Mixed Models Through The Lens of hglm: Applications and Grand Challenges

Xia Shen* Moudud Alam† Lars Rönnegård‡

Abstract

The **hglm** package is a hierarchical-likelihood-based solution for mixed models. Apart from generalized linear mixed models (GLMM), hierarchical generalized linear models (HGLM) can also solve models with non-Gaussian random effects, structured dispersion parameters, and correlated random effects. **hglm** provides a unified approach to various statistical modeling problems. We describe examples in our interdisciplinary research based on the **hglm** package, dealing with large-scale biological data, chemometrics data and geographical data. Thereafter, we discuss some big challenges that empirical scientists desire to solve using mixed models, including modeling high-dimensional interaction effects, having random effects in the mixed model dispersion parameters, joint modeling of spatial and genetic correlations, and multivariate analyses with random effects.

Key Words: R/hglm package, hierarchical generalized linear models, generalized linear mixed models, high-dimensional data, correlated random effects, non-Gaussian random effects.

1. Introduction

Nowadays, random effects modeling is becoming more and more essential for understanding complex data in empirical sciences. The complexity in the "big data" contains grand challenges in terms of size, structure and interpretation, for which random effects models, *a.k.a.* linear mixed models, have great potentials to dissect the problems.

The **hglm** package implements the estimation algorithm for hierarchical generalized linear models (HGLM; Lee and Nelder, 1996). The package fits generalized linear models (GLM; McCullagh and Nelder, 1989) with random effects, where the random effect may come from a conjugate exponential-family distribution (normal, gamma, beta or inverse-gamma). The user may explicitly specify the design matrices both for the fixed and random effects, which means that correlated random effects as well as random regression models can be fitted. Dispersion parameters in the model may also be modeled.

hglm produces estimates of fixed effects, random effects, variance components as well as their standard errors. In the output it also produces diagnostics quantities such as deviances and leverages and related plots.

Generalized linear mixed models (GLMM) have previously been implemented in several R (R Development Core Team, 2011) procedures, such as the `glmer()` function in the **lme4** package and in the `glmmPQL()` function in the **MASS** package. In GLMM, the random effects are assumed to be Gaussian whereas **hglm** allow for other distributions for

*Division of Computational Genetics, Department of Clinical Sciences, Swedish University of Agricultural Sciences, Ulls väg 24E, SE-750 07, Uppsala, Sweden; Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, SE-171 77 Stockholm, Sweden; MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, EH4 2XU Edinburgh, UK

†Statistics, School of Technology and Business Studies, Dalarna University, Röda vägen 3, SE-781 70 Borlänge, Sweden

‡Statistics, School of Technology and Business Studies, Dalarna University, Röda vägen 3, SE-781 70 Borlänge, Sweden; Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Ulls väg 26, SE-750 07, Uppsala, Sweden; Division of Computational Genetics, Department of Clinical Sciences, Swedish University of Agricultural Sciences, Ulls väg 24E, SE-750 07, Uppsala, Sweden

the random effects. **hglm** extends the fitting algorithm of the **dglm** package by including random effects in the linear predictor for the mean. Moreover, the model specification in **hglm** can be given as a formula or alternatively in terms of `y`, `X`, `Z` and `X.disp`, where `y` is the vector of observed responses, `X` and `Z` are the design matrices of the fixed and random effects, respectively, and in the linear predictor for the mean, `X.disp` is the design matrix of the fixed effects for the dispersion parameter. This enables a more flexible modeling of the random effects than specifying the model by an R formula. Consequently, this option is not as user friendly but gives the user a possibility to fit random regression models and random effects with known correlation structure. Such an input feature can be particularly useful when dealing with complex correlated data (*e.g.* spatial data) or high-throughput data (*e.g.* DNA sequencing data). Specifically, the `bigRR = TRUE` option enables **hglm** to interact with another package of ours, **bigRR** (Shen et al., 2013), to efficiently fit high-dimensional data ("$p \gg n$").

Here, we present examples in our interdisciplinary research based on the **hglm** package, dealing with different types of data in biology and chemistry. We highlight some important challenges in these fields that require further investigation in statistical modeling using random effects.

2. Applications using hglm

Although one of the unique features in **hglm** is to fit non-Gaussian random effects, dealing with correlated random effects, modeling dispersion parameters and even handling high-dimensionality appear to be more useful features in **hglm** according to our empirical applications. We provide examples in this section, illustrating different data types and how we have modeled these data using **hglm**.

