Published in final edited form as:

Genet Epidemiol. 2014 September ; 38(0 1): S57–S62. doi:10.1002/gepi.21826.

Genetic Prediction in the Genetic Analysis Workshop 18 Sequencing Data

Andreas Ziegler^{1,2,*}, Nora Bohossian^{3,4}, Vincent P. Diego⁵, and Chen Yao⁶

¹ Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

² Center for Clinical Trials, University of Lübeck, Lübeck, Germany

³ INSERM, UMR1043-CPTP, CHU Purpan, Toulouse, France

⁴ University of Toulouse III – Paul Sabatier, Toulouse, France

⁵ Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas

⁶ Department of Dairy Science, University of Wisconsin, Madison, Wisconsin

Abstract

High-throughput sequencing data can be used to predict phenotypes from genotypes, and this corresponds to establishing a prognostic model. In extended pedigrees the relatedness of subjects provides additional information so that genetic values, fixed or random genetic components, and heritability can be estimated. At the Genetic Analysis Workshop 18 the working group on genetic prediction dealt with both establishing a prognostic model and, in one contribution, comparing standard logistic regression with robust logistic regression in a sample of unrelated affected or unaffected individuals. Results of both logistic regression approaches were similar. All other contributions to this group used extended family data, in general using the quantitative trait blood pressure. The individual contributions varied in several important aspects, such as the estimation of the kinship matrix and the estimation method. Contributors chose various approaches for model validation, including different versions of cross-validation or within-family validation. Within-family validation included model building in the upper generations and validation in later generations. The choice of the statistical model and the computational algorithm had substantial effects on computation time. If decorrelation approaches were applied, the computational burden was substantially reduced. Some software packages estimated negative eigenvalues, although eigenvalues of correlation matrices should be nonnegative. Most statistical models and software packages have been developed for experimental crosses and planned breeding programs. With their specialized pedigree structures, they are not sufficiently flexible to accommodate the variability of human pedigrees in general, and improved implementations are required.

*Correspondence to: Univ.-Prof. Dr. Andreas Ziegler, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, Haus 24, 23538 Lübeck, Germany. Phone: +49 451 500 2780, Fax: +49 451 500 2999, ziegler@imbs.uni-luebeck.de.

Keywords

Decorrelation; genetic prediction; linear mixed model; next-generation sequencing; personalized medicine; prediction; robust regression

Introduction

After more than a decade of frustration and almost stagnation with the use of nonreplicable linkage studies [Risch and Merikangas, 1996], great hope for rapid progress has been raised with the publication of the landmark genome-wide association study by the Wellcome Trust Case Control Consortium [2007]. Indeed, genome-wide association studies (GWAS) have led to a great boost in knowledge about genetic factors that contribute to disease. However, the biology of most of the findings is not well understood, and it might take a long time to unravel the biological function of many of the GWAS hits. For example, the best replicated locus over all GWAS is probably the chromosome 9p21.3 region, a genetic desert, which has been associated with coronary artery disease in several GWAS [Helgadottir et al., 2007; McPherson et al., 2007; Samani et al., 2007]. However, it took four years to understand the function at this locus [Harismendy et al., 2011].

Even if the biology is not well understood for all loci contributing to the genetic disease, “the genetic information” might already tell “people about their risk for various diseases and which medications they should use or avoid” [Ioannidis, 2009: 139]. Although it is clear that the genetic variants that contribute to the pathogenesis of complex genetic diseases need to be distinguished from pharmacogenetic markers that guide treatment selection [Ziegler et al., 2012], terms are sometimes used interchangeably and might lead to confusion. For example, a single-nucleotide polymorphism (SNP) is said to be predictive (predictive SNP) if it forecasts the likely response to treatment. In contrast, predictive genetic tests concern healthy people and aim at determining individuals at risk for a disease, and the process is called genetic prediction. The term *genetic prediction* is used in the same way in animal breeding. If multiple markers are used for disease prediction, the statistical model is termed a predictive model. It is important to note that the term *genetic prediction* is also used when quantitative traits are considered, such as blood pressure or body weight. Here, the aim is to predict the value of the quantitative trait instead of a dichotomous disease phenotype.

