Genetic Heteroscedasticity for Domestic Animal Traits

Majbritt Felleki
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2014

Acta Universitatis agriculturae Sueciae
2014:43

Cover illustration by Sandra and Elliot

ISSN 1652-6880
ISBN (print version) 978- 91-576-8034-1
ISBN (electronic version) 978- 91-576-8035-8

Print: SLU Service/Repro, Uppsala 2014

Genetic Heteroscedasticity for Domestic Animal Traits

Abstract

Animal traits differ not only in mean, but also in variation around the mean. For instance, one sire's daughter group may be very homogeneous, while another sire's daughters are much more heterogeneous in performance. The difference in residual variance can partially be explained by genetic differences. Models for such genetic heterogeneity of environmental variance include genetic effects for the mean and residual variance, and a correlation between the genetic effects for the mean and residual variance to measure how the residual variance might vary with the mean.

The aim of this thesis was to develop a method based on double hierarchical generalized linear models for estimating genetic heteroscedasticity, and to apply it on four traits in two domestic animal species; teat count and litter size in pigs, and milk production and somatic cell count in dairy cows.

The method developed is fast and has been implemented in software that is widely used in animal breeding, which makes it convenient to use. It is based on an approximation of double hierarchical generalized linear models by normal distributions. When having repeated observations on individuals or genetic groups, the estimates were found to be unbiased.

For the traits studied, the estimated heritability values for the mean and the residual variance, and the genetic coefficients of variation, were found in the usual ranges reported. The genetic correlation between mean and residual variance was estimated for the pig traits only, and was found to be favorable for litter size, but unfavorable for teat count.

Keywords: Quantitative genetics, genetic heteroscedasticity of residuals, genetic heterogeneity of environmental variation, genetic heterogeneity of residual variance, double hierarchical generalized linear models, teat count in pigs, litter size in pigs, milk yield in cows, somatic cell count in cows

Author's address: Majbritt Felleki, SLU, Department of Animal Breeding and Genetics, P.O. Box 7023, 750 07 Uppsala, Sweden
E-mail: mfl@du.se

Our understanding of why variances and heritabilities take the levels they do is at best, however, superficial.

William G. Hill and Han A. Mulder

To Jonas

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Rönnegård, L., Felleki, M., Fikse, W.F., Mulder, H.A., Strandberg, E., 2010. Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. *Genetics Selection Evolution* 42(1), 8.

II Felleki, M., Lee, D., Lee, Y., Gilmour, A.R., Rönnegård, L., 2012. Estimation of breeding values for mean and dispersion, their variance and correlation using double hierarchical generalized linear models. *Genetics Research* 94(06), 307–317.

III Rönnegård, L., Felleki, M., Fikse, W.F., Mulder, H.A., Strandberg, E., 2013. Variance component and breeding value estimation for genetic heterogeneity of residual variance in Swedish Holstein dairy cattle. *Journal of Dairy Science* 96(4), 2627–2636.

IV Felleki, M., Lundeheim, N., 2014. Genetic Heteroscedasticity for Teat Count in Pigs. (*manuscript*).

Papers I, II, and III are reproduced with the permission of the publishers.

Abbreviations and symbols

AgeC	Age at calving
DHGLM	Double hierarchical generalized linear models – a class of models and a method for inference
DIM	Days in milk, number of days after calving
htd	Herd-test day
HYM	Herd-year-month
IRWLS	Iterative re-weighted least squares
SCC	Somatic cell count
SCS	Somatic cell score
SE	Standard error, estimated variance of an estimate
VCE	Variance component estimate
ys	Year-season (of calving)
a, a_v	Additive genetic effect of animal
A	Additive genetic relationship matrix
d, d_v	Additive genetic effect of dam
e	Vector of residuals for the mean level
e_v	Vector of residuals for the residual variance level
GCV	Genetic coefficient of variation for the mean $\mathrm{GCV} = \sigma_A/\mu$
GCV_E	Genetic coefficient of variation for the residual variance $\mathrm{GCV}_E = \sigma_{A_v}/\sigma_E^2$
h^2	Mean level heritability $h^2 = \sigma_A^2/\sigma_P^2$
h_v^2	Residual variance level heritability $h_v^2 = \sigma_{A_v}^2/(2\sigma_P^4 + 3\sigma_{S_v}^2)$
$h.b$, $h.b_v$	Effect of herd-birthdate (herd-year-month, HYM)
H	Hessian
I	Identity matrix
m	Additive genetic maternal effect
p, p_v	Permanent environmental effect of animal or dam
q	Vector of hat values, the diagonal of the hat matrix

s, s_v	Additive genetic effect of sire
v	As subscript referring to residual variance level. Sometimes d is used instead of v (Paper II)
X,Z,V,W	Design (incidence) matrices
y	Vector of responses
z_v	Vector of working variables
β	Vector of fixed effects for the mean
β_v	Vector of fixed effects for the residual variance
Γ	Gamma distribution
μ	Estimated mean $\mu = \bar{\hat{y}}$
η_v	Linear predictor for the residual variance, $\log \varphi = \eta_v$
ρ	Genetic correlation
σ^2	Residual variance for the mean level
σ_a^2	Variance component for animal genetic effect for the mean level
$\sigma_{a_v,\text{exp}}^2$	Variance component for animal genetic effect for the residual variance level
σ_A^2	Estimated additive genetic variance, $\sigma_A^2 = \sigma_a^2$ or $\sigma_A^2 = 2(\sigma_s^2 + \sigma_d^2)$
$\sigma_{A_v}^2$	Estimated additive genetic variance for the additively modelled residual variance, $\sigma_{A_v}^2 = \sigma_{S_v}^2 \cdot \sigma_{a_v,\text{exp}}^2 / \sigma_{S_v,\text{exp}}^2$ or $\sigma_{A_v}^2 = \sigma_{S_v}^2 \cdot 2(\sigma_{s_v,\text{exp}}^2 + \sigma_{d_v,\text{exp}}^2)/\sigma_{S_v,\text{exp}}^2$
σ_d^2	Variance component for dam genetic effect for the mean level
$\sigma_{d_v,\text{exp}}^2$	Variance component for dam genetic effect for the residual variance level
σ_E^2	Estimated residual variance, $\sigma_E^2 = \overline{\exp \hat{z}_v}$
$\sigma_{E,\text{exp}}^2$	$\sigma_{E,\text{exp}}^2 = \sigma_E^2 / \exp(\sigma_{S_v,\text{exp}}^2/2)$
σ_p^2	Variance component for permanent environmental effect for the mean level
$\sigma_{p_v,\text{exp}}^2$	Variance component for permanent environmental effect for the residual variance level
σ_P^2	Estimated phenotypic variance
σ_s^2	Variance component for sire genetic effect for the mean level
$\sigma_{s_v,\text{exp}}^2$	Variance component for sire genetic effect for the residual variance level
$\sigma_{S_v}^2$	Transformation of $\sigma_{S_v,\text{exp}}^2$ by $\sigma_{S_v}^2 = \sigma_{E,\text{exp}}^4 \exp\left(2\sigma_{S_v,\text{exp}}^2\right) - \sigma_E^4$
$\sigma_{S_v,\text{exp}}^2$	Sum of all variance components for the residual variance level
σ_v^2	Residual variance for the residual variance level
φ	Vector of residual variances
Φ	Diagonal matrix with diagonal φ

1 Introduction

Domestic animals are under continuous selection for several traits, and the success of increasing them has been tremendous. For instance milk yield in Swedish Holstein has increased from 4,297 kg per cow and year in 1960 to 8,741 kg in 2010 (Swedish Dairy Association, 2011), and the number of live born piglets per litter has increased from 10.9 in 1994 to 13.1 in 2011 (Svenska Pig, 2012).

However, for some traits, it is not only important to improve the mean of the trait, but also to control the variation around the mean. For instance, it would be ideal if sows always had reasonably large litters, to avoid the economically unprofitable small litters, but also to avoid oversized litters that a sow cannot raise.

The variation around the mean, similarly to the mean itself, can be assumed to be influenced by both environmental and genetic factors. For example, by always providing feed of consistently good quality, the variation in milk yield will be reduced. That genetic influence on variation exists is more surprising, but it has been seen in for instance the difference in milk yield variation within daughter groups of sires (Van Vleck, 1968; Clay *et al.*, 1979).

Another phenomenon that has been observed is that the variation might be connected to the mean for a trait, for example a higher average milk yield is associated with higher variation.

Whereas much methodology development has been done for estimation of breeding values and genetic variation for the mean level of traits, not much has been done in the area of estimation of genetic control of variation. One reason is that this kind of estimation is methodologically more challenging. The genetic influence on variation around the trait mean, and its connection to the trait mean, with a primary focus on the estimation process, is therefore the topic of this thesis.

2 Background

2.1 Modelling and estimation of genetic heteroscedasticity of residuals

In quantitative genetic models for animal traits, the residuals are often assumed to be homoscedastic, i.e., the residuals follow the same distribution and thus the variance is the same for all of them. However, evidence exists that both genetic and environmental factors control residual variance. Models, in which genetic or environmental effects or both are included in the residual variance, were introduced during the nineties.

The modelling of the residual variance has been done on different scales. One approach is to assume that fixed and random effects act additively on the residual variance (Mulder *et al.*, 2007) or the standard deviation. However, for these models there is no guarantee that the estimated residual variances or standard deviations will be larger than zero. SanCristobal-Gaudy *et al.* (1998) described a model in which fixed and random effects were assumed to act additively on the logarithm of the residual variances, and the estimated residual variances were thus always larger than zero. This model, called the exponential model, was the one used in this thesis.