2.1 hglm and large-scale "omics" data

One of the ultimate goals in genetics or even the general biology is to understand the link between our *phenotypes* (*i.e.* complex traits and diseases such as height, weight, blood presure, cancer, etc.) and our genetic materials or *genotypes* (*i.e.* information written in the DNA). Modern biotechnology allows us to sequence the DNA of a large number of individuals, which produces "big data" containing information of many positions (*a.k.a.* loci) along the *genome*. Using such data, one could, for instance, test the association between a particular phenotype against each position of the genome, in order to infer functional genes for the phenotype. This strategy is known as genome-wide association study (GWAS).

Since Klein et al. (2005) reported a causal polymorphism of complement factor H that regulates age-related macular degeneration (AMD) in a cover letter of *Science* published nine years ago, more and more genes/loci have been identified via GWAS, regulating for instance, human disease-related traits such as blood pressure (Levy et al., 2009), blood lipids (Aulchenko et al., 2009; Teslovich et al., 2010), coronary heart disease (CARDIoGRAMplusC4D Consortium et al., 2013) and breast cancer (Turnbull et al., 2010), as well as complex traits in other species such as mice (Valdar et al., 2006), maize (Tian et al., 2011), Arabidopsis (Atwell et al., 2010) and so on. Many statistical methods have been developed for GWAS or genomic data analysis (see reviews by Balding, 2006; Cantor et al., 2010), together with quite a few computational tools (*e.g.* Aulchenko et al., 2007; Purcell et al., 2007; Yang et al., 2010).

However, most of the existing methods and studies consider only the genomic information (DNA level), ignoring that it is a complex regulative process from DNA to phenotype. A chain of interconnected information is available nowadays (Figure 1), with data from

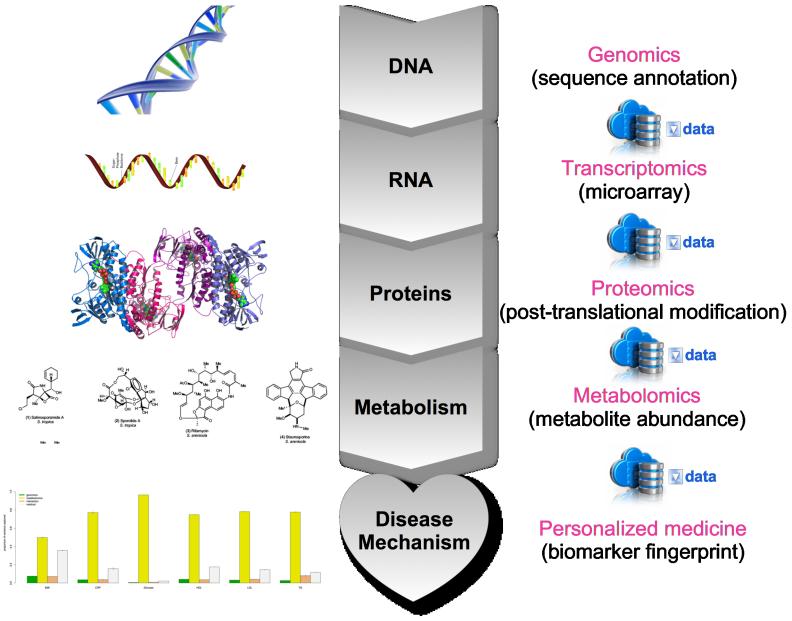

Figure 1: Hierarchical structure from fundamental genomics to personalized treatment for diseases. Each level of omics data contain plenty of information contributes to the ultimate disease incidence. Properly combining all the omics information could improve our understanding of disease mechanism. However, modeling such high-throughput hierarchical data structure is a great challenge both statistically and computationally.

DNA (genomics), RNA (transcriptomics), proteins (proteomics or metabolomics), etc. The hierarchical nature of these *omics* data can hardly be neglected and requires more investigation through statistical modeling.

In our preliminary analysis of approx. 10,000 people (data description not provided due to confidentiality), we considered genomics and metabolomics to be two, potentially interacting, contributors to the phenotype, the following model with multiple correlated random effects was be fitted using **hglm**:

$$\begin{aligned}
\mathbf{y}|\boldsymbol{\beta},\mathbf{g},\mathbf{m},\mathbf{a},\boldsymbol{\theta} &\sim N(\mathbf{X}\boldsymbol{\beta}+\mathbf{g}+\mathbf{m}+\mathbf{a},\mathbf{I}\sigma^2) && (1)\\
\mathbf{g} &\sim N(0,\mathbf{G}\sigma_g^2) && (2)\\
\mathbf{m} &\sim N(0,\mathbf{M}\sigma_m^2) && (3)\\
\mathbf{a} &\sim N(0,\mathbf{G}\circ\mathbf{M}\sigma_{gm}^2) && (4)
\end{aligned}$$