All contributions to our Genetic Analysis Workshop 18 (GAW18) working group on genetic prediction dealt with genetic prediction in the sense just described. Interestingly, only a single research group investigated prediction models for unrelated subjects, and they considered the binary endpoint hypertension [Kesselmeier et al., 2014]. Their approach is discussed first. All other contributions to this working group considered the large pedigree data [Bohossian et al., 2014; Quillen et al., 2014; Yang et al., 2014; Yao et al., 2014], and most contributions used a quantitative trait only.

All investigators working with the family data used a specific linear mixed model (LMM) for analysis. In the following section we introduce the basic LMM and derive the variants of the LMM as used in the individual contributions. We discuss the estimation aims, several approaches to determine the correlation between family members, and some maximization

approaches. Compared with estimating equations for independent individuals, estimating equations for correlated subjects are more difficult to solve, need more computational time, and are less numerically stable. One approach to overcome these obstacles is to decorrelate observations, which means that phenotypes (sometimes also the genotypes) are transformed to make observations uncorrelated. If phenotypes are normally distributed, then family members are independent after decorrelation. Finally, we consider the properties of some of the estimation methods. By investigating the eigenvalues of some of the solutions, we see that not all variances need to be positive. Furthermore, some variances could not be reliably estimated and turned out to be negative. In this case, either parameter estimates need to be bounded to nonnegative values, leading to biased parameter estimates, or restricted estimation approaches are required. The computational effort of a restricted estimation approach is, however, tremendous, and we draw the conclusion that improved implementation of LMMs are required for application in human genetics.

However, before we start to describe the family studies, we consider the single contribution that dealt with unrelated individuals.

Influential Points and Logistic Regression Versus Robust Logistic Regression

Logistic regression is the standard statistical approach for estimating disease probability from independent subjects, although it may be sensitive to model misspecification and outliers. Specifically, few observations may have substantial influence on the parameter estimates. In their contribution, using Cook's distance [Hosmer et al., 1991], a standard logistic regression, Kesselmeier et al. [2014] first showed that several observations had to be termed influential—the investigators termed them outliers—and these observations substantially affected the parameter estimates.

Cantoni and Ronchetti [2001] proposed a new class of robust regression models using M-estimators as an alternative to standard generalized linear models. They showed that their robust logistic regression is robust in the sense that a small amount of contamination at a point x has bounded influence on the asymptotic type I error level and the power of the test. However, Cantoni and Ronchetti did not derive any robustness results in terms of changes in parameter estimates. This might be an explanation for the fact that Kesselmeier et al. [2014] observed similar results for the logistic regression and the robust logistic regression.

Methods

Basic Linear Mixed Model

All remaining contributions to our working group investigated extended pedigrees using the LMM. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})'$ denote the vector of dependent variables of family i ($i = 1, \dots, n$ having T_i family members). Let $\mathbf{X}_i = (x'_{i1}, \dots, x'_{iT_i})'$ be the $T_i \times p$ matrix of covariates, and family member $t = 1, \dots, T_i$. The basic LMM is given by

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{G}_i \boldsymbol{\alpha} + \mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad (1)$$

where $\mathbf{X}_i\boldsymbol{\beta}$ denotes the fixed component of covariates, such as environmental variables, where $\boldsymbol{\beta}$ is the p -dimensional parameter vector of interest; $\mathbf{G}_i\boldsymbol{\alpha}$ is a fixed genetic marker component, where $\boldsymbol{\alpha}$ is the q -dimensional parameter vector of interest and $\mathbf{G}_i$ is of dimension $T_i \times q$; $\mathbf{u}_i \sim N(\mathbf{0}, \sigma_u^2 \mathbf{K}_i)$ is the polygenic component; $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$ is the error term; $\mathbf{K}_i$ reflects the correlation within pedigree i on the polygenic component, and it is twice the expected kinship matrix (i.e., $\mathbf{K}_i = 2\Phi_i$) ($\mathbf{K}_i$ has been termed the genetic relationship matrix or genetic relationship kernel (GRK) by Blangero et al. [2013]); σ_u^2 is the variance of the polygenic component, which is assumed to be homoskedastic; and σ_ε^2 is the variance of the random error, which is assumed to be independent and homoskedastic. With these assumptions $\mathbf{y}_i$ is normally distributed with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}_i = \sigma_u^2 \mathbf{K}_i + \sigma_\varepsilon^2 \mathbf{I}$.

