be the same as the aim of the previously mentioned multivariate segregation analysis, i.e. to infer linkage between genes affecting different traits. The approach of investigating linkage to markers will be more powerful. The approach could also be applied to univariate data.

(3) *Search for functional genes:* The functional gene affecting a trait is identified when the gene-product is known and when mutations resulting in two or more different alleles can be identified at the DNA level. Finding of the actual gene affecting a trait allows unequivocal determination of the effects of such a gene on various other traits, and selection on one of the alleles, for instance for introgression, can be done with maximum efficiency. Also, determination of the physiological mechanism underlying the joint action of the gene on two or more traits will be the ultimate proof for pleiotropic action of such gene and will contribute to the understanding of the regulation of quantitative traits. As shortly discussed, major genes identified are functional genes and segregation analysis could aid in locating such genes by genotyping of the individuals for these genes.

Inferences and validation for use in breeding

For practical breeding, investigation of multivariate effects of the genes identified will generally suffice. For this purpose, the multivariate segregation analysis, possibly including information from markers, could be used. Resolving the existence of pleiotropic genes which cause an unfavourable association between traits would be an important aim for such an analysis. Existence of such pleiotropic genes could seriously

impair selection in the synthetic line. Using linked markers, also a two-step approach could be taken to investigate multivariate effects of genes. Firstly, chromosomal segments could be identified which likely carry one of the major genes identified; then, effects of such a chromosomal segments on various traits could be studied. Identification of the chromosomal segments that likely carry the major genes identified also would be beneficial for selection on the major genes and could indicate possible candidate genes which could be the major genes identified. Use of linked markers could additionally validate presence of the presumed autosomal genes with 2 alleles. Theoretically, the observed pattern of inheritance, for instance, could also have been caused by a two-locus system with interaction and such a situation could not simply be resolved by use of the phenotypic data alone.

An important validation for use in practical breeding also would be to validate effect of the alleles in different genetic backgrounds. In evolutionary genetics (e.g., Dawkins, 1976) it is argued that the effect of a gene depends on the genetic background present, i.e. on the collection of alleles at the same and at other loci. Such a dependency actually implies presence of dominance- and epistatic interactions, which seems plausible for loci affecting complexly regulated quantitative traits. As a result of such interactions, introgression could fail for a major locus affecting a quantitative trait (a QTL). In the context of development of a synthetic line, effect of a major gene could change as a result of selection on background loci, when alleles disappear which were necessary for expression of the gene in the original founder population or in the F_1 or F_2 population. Hence, it would be important to monitor traits and effects of a major gene during selection or introgression in order to avoid the loss of favourable alleles at background loci.

Selection on and use of identified major genes

In Chapter 6 value of the major genes found to affect litter size and backfat thickness were shortly discussed. Selection against the unfavourable recessive alleles of the major genes affecting litter size and backfat will slightly improve the level of the synthetic line. Not discussed was the impact on reduction of phenotypic variation, which could also be of importance for commercial pig-breeding. In the F_2, major genes affecting litter size and backfat accounted for a rough 30% of the phenotypic variation. Other advantages of eradication of the unfavourable alleles would be higher returns from

culled breeding stock when backfat is reduced, increased selection pressure when litter size is increased and higher accuracy of genetic evaluations when variance caused by the major gene is reduced.

As long as major genes segregate in the synthetic line, it will be beneficial for selection to include this knowledge in the model for genetic evaluation. If such a procedure is not used, animals with high merit on polygenes could be discarded because having low merit for the major gene, and this may not be optimal. When a mixed inheritance model is used, polygenic merit of animals and merit on the major locus can be obtained separately, which allows selection on each component independently. The finding of several major genes raises the problem of multivariate genetic evaluations considering several major genes and of optimisation of multivariate selection in such a situation. Such problems have not been addressed yet in theory.

For the major gene affecting intramuscular fat (*MI*), the recessive allele that increased intramuscular fat, here denoted MI^+, can be considered favourable, unless it is associated with, for instance, higher amounts of visible fat. From analyses so-far, *MI* did not appear to be the same as the gene influencing backfat, but this does not exclude that *MI* could have an effect on backfat in another way. When MI^+ is not unfavourable for other traits, maintaining and increasing its frequency in the synthetic line could be interesting. For use of MI^+ to increase intramuscular fat in commercial crossbred slaughter pigs the allele also should be present in a sire-line, because of its recessive nature. Screening of Western breeds used as sire-lines for presence of the MI^+ allele would therefore be interesting. If no such sire-line exists, one could introgress this allele in an existing sire-line, but in that case use of MI^+ will require large investments. If a sire-line would be found, or developed, that also contains the MI^+ allele, crossbred litters of slaughter pigs can be produced containing homozygous MI^+ animals with increased intramuscular fat. Production of crossbred litters containing 100% homozygous MI^+ animals may be difficult. A more interesting approach could be to produce litters containing 50% homozygous MI^+ animals by crossing a MI^+MI^+ boar, e.g. a pure-line boar from a line containing MI^+, with a heterozygous sow, e.g. a hybrid sow with one parent from a Meishan-synthetic line. By use of genetic markers, preferably a marker for the gene itself, homozygous MI^+ animals in such litters could be identified early after birth, and these animals could be placed in a special program to produce extra-tasty quality meat with an increased level of intramuscular fat.