Several approaches have been used for estimation in these models. An expectation-maximization method was used by SanCristobal-Gaudy *et al.* (1998). Mulder *et al.* (2009) developed an iterative bivariate algorithm. Sorensen & Waagepetersen (2003) analyzed data on litter sizes in pigs using a Bayesian Markov chain Monte Carlo (MCMC) algorithm.

Formulas for heritability of residual variance were derived by Mulder *et al.* (2007), and they also came up with formulas for translation of results from the exponential model to models with fixed and random effects additively included in the model for residual variances.

Many terms are used in this relatively new area of research to describe the underlying feature. These are genetic (or genetically structured) heterogeneity of environmental (or residual) variance, genetic control (or genetics) of environmental variation, or genetically structured differences in residual variance. Also the term canalization has been used to describe an evolved genetic buffering that keeps a trait stable around the mean under selection (SanCristobal-Gaudy *et al.*, 1998). This is not to be confused with the term robustness, which means that a trait is stable (unchanged mean) despite environmental changes. A recent term is genetic variance for micro-environmental sensitivity (not to be confused with (macro-environmental) sensitivity, which describes the same feature as robustness) (Mulder *et al.*, 2013). Uniformity has been used as well to informally describe the desired characteristic. In this thesis the term genetic heteroscedasticity was used in the title, because heteroscedasticity is a generally accepted statistical term.

2.2 Double hierarchical generalized linear models (DHGLM)

The term hierarchical generalized linear models is used for both a class of models and a tool for estimation (Lee & Nelder, 1996). It is an extension of the mixed model equations (Henderson, 1953), restricted maximum likelihood, REML (Thompson, 1962; Patterson & Thompson, 1971) and generalized linear models (Nelder & Wedderburn, 1972).

Double hierarchical generalized linear models (DHGLM) is an expansion of the hierarchical generalized linear models to also include a structure for one or more variance components and/or the residual variance (Lee & Nelder, 2006). The structures can contain both fixed and random effects, and several distributions of the traits and random effects, as well as link functions between the parameters to be structured and the additively included effects, can be used.

The estimation tool builds on the joint likelihood of the fixed and random effects, called the h-likelihood (Lee & Nelder, 1996). As estimation moves down in the hierarchy from the mean level to the levels of the residual variance and variance components, the h-likelihood is modified in one or more steps to be adjusted profile likelihoods not containing the parameters already estimated.

The theoretical estimation of parameters, and the implementation for estimation, turn out to be straight-forward in many cases. Estimates are in general found to be unbiased, even for complicated binary traits (Lee *et al.*, 2006), for which penalized quasi-likelihood, PQL (Breslow & Clayton, 1993) has been shown to fail.

DHGLM is a recently developed tool, but further applications in animal breeding are expected because of the richness of models, the easiness of

implementation, and the speed of fitting the models, which altogether makes the method suitable for the large data sets often collected in animal breeding (Rönnegård & Lee, 2013).

3 Aim of the thesis

The aim of the thesis was to develop a DHGLM-based method that can be used for estimation in models with genetically structured heterogeneity of residual variance for large data sets, and to apply it for some domestic animal traits; teat count and litter size in pigs, and milk production and somatic cell count in dairy cows.

4 Summary of performed studies

Papers are referred to by numbers I-IV.

4.1 Material

Three data sets were used for the studies; one set on litter size in pigs (Papers I-II), one set on teat count in pigs (Paper IV), and one set containing two traits in dairy cows that were milk yield and somatic cell count (Paper III). A summary of the size of the data sets and the mean, median, variance, and standard deviation of the traits are found below (Table 1).

Table 1. *Size of data sets and the mean, median, variance, and standard deviations of trait values*

	Records	Animals with records	Animals in pedigree	Mean	Median	Variance	Standard deviation
Litter size I-II	10,060	4,149	6,437	10.29	10	9.91	3.1
Teat count IV	118,267	118,267	121,872	14.53	14	0.84	0.9
Milk yield (l/d) III	1,693,154	177,411	466,720	29.13	29.2	45.5	6.7
SCS* III	1,693,154	177,411	466,720	2.36	2.05	2.76	1.7

*Somatic Cell Score, transformation of somatic cell count (SCC, count/ml) by $SCS = \log_2(SCC/100{,}000) + 3$.

Simulation studies were performed in Papers I and II. Only part of the simulation study from Paper II will be summarized here.

4.1.1 Litter size in pigs (Papers I-II)

The data on litter size in pigs was from Sorensen & Waagepetersen (2003) and contained for each litter size the identity of sow (4,149 sows), parity (9 levels), season (4 levels), herd (82 levels), and type of insemination (2 levels). The data

was highly imbalanced; 13 herds contained five observations or less, and the ninth parity contained nine observations only.

4.1.2 Simulated data (Paper II)

For simulation of data, the pedigree of the pig litter size data was used, and the number of sows with records was fixed as in the original dataset. The total number of observations on litter sizes was either kept (N = 10,060), or increased by changing the number of repeated records per sow (parities) to 4 ($N = 4 \cdot 4{,}149 = 16{,}596$) or 9 ($N = 9 \cdot 4{,}149 = 37{,}341$). A fixed effect of insemination type was simulated. The values of variance components for the simulation were taken from results by Sorensen & Waagepetersen (2003).

4.1.3 Milk yield and somatic cell count in cows (Paper III)

The data on dairy cow traits contained observations on milk yield (l/day) and somatic cell count concentration (SCC, count/ml), and the identity of the cow (177,411 cows). Further, the variables herd (1,759 levels), herd-test day (htd, 21,570 levels), year-season of calving (ys, 32 levels), age at calving (AgeC, continuous), and days in milk (DIM, continuous) were given.

The somatic cell count was transformed to somatic cell score by $\mathrm{SCS} = \log_2(\mathrm{SCC}/100{,}000) + 3$ (Ali & Shook, 1980).

4.1.4 Teat count in pigs (Paper IV)

Observations on teat counts were connected with the pig identity (118,267 pigs). The sire identity (586 sires) and dam identity (7,813 dams) were also added to the data. Effects considered were sex (2 levels), herd (17 levels), birthdate (year-month, 52 levels), and herd-birthdate (HYM).

Figure 1 illustrates the unfavorable linear increase of the phenotypic variance with increase of the mean for paternal half sib groups.

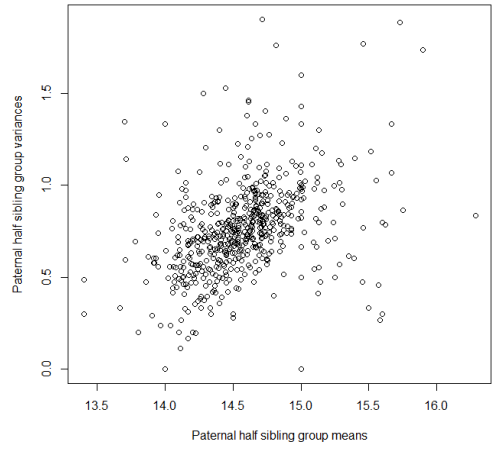

Figure 1. Comparison of the paternal half sib group means and variances for teat count observations. This illustrates unfavorable linear increase of the variance with increase of the mean.

4.2 Methods

4.2.1 Models for litter sizes, somatic cell score, and milk yield

Models with heteroscedastic residuals were used for the analysis of the traits litter size, somatic cell score, and milk yield (Papers I-III). Fixed and random effects were included in both the mean and residual variance levels of the models.

Random effects are listed in Table 2. The models were similarly structured; the genetic effect of animal and the permanent environmental effect of animal were random effects for both levels. Fixed effects are given in Table 3. The models also included an intercept for both levels.

Let y be the vector of observations. It was assumed that $y = X\beta + Za + Wp + e$, where β was the vector of fixed effects including an intercept, a was the vector of the animal genetic effects (animal identity), p was the vector of permanent environmental effects (animal identity), e was the vector of residuals, and X, Z, and W were known incidence matrices.

The distribution of the animal additive genetic effects was $a \sim \mathrm{N}(0, \sigma_a^2 A)$, where A was the additive genetic relationship matrix, and the distribution of the permanent environmental effects was $p \sim \mathrm{N}(0, \sigma_p^2 I)$, where I was the identity matrix.

The distribution of the residuals was assumed to be $e \sim \mathrm{N}(0, \Phi)$, where Φ was a diagonal matrix with diagonal φ. It was moreover assumed that the residual variance was structured by $\log \varphi = X_v\beta_v + Za_v + Wp_v$, where symbols were the same as for the mean level above.

The permanent environmental effects $p \sim \mathrm{N}(0, \sigma_p^2 I)$ and $p_v \sim \mathrm{N}(0, \sigma_{p_v,\mathrm{exp}}^2 I)$ were assumed independent. The animal genetic effects $a \sim \mathrm{N}(0, \sigma_a^2 A)$ and $a_v \sim \mathrm{N}(0, \sigma_{a_v,\mathrm{exp}}^2 A)$ were assumed independent in Papers I and III but dependent in Paper II,

$$\begin{pmatrix} a \\ a_v \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} \sigma_a^2 A & \rho\sigma_a\sigma_{a_v,\mathrm{exp}} A \\ \rho\sigma_a\sigma_{a_v,\mathrm{exp}} A & \sigma_{a_v,\mathrm{exp}}^2 A \end{pmatrix}\right).$$

The subscript $_{\mathrm{exp}}$ was used because the residual variance was modelled on a logarithmic scale.