A few remarks on this LMM are required. First, the independence of the errors within a family states that $\mathbf{X}_i\boldsymbol{\beta}$ models all environmental factors and that $\mathbf{G}_i\boldsymbol{\alpha} + \mathbf{u}_i$ captures all genetic information. Second, the homoskedasticity of the polygenic component means that the variance of the polygenes is constant over generations and even within nuclear families of the same large family. Third, the polygenic component captures the additive genetic variance only; dominance and epistasis variance caused by polygenic components are neglected in this model. Fourth, the fixed genetic component $\mathbf{G}_i\boldsymbol{\alpha}$ captures all effects from genetic markers using standard coding. For example, for a SNP j with alleles A and a and additive coding, g_{ijt} equals 0, 1, or 2 if subject t of family i has 0, 1, or 2 a alleles. Other codings can be used, such as recessive, dominant, or heterozygous advantage [Ziegler and König, 2010], and coding schemes for rare variants [Dering et al., 2011] may also be used. Fifth, the model is identifiable only if polygenic effects $\mathbf{u}_i$ and the error term $\boldsymbol{\varepsilon}_i$ are uncorrelated. The final aspect is that the LMM of Eq. (1) does not include a gene by environment interaction component, although this can be easily modeled through $\mathbf{X}_i$ and $\mathbf{G}_i$.

Different modifications of the basic LMM of Eq. (1) have been used in our working group. For example, Bohossian et al. [2014] used a random genetic marker $\mathbf{G}_i\boldsymbol{\alpha}_i$ with $\boldsymbol{\alpha}_i \sim N(\mathbf{0}, \sigma_\alpha^2 \mathbf{I})$. Yao et al. [2014] not only examined cross-section data but also included the longitudinal aspect of the data. To this end, they considered the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{T}_i\mathbf{c}_i + \boldsymbol{\varepsilon}_i, \quad (2)$$

where $\mathbf{c}_i \sim N(\mathbf{0}, \sigma_c^2 \mathbf{I})$ is the random effect for repeated measurements. In Eq. (2) the polygenic component $\mathbf{u}_i$ is combined with a design matrix, sometimes termed the incidence matrix, $\mathbf{Z}_i$, to capture structure from the repeated nature of the measurements.

Finally, the simplified model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \boldsymbol{\varepsilon}_i \quad (3)$$

without a fixed genetic effect has also been used in our working group.

Likelihood, Decorrelation, and Choice of the Genetic Relationship Kernel

Estimation of the model from Eq. (1) is burdensome because it invariably requires repeated inversion of a potentially large covariance matrix, and it is reasonable to make use of the eigenvalue decomposition (EVD) of the covariance matrix $\Sigma_i = \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^{-1}$ into a diagonal matrix $\mathbf{\Lambda}_i = \text{diag}(\lambda_{ii})$ of eigenvalues and an orthogonal matrix $\mathbf{U}_i$ of eigenvectors [Gentle, 2007: 120]. The EVD is an orthogonal transformation of the data vector with the aim to linearly transform the vector of nonindependent observations $\mathbf{y}_i$ to a vector of transformed but independent observations $\mathbf{\tilde{y}}_i$. The EVD simplifies the likelihood because transformed observations are independent, and it thus allows simplified calculations. For example, Yang et al. [2014] rewrote the log-likelihood of Eq. (1) as

$$L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_u^2, \sigma_\varepsilon^2) = -\frac{1}{2} \left[n \ln(2\pi\sigma_u^2) + \ln \det \left(\mathbf{\Lambda}_i + \frac{\sigma_\varepsilon^2}{\sigma_u^2} \mathbf{I} \right) + \frac{1}{\sigma_u^2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{G}_i \boldsymbol{\alpha}_i)' \mathbf{U}_i' \left(\mathbf{\Lambda}_i + \frac{\sigma_\varepsilon^2}{\sigma_u^2} \mathbf{I} \right)^{-1} \mathbf{U}_i (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{G}_i \boldsymbol{\alpha}_i) \right], \quad (4)$$

to which they added some penalty so that the penalized log-likelihood was given by

$$PL_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_u^2, \sigma_\varepsilon^2, \kappa) = L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_u^2, \sigma_\varepsilon^2) - \kappa p(\boldsymbol{\alpha}) \quad (5)$$

for some penalty function $p(\boldsymbol{\alpha})$ and the Lagrange multiplier κ .