References

Archibald AL, Brown JF, Couperwhite S, McQueen HA, et al. (1995) The PiGMaP consortium linkage map of the pig (Sus scrofa). Mamamalian Genome 6: 57-175

Box GEP, Tiao GC (1973) Bayesian statistical inference in statistical analysis. Addision-Wesley, Reading

Dawkins R (1976) The selfish gene. Oxford Univ Press, Oxford

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 4: 457-472

Gianola D, Foulley JL (1990) Variance estimation from integrated likelihoods (VEIL). Genet Sel Evol 22: 403-417

Harville DA, Carriquiry AL (1992) Classical and Bayesian prediction as applied to an unbalanced mixed linear model. Biometrics 48: 987-1003

Hobert JP, Casella G (1994) Gibbs sampling with improper prior distributions. Technical report BU-1221-M, Biometrics Unit, Cornell University

Hoeschele I, Tier B (1995) Estimation of variance components of threshold characters by marginal posterior modes and means via Gibbs sampling. Genet Sel Evol 27: 519-540

Janss LLG, Van der Werf JHJ (1991) Identification of a major gene in F_2 data when alleles are fixed in the parental breeds (abstract). Proc 42nd meeting of the Europ Assoc Anim Prod, september 1991, Berlin, Germany, vol 1: 121

Kinghorn BP, Kennedy BW, Smith C (1993) A method of screening for genes of major effect. Genetics 134 : 351-360

Lin S, Thompson E, Wijsman E (1993) Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. IMA J Math Med & Biol 10: 1-17

Sheehan N, Thomas A (1993) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics 49: 163 - 175

Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related markov chain monte carlo methods. J Roy Stat Soc B 55: 3-24

Stricker C, Fernando RL, Elston RC (1995) An algorithm to approximate the likelihood for pedigree data with loops by cutting. Theor Appl Genet 91: 1054-1063

Tanner MA (1993) Tools for statistical inference. Springer-verlag, New York

General discussion 2. Change of genetic variance in crosses and in selected (synthetic) lines

Chapter 8

Understanding the changes of genetic variance in crosses and in synthetic lines derived thereof could be an aid to optimise (multivariate) selection in such a synthetic line. Expected changes are an increase of genetic variance in the F_2, and a decrease of genetic variance relative to the F_2 in the later generations. Some indicative quantifications of these variance changes are made. To model and possibly extrapolate changes of genetic variance in a synthetic line due to selection, a finite locus model is proposed.

When a major gene is identified, selection on background genes will remain important in development of synthetic lines and likely also in approaches to introgress QTL's. Optimising selection schemes, especially when multiple traits are considered, will require knowledge on heritabilities of the traits and genetic and environmental correlations between the traits considered. In synthetic lines a complication arises, because genetic variances and covariances could change relatively quickly in the first generations after crossing of the founder lines, due to various effects: (1) change of allele frequencies in the cross to the average of allele frequencies in the founder lines; (2) change of genotype frequencies from non-Hardy-Weinberg proportions in the F_1 to Hardy-Weinberg proportions in the F_2; (3) change of allele frequencies in the synthetic line under selection; (4) introduction of linkage disequilibrium in the F_2, which is slowly broken down in the subsequent generations.

We define here 'true' change of genetic variance as the change in genetic variance caused by change of allele frequencies. The effect mentioned under (1) will then result in such true change of genetic variance in the cross relative to genetic variances in the founder lines and the effect mentioned under (3) will result in true change of genetic variance in the synthetic line. True change of genetic variance is excluded in the infinitesimal genetic model, in which a large number of genes, each with small effect, is assumed. However, in crosses between extreme lines and in synthetic lines derived from such crosses, one or a few genes with large effect could

segregate. In that case, ignoring changes in allele frequencies and ignoring true change of genetic variance is hardly tenable.

In the sequel, the above mentioned changes in genetic variances are described in more detail. In some cases, attempts are made to quantify these changes, although these quantifications largely remain indicative.

Change of genetic variance in the F_1 and F_2 of a cross

In the F_1 of a cross allele frequency will change to the average values of allele frequencies in the founder lines and genotype frequencies in the F_1 will show a departure from Hardy-Weinberg proportions. Linkage disequilibrium is not present in the F_1 because gametes of founder individuals that formed F_1 individuals can be assumed to have been in linkage equilibrium. The departure from Hardy-Weinberg proportions generally counterbalances the effect of allele frequency change and the F_1 has similar genetic variance as the average of genetic variances in the founder lines; under additivity of gene-effects, genetic variance in the F_1 actually is equal to the average of genetic variances in the founder lines, as follows, e.g., from Lande (1981).

In the F_2 population more marked changes will occur. Here, assuming random mating, genotypes will be in Hardy-Weinberg proportions. This allows the effect of allele frequency change that appeared in the F_1 to become apparent. This always causes an increase of genetic variance. Secondly, in F_2, a linkage disequilibrium will be created: for two loci on the same chromosome in an F_1 gamete, the probability for locus 1 to carry an allele that is more prevalent in, say, the paternal founder line depends on whether locus 2 carries an allele that is more or less prevalent in the paternal founder line.

Genetic variance in the F_2

Variance change due to allele frequency change, as becomes apparent in the F_2, equals, for each locus i, $\frac{1}{8}d_i^2$, where d_i is the difference between founder lines explained by locus i (e.g. Lande, 1981). Then, for a number of loci n explaining the total difference between founder lines D, variance increase in the F_2 will depend on n and on the variation of d_i values (Lande, 1981, eq. 5). In general, when n is smaller and $\mathrm{Var}(d_i)$ is larger, variance increase in F_2 will become larger. For n approaching infinity, variance increase in the F_2 will approach zero. When there is reasonable variation in

d_i values, the few loci with largest d_i generally account for a large portion of the variance increase. When, for instance, n=10 and d_i's are 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, ..., $\frac{1}{10}$, the two loci with largest d_i's account for 81% of the total increase of genetic variance arising in the cross. When n=10 and d_i's are 1, $\frac{1}{9}$, $\frac{1}{8}$, $\frac{1}{7}$, ..., $\frac{1}{10}$, the two loci with largest d_i's account for only 47% of the total increase of genetic variance arising in the cross. In this second case, Var(d_i) is smaller relative to the total difference between founder lines, and there is no markedly 'major' locus, because the second largest locus explains about the same difference as the largest locus.