where y is the phenotypic response vector, $\boldsymbol{\beta}$ the fixed effects (sex and age) with design matrix $\mathbf{X}$, $\mathbf{g}$ the random genomic effects, $\mathbf{m}$ the random metabolomic effects and $\mathbf{a}$ the random interaction effects between the genome and metabolome. $\boldsymbol{\theta}$ denotes the vector of variance components $(\sigma_g^2,\sigma_m^2,\sigma_{gm}^2,\sigma^2)'$. Each of the random effects terms are correlated due to the genomic relationship (given by $\mathbf{G}$) and similarity in metabolomic profiles (given by $\mathbf{M}$) of the individuals. The correlation structure for the interaction between the two is constructed as the Hadamard product $\mathbf{G}\circ\mathbf{M}$ as described earlier in relation to modeling of random epistatic effects (Rönnegård et al., 2008). This is a direct polygenic-poly-metabolite way to test the contribution of each level of omics data using a linear mixed model. After computing the matrices `G` and `M` in R, model (4) can be fitted by the following commands:

```
svd.G <- svd(G) ## SVD is used here and onwards
                ## because it's numerically more stable
```

```
svd.M <- svd(M)
svd.GM <- svd(G*M)
Z.G <- svd.G$u %*% diag(sqrt(svd.G$d)) ## calculating design matrix for
                                       ## genetic effects
Z.M <- svd.M$u %*% diag(sqrt(svd.M$d)) ## calculating design matrix for
                                       ## metabolites effects
Z.GM <- svd.GM$u %*% diag(sqrt(svd.GM$d)) ## calculating design matrix for
                                          ## the interaction effects
require(hglm)
model1 <- hglm(y = y, ## response vector
               X = model.matrix(~ sex + age), ## fixed effects design matrix
               Z = cbind(Z.G, Z.M, Z.GM), ## 3 random effects design matrices
               RandC = c(ncol(Z.G), ncol(Z.M), ncol(Z.GM)) ## column numbers
               )
```

The model fitting results indicated that omics data beyond genomics, such as metabolomics, could explain, *e.g.* approx. 5 times more observed variance of body mass index (BMI) than the genome itself. Although this is a rather crude way of incorporating different types of omics data, one can realize the potential of modeling "big data" in the biological regulative hierarchy.

Another strategy that we have investigated is to "weight" different genomic markers (predictors) differently, in order to obtain a model with better predictive performance. Without information beyond the genome, a double HGLM (DHGLM) can be fitted by iterating between two HGLMs, one for the mean, the other for the dispersion parameter of the random effects (Shen et al., 2011). For a complex trait, the phenotype $\mathbf{y}$ ($n \times 1$ vector) is postulated as a random effect model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \tag{5}$$

where $\mathbf{g} \sim N(\mathbf{0}, \mathrm{diag}(\boldsymbol{\lambda}))$ are the effects of genomic markers, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)'$ are the variances of the SNP effects, and the residuals $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$. The fixed effects $\boldsymbol{\beta}$ included an intercept and other fixed effects to reduce the residual errors. The variance components $\boldsymbol{\lambda}$ are modeled as

$$\log \boldsymbol{\lambda} = \mathbf{1}a + \mathbf{b} \tag{6}$$

with an intercept a as fixed effect and normally distributed random effects $\mathbf{b}$. When $\mathbf{b}$ is i.i.d., the above model is equivalent to the well-known "BayesA" model (Meuwissen et al., 2001) in the genomic prediction area. Such a model, even though it uses nothing else but the genomic information, applies variable shrinkage to the loci effects and therefore better predictive power. However, due to the high-dimensionality in the big genomic data, fitting this DHGLM can be computationally very heavy or even impossible within a reasonable time frame. We showed that such DHGLM can be simplified as a 2-step generalized ridge regression, which is much more efficient to compute, without losing its predictive performance (Shen et al., 2013). Our generalized ridge regression model was named "heteroscedastic effects model" (HEM) and implemented in the **bigRR** package. With the data $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{Z}$ loaded, the above model can be fitted as:

```
require(bigRR)
RR <- bigRR(y = y, X = X, Z = Z)  ## fitting a big ridge regression
                                  ## without b in eq. (6)
HEM <- bigRR_update(RR, Z = Z)  ## fitting heteroscedastic effects model
```

Our algorithm and implementation in R are quite efficient especially when the number of observations is not very big, *e.g.* a HEM with 100 observations and 1,000,000 predictors can be fitted in approx. 3 minutes on an ordinary laptop. It would be ideal that one can

incorporate other types of omics data or biological information of different genomic markers into the dispersion part of the model, however, no successful improvement in prediction has been found in our trials, and this topic remains a challenge to genetisists.