Decorrelation, also termed whitening, leads to substantial simplifications in the estimating equations. The estimation process is simpler, faster, and numerically more stable. It is, however, questionable how the GRK $\mathbf{K}_i$ should be determined. In our working group three different approaches were used (Table 1). All research groups used the kinship coefficient without genetic marker information, which is based on the reported pedigree relationships. Improvements can be obtained by using estimated relationships instead of reported ones. To this end, Diego et al. [unpublished data] used pathway-specific genetic-marker-based identity-by-descent estimates, and both Quillen et al. [2014] and Yao et al. [2014] used genome-wide marker-based estimates. Interestingly, Yao and colleagues based their GRK on the observed noncentered moment matrix (which corresponds to identity-by-state values), not on estimated identity-by-descent values. Other methods are available as well (see, e.g., VanRaden [2008]).

Bohossian et al. [2014], following the work of Aulchenko et al. [2007], used a different decorrelation approach. In their approach, termed GRAMMAR, the linear model given by Eq. (3) is fitted in the first step, and then the estimated residual $\hat{\varepsilon}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} - \mathbf{G}_i \hat{\boldsymbol{\alpha}}_i$ is used in the second step for genetic marker analyses.

Eigenvalues and Heritability

Another simplification can be obtained from Eq. (3) under the assumption that phenotypes are standardized, that is,

$$\text{Var}(y_{it}) = \sigma_y^2 = \sigma_u^2 + \sigma_\varepsilon^2 = 1 \quad (6)$$

[Blangero et al., 2013]. In this case the covariance matrix $\Sigma_i = \text{Var}(\mathbf{y}_i)$ is identical to $h^2 \mathbf{K}_i + (1 - h^2) \mathbf{I}$, where $h^2 = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$ denotes the narrow sense heritability of the polygenic component, and the log-likelihood function reduces to

$$L_i = -\frac{1}{2} \sum_{t=1}^{T_i} \left[1 + h^2 (\lambda_{it} - 1) \right] - \frac{1}{2} \sum_{t=1}^{T_i} \frac{\tilde{y}_{it}^2}{1 + h^2 (\lambda_{it} - 1)}, \quad (7)$$

where $\tilde{\mathbf{y}}_i = \mathbf{U}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})$ is the vector of transformed residuals having components $\tilde{y}_{it}$. By noting that $\mathbf{U}_i \mathbf{I} \mathbf{U}_i' = \mathbf{I}$, we see that the λ_{it} are obtained by applying the EVD to the GRK $\mathbf{K}_i = \tilde{\mathbf{U}}_i \tilde{\mathbf{\Lambda}}_i \tilde{\mathbf{U}}_i^{-1}$ additive genetic eigenvalues, and $\tilde{\mathbf{\Lambda}}_i = \text{diag}(\lambda_{it})$.

Blangero et al. [2013] and Diego et al. [unpublished data] have shown that for this model the expected likelihood ratio statistic (LRT), when the estimate $\hat{h}^2$ exceeds 0, is given by

$$E(LRT) \Big|_{\hat{h}^2 > 0} = - \sum_{t=1}^{T_i} \ln \left[1 + h^2 (\lambda_{it} - 1) \right]. \quad (8)$$

This means that the expected value of the test statistic depends on the eigenvalues λ_{it} of the GRK, that is, the kinship matrix. Eigenvalues of selected relative pairs are displayed in Table 2; for other relative pairs and expected LRT values for varying heritabilities, see Blangero et al. [2013]. It is important to note that the eigenvalues for a monozygotic twin pair are either 0 or 2. Consequently, when the heritability is exactly equal to 1, the expected LRT (and consequently the likelihood function) becomes degenerate. However, this can be dealt with by bounding the heritability to slightly less than 1. The value of the expected LRT is such that it can be used to compute power in a likelihood analysis [Brown et al., 1999]. Diego et al. [unpublished data] exploited this utility of the expected LRT to show that different GRK algorithms can be compared in terms of their respective power functions and asymptotic relative efficiencies [Blangero et al., 2013].