Variance increase arising in the F_2 also conveniently can be described by use of a conceptual 'effective number of loci', n_e, defined to explain equal differences between the founder lines (Castle, 1921). For such an effective number, variance increase in the F_2 is described as $\frac{1}{8}D^2/n_e$, and n_e is a parameter that describes the proneness of genetic variance to change (increase) in a cross for a particular trait. For the above example with d_i's 1, $\frac{1}{2}$, ..., n_e=5.5, indicating reasonable proneness of genetic variance to change, and for the example with d_i's 1, $\frac{1}{9}$, ..., n_e=7.9, indicating less proneness of genetic variance to change. In these comparisons, change of genetic variance is relative to D^2. For practical application to crosses between outbred lines, such as the Meishan x Western cross, increases of genetic variance in the F_2 generally corresponding to n_e values between 5 and 10 (Lande, 1981), and therefore could typically range from $D^2/40$ to $D^2/80$. However, presence of a major gene which explains a large portion of a difference between two extreme lines could make the variance increase much larger, while for crosses between relatively similar lines variance increases could be much smaller. It should be noted that the variance increases reported by Lande (1981) are totals of variance increase as commonly observed in practice, which will include effects of change of allele frequencies as well as effects of linkage disequilibrium.

A second effect contributing to change of genetic variance in the F_2 is the mentioned linkage disequilibrium. In the F_2, linkage disequilibrium creates a covariance c_{ij} between two loci i and j on the same chromosome of $c_{ij}=\frac{1}{8}(1-2r_{ij})d_i d_j$, where r_{ij} is the recombination rate between locus i and locus j, d_i the difference between founder lines explained by locus i and d_j the difference between founder lines explained by locus j (Lande, 1981). When considering loci i and j to affect the same trait, variance of this trait is increased with twice the given covariance. The given

covariance can be positive as well as negative, dependent on the signs of d_i and d_j. In a cross d_i and d_j will tend to be more frequently of the same sign, such that the overall effect of created linkage disequilibrium will be an increase of genetic variance.

A value for the expectation of the minimum increase of genetic variance due to linkage disequilibrium can be obtained using an effective number concept as $(1\text{-}2\bar{r})D^2/8$. Here $\bar{r}$ is an average recombination rate between all loci over chromosomes. This minimum expected value does not depend on the (effective) number of loci n_e. Therefore, this variance increase due to linkage does not approach zero when the number of loci approaches infinity, as was the case for the increase of genetic variance due to change of allele frequencies. For pigs, assuming randomly placed loci, $\bar{r}$ is ≈0.48. This value is close to 0.5 because most loci will not be on the same chromosome, having r=0.5. A value of 0.48 for $\bar{r}$ results in a minimum expected increase of genetic variance due to linkage of $D^2/200$. Because this is an expected minimum, actual variance increase due to linkage could well be larger and could be a non-negligible proportion of the total increase of genetic variance in the F_2. This is generally noted; for instance, Zeng et al. (1990) considered linkage to be a significant disturbance for estimation of the effective number of loci.

Change of genetic covariances

Above mentioned effects on genetic variances, also can affect genetic covariances. Change of genetic covariances due to change of allele frequencies can arise for pleiotropic loci affecting traits. When such pleiotropic loci segregated in founders, these loci also would have caused genetic covariances within the founder lines, but at average smaller due to differences in allele frequencies. Pleiotropic loci that did not segregate in founders would not have caused genetic covariances within the founder lines, but only genetic covariances between founder lines. Such mechanism of change of allele frequencies at pleiotropic loci is one mechanism that can introduce genetic covariances in the cross formerly not (strongly) present within the founder lines. Linkage disequilibrium also can cause change of genetic covariances when loci affecting two traits are linked. For covariances between loci affecting different traits, d_i and d_j values will tend to be of the same sign as the between founder line covariance. For the cross between Meishan and Western lines, for instance, a positive covariance exists between founder lines for fertility and fatness, so that linkage

disequilibrium is expected to introduce a similar genetic association in Meishan-Western crossbreds. This mechanism of linkage between loci affecting two traits is a second mechanism that can introduce genetic covariances in the cross formerly not present within the founder lines. Both mechanisms will introduce genetic covariances that, in most practical applications, will be unfavourable.

Genetic variances in generations after the F_2

No selection

When no selection would be practised, allele frequencies will not change except by random drift, and increased genetic variance in F_2 caused by change of allele frequencies will remain in the generations following the F_2. Also increased genetic covariances caused by change of allele frequencies at pleiotropic loci will remain in the generations following the F_2. Linkage disequilibrium is expected to reduce due to recombination of chromosome segments. Due to this break-down of linkage disequilibrium, the parts of increased genetic variances and covariances caused by linkage will gradually vanish. The rate of linkage break-down for each locus will be geometric in the covariances, i.e. c_{ij}^{t}/c_{ij}^{t-1} will be constant. But, the rate of break-down for the total of all covariances will not be geometric, and this rate will decrease: in the first generations, reduction of covariance is mainly caused by the loosely linked loci, with a fast rate of break-down, but in the later generations loosely linked loci will approach equilibrium and the rate of break-down will be determined more by the tightly linked loci. The fact that the rate of break-down in a given generation will be at most equal to the rate of break-down in the preceding generation would be valuable, but the absolute level of this rate remains difficult to determine because it requires knowledge on the proportion of variance increase in the F_2 that can be attributed to the effect of linkage disequilibrium.