2.2 hglm and high-dimensional chemometric analysis

Prediction problems using high-dimensional data exist in various fields of research, *e.g.* analytical chemistry or *chemometrics*. One aim in chemometrics is to infer the contents of a chemical mixture according to a spectrum profile from *e.g.* Fourier transform infrared spectroscopy (FTIR) containing signals at different wavelengths. The major reason for performing such prediction analysis is that determination of chemical compounds in a mixture through separation methods such as high-performance liquid chromatography (HPLC) is very costly. Successful inference of the mixture contents via a unified FTIR spectrum would reduce the cost and also improve efficiency. Different from the genomic prediction problem in biology, the predictors from FTIR (*i.e.* the $\mathbf{Z}$ matrix using the notation above) are actually determined by the response variable $\mathbf{y}$, but the mathematical association between $\mathbf{y}$ and $\mathbf{Z}$ can be treated similarly as in genetics.

Partial least squares (PLS) regression has been a classic and popular method to conduct chemometric prediction. While recently, we showed that in our real experimental silage samples, HEM, as a representative of DHGLM, out-performed PLS and ridge regression (RR) in terms of predictive power (Shen et al., 2014). For about 70% of the cases, HEM, a linear mixed model with structured dispersion of random effects, predicted better than PLS, and for approx. 90% of the cases, HEM dominated RR. One reason why such a linear mixed model with re-weighted random effects could perform well, is that the architecture inherited in the data fits the intermediate shrinkage of the model. In both genetics and chemometrics, one can find that only a small number of the predictors have relatively large effects, whereas the rest of the predictors cannot be ignored either since the sum of their small effects contributes significantly to the prediction. HEM or similar DHGLMs allow us to properly penalize the random effects differently according to their contributions to the variance of the response variable.

2.3 Spatial modeling using hglm

In **hglm**, dealing with correlated random effects through flexible user-defined design matrices is a useful feature that most mixed model packages misses. The reason why we implemented the package in this way initially was that many of our modeling problems in practice require correlated structure in the random effects. Besides relatedness among individuals due to genetics introduced above, spatial relatedness due to *e.g.* different sampling locations could also be considered. From version 2.0 of **hglm**, we developed and included a new algorithm for fitting spatially correlated random effects (Alam et al., 2014). A new random effects family `CAR` is included for fitting such spatial generalized linear models with conditionally autoregressive (CAR; Besag, 1974) random effects.

Using this new feature, we carried out a novel analysis, trying to incorporate both the genetic and spatial correlation structure in approx. 600 Scots pines sampled in northern Sweden by The Forestry Research Institute of Sweden, SkogForsk. The trees' height $\mathbf{y}$ was modeled by a linear mixed model:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{s} + \mathbf{e} \tag{7}$$

where $\boldsymbol{\mu}$ is an intercept term, $\mathbf{Z}$ and $\mathbf{W}$ incidence matrices connecting the individual random effects with the observed phenotypes, $\mathbf{a}$ normally distributed additive genetic effects

with variance-covariance matrix $\mathbf{A}\sigma_a^2$ with $\mathbf{A}$ being the additive relationship matrix (*e.g.* see pp. 442-444 of Pawitan, 2001), $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ the residuals, and $\mathbf{s}$ are the random spatial effects that follow a CAR model, *i.e.* the inverse of the variance-covariance matrix of $\mathbf{s}$ is given by

$$\mathbf{\Sigma}^{-1} = \frac{1}{\tau}(\mathbf{I} - \rho\mathbf{D}) \tag{8}$$

where τ is a spatial variance parameter, ρ a spatial correlation parameter, $\mathbf{I}$ the identity matrix, and $\mathbf{D}$ is a neighborhood matrix having elements 0's and 1's indicating which seedlings are adjacent to each other. In **hglm** 2.0, with data `y`, `A` and `D` loaded, the above linear mixed model including two random effects terms with different distributions can be fitted as:

```
svd.A <- svd(A)
Z.A <- svd.A$u %*% diag(sqrt(svd.A$d))
n <- length(y)
require(hglm)
model2 <- hglm(y = y, ## response vector
               X = matrix(1, n, 1), ## intercept-only fixed effect
               Z = cbind(diag(n), Z.A), ## 2 random effects design matrices
               rand.family = list(CAR(D = D), gaussian()), ## different families
               RandC = c(100, 20) ## column numbers specified
               )
```

The estimated spatial correlation parameter was $\hat{\rho} = 0.126$. Compared to a linear mixed model with only the genetic effects, including the spatial effects reduced the genetic variance estimate by a quarter and the residual variance estimate by a third. These results confirmed the importance of including spatial effects in plant trials.

3. Grand challenges

Besides **hglm**, there are other R packages that solve linear mixed models, and together as a R-based framework, we face quite a few practical challenges for which the current mixed model packages have not yet been well established. We propose the future developments to handle the following issues based on linear mixed models. We foresee the great importance of such contributions to empirical scientific research.