Estimation and Negative Eigenvalues

The fixed effects can be estimated by using a best linear unbiased estimation (BLUE), whereas random effects can be estimated by using a best linear unbiased prediction (BLUP). BLUE and BLUP can be computed simultaneously by solving the so-called mixed-model equations, given, for example, by Searle et al. [1992], and several alternative approaches are available [VanRaden, 2008]. Other estimation approaches are also available, such as the maximum-likelihood (ML), restricted ML (REML), penalized ML, and gradient methods, and they are implemented in various software packages used by group members (see Table 1).

The estimator $\hat{\mathbf{u}}_i$ from Eq. (1) is called the estimated genetic value (EGV) in human genetics, and it is the equivalent to the concept of the estimated breeding value (EBV) in animal and

plant genetics. If a model with a fixed incidence matrix $\mathbf{Z}_i$ is used, estimates of $\hat{\mathbf{u}}_i$ can be obtained in this case [VanRaden, 2008].

The GRK reflects a correlation matrix, and its eigenvalues should be nonnegative. In fact, one eigenvalue of a monozygotic twin pair is expected to be 0. However, negative eigenvalues can be observed when the ML is used for estimation, as implemented in such software packages as PLINK, GCTA, and KINGr, and one solution is to bound these eigenvalues to the admissible parameter space [Quillen et al., 2014]. If some eigenvalues are estimated to be substantially smaller than 0, then others will be estimated to be substantially larger than 1 (for balancing), and this results in extreme heritability estimates (Eq. (8)) [Blangero et al., 2013]. Investigators should therefore carefully choose a proper algorithm for estimating the GRK, and some animal and plant breeders have clearly recommended the use of the REML instead of ML; see, for example, Piepho et al. [2008].

Furthermore, the assumption of a homogeneous variance σ_u^2 for the genetic effect is not justifiable because $\text{Var}(\hat{\mathbf{u}}_i)$ is a function of kin, and the more relatives in the pedigree, the lower the estimated standard error of $\hat{\mathbf{u}}_i$.

Model Validation

The members of our working group used various approaches for model validation (see Table 1). Kesselmeier et al. [2014], who dealt with the case-control data in unrelated individuals, used a leave-one-out cross-validation (CV) approach. Yang et al. [2014] ran a tenfold CV 20 times. Bohossian et al. [2014] used the first replicate for training and the other 199 replicates for testing. This approach was criticized in the group discussions because sequences, and thus genotypes, were fixed for all replicates and not generated anew for each replicate. An interesting validation approach was taken by Yao et al. [2014]. In one validation approach, they used the first three generations of each family for model development and the fourth and fifth generations from the same families for model validation. However, this approach can suffer if there is a high rate of missing data in the founder generations, and there could be secular trends between generations. Finally, for late-onset diseases testing the model in the younger generation may lack power. In another validation approach, Yao and colleagues used a fivefold CV. However, CV samples were not generated randomly; instead, families were ordered by family size, and every five families were randomly assigned to five different folds from the CV.

Discussion

The discussions of the working group on genetic prediction were intense and different from the previous experience of group members. Specifically, the LMM used by many contributors to this group were technically challenging, and most of the discussion time was dedicated to details of the models and technical details, such as implementation. The results from the actual analysis of the GAW18 data were secondary. At previous workshops most of the discussion was usually focused on comparing results, not just statistical methodology. One reason was the late data delivery combined with the high computational burden of some of the models used.

For example, when the decorrelation approach was used, the analysis of 60,000 SNPs from a single chromosome using REML with iterations took about half an hour on a desktop PC with a single core. In contrast, the BayesC π and GBLUP (genomic BLUP), required 2 days for 12,000 Markov chain Monte Carlo iterations on a cluster when all 960,000 SNPs from all available chromosomes were analyzed. Specifically, the BayesC π model treats the SNP effects and their variance as random variables. Furthermore, the effects of only a fraction π of the SNPs are assumed to follow the same probability distribution.