Selection

In order to describe the change of genetic variance under selection, here a finite locus model based on the 'effective number' concept is introduced. For a finite number of loci, selection will change allele frequencies, and hence will change the 'true' genetic variance. Such a possible change is considered here, because in a synthetic line a few genes with relatively large effect may segregate, such that genetic variance may indeed

be prone to change under selection. Proneness of genetic variance to change under selection in the synthetic is intuitively related to proneness of genetic variance to change from founder lines to the F_2, as could be described by the effective number of loci n_e. The finite locus model introduced here similarly uses n_e to describe proneness of genetic variance to change under selection. Using the concept of an effective number of loci, the relationship between change in genetic mean μ_g and change in genetic variance V_g can be expressed by the differential equation (Park, 1977):

$$n_e dV_g = -\mu_g d\mu_g \qquad (1)$$

Derivation of this equation assumes additivity and equal effects and equal frequencies of the 'effective loci' and the obtained relationship depends on the genetic mean μ_g, which should generally be considered unknown. From equation (1) many known estimators for the effective number of loci from crossbred data can be deduced (see Park, 1977). Also interesting is that for one particular estimator of the effective number of loci from crossbred data, Ollivier and Janss (1993) showed an extension to include dominance effects. When such extension also would be possible for the general equation (1), this could supply means to model inbreeding depression with a finite locus model. To describe variance change under selection using (1), consider for a generation 1 to a generation 2 changes in genetic mean from $\mu_{g,1}$ to $\mu_{g,2}$ and changes in genetic variance from $V_{g,1}$ to $V_{g,2}$. Then, integrating equation (1) between the bounds set by these changes, leads to:

$$n_e(V_{g,2}-V_{g,1}) = \tfrac{1}{2}(\mu_{g,2}^2 - \mu_{g,1}^2)$$

and on substituting $\mu_{g,2}=\mu_{g,1}+R$, where R is the selection response,

$$V_{g,2}-V_{g,1} = -\mu_{g,1}R/n_e - \tfrac{1}{2}R^2/n_e \qquad (2)$$

Equation (2) shows that using the concept of an effective number, change in genetic variance is a function of the squared response with an additional linear term in response dependent on the genetic mean. Use of three or more generations in which genetic variances and population means (transformed to responses) are measured,

allows to fit the second order linear function suggested by (2) and allows estimation of the effective number. For an application to predict variance changes in a synthetic line derived from an F_2, an initial crude approximation could use $\mu_{g,1}=0$, which corresponds to assuming all 'effective loci' to have allele frequencies of $\frac{1}{2}$. When some generations of the synthetic line are obtained, prediction of further changes can be improved by also estimating $\mu_{g,1}$. Hence, n_e can be used to describe proneness of genetic variance to change in the F_2 of a cross, as well as proneness of genetic variance to change under selection for a particular trait in a particular population. In some limited simulation studies, equation (2) was found quite apt to model variance change and to extrapolate such change even up to the selection limit. In these simulations, a number of additive loci with variable effects and frequencies was used.

Selection would also interfere with the break down of linkage disequilibrium. Variance increase due to linkage disequilibrium is based on favourable alleles originating from one of the founder lines to remain coupled in the cross, and selection will favour individuals in which this coupling is still present. For such loci, therefore, linkage break down will be retarded. For linked loci that cause an unfavourable genetic covariance between two traits, however, recombinants, which no longer show the unfavourable association, would be favoured in selection. Hence, in such a case, selection would speed up linkage break-down and would more rapidly reduce unfavourable covariances then expected under random mating.

Conclusions

Knowledge of the changes of genetic variance that occur when two lines are crossed and when a line is selected could aid to optimise selection in a synthetic line. In the F_2, increased genetic variances and covariances are generally expected. For univariate selection, the increased genetic variance could be used to advantage, but increased covariances will generally be unfavourable and will limit selection pressure to be applied on each single trait. As part of these increases are caused by linkage, the generally unfavourable covariances will reduce, offering better opportunities for selection in later generations. Because of this, Bidanel et al. (1991) suggested random breeding of a few generations before start of selection in the synthetic line. Use of a large population and a mild selection pressure could also be a valuable strategy to increase the chances of favourable recombinants appearing and to maintain individuals

with a favourable recombination for breeding.

Part of the increased variances and covariances, however, may have a more permanent nature, caused by change of allele frequencies. Change of allele frequencies at pleiotropic loci could introduce unfavourable genetic covariances which form an impediment for selection in the synthetic line. It seems very difficult, but also very important, to determine whether an unfavourable genetic covariance is caused by linkage and will gradually reduce, or whether an unfavourable genetic covariance is caused by segregation at pleiotropic loci and will not reduce. Search for (major) genes affecting the traits could be useful to better understand the genetic covariances and to determine whether selection in the synthetic line can be successful. In this context, future molecular genetic research on the Meishan crossbreds could aid to determine the value of a Meishan synthetic line for commercial pig-breeding.

To optimise multivariate selection, knowledge on genetic variances and covariances will be important. As these parameters may change in a synthetic line, monitoring of these parameters is useful. A finite locus model was introduced which could be valuable to extrapolate trends in genetic variances and covariances, although in practice the estimation of such trends could suffer from large inaccuracy. This finite locus model could also be of use to describe change of genetic variance under selection in general outbred populations.