4.2.2 Model for teat counts

For the teat counts in Paper IV, three additive genetic and two environmental effects were included in the mean level and two additive genetic and two environmental effects were included in the residual variance level. The effects are found in Table 2. The additive genetic structure ‘Full sib’ means that both

the sire and the dam effects were included, but their estimated variance components were forced to be equal.

The mean model for teat count was $y = X\beta + (Z_s s + Z_d d) + Z_m m + Vh.b + Wp + e$, and the residual variance model was $\log \varphi = X\beta_v + (Z_s s_v + Z_d d_v) + Vh.b_v + Wp_v$.

The random effects were sire s and s_v, dam d and d_v, genetic maternal m, herd-year-month of birth $h.b$ and $h.b_v$, and permanent environmental maternal p and p_v. Following Canario *et al.* (2010) distributions of random effects were assumed to be (s and d were independent identically distributed, and so were s_v and d_v)

$$\begin{pmatrix} s \\ m \\ s_v \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} \sigma_s^2 A & \rho_{sm}\sigma_s\sigma_m A & \rho\sigma_s\sigma_{s_v,\exp} A \\ \rho_{sm}\sigma_s\sigma_m A & \sigma_m^2 A & 0 \\ \rho\sigma_s\sigma_{s_v,\exp} A & 0 & \sigma_{s_v,\exp}^2 A \end{pmatrix}\right),$$

$$\begin{pmatrix} h.b \\ h.b_v \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} \sigma_{h.b}^2 I & \rho_{h.b}\sigma_{h.b}\sigma_{h.b_v,\exp} I \\ \rho_{h.b}\sigma_{h.b}\sigma_{h.b_v,\exp} I & \sigma_{h.b_v,\exp}^2 I \end{pmatrix}\right),$$

$$\begin{pmatrix} p \\ p_v \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} \sigma_p^2 I & \rho_p\sigma_p\sigma_{p_v,\exp} I \\ \rho_p\sigma_p\sigma_{p_v,\exp} I & \sigma_{p_v,\exp}^2 I \end{pmatrix}\right).$$

Table 2. *Random effects included in the models (Papers I-IV)*

	Genetic for mean	Environmental for mean	Genetic for residual variance	Environmental for residual variance	Genetic correlation
Litter size I	Animal	Identity	Animal	Identity	No
Litter size II	Animal	Identity	Animal	Identity	Yes
Milk yield III	Animal	Identity	Animal	Identity	No
SCS III	Animal	Identity	Animal	Identity	No
Teat count IV	Full sib and maternal	HYM and maternal	Full sib	HYM and maternal	Yes

Table 3. *Fixed effects included in the models (Papers I-IV)*

	Fixed effects for the mean	Fixed effects for the residual variance
Litter size I+II	herd, season, insemination type, parity	insemination type, parity
Milk yield and SCS III	htd, ys, AgeC, $(\text{AgeC})^2$, $(\text{AgeC})^3$, DIM, exp(-0.05*DIM) (all except htd and ys continuous)	herd, ys, AgeC, $(\text{AgeC})^2$, DIM, $(\text{DIM})^2$ (all except herd and ys continuous)
Teat count IV	sex, herd, year-month of birth	sex, herd, year-month of birth

4.2.3 Estimation using DHGLM

For notation simplicity, estimation using DHGLM is considered for the model for pig litter size data that include correlation between additive genetic effects for the mean and the residual variance (Paper II). The theory behind the estimation is found in the appendix in Paper II.

The algorithm used is:

1. Initiate weights $1/\varphi$ for the mean level and $(1-q)/2$ for the residual variance level (q is the hat value, that is the diagonal of the matrix $[X\ Z\ W]H^{-1}[X\ Z\ W]'\Phi^{-1}$, where H is the Hessian). Initiate the working variables z_v for the residual variance level.
2. Fit normal distribution

$$\begin{pmatrix} y \\ z_v \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & X_v \end{pmatrix}\begin{pmatrix} \beta \\ \beta_v \end{pmatrix} + \begin{pmatrix} Z & 0 \\ 0 & Z \end{pmatrix}\begin{pmatrix} a \\ a_v \end{pmatrix} + \begin{pmatrix} W & 0 \\ 0 & W \end{pmatrix}\begin{pmatrix} p \\ p_v \end{pmatrix} + \begin{pmatrix} e \\ e_v \end{pmatrix}$$

   with $p \sim \mathrm{N}(0, \sigma_p^2 I)$, $p_v \sim \mathrm{N}(0, \sigma_{p_v,\exp}^2 I)$, $e \sim \mathrm{N}(0, \sigma^2\Phi)$, $e_v \sim N(0, \sigma_v^2 \mathrm{diag}(1-q)/2)$, and (a, a_v) all being independent of each other, but a and a_v correlated.
3. Update the residual variance level with new weights $(1-q)/2$ and new working variables $z_v = \log\varphi + \Phi^{-1}(\hat{e}^2/(1-q) - \varphi)$.
4. Identical to step 2.
5. Update the mean level with new weights $1/\varphi = 1/(\exp\widehat{z_v})$.
6. Run step 2.-5. until convergence ($\sigma^2 = 1$).

4.2.4 Phenotypic variance and heritability

To be able to find the heritability values, an estimate of the phenotypic variation and therefore the residual variance was needed. The residual variance was found as the average of the estimated residual variances, $\sigma_E^2 = \overline{\exp\widehat{z_v}}$.

The estimated phenotypic variance was $\sigma_P^2 = \sigma_a^2 + \sigma_p^2 + \sigma_E^2$ for the litter size, milk yield, and somatic cell count data, and $\sigma_P^2 = \sigma_s^2 + \sigma_d^2 + 2\sqrt{2}\rho_{sm}\sigma_s\sigma_m + \sigma_m^2 + \sigma_{h.b}^2 + \sigma_p^2 + \sigma_E^2$ for the teat count data. The inclusion of the covariance was handled by Willham (1972) in the case of the direct genetic effect of animal together with the maternal genetic effect for which $\mathrm{cor}(a, m) = 1/2$. For the full sib model the theoretical correlation is $\mathrm{cor}(s + d, m) = 1/\sqrt{2}$, and therefore $2\sqrt{2}\rho_{sm}\sigma_s\sigma_m$ is included.

Heritability values were defined as $h^2 = \sigma_A^2/\sigma_P^2$, where the additive genetic variance component was estimated as $\sigma_A^2 = \sigma_a^2$ for the litter size, milk yield, and somatic cell count data, and as $\sigma_A^2 = 4\sigma_s^2$ for the teat count data.

The average of the predicted values was $\mu = \bar{\hat{y}}$, and the genetic coefficient of variation was $\mathrm{GCV} = \sigma_A/\mu$.

4.2.5 Residual variance heritability

The residual variance heritability was derived by Mulder *et al.* (2007) and in Paper IV these equations were extended to include permanent environmental effects in the residual variance level.

The additive genetic variance component $\sigma^2_{a_v,\text{exp}}$ for the residual variance on the logarithmic scale, was substituted by the additive genetic variance component $\sigma^2_{A_v}$ for the residual variance on the additive scale. The heritability for residual variance was $h^2_v = \sigma^2_{A_v}/(2\sigma^4_P + 3\sigma^2_{S_v})$, where $\sigma^2_{S_v}$ was the sum of all variance components on the additive scale (Mulder *et al.*, 2007; Paper IV).

Corresponding to the mean GCV, the genetic coefficient of variation for residual variance was $\text{GCV}_E = \sigma_{A_v}/\sigma^2_E$. Note that

$$\text{GCV}^2_E = \frac{\sigma^2_{A_v}}{\sigma^4_E} = \frac{\sigma^4_E(\exp(\sigma^2_{a_v,\text{exp}})-1)}{\sigma^4_E} \approx \sigma^2_{a_v,\text{exp}},$$

when $\sigma^2_{a_v,\text{exp}}$ is small (<0.2), thus GCV_E can be found directly from parameter estimates.

4.3 Results

4.3.1 Results from the analysis of data sets

The variance component estimates (VCEs) from all studies are collected in Table 4. The VCEs for Litter size I and II differed because the genetic correlation ρ was not estimated for Litter size I. The additive genetic VCEs σ^2_a and $\sigma^2_{a_v,\text{exp}}$ were larger, and the permanent environmental VCEs σ^2_p and $\sigma^2_{p_v,\text{exp}}$ smaller for Litter size II.

The genetic correlation between the additive genetic effects for the mean and the residual variance was found to be favorable for litter size (negative), but unfavorable for teat count (positive). Both correlations were significantly different from zero.

For teat count the genetic maternal VCE was σ^2_m = 0.01 (SE 0.003), and the correlation between the maternal and the sire-dam effect was ρ_{sm} = -0.10 (0.063) thus not significant. The mean and residual variance correlations for effects HYM and permanent environmental maternal were $\rho_{h.b}$ = 0.47 (0.062) and ρ_p = 0.66 (0.086), respectively.

A sire model for teat count was also fitted in Paper IV, and the results were similar to those from the sire-dam model.

Estimated heritability and genetic coefficients of variation are found in Table 5, and formulas are given below the table.