3.1 Challenge I: high-dimensionality

"Big data", as a fashionable term, has been emphasized in various disciplines. Generally speaking, "big data" are large, population-scale data sets, typically from innovations in high-throughput technology, *e.g.* genome sequencing, internet technology, image processing, etc. These "big data" are so large and so high-dimensional that even storing, transferring and visualizing can be quite difficult. The data are usually collected by many different ways and different groups and require advanced computational tools and new statistical methods to analyze. Therefore, the first challenge we normally face in practical data analysis is to handle high-dimensional data, by fitting a feasible high-dimensional mixed model for instance, on an accessible computer. For example, our HEM method described above is a reduced DHGLM, which is capable of estimating hundreds of thousands of effects, although sacrificed the full likelihood. The size of "big data" could be the first practical concern in the future developments of mixed model packages.

3.2 Challenge II: correlated random effects & multicollinearity

The high-dimensional random effects or predictors usually show another characteristic in real data, *i.e.* some of the variables, usually adjacent to each other according a sequential

order, can be highly correlated (see Figure 1 in Shen et al., 2014, for a visualization of such correlation in genomic and FTIR spectrum data). Although it is not a computational problem for fitting linear mixed models or ridge regression in such data, properly modeling such multicollinearity structure, *e.g.* in the dispersion parameter of the random effects, could dramatically improve the estimation and predictive power.

3.3 Challenge III: hierarchical data structure

As we have introduced in Figure 1, it has been a very challenging problem to model hierarchical omics data structure in biology. Superficially, we obtained a collection of different types of explanatory variables or predictors, but the underlying biological meaning of the data told us that one set of these predictors may have effects on another. A naive model for such structure underlying a complex trait $\mathbf{y}$ could be:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{u} + \mathbf{e} \tag{9}$$

where $\mathbf{a} \sim N(0, \mathbf{G}\sigma_g^2)$ are the genetic effects, $\mathbf{u} \sim N(0, \mathbf{M}\sigma_m^2)$ the higher-level omic effects such as metabolomic effects, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ the fixed effects and $\mathbf{e}$ the residuals. Note that both $\mathbf{a}$ and $\mathbf{u}$ are *individual* effects, having the same size as $\mathbf{y}$, which means both $\mathbf{Z}$ and $\mathbf{W}$ are square matrices. In this way, the omics effects could be further modeled by the genetic effects according to the hierarchy:

$$\mathbf{u} = \boldsymbol{\mu}' + \mathbf{Z}\mathbf{a}' + \mathbf{e}' \tag{10}$$

where $\mathbf{a}' \sim N(0, \mathbf{G}\sigma_g'^2)$ are the genetic effects, $\boldsymbol{\mu}' = \mathbf{X}\boldsymbol{\beta}'$ the fixed effects and $\mathbf{e}'$ the residuals. Thus, we obtain a unified mixed model:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{W}\boldsymbol{\mu}' + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{Z}\mathbf{a}' + \mathbf{W}\mathbf{e}' + \mathbf{e} \tag{11}$$

where the term $\mathbf{W}\mathbf{Z}\mathbf{a}'$ should explain the variance in $\mathbf{y}$ that is due to the "chain" of DNA - protein - phenotype, rather than the direct genetic effects $\mathbf{Z}\mathbf{a}$ and omics effects $\mathbf{W}\mathbf{e}'$. Such a model with multiple correlated random effects will be our next investigation, and in general, modeling hierarchical data structure is a strongly needed topic that requires more theoretical and empirical investigations.

3.4 Challenge IV: structured dispersion

We have described examples of modeling dispersion parameters as DHGLM, however, a general and flexible framework for including fixed and random effects in different dispersion parameters is still far from well established. An ideal mixed model toolbox should contain a mixed model module that can be adopted in different parts of a model - the mean and any dispersion parameter. This is particularly important since more and more empirical evidence of variance heterogeneity has been found in genetics (*e.g.* Shen et al., 2012; Geiler-Samerotte et al., 2013). Dispersion parts of linear mixed models would become as essential as the regular linear mixed model for the mean (Lee and Nelder, 2006; Lee et al., 2006).

3.5 Challenge V: high order corrections for discrete responses

A major drawback of **hglm** is the use of extended quasi-likelihood (`EQL`) method which might produce biased results, especially for discrete responses such as binary or Poisson data. In the latest version of **hglm**, we have implemented `HL11` correction to reduce this

problem. The `HL11` method gives improved estimates compared to `EQL` for a Poisson GLMM when the number of levels in the random effect are large and i.i.d. (see **hglm** vignette for simulation results). The implementation follows the Appendix in Lee and Lee (2012). However, higher order corrections may be more useful, but they would make the estimation procedure considerably slow. Implementation of efficient higher order corrections for discrete data is challenging and would benefit the use of mixed models.