References

Bidanel JP, Caritez JC, Legault C (1991) Ten years of experiments with Chinese pigs in France. 2. Utilisation in crossbreeding. Pig News Inform 12: 239-243

Lande R (1981) The minimum number of genes contributing to quantitative variation between and within populations. Genetics 99: 541-553

Ollivier L, Janss LLG (1993) A note on the estimation of the effective number of additive and dominant loci contributing to quantitative variation. Genetics 135: 907-909

Park YC (1977) Theory for the number of genes affecting quantitative characters. I. Estimation of and variance of the estimation of gene number for quantitative traits controlled by additive genes having equal effect. Theor Appl Genet 50: 153-161

Zeng ZB, Houle D, Cockerham CC (1990) How informative is Wright's estimator of the number of genes affecting a quantitative character? Genetics 126: 235-247

Summary

Litter size is an important characteristic in pig breeding. Apart from selection within available lines, also the development of a synthetic line with the Chinese Meishan breed could be an interesting approach to obtain a line with an increased level of litter size. To investigate genetic aspects of traits of interest in such a synthetic line, Dutch pig breeding companies have produced F_1 and F_2 Meishan x Western crossbreds. This thesis focusses on one important genetic aspect, the presence of major genes. In Chapters 2 to 4, statistical methodology to model a major gene inheritance is investigated and developed; Chapters 5 and 6 consider analysis of data collected on the produced Meishan crossbreds for presence of major genes. To develop a synthetic line with Meishan, presence of major genes affecting litter size, growth and fatness is of interest. Additionally, the presence of major genes is investigated for meat quality traits.

Statistical methodology

In Chapter 2, the possibility to detect major genes by use of F_1 and F_2 is investigated. Here, special attention is paid to the situation where alleles at the major locus are fixed in the founder populations. Using 1000 F_2 observations, the power to detect major genes reaches more than 95% for additive and completely dominant effects (difference between homozygotes) of 4 and 2 residual standard deviations, respectively. When F_1 data is included, any increase in variance from F_1 to F_2 biases parameter estimates and leads to putative detection of a major gene. Also when in reality alleles at the major locus segregate in the founder populations, parameter estimates become biased, unless the average allele frequency in the founder populations is exactly 0.5. Use of data and use of a model in which alleles segregate in parents, e.g. F_3 data, is concluded to give better robustness and larger power. The latter is confirmed in a separate study, as referenced in Chapter 7, which shows that effects up to 4 times as small can be detected when alleles at the major locus segregate in the founder lines. Based on the findings in Chapter 2, Chapters 3 and 4 focus on the development of general models for a mixed inheritance. Use of such models is referred to as 'segregation analysis'.

In Chapter 3, an advancement is made for use of analytical approaches to segregation analysis. It is noted that animal breeding pedigrees, as opposed to human

pedigrees, generally contain many loops, such that exact computation of likelihoods is infeasible. Loops in animal breeding pedigrees arise due to multiple matings, i.e. sires are generally mated to several dams, and due to inbreeding. Multiple matings generally already create many loops when considering 3-generation pedigrees. In Chapter 3, 'iterative peeling' is introduced, a method equivalent to the traditional recursive peeling method to compute exact likelihoods in non-looped pedigrees, but which also can be used to obtain approximate likelihoods in looped pedigrees. In simulations, hypothesis testing and parameter estimation are compared based on approximated likelihoods in looped pedigrees and exact likelihoods in non-looped pedigrees. This shows that no biases are introduced by the approximation in looped pedigrees. Iterative peeling is developed and investigated using a monogenic model, but could be extended to compute likelihoods for a mixed inheritance model. Such extension, however, was not made because an alternative non-analytical approach became available and was developed in Chapter 4.

In Chapter 4, the application of Gibbs sampling is considered for inference in a mixed inheritance model. Gibbs sampling is a Markov chain Monte Carlo procedure which does not require analytical approximation. The approximation in such an approach is of a different nature: a marginal posterior distribution, or a feature thereof, is estimated based on a finite sample from the true posterior distribution. To generate such a sample, a Markov chain is constructed with an equilibrium distribution equal to the posterior distribution to be approximated. For application of Gibbs sampling to a mixed inheritance model, an implementation on scalar components, as used for human populations, appears not efficient because mixing of parameters in the Markov chain is slow. Therefore, an approach with blockwise sampling of genotypes is proposed for use in animal populations. The blockwise sampling, by which genotypes of a sire and its final progeny were sampled jointly, is effective to improve mixing. In Chapter 4 it is concluded that further measures to improve mixing could be looked for. In later Chapters such a further improvement is found in the additional use of a relaxation technique. In Chapter 4, inferences are made from a single Gibbs chain. In later Chapters, this approach is improved by use of multiple chains from which convergence of the Gibbs sampler is assessed by comparison of between- and within chain variances in an analysis-of-variance. The use of Bayesian estimators, which is feasible when using Gibbs sampling, is found preferable over the use of classical

maximum likelihood estimators. In Chapter 7, it is discussed that the use of Bayesian procedures fits in a general trend to better account for uncertainty in statistical estimation procedures.

Analysis of data

In Chapters 5 and 6, analysis of data obtained on the Meishan crossbreds is presented. In Chapter 5, presence of major genes affecting meat quality traits is investigated using data from F_2 individuals. Cooking loss, drip loss, two pH measurements, intramuscular fat, shearforce and back-fat thickness (by HGP measurement) are found to be likely influenced by a major gene. In all cases, a recessive allele is found, which originates from one of the founder lines, likely the Meishan breed. By studying associations between genotypes for major genes affecting the various traits, it is concluded that cooking loss, two pH measurements and possibly backfat thickness are influenced by one gene, and that a second gene influences intramuscular fat and possibly shearforce and drip loss. The statistical findings are supported by demonstrating marked differences in variances of families of fathers inferred as carriers and families of fathers inferred as non-carriers.