Table 4. *Variance component estimates and standard errors for the traits studied (Papers I-IV)*

	σ_a^2	σ_p^2	$\sigma_{a_v,\text{exp}}^2$	$\sigma_{p_v,\text{exp}}^2$	ρ
Litter size I	1.35 (0.18)	0.44 (0.14)	0.09 (0.02)	0.06 (0.02)	
Litter size II	1.61 (0.18)	0.28 (0.13)	0.15 (0.03)	0.05 (0.02)	-0.52 (0.07)
Milk yield III	8.78 (0.21)	12.40 (0.14)	0.049 (0.0034)	0.37 (0.0031)	
SCS III	0.28 (0.011)	1.03 (0.0085)	0.046 (0.0038)	0.61 (0.0040)	
Teat count IV	0.15 (0.009)*	0.02 (0.002)**	0.12 (0.009)*	0.10 (0.007)**	0.79 (0.025)

* Full sib additive genetic VCE, $\sigma_s^2 + \sigma_d^2 = 2\sigma_s^2$.
** Sum of environmental VCEs, $\sigma_{h.b}^2 + \sigma_p^2$.

Table 5. *Estimated heritability and genetic coefficients of variation (Papers I-IV)*

	μ	σ_E^2	σ_P^2	$\sigma_{A_v}^2$	h^2	GCV	h_v^2	GCV_E
Litter size I	10.3	6.74	8.53	4.64	0.16	0.11	0.03	0.32
Litter size II	10.4	6.69	8.58	7.53	0.19	0.12	0.04	0.41
Milk yield III	29.2	9.36	30.5	5.32	0.29	0.10	0.003	0.25
SCS III	2.34	1.16	2.47	0.09	0.11	0.23	0.006	0.26
Teat count IV	14.5	0.64	0.81	0.10	0.36	0.04	0.07	0.51

$\mu = \bar{\hat{y}}$
$\sigma_E^2 = \overline{\exp \hat{z}_v}$
$\sigma_P^2 =$ sum of mean level VCEs $+ \sigma_E^2$
For litter size, milk yield and somatic cell score $\sigma_{A_v}^2 = \sigma_{S_v}^2 \cdot \sigma_{a_v,\text{exp}}^2 / \sigma_{S_v,\text{exp}}^2$, $\sigma_{S_v}^2 = \sigma_{E,\text{exp}}^4 \exp\left(2\sigma_{S_v,\text{exp}}^2\right) - \sigma_E^4$, $\sigma_{S_v,\text{exp}}^2 =$ sum of residual variance level VCEs , $\sigma_{E,\text{exp}}^2 = \sigma_E^2 / \exp(\sigma_{S_v,\text{exp}}^2/2)$. For teat count $\sigma_{A_v}^2 = 2\sigma_{S_v}^2 \cdot (\sigma_{S_v,\text{exp}}^2 + \sigma_{d_v,\text{exp}}^2)/\sigma_{S_v,\text{exp}}^2$
$h^2 = \sigma_A^2/\sigma_P^2$, for litter size, milk yield and somatic cell score $\sigma_A^2 = \sigma_a^2$, for teat count $\sigma_A^2 = 2(\sigma_s^2 + \sigma_d^2)$
$\text{GCV} = \sigma_A/\mu$
$h_v^2 = \sigma_{A_v}^2/(2\sigma_P^4 + 3\sigma_{S_v}^2)$, $\text{GCV}_E = \sigma_{A_v}/\sigma_E^2$

4.3.2 Results from the simulation study

Some results from the simulation study (Paper II) are given in Table 6. For all simulation settings, the averages of the VCEs for the additive genetic effects σ_a^2 and $\sigma_{a_v,\text{exp}}^2$, as well as the average of the genetic correlations ρ were well in agreement with the true values. The averages of the VCEs for the permanent environmental effects σ_p^2 and $\sigma_{p_v,\text{exp}}^2$ were not near the true value in the original parity setting and in the four-parity setting. The mean level permanent environmental variance component was under-estimated, and the residual variance level variance component was over-estimated. In the nine-parity setting, the averages of the permanent environmental VCEs were close to the true values.

Table 6. *Averages and standard errors of estimated variance components for simulated data with same pedigree as the litter size data (Papers I+II). The left hand column contains the simulated data structure*

	σ_a^2	σ_p^2	$\sigma_{a_v,\text{exp}}^2$	$\sigma_{p_v,\text{exp}}^2$	ρ
True values	1.62	0.60	0.09	0.06	-0.62
Original distrib.*	1.56(0.017)	0.24(0.016)	0.08(0.003)	0.13(0.004)	-0.61(0.012)
Four parities	1.65(0.017)	0.51(0.012)	0.09(0.002)	0.15(0.003)	-0.64(0.008)
Nine parities	1.62(0.013)	0.60(0.008)	0.09(0.001)	0.09(0.001)	-0.64(0.005)

*Original parity distribution. In this setting twenty-seven out of 100 replicates did not converge. Estimates are for all replicates (with minor differences in results if these 27 replicates were included or not)

4.4 News of the studies

4.4.1 Estimation in animal models with genetic heteroscedastic residuals can be done using DHGLM

Papers I and II consecutively show that DHGLM can be used for estimation in animal models with genetically structured residual variance heterogeneity.

In the DHGLM setting correlations between the random effects for the *same* level had already been implemented (Lee *et al.*, 2006). In Paper II, DHGLM was extended to include models with correlations between random effects for different levels, the mean level and the residual variance level.

The algorithm was implemented in ASReml 4.0 (Gilmour *et al.*, 2009), which is a common software used for animal models, and it became very fast and easy to use.

Data of pig litter size, previously analyzed using the Markov chain Monte Carlo method (Sorensen & Waagepetersen, 2003), was re-analyzed using the algorithm.

Simulation studies were done to study the performance of the algorithm with respect to bias and precision. It was found that the estimates were unbiased in the case of a (yet unspecified) number of repeated observations on individuals.

4.4.2 DHGLM can be used for large data sets

In Paper III the algorithm from Paper I was used for a large data set on milk yield and somatic cell count. Over 1.5 million observations on more than 170,000 cows related through a pedigree of more than 400,000 animals were analyzed. This was an example that the algorithm can be used for large data sets. For the analysis a week was required to obtain convergence of VCEs.

4.4.3 DHGLM can be used for data without repeated observations

The data used for Papers I-III all contained repeated observations on individuals. In Paper IV, a data set on teat count in pigs was used, hence no repeated observations on individuals. Even though in some cases it would be possible to use the algorithm anyway, results from fitting a model on individuals would be biased. Therefore half sib and full sib analysis were performed.

It was found that the results from the heteroscedastic analysis of the teat count data, were similar for the genetic half sib and full sib structure, and that the mean level heritability was the same as found by fitting a model with homoscedastic residuals. Therefore any of the structures could be used considering the teat count data.

5 General discussion

5.1 Discussion of the results from Papers I-IV

5.1.1 Comparison of the heritability values

The estimated mean heritability values (Papers I-IV) were largest for teat count ($h^2 = 0.36$), followed by milk yield (0.29), litter size (0.19), and SCS (0.11) (Table 5). This reflects the common statement that morphological traits like teat counts are more heritable than fitness and health traits such as litter size and SCS (Falconer & Mackay, 1996).

The genetic coefficient of variation for mean was in opposite order with the largest value for SCS (GCV = 0.23), and thereafter litter size (0.12), milk yield (0.10) and teat count (0.04). The implication is that if the mean is changed by one genetic standard deviation, this change will correspond to a larger relative change for litter size than for teat count. The genetic coefficient of variation for SCS is difficult to interpret because of the logarithmic scale.

The order of heritability values for the residual variance, and that of the genetic coefficient of variation for residual variance, were the same. Teat count represented the largest values ($h_v^2 = 0.07$, $\mathrm{GCV}_E = 0.51$), followed by litter size (0.04, 0.41), SCS (0.006, 0.26), and milk yield (0.003, 0.25). This was almost the same order as the mean heritability values, however, milk yield had the lowest value of residual variance heritability. For litter size, these values were well in agreement with previously published heritability values (0.021 to 0.048) and genetic coefficients of variation (0.27 to 0.51) for residual variance in several species (Hill & Mulder, 2010). Genetic control of residual variance for teat count, milk yield and somatic cell count has not been analyzed previously.

The genetic correlation between the mean and residual variance levels (Table 4) was favorable for litter size (-0.52), but unfavorable for teat count (0.79). For teat count the numerically positive estimate of the genetic

correlation indicates that selection for increased mean will lead to increased variation in the number of teats. Contrary, for litter size the variation is expected to decrease as the mean increases, but the sign of the genetic correlation has been shown to be dependent on the scale (Yang *et al.*, 2011).

5.1.2 Alternatives for the calculation of phenotypic variance

The value of the phenotypic variance has impact on the heritability values, and is therefore important. The phenotypic variance σ_P^2 depends on the residual variance σ_E^2, and the residual variance in models with heteroscedastic residuals can be calculated in several ways.

A method previously used (Mulder et al., 2007) was to fit a model with homoscedastic residuals and to use the estimated phenotypic variance from that model for finding the heritability values for the mean and residual variance. However, the phenotypic variance from a model with heteroscedastic residuals is smaller than that from a model with homoscedastic residuals, because more variation is explained in the latter by the fixed effects for the residual variance.

The residual variance σ_E^2 can be found as the average of the expected residual variances, $\sigma_E^2 = \overline{\exp(X_v\beta_v)}\exp\left(\sigma_{S_v,\text{exp}}^2/2\right)$ (Mulder *et al.*, 2007; Paper IV), but it is easier and more correct to calculate the average of the estimated residual variances, $\sigma_E^2 = \overline{\exp \widehat{z_v}}$, which has been used in this thesis. This is similar to the average of the estimated mean values of the observations, $\mu = \bar{\hat{y}}$.