This difference in computational speed illustrates that decorrelation approaches greatly reduce the computational burden. It is, however, still unclear how to best choose the GRK because the genome-wide estimation of identity-by-descent values is challenging. However, it would be even better if local identity-by-descent information (i.e., from smaller chromosomal regions) could be used for estimation [McVean, 2009].

An important aspect of why most of the discussions focused on the models is the overall negative findings in the group. For example, Bohossian et al. [2014] concluded that their approach had poor ability to identify associated loci in the provided pedigree data. Similar conclusions were drawn by Yang et al. [2014] and Yao et al. [2014]. This consistent finding was reported by one of the reviewers during the review of the individual contributions.

Nevertheless, Quillen et al. [2014] found that the adjustment for the effects of environmental factors (i.e., the extra variability caused by environmental variables) shows promise for increasing the power of associations. This finding seems to be somewhat contradictory to the results from genome-wide association studies, where it has been found that the power of identifying an association may be substantially reduced if adjustments are made for covariates [Pirinen et al., 2012]. However, as discussed in detail by Mefford and Witte [2012], the power of the analysis depends on the ascertainment model, the disease prevalence, and effect sizes.

An important aspect in addition to computational burden is the validity of results. It has been shown analytically that some eigenvalues are 0 if monozygotic twin pairs are included in the model. As a result, the normal distribution is degenerate even if the reported pedigree structure is used for determining the GRK. However, when the GRK was estimated, negative eigenvalues were obtained by some software packages using ML estimation. Most of these programs, which allow for other estimation approaches, such as REML, have been developed for experimental crosses and planned breeding programs. With their specialized pedigree structures, they are not sufficiently flexible to accommodate the variability of human pedigrees in general. Improved implementation of the LMM for human pedigrees is urgently required.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We have learned much from all participants in the group discussions. The colleagues participating in the discussions were John Blangero, Nora Bohossian, Vincent Diego, Wei Guo, Guan-Hua Huang, Miriam Kesselmeier, Kristine Lee, Maria Martinez, Ellen Quillen, Saroja Voruganti, Can Yang, and Yin Yao Shugart. The Genetic Analysis Workshops are funded by the National Institutes of Health through grant R01 GM031575. The work of AZ was funded by the German Ministry of Education and Research (through grant 0315536F).

References

- Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*. 2007; 177(1):577–585. [PubMed: 17660554]
- Blangero J, Diego VP, Dyer TD, Almeida M, Peralta J, Kent JW Jr, Williams JT, Almasy L, Goring HH. A kernel of truth: Statistical advances in polygenic variance component models for complex human pedigrees. *Adv Genet*. 2013; 81:1–31. [PubMed: 23419715]
- Bohossian N, Saad M, Legarra A, Martinez M. Single-marker and multi-marker mixed models for polygenic score analysis in family-based data. *BMC Proc*. 2014; 8(suppl 2):S63. [PubMed: 25519337]
- Brown BW, Lovato J, Russell K. Asymptotic power calculations: Description, examples, computer code. *Stat Med*. 1999; 18(22):3137–3151. [PubMed: 10544312]
- Cantoni E, Ronchetti E. Robust inference for generalized linear models. *J Am Stat Assoc*. 2001; 96(455):1022–1030.
- Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: An overview of collapsing methods. *Genet Epidemiol*. 2011; 35(suppl 1):S12–S17. [PubMed: 22128052]
- Gentle, JE. *Matrix Algebra*. Springer; New York: 2007.
- Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature*. 2011; 470(7333):264–268. [PubMed: 21307941]
- Helgadóttir A, Thorleifsson G, Manolescu A, Gretarsdóttir S, Blondal T, Jonasdóttir A, Jonasdóttir A, Sigurdsson A, Baker A, Palsson A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*. 2007; 316:1491–1493. [PubMed: 17478679]
- Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: A case study. *Am J Public Health*. 1991; 81(12):1630–1635. [PubMed: 1746660]
- Ioannidis JP. Personalized genetic prediction: Too limited, too expensive, or too soon? *Ann Intern Med*. 2009; 150(2):139–141. [PubMed: 19153414]
- Kesselmeier M, Legrand C, Peil B, Kabisch M, Fischer C, Hamann U, Lorenzo Bermejo J. Practical investigation of the performance of robust logistic regression to predict the genetic risk of hypertension. *BMC Proc*. 2014; 8(suppl 2):S65. [PubMed: 25519338]
- McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, et al. A common allele on chromosome 9 associated with coronary heart disease. *Science*. 2007; 316:1488–1491. [PubMed: 17478681]
- McVean G. A genealogical interpretation of principal components analysis. *PLoS Genet*. 2009; 5(10):e1000686. [PubMed: 19834557]
- Mefford J, Witte JS. The covariate's dilemma. *PLoS Genet*. 2012; 8(11):e1003096. [PubMed: 23162385]
- Piepho HP, Möhring J, Melchinger AE, Büchse A. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*. 2008; 161(1–2):209–228.
- Pirinen M, Donnelly P, Spencer CC. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet*. 2012; 44(8):848–851. [PubMed: 22820511]
- Quillen EE, Voruganti VS, Chittoor G, Rubicz R, Peralta JM, Almeida MAA, Kent JW Jr, Dyer TD, Comuzzie AG, Goring HHH, et al. Evaluation of estimated genetic values and their application to