In Chapter 6, presence of major genes is investigated for two growth traits, backfat thickness (by ultrasonic measurement) and litter size at first and second parity, using data from F_1 and F_2 crossbreds. Here, two analyses are performed for each trait. In a first analysis, joint analysis of F_1 and F_2 crossbred data is performed, in which different error variances are fitted for F_1 and F_2 observations. In this first analysis, significant contributions of major-gene variance are found for the two growth traits, for backfat, and for litter size at first parity. In a second analysis, analysis of F_2 data only is performed to check whether no biases are introduced in the joint analysis of F_1 and F_2 data. In the second analysis, no major genes are found for growth traits. Major genes affecting backfat and litter size at first parity are confirmed. Effects of the gene affecting backfat are similar to the effects of the gene affecting backfat identified in Chapter 5, and this likely is the same gene. The major genes affecting backfat and litter size are dominant genes, of which the recessive alleles can be considered unfavourable: the recessive alleles of these genes cause an increase of backfat and a decrease of litter size.

General results from the statistical analyses indicate that further molecular

genetic research effort to map these genes will have a high probability of success. In Chapter 7 benefits are discussed from selection against the recessive alleles of the genes influencing backfat and litter size, as well as use of the gene affecting intramuscular fat to produce extra-tasty quality meat.

Conclusions

In this thesis, segregation analysis (SA) is made applicable for use in animal populations. SA will be a valuable addition to linkage analysis, where SA will be more typically applied to large amounts of data which are routinely collected. In the search for genes affecting quantitative traits, SA can directly identify functional genes, and can estimate genotypes of animals for such a functional gene. In combination with linkage analyses, this could supply important aids for molecular geneticists to locate functional genes. In this thesis, a number of major genes was identified to affect traits in the Meishan crosses. Further genetic analyses could generate more knowledge on the regulation of the quantitative traits involved and will aid in assessing the value of these genes for practical breeding. Chapter 8 additionally describes expected variance changes in a synthetic line, which could aid to optimise selection in such a line.

Samenvatting

Inleiding

De veredeling van varkens in Nederland wordt gedaan door gespecialiseerde fokkerij-organisaties. Zulke fokkerij-organisaties verkopen, onder andere, jonge moederdieren die op vermeerderingsbedrijven gebruikt worden voor de produktie van slachtvarkens. De ideale moederdieren moeten veel biggen werpen die vitaal zijn en die goede eigenschappen hebben voor de mesterij. De worpgrootte van de moederdieren is deels genetisch bepaald, en daarbij van grote economische waarde, zodat worpgrootte een van de belangrijke aandachtspunten is in de veredeling van de zogenaamde moederrassen of -lijnen.

Worpgrootte zou verbeterd kunnen worden door het benutten van genetische variatie binnen een lijn middels selectie. Een tweede mogelijkheid is het benutten van genetische variatie tussen lijnen of rassen door kruising. Gekruiste dieren kunnen dan gebruikt worden als stamouders voor een nieuwe zogenaamde 'synthetische' lijn. Afhankelijk van de genetische achtergronden van de belangrijke kenmerken kunnen in een dergelijke synthetische lijn de eigenschappen van de uitgangslijnen mogelijk gecombineerd worden. In Nederland hebben 5 fokkerij-organisaties een kruisings-experiment uitgevoerd om te onderzoeken of verbetering van de toomgrootte mogelijk is door zulk een synthetische lijn te ontwikkelen. Hierbij werden kruislingen (eerste generatie kruislingen of F_1's, en tweede generatie kruislingen of F_2's) geproduceerd tussen het zeer vruchtbare Chinese Meishan ras en lokale Europese rassen. Waarnemingen aan deze Meishan kruislingen werden gebruikt voor statistisch-genetische analyses om de genetische achtergrond van belangrijke kenmerken te onderzoeken en zodoende de haalbaarheid voor de vorming van een synthetische lijn te bepalen.

Onderzoek naar de genetische achtergrond van kenmerken in de Meishan kruislingen is in dit proefschrift toegespitst op één onderwerp, de mogelijke aanwezigheid van hoofdgenen. Een hoofdgen is een enkel gen dat in belangrijke mate, maar niet geheel, de vererving van een kenmerk bepaalt. De genen welke het resterende deel van de overerving bepalen worden achtergrondgenen genoemd. De vererving van een kenmerk dat beïnvloed wordt door een hoofdgen is daarom deels discreet en deels continu: het hoofdgen zorgt voor een discrete overerving waarbij doorgaans slecht 2 of 3 verschillende genetische varianten of 'genotypen' bestaan; de achtergrondgenen zorgen

voor een continue overerving waarbij er een continuüm van genetische varianten bestaat. Het genetische model waarbij uitgegaan wordt van een hoofdgen en achtergrondgenen, wordt dan ook wel een gemengd overervingsmodel genoemd. In de Hoofdstukken 2 tot en met 4 van dit proefschrift worden statistische methodes voor het modelleren van een gemengde overerving onderzocht en ontwikkeld. In de Hoofdstukken 5 en 6 worden vervolgens analyses gepresenteerd van waarnemingen aan de Meishan kruislingen om te onderzoeken of hoofdgenen inderdaad een rol spelen bij de overerving van bepaalde kenmerken. Voor de ontwikkeling van een synthetische lijn met Meishan is de aanwezigheid van hoofdgenen voor worpgrootte, groei en vetheid interessant (Hoofdstuk 6). Daarnaast worden er ook analyses van aantal vleeskwaliteitskenmerken gepresenteerd (Hoofdstuk 5).