The above formula for σ_E^2 is however used to find $\sigma_{E,\text{exp}}^2 = \sigma_E^2 / \exp\left(\sigma_{S_v,\text{exp}}^2/2\right)$, which is needed to find the additive genetic VCE $\sigma_{A_v}^2$ for the residual variance, corresponding to an additively modelled residual variance (Mulder *et al.*, 2007; Paper IV). This is used for finding the heritability of the residual variance, which is the regression of a_v on y^2. The regression corresponds to the regression of a on y for finding the mean heritability value.

5.2 Results when the dispersion in the Gamma distribution for the residual variance level is fixed

In this section the dispersion in the Gamma distribution of the residual variances is discussed. Previous results (Papers I-IV) were from a Gamma distribution with under- or over-dispersion included. Here some results with fixed dispersion will be given (Table 7 and 8).

The fitting of the Gamma distribution

$$\frac{\hat{e}^2}{1-q} \sim \Gamma\left(\frac{1-q}{2}, \frac{1-q}{2\varphi}\right), \log\varphi = \eta_v$$

is done by iteratively fitting and updating φ in

$z_v \sim N\left(\eta_v, \frac{2}{1-q}\right)$, $z_v = \log\varphi + \Phi^{-1}\left(\frac{\hat{e}^2}{1-q} - \varphi\right)$

but if the dispersion is not fixed, the residual variance in the normal distribution is $2/(1-q)\sigma_v^2$ where σ_v^2 reflects under- or over-dispersion in the Gamma distribution.

The litter size data was previously analyzed by Sorensen & Waagepetersen (2003), and their results were σ_a^2 = 1.62 (SE 0.213), σ_p^2 = 0.60 (0.155), $\sigma_{a_v,\text{exp}}^2$ = 0.09 (0.018), $\sigma_{p_v,\text{exp}}^2$ = 0.06 (0.010), and ρ = -0.62 (0.093). The VCEs when fixing the dispersion in the residual variance level (Table 7) are much more similar to these estimates, than those obtained by letting the dispersion vary freely (Table 4). This gives an indication that the best fit is obtained by fixing the residual variance level dispersion.

However, while the likelihood function of the analysis Litter size II (Paper II) converged, for Litter size I (Paper I) and Teat count (Paper IV) it did not converge, but the parameter estimates converged.

For the litter size data, the permanent environmental VCE for the residual variance could not be estimated, neither with nor without the genetic correlation.

Table 7. *Variance component estimates and standard errors for the litter size and teat count data with no under- or over-dispersion in the residual variance Gamma distribution*

	σ_a^2	σ_p^2	$\sigma_{a_v,\text{exp}}^2$	$\sigma_{p_v,\text{exp}}^2$	ρ
Litter size I	1.36(0.196)	0.69(0.157)	0.02(0.012)	0.00(0.000)	
Litter size II	1.60(0.202)	0.54(0.153)	0.05(0.013)	0.00(0.000)	-0.63(0.113)
Teat count IV	0.15(0.009)	0.02(0.002)	0.14(0.011)	0.20(0.008)	0.79(0.024)

Table 8. *Estimated heritability and genetic coefficients of variation for the litter size and teat count data with no under- or over-dispersion in the residual variance Gamma distribution*

	μ	σ_E^2	σ_P^2	$\sigma_{A_v}^2$	h^2	GCV	h_v^2	GCV_E
Litter size I	10.3	7.15	9.21	1.24	0.15	0.11	0.01	0.16
Litter size II	10.4	7.08	9.21	2.34	0.17	0.12	0.01	0.22
Teat count IV	14.5	0.64	0.81	0.07	0.36	0.04	0.08	0.58

Formulas are found in the footnotes of Table 5.

5.3 Genetic effects other than the animal genetic effect

5.3.1 Animal genetic effect for the mean level, grouped genetic effect for the residual variance level

It is not always possible to fit a model that includes the animal additive genetic effect both for the mean and the residual variance level. Moreover, when

repeated observations on individuals are not present, even if estimation is possible, the estimates will probably be biased (Paper II).

One alternative is the full or half sib family model as suggested in Paper IV. Another alternative is to include an animal genetic effect for the mean level, and a full sib (sire-dam) or half sib (sire) effect for the residual variance level. Results from such models, however, disagreed with results from models with the same genetic effect for mean and residual variance, and the mean heritability values were too small compared with those from the models with homoscedastic residuals. This has also been observed by Sonesson *et al.* (2013).

5.3.2 Correcting the residual variance to remove additive genetic variance

When an animal genetic effect is included in the mean level of a model with heteroscedastic residuals, the residual variance is truly an environmental variance under the assumption that no non-additive genetic variance is present. However, when a sire effect or a sire-dam effect is included in the mean level, the residual variance also contains three quarters or a half of the additive genetic variance, respectively.

Therefore the residual variance model is not a model of environmental variation, and a correction is needed. Mulder *et al.* (2013) developed such a correction in the case of a paternal half sib (sire) model.

The residuals are corrected by multiplication by $\sqrt{\bar{\varphi} - 3\sigma_s^2}/\sqrt{\bar{\varphi}}$, where φ^{-1} are the weights for the mean level. Exponentials of estimated responses from the residual variance level will be environmental residual variances for the mean. These have to be back-corrected by adding the additive genetic variance previously subtracted before using them as weights for the mean level.

While it is obvious that the residual variance for the mean level must be corrected to only contain environmental variance, it is not that clear if the residual variance level must be corrected, because it is not obvious what effects to include in the residual variance. Fixing the dispersion to 1 might solve the problem.

5.4 Scale

Yang *et al.* (2011) simultaneously estimated Box-Cox transformation ($f_\lambda(y) = (y^\lambda - 1)/\lambda, \lambda \neq 0; \log y, \lambda = 0$) to achieve conditional normality of litter size data, and fitted a model with heteroscedastic residuals, using a Bayesian Markov chain Monte Carlo approach. Comparing results for the untransformed and transformed data, surprisingly the estimate of genetic correlation was altered from being significantly smaller than zero (untransformed data), to

being significantly larger than zero (transformed data). This illustrates the importance of considering skewness of data. No scaling effect was found in Paper III, and Sonesson *et al.* (2013) found no scaling effect by comparing estimated variance components of untransformed and transformed weights in salmon. In these papers, however, the genetic correlation was not estimated.

The common assumption of quantitative genetics is that trait values are sums of many small loads (Fisher, 1918), and therefore, by the law of large numbers, normal distributed. This is in practice not true for all traits, and most likely not when strong selection is involved. Box-Cox transformation might be a solution for this, but back-transformation of the estimates to the original scale of interest is not straight-forward for the genetic correlation.

The transformation of somatic cell count into somatic cell score also alters the estimated parameter values, but it has not been considered what the difference will be for the genetic correlation (which was not estimated in Paper III).

5.5 DHGLM and the approximation

DHGLM has been criticized by several authors, and defended by the creators (Lee & Nelder, 1996; Lee *et al.*, 2007; Lee & Nelder, 2009a; Louis, 2009; Molenberghs *et al.*, 2009; Meng, 2009; Lee & Nelder, 2009b; Lee *et al.*, 2006). To go through the criticism is outside the scope of this thesis, but a thorough summary can be found in Rönnegård *et al.* (2014).

The algorithm used in Papers I-IV is an approximation of DHGLM (Paper II). The approximation in terms of iterative weighted least squares (IRWLS, explained well by Pawitan (2001)) was done to make it possible to use standard software for animal trait models, and to make it possible to fit a model with genetically structured residual variance heterogeneity to large data sets. This corresponds to penalized quasi-likelihood tools, PQL (Breslow & Clayton, 1993) for generalized linear mixed models.

How much bias the IRWLS approximation, and the DHGLM method itself, add to parameter estimates can be studied by simulations corresponding to the data set of interest, as done in Paper II.

5.6 Evidence for genetic control of environmental variation

The evidence for genetic control of environmental variation can be considered from different perspectives (Table 9).

The first perspective is if selection can be done to reduce variance, but so far no studies have revealed convincing evidence for the possibility to select

for reduced residual variance (Hill & Mulder, 2010), contrary to the means of traits, which have been increased by selection even for traits expressing small heritability values (Nielsen *et al.*, 2013).

There is an interesting connection between mean level selection and the variance. In theory the genetic variance should decrease as a consequence of threshold selection of the mean, but in practice the phenotypic variance often increases (Falconer & Mackay, 1996). One explanation could be that homozygotes are more sensitive to the environment, because they will only have one enzyme as a product of the gene in question, and not the flexibility of two different enzymes. With time the environmental conditions also change, which could contribute to an increased variation and that different genes become involved.

A difference between mean level selection and residual variance selection is that the mean is often selected upwards, while the residual variance is selected downwards. It might be that upward selection is easier than downward, because there is a downward limit (zero), but no upward limit.

Table 9. *Support for genetic control of mean μ and residual variance σ^2 from selection, quantitative genetics (QG), and association studies (QTL/GWAS)*

	Selection	QG	QTL/GWAS
μ	Response[1]	Breeding values and heritabilities[1]	Some, but most heritability is missing[2]
σ^2	No convincing response[3]	Breeding values and heritabilities[3]	Some[4]

[1](Falconer & Mackay, 1996) [2](Maher, 2008) [3](Hill & Mulder, 2010) [4](Rönnegård & Valdar, 2011, 2012; Shen *et al.*, 2012)

Looking from a quantitative genetics perspective, as used in this thesis, evidence for the possibility to select for both the mean and the residual variance has been found. For the residual variance, heritability values have in general been found to be smaller than 0.1 (Hill & Mulder, 2010), but the genetic coefficients of variation have been found to be moderate.