References

M. Alam, L. Rönnegård, and X. Shen. Fitting spatial models in hglm. *Submitted*, 2014.

S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, R. Jiang, N. W. Muliyati, X. Zhang, M. A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J. R. Ecker, N. Faure, J. M. Kniskern, J. D. G. Jones, T. Michael, A. Nemri, F. Roux, D. E. Salt, C. Tang, M. Todesco, M. B. Traw, D. Weigel, P. Marjoram, J. O. Borevitz, J. Bergelson, and M. Nordborg. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298): 627–631, June 2010.

Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn. GenABEL: an R library for genome-wide association analysis. *Bioinformatics (Oxford, England)*, 23(10):1294–1296, May 2007.

Y. S. Aulchenko, S. Ripatti, I. Lindqvist, D. Boomsma, I. M. Heid, P. P. Pramstaller, B. W. J. H. Penninx, A. C. J. W. Janssens, J. F. Wilson, T. Spector, N. G. Martin, N. L. Pedersen, K. O. Kyvik, J. Kaprio, A. Hofman, N. B. Freimer, M.-R. Jarvelin, U. Gyllensten, H. Campbell, I. Rudan, A. Johansson, F. Marroni, C. Hayward, V. Vitart, I. Jonasson, C. Pattaro, A. Wright, N. Hastie, I. Pichler, A. A. Hicks, M. Falchi, G. Willemsen, J.-J. Hottenga, E. J. C. de Geus, G. W. Montgomery, J. Whitfield, P. Magnusson, J. Saharinen, M. Perola, K. Silander, A. Isaacs, E. J. G. Sijbrands, A. G. Uitterlinden, J. C. M. Witteman, B. A. Oostra, P. Elliott, A. Ruokonen, C. Sabatti, C. Gieger, T. Meitinger, F. Kronenberg, A. Döring, H.-E. Wichmann, J. H. Smit, M. I. McCarthy, C. M. van Duijn, L. Peltonen, and ENGAGE Consortium. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature Genetics*, 41(1):47–55, Jan. 2009.

D. J. Balding. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10): 781–791, Oct. 2006.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 1974.

R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American journal of human genetics*, 86:6–22, 2010.

CARDIoGRAMplusC4D Consortium, P. Deloukas, S. Kanoni, C. Willenborg, M. Farrall, T. L. Assimes, J. R. Thompson, E. Ingelsson, D. Saleheen, J. Erdmann, B. A. Goldstein, K. Stirrups, I. R. König, J.-B. Cazier, A. Johansson, A. S. Hall, J.-Y. Lee, C. J. Willer, J. C. Chambers, T. Esko, L. Folkersen, A. Goel, E. Grundberg, A. S. Havulinna, W. K. Ho, J. C. Hopewell, N. Eriksson, M. E. Kleber, K. Kristiansson, P. Lundmark, L.-P. Lyytikäinen, S. Rafelt, D. Shungin, R. J. Strawbridge, G. Thorleifsson, E. Tikkanen, N. Van Zuydam, B. F. Voight, L. L. Waite, W. Zhang, A. Ziegler, D. Absher, D. Altshuler, A. J. Balmforth, I. Barroso, P. S. Braund, C. Burgdorf, S. Claudi-Boehm, D. Cox, M. Dimitriou, R. Do, DIAGRAM Consortium, CARDIOGENICS Consortium, A. S. F. Doney, N. El Mokhtari, P. Eriksson, K. Fischer, P. Fontanillas, A. Franco-Cereceda, B. Gigante, L. Groop, S. Gustafsson, J. Hager, G. Hallmans, B.-G. Han, S. E. Hunt, H. M. Kang, T. Illig, T. Kessler, J. W. Knowles, G. Kolovou, J. Kuusisto, C. Langenberg, C. Langford, K. Leander, M.-L. Lokki, A. Lundmark, M. I. McCarthy, C. Meisinger, O. Melander, E. Mihailov, S. Maouche, A. D. Morris, M. Müller-Nurasyid, MuTHER Consortium, K. Nikus, J. F. Peden, N. W. Rayner, A. Rasheed, S. Rosinger, D. Rubin, M. P. Rumpf, A. Schäfer, M. Sivananthan, C. Song, A. F. R. Stewart, S.-T. Tan, G. Thorgeirsson, C. E. van der Schoot, P. J. Wagner, Wellcome Trust Case Control Consortium, G. A. Wells, P. S. Wild, T.-P. Yang, P. Amouyel, D. Arveiler, H. Basart, M. Boehnke, E. Boerwinkle, P. Brambilla, F. Cambien, A. L. Cupples, U. de Faire, A. Dehghan, P. Diemert, S. E. Epstein, A. Evans, M. M. Ferrario, J. Ferrières, D. Gauguier, A. S. Go, A. H. Goodall, V. Gudnason, S. L. Hazen, H. Holm, C. Iribarren, Y. Jang, M. Kähönen, F. Kee, H.-S. Kim, N. Klopp, W. Koenig, W. Kratzer, K. Kuulasmaa, M. Laakso, R. Laaksonen, J.-Y. Lee, L. Lind, W. H. Ouwehand, S. Parish, J. E. Park, N. L. Pedersen, A. Peters, T. Quertermous, D. J. Rader, V. Salomaa, E. Schadt, S. H. Shah, J. Sinisalo, K. Stark, K. Stefansson, D.-A. Trégouët, J. Virtamo, L. Wallentin, N. Wareham, M. E. Zimmermann, M. S. Nieminen, C. Hengstenberg, M. S. Sandhu, T. Pastinen, A.-C. Syvänen, G. K. Hovingh, G. Dedoussis, P. W. Franks, T. Lehtimäki, A. Metspalu, P. A. Zalloua,