Statistische methoden

In Hoofdstuk 2 wordt onderzocht of het mogelijk is de aanwezigheid van een hoofdgen te bepalen door gebruik te maken van waarnemingen aan de F_1 en F_2 kruislingen. Hierbij is speciaal aandacht gegeven aan een situatie waarbij de verschillende allelen van het hoofdgen gefixeerd waren in de stamlijnen. Door gebruik te maken van 1000 waarnemingen aan de F_2's is de kans meer dan 95% om aanwezigheid te detecteren van een additief hoofdgen met een effect (verschil tussen homozygoten) van 4 residuele standaard deviaties of een dominant hoofdgen met een effect van 2 residuele standaard deviaties. Wanneer ook waarnemingen aan de F_1 kruislingen worden gebruikt leidt elke verhoging van variantie tussen de F_1's en F_2's tot onzuiverheid van de schatting van met name het effect van het hoofdgen en tot een mogelijke abusievelijke detectie van een hoofdgen. Ook wanneer in werkelijkheid de allelen van het hoofdgen segregeren in de stamlijnen worden effecten onzuiver geschat, tenzij de gemiddelde allelfrequentie in de stamlijnen exact 0.5 is. Er wordt geconcludeerd dat hoofdgenen beter te detecteren zijn door waarnemingen te gebruiken waarbij allelen van het hoofdgen segregeren in de stamlijnen, of door waarnemingen van een F_3 te gebruiken. Dit is bevestigd in een aparte studie, aangehaald in Hoofdstuk 7, waarin wordt getoond dat hoofdgenen met een tot 4 keer kleiner effect detecteerbaar zijn wanneer allelen van het hoofdgen segregeren in de stamlijnen. Gebaseerd op de conclusies uit Hoofdstuk 2 is in Hoofdstukken 3 en 4 de aandacht gericht op het ontwikkelen van algemene modellen voor het beschrijven van een gemengde overerving. Het gebruik van zulke modellen

en detectie van een hoofdgen op basis van zulke modellen wordt aangeduid als segregatie analyse.

In Hoofdstuk 3 wordt een bijdrage geleverd voor een analytische toepassing van segregatie analyse. Een analytische toepassing is erg moeilijk omdat de benodigde berekeningen onuitvoerbaar worden wanneer in een populatie afstammingslussen voorkomen. Afstammingslussen worden veroorzaakt door meervoudige paringen en door inteelt en blijken in populaties van landbouwhuisdieren veelvuldig voor te komen. Daarom is een iteratieve peeling-methode voorgesteld die equivalent is aan de traditionele recursieve peeling-methode voor het exact berekenen van waarschijnlijkheden wanneer afstammingslussen niet voorkomen, maar die ook bruikbaar is om een benaderde waarschijnlijkheid te berekenen wanneer afstammingslussen wel voorkomen. Met simulatiestudies worden statistische toetsen en parameterschattingen vergeleken gebaseerd op exact berekende waarschijnlijkheden bij afwezigheid van afstammingslussen en gebaseerd op benaderde waarschijnlijkheden bij aanwezigheid van afstammingslussen. Hieruit blijkt dat er geen onzuiverheid geïntroduceerd wordt door het gebruik van de benadering bij aanwezigheid van afstammingslussen. De iteratieve peeling-methode is ontwikkeld en onderzocht voor een monogene overerving (een enkel gen zonder additionele achtergrondgenen), maar zou uitgebreid kunnen worden naar een gemengde overerving. Een dergelijke uitbreiding echter is niet gemaakt omdat een alternatieve niet-analytische benadering bekend werd en ontwikkeld is in Hoofdstuk 4.

In Hoofdstuk 4 is de toepassing van Gibbs-sampling beschreven om parameters van een gemengd overervingsmodel te kunnen schatten. Gibbs sampling, in tegenstelling tot de hiervoor beschreven analytische benadering, is een Monte-Carlo-Markov-keten benadering. Deze benadering heeft een geheel ander karakter: met zulk een benadering wordt een marginale a-posteriori verdeling, of een kenmerk daarvan, geschat middels een beperkt aantal trekkingen die gegenereerd worden uit de werkelijke marginale a-posteriori verdeling. Om de gewenste trekkingen te genereren wordt een Markov keten geconstrueerd waarvan de evenwichtsverdeling de te benaderen a-posteriori verdeling is. Een toepassing van Gibbs sampling waarbij de parameters van het gemengde overervingsmodel als scalairen worden behandeld, blijkt te resulteren in een slecht mengen van de parameters in de Markov keten. Daarom is een toepassing voorgesteld waarbij genotypen in blokken worden behandeld, hetgeen resulteert in een

beter mengen van de parameters. In latere hoofdstukken is nog een verdere verbetering van het mengen van de parameters in de Markov keten bereikt door ook een relaxatie-techniek te gebruiken. Parameterschattingen in Hoofdstuk 4 zijn gebaseerd op een enkele Markov keten. In latere hoofdstukken is deze procedure verbeterd door meerdere Markov ketens te gebruiken waarbij convergentie van de Gibbs-sampler bepaald is door de vergelijking van binnen- en tussen keten-varianties in een variantie-analyse. Het gebruik van Bayesiaanse schatters, wat mogelijk is wanneer Gibbs-sampling wordt gebruikt, wordt geprefereerd boven het gebruik van klassieke hoogste-waarschijnlijkheids-schatters. In Hoofdstuk 7 wordt bediscussieerd dat het gebruik van zulke Bayesiaanse schatters past in een algemene trend om beter rekening te houden met onnauwkeurigheid in statistische schattingsprocedures.