Finally, from the perspective of studies using molecular genetic information (e.g., genome wide association studies, GWAS), some evidence for additive genetic control of trait values have been found, but most of the heritability previously estimated is unexplained (Maher, 2008). This is an interesting topic, but outside the scope of this thesis. For residual variance, genetic control has also been found (Shen *et al.*, 2012).

5.7 What genetically structured residual variance heterogeneity reveals about nature

Genetic heterogeneity of residual variance is interpreted as a reaction on small differences in environment, sometimes called micro-environmental changes. This is illustrated in Figure 2. Two genotypes maintained in a range of environments may express a reaction norm (different inclinations) on the environments, but this is not what is modeled by including genetically structured residual variances. The modelled difference in residual variance is illustrated by different lengths of vertical lines at a given value of the environment.

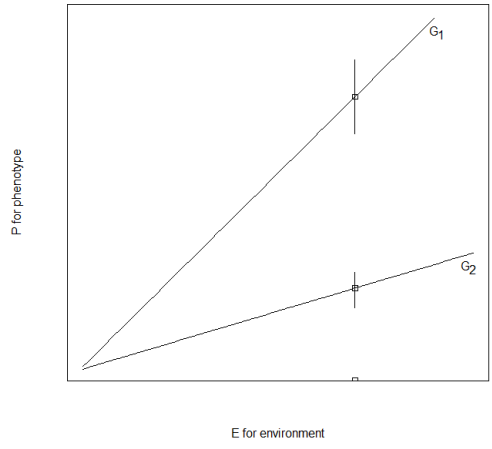

Figure 2. A difference in residual variance is illustrated by different lengths of vertical lines at a given value of the environment.

The inclusion of reaction norms (reactions on macro-environmental changes) in models with genetically structured residual variance heterogeneity has been studied by Mulder *et al.* (2013), who found by simulation that reaction norms and genetic heterogeneity of residual variance could be separated using DHGLM.

Reaction norms are intuitively easier to interpret than genetically structured residual variance heterogeneity. The latter can be observed and modelled, but what the underlying mechanisms are, is hard to grasp. However, if reaction norms, dominance, epistasis, epigenetics, or generally, all sorts of interplays between genes and environment adapting to and altering themselves and each other, are present in the data but not modelled, these phenomena will end up in the residual variance, and hence create a genetic structure in the residual variance. This is probably a part of what genetically structured residual variance heterogeneity explains about nature.

6 Conclusions

An algorithm building on DHGLM, fast and feasible for large data sets on animal traits with pedigrees was developed and was found to be capable of giving unbiased estimates. However, several repeated observations per individual were necessary to obtain unbiased estimates. In the case of a single observation per individual, the analysis could be done for genetic groups such as half or full sib groups.

The algorithm can be used to find genetic control of environmental variation, and to find genetic correlation between the mean and the residual variance of a trait.

7 Future research

7.1 Multiple traits and genetic heterogeneity

Lundeheim *et al.* (2013) and Chalkias *et al.* (2013) analyzed data on several pig traits modeled with homoscedastic residuals with the aim to find genetic correlations among these traits. To include genetic correlation between traits, for instance between litter size and teat count, or between milk yield and SCS is also possible for both levels of a model with heteroscedastic residuals.

The simultaneous fitting of litter size and teat count traits in a model with heteroscedastic residuals, might be problematic because all animals will have a single observation of teat count, and only a few animals will have one or more observations of litter size. A possible solution used by Lundeheim *et al.* (2013) is to analyze teat count together with litter size of the first parity, litter size of the second parity, and so on. Then all traits (teat count, size of first litter, and size of second litter) come with a single observation per individual.

There will probably be only a few observations of litter size per half or full sib group, because altogether few individuals will have observations of litter size. Therefore, even when including sire genetic effects instead of the animal genetic effects for both levels of a model with heteroscedastic residuals, the estimates using DHGLM might be biased.

Milk yield and somatic cell score, on the other hand, are traits very suitable for simultaneous analysis, because repeated observations of both traits are given for each individual observed. A natural extension of the study done in Paper IV would be to first include the genetic correlation between the mean and the residual variance within traits, and thereafter to analyze the traits simultaneously with genetic correlations between them, at least for the mean level.

Note that for milk yield and somatic cell count, inclusion of a sire genetic effect instead of the animal genetic effect could be a way to increase

computing speed, but inclusion of a sire-dam genetic effect will only slightly reduce complexity, because few cows are full sibs.

7.2 Simulation study for fixed residual variance level dispersion

The agreement between the results for the litter size data analyzed with fixed dispersion ($\sigma_v^2 = 1$) in the Gamma distribution of the squared residual variances (Table 7), and the results obtained by Sorensen & Waagepetersen (2003), indicates that fitting the DHGLM with fixed dispersion for the residual variance level might give better estimates than when the dispersion is allowed to vary freely. A simulation study comparing DHGLM with and without fixed dispersion for the residual variance level could provide more insight.

For the three models and data sets analyzed with fixed dispersion (Table 7), convergence was only obtained for one of them, and another aim of a simulation study could be to compare the frequency of convergence between DHGLM with fixed dispersion for the residual variance level, and the DHGLM algorithm as used in the Papers I-IV (under- or over-dispersion allowed).

7.3 Sire and sire-dam genetic effects and different effects for the mean and residual variance level

The correction described in section 5.3.2 should be implemented in the heteroscedastic analysis of teat count (Paper IV), because the residual variance contains additive genetic variance. The correction could be implemented as a standardized tool in a future project.

Future research on how to correct for different genetic effects for the mean and residual variance would also be interesting. The heteroscedastic model with the animal genetic effect for the mean level, and the sire or the sire-dam effect for the residual variance level is intuitively attractive, and the residual variance is purely environmental, but the estimates of the genetic variance components and the genetic correlation were severely biased in most examples studied. A correction of this would be desirable.

7.4 Genetic effects included in any variance component

Hill & Mulder (2010) suggested to include a genetic effect in both the residual variance and in the variance component of the permanent environmental effect.

In this thesis only the residuals are assumed to be heteroscedastic, but in a DHGLM setting there are no theoretical limitations to structuring also other variance components. In addition to the suggestion on structuring the variance

component for a permanent environmental effect, an interesting approach could be to model the additive genetic effects to be heteroscedastic by structuring the additive genetic variance σ_a^2 with genetic and environmental effects.

It is possible to do the structuring of any variance component in a small scale (factors with ten or fewer levels) using the software GenStat. If GenStat were extended to include sparse matrices and tools for pedigree handling, mainly computing power together with the size and structure of the data would limit the possibility to include random effects in the residual variance and other variance components.

7.5 DHGLM without approximations

The DHGLM used in Papers I-IV is approximated by normal distributions (Paper II). This was necessary to use the algorithm on large data, and to implement the algorithm in ASReml.

When repeated observations are too few (per animal or per genetic group, 'too few' has not been derived as a specific number or fraction, this could also be done in future research), it has been seen that estimates are biased. It might be that the normal distribution approximation causes this bias, and that the bias will disappear if the DHGLM is not approximated.

Future research could look into this. For instance it would be an option that GenStat was adapted to handle pedigrees, because choosing higher order Laplace approximations is already possible in GenStat. However, the higher order approximations require more computer power and time, so the fit of DHGLM on large data will probably not be feasible until calculation efficiency has been increased.

7.6 Other trait distributions than normal

Some of the traits considered in this thesis should intuitively be modelled by a Poisson distribution (litter size, teat count and somatic cell count), while the normal distribution modelling of milk yield is intuitively correct.

The transformation of somatic cell count into somatic cell score solves the problem of normality for SCC. For the litter size and teat count traits, arguing for a normal distribution approximation of the Poisson distributions is reasonable in view of the size of the data, and also because of the quantitative genetic assumption that a trait value is the sum of many loads (Fisher, 1918).

Future research could however include other models for trait values. Many distributions are accessible in the DHGLM setting, those are for instance normal, Poisson, binomial, Gamma, and negative binomial.

As described in the previous section, it is not yet feasible to fit a DHGLM without approximation for large data sets. Hence, using other models than the normal, makes estimation difficult, and solving the estimation problems makes estimates biased, but still these barriers could perhaps be overcome.

7.7 Other residual variance models than the exponential

Figure 1 illustrates a linear connection between the mean and the residual variance of teat count for paternal half sib groups. This questions the use of the exponential structure of residual variance.

Other models for residual variance have been suggested. Two of them are the additive residual variance model in which φ itself is considered additive, and the standard deviation model in which $\sqrt{\varphi}$ is considered additive. The problem of both of these, is the requirement for both φ and $\sqrt{\varphi}$ to be larger than zero, a problem that is solved by additive modelling of $\log \varphi$.

Future research could include alternative links for φ. DHGLM has the tools to handle many links (identity, log, inverse, logit (for estimation of probabilities), probit (similar to logit, threshold by cumulative normal distribution), and complementary log log (also complement to logit)), but not the square root, which therefore should be implemented if possible (Lee *et al.*, 2006).

7.8 Model assessment

Model assessment has been studied by Mulder *et al.* (2013), and future research should include and develop these tools, preferable implement standardized tools for selection of models.

7.9 Other uses of DHGLM

7.9.1 Genome wide association studies

In this thesis, analysis of genetically structured residual variance heterogeneity by DHGLM was used in a quantitative genetic setting. In other studies DHGLM has been applied to single nucleotide polymorphism (SNP) marker data to perform genome wide association studies (Shen *et al.*, 2011). In usual analysis of SNP data, p-values are compared to find important loci. Using

DHGLM the estimated variance components are compared among loci to find the controlling genes.