A. Siegbahn, S. Schreiber, S. Ripatti, S. S. Blankenberg, M. Perola, R. Clarke, B. O. Boehm, C. O'Donnell, M. P. Reilly, W. März, R. Collins, S. Kathiresan, A. Hamsten, J. S. Kooner, U. Thorsteinsdottir, J. Danesh, C. N. A. Palmer, R. Roberts, H. Watkins, H. Schunkert, and N. J. Samani. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics*, 45(1):25–33, Jan. 2013.

K. Geiler-Samerotte, C. Bauer, S. Li, N. Ziv, D. Gresham, and M. Siegal. The details in the distributions: why and how to study phenotypic variability. *Current Opinion in Biotechnology*, 24(4):752–759, Aug. 2013.

R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308:385–388, 2005.

W. Lee and Y. Lee. Modifications of REML algorithm for HGLMs. *Statistics and Computing*, 2012.

Y. Lee and J. A. Nelder. Double hierarchical generalized linear models (with discussion). *Applied Statistics*, 55:139–185, 2006.

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

Y. Lee, J. A. Nelder, and Y. Pawitan. *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall/CRC, 2006.

D. Levy, G. B. Ehret, K. Rice, G. C. Verwoert, L. J. Launer, A. Dehghan, N. L. Glazer, A. C. Morrison, A. D. Johnson, T. Aspelund, Y. Aulchenko, T. Lumley, A. Köttgen, R. S. Vasan, F. Rivadeneira, G. Eiriksdottir, X. Guo, D. E. Arking, G. F. Mitchell, F. U. S. Mattace-Raso, A. V. Smith, K. Taylor, R. B. Scharpf, S.-J. Hwang, E. J. G. Sijbrands, J. Bis, T. B. Harris, S. K. Ganesh, C. J. O'Donnell, A. Hofman, J. I. Rotter, J. Coresh, E. J. Benjamin, A. G. Uitterlinden, G. Heiss, C. S. Fox, J. C. M. Witteman, E. Boerwinkle, T. J. Wang, V. Gudnason, M. G. Larson, A. Chakravarti, B. M. Psaty, and C. M. van Duijn. Genome-wide association study of blood pressure and hypertension. *Nature Genetics*, May 2009.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications, 2001.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American journal of human genetics*, 81(3):559–575, 2007.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

L. Rönnegård, R. Pong-Wong, and O. Carlborg. Defining the assumptions underlying modeling of epistatic QTL using variance component methods. *J. Hered.*, 99(4):421–425, July 2008.

X. Shen, L. Rönnegård, and O. Carlborg. Hierarchical likelihood opens a new way of estimating genetic values using genome-wide dense marker maps. *BMC proceedings*, 5 Suppl 3:S14, 2011.

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

X. Shen, M. Alam, F. Fikse, and L. Rönnegård. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, Apr. 2013.

X. Shen, Y. Li, L. Rönnegård, P. Udén, and O. Carlborg. Application of a genomic model for high-dimensional chemometric analysis. *Journal of Chemometrics*, 28:548–557, 2014.