- genome-wide investigation of systolic blood pressure. *BMC Proc.* 2014; 8(suppl 2):S66. [PubMed: 25519398]
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science.* 1996; 273(5281):1516–1517. [PubMed: 8801636]
- Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann H-E, et al. Genomewide association analysis of coronary artery disease. *New Engl J Med.* 2007; 357(5):443–453. [PubMed: 17634449]
- Searle, SR.; Casella, G.; McCulloch, CE. *Variance Components.* Wiley; New York: 1992.
- VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci.* 2008; 91(11):4414–4423. [PubMed: 18946147]
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447(7145):661–678. [PubMed: 17554300]
- Yang C, Li C, Chen M, Chen X, Hou L, Zhao H. A penalized linear mixed model for genomic prediction using pedigree structures. *BMC Proc.* 2014; 8(suppl 2):S67. [PubMed: 25519399]
- Yao C, Leng N, Weigel KA, Lee KE, Engelman CD, Meyers KJ. Prediction of genetic contributions to complex traits using whole-genome sequencing data. *BMC Proc.* 2014; 8(suppl 2):S68. [PubMed: 25519339]
- Ziegler A, Koch A, Krockenberger K, Großhennig A. Personalized medicine using DNA biomarkers: A review. *Hum Genet.* 2012; 131(10):1627–1638. [PubMed: 22752797]
- Ziegler, A.; König, IR. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications.* 2nd edition. Wiley-VCH; Weinheim: 2010.

Table 1
Characteristics of the contributions using extended pedigrees to the working group on genetic prediction

Contribution	Linear mixed model	Focus of estimation	Decorrelation	GRK	Estimation method	Implementations used	Model validation
Bohossian et al. [2014]	Eq. (1) with random genetic marker effect	Random genetic component α	GRAMMAR	Kinship	BLUP	GS3	Replicate-based (training replicate 1; test replicates 2-200)
Diego et al. * (unpublished)	Eq. (1)	Heritability based on σ_u^2	Eigenvalue decomposition	Kinship; pathway-specific marker-based	ML	GCTA, KING (KINGh, KINGr), PLINK	None
Quillen et al. [2014]	Eq. (1)	EGV	Eigenvalue decomposition	Kinship genome-wide marker-based	ML	SOLAR GAUSS scripts	None
Yang et al. [2014]	Eq. (1)	Fixed genetic component α	Eigenvalue decomposition	Kinship; genome-wide marker-based	Penalized ML	Customized implementation	20 times tenfold CV
Yao et al. [2014]	Eq. (2)	EGV	None	Kinship	Stage I: REML; stage II: GBLUP, BayesCn	Stage I: Pedigreemm (R package); stage II: MCMC	Fivefold CV with ordering by family size; within-family validation (training generations 1-3; test generations 4-5)

* Findings not submitted for GAW18 proceedings.

Table 2

Eigenvalues for selected relative pairs

Relationship	Eigenvalue 1	Eigenvalue 2
Monozygotic twins	2	0
Sib pair	3/2	1/2
Sibship	$(n_s + 1)/2$	$(n_s - 1)/2$

n_s is the sibship size.