Rönnegård & Valdar (2011) took this a step forward and used DHGLM to study variance controlling genes. Associations were estimated on both the mean and residual variance level for the trait studied.

Future use of DHGLM for molecular genetic data is an area to develop.

7.9.2 Other uses of DHGLM with correlation between effects for the mean and residual variance level

Other uses of DHGLM with correlation between effects for the mean and the residual variance level include, among many possible topics, finance data (Lee & Nelder, 2006) and spatial modelling.

7.10 Teat count in pigs as a model trait

Intuitively, teat count in pigs should be highly genetic. It is difficult to imagine any environmental influence on the trait other than for instance hormonal states, disease, or stress in the sow as the fetus develops. When the teats have been developed, the number of them will not be changed.

This is contrary to the traits litter size, milk yield, and somatic cell count, where more environmental factors can be in continuous action during a lifetime, for instance feed quality, feed intake, stress, bacteria, and disease.

The teat count could therefore be an exceptional model trait (similar to the example of abdominal bristles in flies (Falconer & Mackay, 1996)) for studying genetic variation. Not only it is intuitively genetic, it also takes a span of trait values, which makes it easier to model than a binomial trait.

An intriguing topic for future research would be to combine pedigree, sequence information, and teat count observations to obtain knowledge of the fancy e-topics, such as (gene-) environment interaction, (gene) expression, epistasis, and epigenetics.

8 Sammanfattning på svenska

Inom husdjursavel försöker man ständigt öka produktiviteten genom att öka djurens produktionsegenskaper – till exempel större kullar och ökad tillväxthastighet hos grisar, och ökad mjölkproduktion hos kor. Samtidigt vill man ha uniformitet; kullarna ska vara lika stora, grisarna ska växa lika snabbt och så vidare.

Egenskaper i dottergrupper från olika tjurar kan ha olika variation inom grupperna, trots att gruppen av mödrar borde vara i stort sett identiska. Därför menar man att gener också bidrar till kontroll av variation. Miljöskillnaderna för husdjur är marginella, särskilt inom samma land eller produktionssystem, varför man tänker på skillnaden i variation inom grupper som reaktioner på oidentifierade miljöskillnader. Om det är så att gener kontrollerar uniformiteten, bör man kunna selektera för uniformitet.

Det verkar också finnas samband mellan väntevärde och variation. I så fall, om sambandet är att variationen ökar om väntevärdet ökar, finns en risk med den intensiva aveln som bedrivs att just uniformiteten kan äventyras när väntevärdet ökar.

Modeller som inkluderar genetiska komponenter i både väntevärdet och i residualvariansen, samt korrelationen mellan de två genetiska komponenterna, kan användas för att svara på frågorna. Om den genetiska delen av variationen i residualvariansen är betydande, kan vi kanske selektera för uniformitet. Korrelationen mellan den genetiska komponenten i väntevärdet och den genetiska komponenten i residualvariansen kan ge en indikation på hur uniformiteten påverkas av selektion för ökat väntevärde.

Skattning av modellerna kan göras med Bayesianska metoder, men det tar lång tid, och därför används i denna avhandling istället en metod baserad på teorin för dubbla hierarkiska generaliserade linjära modeller (DHGLM). En algoritm har härletts utifrån DHGLM, och för att kunna använda den på stora datasett, har den approximerats med normalfördelningar. På så sätt kan den användes i standard programvara för husdjursavel, och den har implementerats

i ASReml 4.0. Algoritmen är snabb, och en simuleringsstudie har visat att den leder till bra skattningar när det finns tillräckligt många upprepade observationer per individ eller grupp.

I avhandlingen har algoritmen använts på egenskaperna mjölkproduktion och celltal hos kor samt kullstorleker och spenantal på grisar. De estimerade arvbarhetsvärdena ligger inom de interval som tidigare har rapporterats för båda medelvärdet och residualvariansen. Den genetiska korrelationen mellan medelvärdet och residualvariansen blev endast estimerat för kullstorleker och spenantal. För kullstorleker var den gynnsam, men för spenantal var den ogynnsam. Detta betyder att för spenantal kan det vara så att residualvariansen ökar när medelvärdet ökar, till exempel på grund av selektion för ökat produktion.

References

Ali, A. K. A. & Shook, G. E. (1980). An Optimum Transformation for Somatic Cell Concentration in Milk. *Journal of Dairy Science* 63(3), 487–490.

Breslow, N. E. & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88(421), 9–25.

Canario, L., Lundgren, H., Haandlykken, M. & Rydhmer, L. (2010). Genetics of growth in piglets and the association with homogeneity of body weight within litters. *Journal of Animal Science* 88(4), 1240–1247.

Chalkias, H., Rydhmer, L. & Lundeheim, N. (2013). Genetic analysis of functional and non-functional teats in a population of Yorkshire pigs. *Livestock Science* 152(2–3), 127–134.

Clay, J. S., Vinson, W. E. & White, J. M. (1979). Heterogeneity of Daughter Variances of Sires for Milk Yield. *Journal of Dairy Science* 62(6), 985–989.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Harlow: Longman. ISBN 0-582-24302-5.

Fisher, R. A. (1918). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52(02), 399–433.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. & Thompson, R. (2009). ASReml User Guide Release 3. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK. Available from: www.vsni.co.uk. [Accessed 2014-03-26].

Henderson, C. R. (1953). Estimation of Variance and Covariance Components. *Biometrics* 9(2), 226–252.

Hill, W. G. & Mulder, H. A. (2010). Genetic analysis of environmental variation. *Genetics Research* 92(5-6), 381–395.

Lee, Y. & Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(4), 619–678.

Lee, Y. & Nelder, J. A. (2006). Double Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 55(2), 139–185.

Lee, Y. & Nelder, J. A. (2009a). Likelihood Inference for Models with Unobservables: Another View. *Statistical Science* 24(3), 255–269.

Lee, Y. & Nelder, J. A. (2009b). Rejoinder: Likelihood Inference for Models with Unobservables Another View. *Statistical Science* 24(3), 294–302.

Lee, Y., Nelder, J. A. & Noh, M. (2007). H-likelihood: problems and solutions. *Statistics and Computing* 17(1), 49–55.

Lee, Y., Nelder, J. A. & Pawitan, Y. (2006). *Generalized linear models with random effects : unified analysis via H-likelihood*. Boca Raton: Chapman & Hall/CRC. (Monographs on statistics and applied probability, 0960-6696 ; 106). ISBN 1-58488-631-5 (inb.).

Louis, T. A. (2009). Discussion of Likelihood Inference for Models with Unobservables: Another View. *Statistical Science* 24(3), 270–272.

Lundeheim, N., Chalkias, H. & Rydhmer, L. (2013). Genetic analysis of teat number and litter traits in pigs. *Acta Agriculturae Scandinavica, Section A - Animal Science* 63(3), 121–125.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News* 456(7218), 18–21.

Meng, X.-L. (2009). Decoding the H-likelihood. *Statistical Science* 24(3), 280–293.

Molenberghs, G., Kenward, M. G. & Verbeke, G. (2009). Discussion of Likelihood Inference for Models with Unobservables: Another View. *Statistical Science* 24(3), 273–279.

Mulder, H. A., Bijma, P. & Hill, W. G. (2007). Prediction of Breeding Values and Selection Responses With Genetic Heterogeneity of Environmental Variance. *Genetics* 175(4), 1895–1910.

Mulder, H. A., Hill, W. G., Vereijken, A. & Veerkamp, R. F. (2009). Estimation of Genetic Variation in Residual Variance in Female and Male Broiler Chickens. *Animal* 3(12), 1673–1680.

Mulder, H. A., Rönnegård, L., Fikse, W. F., Veerkamp, R. F. & Strandberg, E. (2013). Estimation of genetic variance for macro- and micro-environmental sensitivity using double hierarchical generalized linear models. *Genetics Selection Evolution* 45(1), 23.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* 135(3), 370.

Nielsen, B., Su, G., Lund, M. S. & Madsen, P. (2013). Selection for increased number of piglets at d 5 after farrowing has increased litter size and reduced piglet mortality. *Journal of Animal Science* 91(6), 2575–2582.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3), 545–554.

Pawitan, Y. (2001). *In all likelihood : statistical modelling and inference using likelihood*. Oxford: Clarendon. (Oxford Science publications). ISBN 0-19-850765-8.

Rönnegård, L. & Lee, Y. (2013). Exploring the potential of hierarchical generalized linear models in animal breeding and genetics. *Journal of Animal Breeding and Genetics* 130(6), 415–416.

Rönnegård, L., Shen, X. & Alam, M. (2014). The hglm Package (Version 2.0). R Package Vignette. Available from: http://r.meteo.uni.wroc.pl/web/packages/hglm/vignettes/hglm.pdf. [Accessed 2014-04-16].

Rönnegård, L. & Valdar, W. (2011). Detecting Major Genetic Loci Controlling Phenotypic Variability in Experimental Crosses. *Genetics* 188(2), 435–447.

Rönnegård, L. & Valdar, W. (2012). Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genetics* 13(1), 63.

SanCristobal-Gaudy, M., Elsen, J.-M., Bodin, L. & Chevalet, C. (1998). Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. *Genetics Selection Evolution* 30(5), 423.