T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, C. T. Johansen, S. W. Fouchier, A. Isaacs, G. M. Peloso, M. Barbalic, S. L. Ricketts, J. C. Bis, Y. S. Aulchenko, G. Thorleifsson, M. F. Feitosa, J. Chambers, M. Orho-Melander, O. Melander, T. Johnson, X. Li, X. Guo, M. Li, Y. S. Cho, M. J. Go, Y. J. Kim, J.-Y. Lee, T. Park, K. Kim, X. Sim, R. T.-H. Ong, D. C. Croteau-Chonka, L. A. Lange, J. D. Smith, K. Song, J. H. Zhao, X. Yuan,

J. Luan, C. Lamina, A. Ziegler, W. Zhang, R. Y. L. Zee, A. F. Wright, J. C. M. Witteman, J. F. Wilson, G. Willemsen, H.-E. Wichmann, J. B. Whitfield, D. M. Waterworth, N. J. Wareham, G. Waeber, P. Vollenweider, B. F. Voight, V. Vitart, A. G. Uitterlinden, M. Uda, J. Tuomilehto, J. R. Thompson, T. Tanaka, I. Surakka, H. M. Stringham, T. D. Spector, N. Soranzo, J. H. Smit, J. Sinisalo, K. Silander, E. J. G. Sijbrands, A. Scuteri, J. Scott, D. Schlessinger, S. Sanna, V. Salomaa, J. Saharinen, C. Sabatti, A. Ruokonen, I. Rudan, L. M. Rose, R. Roberts, M. Rieder, B. M. Psaty, P. P. Pramstaller, I. Pichler, M. Perola, B. W. J. H. Penninx, N. L. Pedersen, C. Pattaro, A. N. Parker, G. Paré, B. A. Oostra, C. J. O'Donnell, M. S. Nieminen, D. A. Nickerson, G. W. Montgomery, T. Meitinger, R. McPherson, M. I. McCarthy, W. McArdle, D. Masson, N. G. Martin, F. Marroni, M. Mangino, P. K. E. Magnusson, G. Lucas, R. Luben, R. J. F. Loos, M.-L. Lokki, G. Lettre, C. Langenberg, L. J. Launer, E. G. Lakatta, R. Laaksonen, K. O. Kyvik, F. Kronenberg, I. R. König, K.-T. Khaw, J. Kaprio, L. M. Kaplan, A. Johansson, M.-R. Jarvelin, A. C. J. W. Janssens, E. Ingelsson, W. Igl, G. K. Hovingh, J.-J. Hottenga, A. Hofman, A. A. Hicks, C. Hengstenberg, I. M. Heid, C. Hayward, A. S. Havulinna, N. D. Hastie, T. B. Harris, T. Haritunians, A. S. Hall, U. Gyllensten, C. Guiducci, L. C. Groop, E. Gonzalez, C. Gieger, N. B. Freimer, L. Ferrucci, J. Erdmann, P. Elliott, K. G. Ejebe, A. Döring, A. F. Dominiczak, S. Demissie, P. Deloukas, E. J. C. de Geus, U. de Faire, G. Crawford, F. S. Collins, Y.-d. I. Chen, M. J. Caulfield, H. Campbell, N. P. Burtt, L. L. Bonnycastle, D. I. Boomsma, S. M. Boekholdt, R. N. Bergman, I. Barroso, S. Bandinelli, C. M. Ballantyne, T. L. Assimes, T. Quertermous, D. Altshuler, M. Seielstad, T. Y. Wong, E.-S. Tai, A. B. Feranil, C. W. Kuzawa, L. S. Adair, H. A. Taylor, Jr, I. B. Borecki, S. B. Gabriel, J. G. Wilson, H. Holm, U. Thorsteinsdottir, V. Gudnason, R. M. Krauss, K. L. Mohlke, J. M. Ordovas, P. B. Munroe, J. S. Kooner, A. R. Tall, R. A. Hegele, J. J. P. Kastelein, E. E. Schadt, J. I. Rotter, E. Boerwinkle, D. P. Strachan, V. Mooser, K. Stefansson, M. P. Reilly, N. J. Samani, H. Schunkert, L. A. Cupples, M. S. Sandhu, P. M. Ridker, D. J. Rader, C. M. van Duijn, L. Peltonen, G. R. Abecasis, M. Boehnke, and S. Kathiresan. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466:707–713, Aug. 2010.

F. Tian, P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics*, 43(2):159–162, Feb. 2011.

C. Turnbull, S. Ahmed, J. Morrison, D. Pernet, A. Renwick, M. Maranian, S. Seal, M. Ghoussaini, S. Hines, C. S. Healey, D. Hughes, M. Warren-Perry, W. Tapper, D. Eccles, D. G. Evans, Breast Cancer Susceptibility Collaboration (UK), M. Hooning, M. Schutte, A. van den Ouweland, R. Houlston, G. Ross, C. Langford, P. D. P. Pharoah, M. R. Stratton, A. M. Dunning, N. Rahman, and D. F. Easton. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics*, 42(6):504–507, June 2010.

W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38(8):879–887, Aug. 2006.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: A Tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics*, 88(1):76–82, 2010.