Data-analyse

In Hoofdstukken 5 en 6 worden analyses gepresenteerd van de waarnemingen aan de Meishan kruislingen. In Hoofdstuk 5 wordt de aanwezigheid van hoofdgenen onderzocht voor vleeskwaliteitskenmerken, gebruikmakend van waarnemingen aan F_2 kruislingen. Hier wordt aangetoond dat kookverlies, vochtverlies, pH, intramusculair vet en snijweerstand van het vlees, alsmede de rugspekdikte (middels een HGP-meting), waarschijnlijk door een hoofdgen worden beïnvloed. Voor al deze kenmerken is een recessief allel gevonden dat afkomstig is van een van de uitgangslijnen, waarschijnlijk het Meishan ras. Door de associaties te bestuderen tussen de genotypen van dieren voor de hoofdgenen voor elk van de kenmerken, is geconcludeerd dat kookverlies, pH en mogelijk rugspekdikte beïnvloed worden door hetzelfde gen, terwijl een tweede gen intramusculair vet en mogelijk snijweerstand en vochtverlies beïnvloedt. Deze bevindingen worden ondersteund door het aantonen van markante variantieverschillen tussen families van vaders die geïdentificeerd zijn als drager van het recessieve allel en families van vaders die geïdentificeerd zijn als niet-drager van het recessieve allel. In Hoofdstuk 6 wordt de aanwezigheid van hoofdgenen onderzocht voor twee groeikenmerken, rugspekdikte (middels een ultrasone meting) en worpgrootte bij eerste en tweede pariteit, gebruikmakend van waarnemingen aan F_1 en F_2 kruislingen. Hier worden twee analyses per kenmerk beschreven. In een eerste analyse zijn waarnemingen van F_1 en F_2 kruislingen gezamenlijk geanalyseerd, waarbij voor de F_1 en F_2 verschillende residuele varianties worden gemodelleerd. In deze eerste

analyse worden significante bijdragen van een hoofdgen gevonden voor de groeikenmerken, rugspekdikte en voor de worpgrootte bij eerste pariteit. Vervolgens wordt voor elk kenmerk een tweede analyse beschreven waarbij alleen waarnemingen van de F_2 kruislingen worden gebruikt om te controleren of er in de eerste gezamenlijke analyse geen onzuiverheden in de schattingen geïntroduceerd zijn. In deze tweede analyse worden geen hoofdgenen gevonden voor de groeikenmerken. Invloed van een hoofdgen op rugspekdikte en op toomgrootte wordt bevestigd. De effecten van het hoofdgen dat rugspekdikte beïnvloedt komen overeen met de effecten van het gevonden hoofdgen voor rugspekdikte in Hoofdstuk 5, en deze genen zijn waarschijnlijk dezelfde. De hoofdgenen voor rugspekdikte en worpgrootte hebben recessieve allelen die als ongunstig aangemerkt kunnen worden: respectievelijk rugspek verhogend en worpgrootte verlagend. De algemene resultaten van de statistische analyses tonen aan dat moleculair-genetisch onderzoek om de geïdentificeerde genen op het genoom te lokaliseren een goede kans van slagen heeft. In Hoofdstuk 7 worden de voordelen bediscussieerd van selectie tegen de ongunstige recessieve allelen van de hoofdgenen voor rugspekdikte en worpgrootte, alsmede de mogelijkheid om extra smakelijk kwaliteitsvlees met een verhoogd gehalte intramusculair vet te produceren.

Conclusies

In dit proefschrift is segregatie analyse (SA) algemeen toepasbaar gemaakt voor gebruik in populaties van landbouwhuisdieren. SA kan een nuttige aanvulling zijn op koppelingsanalyse, waarbij SA typisch toepasbaar is voor grote aantallen waarnemingen die routinematig zijn verzameld. In onderzoek naar genen welke kwantitatieve kenmerken beïnvloeden kan SA direct een functioneel gen identificeren en kunnen genotypen van dieren voor een dergelijk functioneel gen geschat worden. In combinatie met koppelingsanalyse kan dit een belangrijk hulpmiddel zijn voor moleculaire genetici bij het lokaliseren van functionele genen. In dit proefschrift worden een aantal hoofdgenen geïdentificeerd die kenmerken in Meishan-kruislingen beïnvloeden. Verder genetisch onderzoek kan de kennis over de regulatie van de onderzochte kenmerken vergroten en zal van belang zijn in het bepalen van de waarde van deze genen voor de praktische veredeling. Als toevoeging is in Hoofdstuk 8 beschreven hoe de genetische variantie kan veranderen in kruislingen en in een synthetische lijn, wat behulpzaam kan zijn voor de optimalisatie van een selectieprogramma in een synthetische lijn.

About the author

Lucas Lodewijk Gustave (Luc) Janss was born on December 9, 1965 in 's-Hertogenbosch, The Netherlands. After classical grammar-school, he studied Animal Science at the Wageningen Agricultural University (WAU), starting in 1985 and graduating in 1990 with a major in Animal Breeding/Quantitative Genetics and a minor in Statistics. Research for his major in Animal Breeding was conducted at INRA, Jouy-en-Josas, France, working on threshold models for the genetic analysis of difficult births in cattle. From 1990, he was a PhD-research-fellow at the Department of Animal Breeding of WAU, with the assignment to analyse data from crosses with Chinese pigs for the presence of major genes. He succeeded in doing so by development of a novel Bayesian approach to segregation analysis using Gibbs sampling. Part of the research on the application of segregation analysis was conducted at the AFRC Roslin Institute near Edinburgh, Scotland. During his PhD-research-fellowship he also was involved in diverse other genetic analyses, ranging from mortality in chickens and in piglets to milk production in cattle. Interrupted by a military service, where he was trained in First Aid and acted as an ambulance-driver, he finished his thesis in 1996, describing the research on major genes. At present, he is appointed as a post-doc research-fellow, again at the Department of Animal Breeding of WAU with a free assignment to further investigate genetics.