Shen, X., Pettersson, M., Rönnegård, L. & Carlborg, Ö. (2012). Inheritance Beyond Plain Heritability: Variance-Controlling Genes in Arabidopsis thaliana. *PLoS Genet* 8(8), e1002839.

Shen, X., Rönnegård, L. & Carlborg, Ö. (2011). Hierarchical likelihood opens a new way of estimating genetic values using genome-wide dense marker maps. *BMC Proceedings* 5(Suppl 3), S14.

Sonesson, A. K., Ødegård, J. & Rönnegård, L. (2013). Genetic heterogeneity of within-family variance of body weight in Atlantic salmon (Salmo salar). *Genetics Selection Evolution* 45(1), 41.

Sorensen, D. & Waagepetersen, R. (2003). Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genetics Research* 82(03), 207–222.

Swedish Dairy Association (2011). Cattle statistics. Swedish Dairy Association. Available from: http://www.svenskmjolk.se/Global/Dokument/Dokumentarkiv/Statistik/Husdjursstatistik%202011%20-%20webb.pdf. [Accessed 2014-03-27].

Svenska Pig (2012). Sugg medel utveckling. Svenska Pig. Available from: http://www.pigwin.se/medeltal-sugg. [Accessed 2014-03-27].

Thompson, W. A. (1962). The Problem of Negative Estimates of Variance Components. *The Annals of Mathematical Statistics* 33(1), 273–289.
Willham, R. L. (1972). The Role of Maternal Effects in Animal Breeding: III. Biometrical Aspects of Maternal Effects in Animals. *Journal of Animal Science* 35(6), 1288–1293.
Van Vleck, L. D. (1968). Variation of Milk Records Within Paternal-Sib Groups. *Journal of Dairy Science* 51(9), 1465–1470.
Yang, Y., Christensen, O. F. & Sorensen, D. (2011). Analysis of a genetically structured variance heterogeneity model using the Box–Cox transformation. *Genetics Research* 93(01), 33–46.

Acknowledgements

The studies included in this thesis were performed at the School of Technology and Business Studies, Dalarna University (DU) and at the Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences (SLU) with funding equally provided by both institutions. A few months were also financed by the RobustMilk project, which was financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. One conference was financed by Knut and Alice Wallenberg's fund. Several conferences and one course were financed by the research profile Complex Systems – Microdata Analysis, Dalarna University. Another course was financed by European Cooperation in Science and Technology (COST). Data were provided by Daniel Sorensen and the Danish database of the national breeding program, the Swedish pig breeding organization Nordic Genetics, and the Swedish Dairy Association.

While funding is necessary, it is not sufficient. Several people have carried the burden and shared the joy with me.

My supervisors have done a tremendous amount of work. Lars Rönnegård, thank you for introducing me to the world of hierarchical generalized linear models, for unselfishly sharing your ideas and plans with me, and for walking the first part of the journey while I went on maternity leave. Thank you for teaching, especially for volunteering on the scientific writing course. I have attended several conferences financed from your own pocket of research money. Many hours were spent travelling together, hours well used for interesting discussions. The warm welcome to Dalarna at your home was a blessing, as well as your visit after my daughter came into the world. Thank you for allowing me to do popular science during working hours, that gave me a boost in understanding the world of genetics. That you even shared office for a short period of time is something few professors-to-be would have done. Finally, among everything to thank you for, the words patience, belief and involvement comes to my mind.

Erling Strandberg, thank you for coming up with the idea of considering genetic heterogeneity, for letting a mathematician enter the PhD program at SLU, and for the involvement in the contract of a shared PhD student position with DU. Our meetings and your advice have always given new energy to the project, and your way to do things is a perfect example. During the concluding phase, I was privileged and very happy to have many opportunities to listen to your thoughts, and I did not only take advantage of your insight, but also appreciated to get to know you a little better. Thank you for maintaining the structure of my PhD education at SLU, surely with many parts I never even noticed.

I also want to thank people who helped me with administration; those are Jan-Ove Netsby, Ulrika Eriksson, Anna Maria Gylling, Helena Pettersson, Harriet Staffans, Monica Jansson, Anniqua Melin, Jörgen Sahlin, and surely somebody who I never saw, you know who you are. Computer issues were solved by Dan Englund and David Hammarbäck. Thanks also to study directors Anne Lundén and Susanne Eriksson.

The list of people without whose effort there would be no doctoral degree is long.

Anders Forsman, thank you for the open door into your office welcoming me and everybody to get wise thoughts on every aspect of life, humbly sharing your own experiences.

Dirk-Jan De Koning, thank you for inviting me to climb the trees, for advice, nice kick offs, and for the house in Edinburgh. I have always felt fully included in the Quantitative Genetics group, even though working on a distance.

Without co-authors, no papers, no thesis. Nils Lundeheim, thank you, my emails were answered quickly, and manuscripts were returned with clear comments and suggestions. Youngjo Lee, thank you also for skiing together in Orsa. Dongwhan Lee, thank you for lending out your mathematical skills. Arthur Gilmour, thank you for the tremendous job you have done answering my emails and implementing new commands in ASReml, thank you for your advice, and for being a native English speaker, and, I am happy that we even share the same faith. Han Mulder, thank you for many interesting discussions and answers. Freddy Fikse, thank you for answering questions and knowing the most. Helena Chalkias, thank you for introducing me to teats in pigs.

Teachers are always judged hard, but mine were all of the kind to be put on pedestals. Hossein Jorjani, your clever words will ring in my brain for the rest of my life. Moudud Alam, thank you for discussions in which your clever inputs were always present, and for knowing the foundations of statistics.

Göran Andersson, I appreciated your advice and your ability to run away on side tracks.

My colleagues at Dalarna University have made everyday life a pleasure. Thanks to Richard Stridbeck (also for the all of the coffee breaks), Mark Dougherty (for your skiing lessons as well), and Changli He for the experience of teaching together. Ola Nääs, your friendly good morning is a lovely sound in the corridors, making mornings good. Thanks to Dao Li, Mengjie Han, Kristin Svensson, and Daniel Wikström for sharing office. Yujiao Li, thanks for being an encouraging student and later on a colleague. Thanks to you and also to Xiaoyun Zhao and Catia Cialani for partnering at the medicine ball exercises, making me smile and laugh through it all. Xiangli Meng, thank you for always telling good stories, and Xia Shen, thank you for the example of well-prepared lectures (like art). Kenneth Carling, thank you for waiting patiently in Borlänge for the interview while I tried to find you in Falun. Thank you for financing conferences and trips. I have moreover benefited a lot from the well prepared medicine ball exercises, and in addition you came up with a food suggestion to achieve maximal strength and fitness.

I always felt very welcome in Uppsala at Swedish University of Agricultural Sciences, every time finding people who cared about a stranger and who would make a space available at the lunch table. Especially thanks to people in the division of Quantitative Genetics for letting me belong to the group, for input on my research, and for keeping an office space available for me. Elisabeth 'Lisa' Jonas, thank you for the wonderful trip to New Zealand, and for letting med stay overnight at your place and use your kitchen several times. Thanks for sharing room in the house in Edinburgh, and I enjoyed getting to know you, Anna Maria Johansson, there as well. Jessica Franzén, your short entrance into my life inspired me a lot. Örjan Carlborg, thank you for letting me be a part of the Computational Genetics group at occasions. I am very thankful to all course mates, particularly Sofia Nyman (thanks also for the accommodation offer), Helena Eken Asp, and Emelie Zonabend, for letting me participate in interesting discussions on Skype.

Several people outside the universities made it all possible. Thanks to Dan Kristensson, Agno Kristensson, David Hammarbäck, and Bruno Hopstadius for traveling together to Borlänge and during the journeys to listen to both frustrations and celebrations, especially Dan for persistence through it all. Without you to pick me up or to expect my arrival early morning I do not know if there would be any dissertation at all.

Britt-Marie Laurell, you solved the most difficult part of the project. I will be forever grateful to you for all you have done. Surely my epigenetics is totally changed because of you. Thanks also to Åke Laurell for goodness

through all years. Magda Laurell, your home was always open for my children (or did they just walk in?), giving them almost a second home during intensive periods.

Many friends are far away in distance but always close to my heart. Kerstin Strandanger, you never believed that I would be a full time house wife. Bende Sylvan, Bolette Skamling, Dorthe Lykke Jensen, Esther Bjerg Dalgaard, Lea Barslund Lauridsen, Lene Tvilling and Jeff Rømer with children in Copenhagen, the Johnston, Mulligan, and Smelt family in New Zealand, several families in Santa Barbara, thank you for being in my life!

My extended family in the church in Siljansnäs adds a substantial part to everyday happiness and therefore efficiency. Mentioning some is risky, but mentioning none feels empty. I am thankful to both small and grown in families Back, Baesch, Dannvik, Demasure, Edenius, Ekman, Enqvist, Fridborg, Furingsten, Hammarbäck, Hopstadius, Jonsson, Kristensson, Kyller, Laurell, Lind, Lindberg, Matsson, Matsson, Milutin, Måhl, Olsson, Plars, Sandström, Santana, Sundkvist, Sved, Wilhelmsson, and Zetterlund. Thanks to the Lord for carefully puzzling my life together, for sending opportunities just at the right time, and for guidance through it all.

Jonas, thank you for everything you have done to make it possible for me to do this journey. Thank you for letting me travel all over the world and for meanwhile keeping everything at home running perfectly. Your support and encouragement did never fail, and you have believed in the project and my brilliance all way through. Sandra and Elliot, thank you for making me laugh, and for being fully present in the important things of life when I became too focused on working. You make every day exciting, unpredictable, and meaningful. Love you!