RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Fast Genome-Wide QTL Association Mapping on Pedigree and Population Data

Hua Zhou,[1] John Blangero,[2] Thomas D. Dyer,[2] Kei-hang K. Chan,[3,4] Kenneth Lange,[3,5,6] and Eric M. Sobel[3]*

[1] *Department of Biostatistics, University of California, Los Angeles, California, United States of America;* [2] *South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Texas, United States of America;* [3] *Department of Human Genetics, University of California, Los Angeles, California, United States of America;* [4] *Department of Epidemiology, University of California, Los Angeles, California, United States of America;* [5] *Department of Biomathematics, University of California, Los Angeles, California, United States of America;* [6] *Department of Statistics, University of California, Los Angeles, California, United States of America*

Received 12 August 2015; Revised 2 May 2016; accepted revised manuscript 8 May 2016.
Published online 12 December 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21988

ABSTRACT: Since most analysis software for genome-wide association studies (GWAS) currently exploit only unrelated individuals, there is a need for efficient applications that can handle general pedigree data or mixtures of both population and pedigree data. Even datasets thought to consist of only unrelated individuals may include cryptic relationships that can lead to false positives if not discovered and controlled for. In addition, family designs possess compelling advantages. They are better equipped to detect rare variants, control for population stratification, and facilitate the study of parent-of-origin effects. Pedigrees selected for extreme trait values often segregate a single gene with strong effect. Finally, many pedigrees are available as an important legacy from the era of linkage analysis. Unfortunately, pedigree likelihoods are notoriously hard to compute. In this paper, we reexamine the computational bottlenecks and implement ultra-fast pedigree-based GWAS analysis. Kinship coefficients can either be based on explicitly provided pedigrees or automatically estimated from dense markers. Our strategy (a) works for random sample data, pedigree data, or a mix of both; (b) entails no loss of power; (c) allows for any number of covariate adjustments, including correction for population stratification; (d) allows for testing SNPs under additive, dominant, and recessive models; and (e) accommodates both univariate and multivariate quantitative traits. On a typical personal computer (six CPU cores at 2.67 GHz), analyzing a univariate HDL (high-density lipoprotein) trait from the San Antonio Family Heart Study (935,392 SNPs on 1,388 individuals in 124 pedigrees) takes less than 2 min and 1.5 GB of memory. Complete multivariate QTL analysis of the three time-points of the longitudinal HDL multivariate trait takes less than 5 min and 1.5 GB of memory. The algorithm is implemented as the Ped-GWAS Analysis (Option 29) in the Mendel statistical genetics package, which is freely available for Macintosh, Linux, and Windows platforms from http://genetics.ucla.edu/software/mendel.**

Genet Epidemiol 41:174–186, 2017.

KEY WORDS: genome-wide association study; pedigree; kinship; score test; fixed-effects models; multivariate traits

Introduction

Genome-wide association studies (GWAS) are now at a crossroads. After the discovery of thousands of genes influencing hundreds of common traits [Hindorff et al., 2009], much of the low-hanging fruit has been plucked [Ku et al., 2010, Visscher et al., 2012]. Because of the enormous sample sizes of current studies, new trait genes are still being uncovered. Unfortunately, most entail small effects. Is it possible that inheritance is predominantly polygenic, and a law of diminishing returns has set in? The push to exploit rare variants is one response to this dilemma. The previous generation of geneticists relied on linkage to map rare variants. Linkage mapping fell from grace because of its poor resolution. Reducing a genome search to a one or two megabase region leaves too large an expanse of DNA to sift. The real gold of linkage mapping may well be its legacy pedigrees [Ott et al., 2011]. Pedigree data is particularly attractive in association studies because it permits control of population substructure and study of parent-of-origin effects. Related affecteds are also more likely to share the same disease predisposing gene than unrelated affecteds. Even in population-based association studies, taking into account estimated identity-by-descent (IBD) information is apt to reduce false positives and increases power. The recent availability of dense marker data from genotyping chips enables quick and accurate estimation of global and even local IBD [Day-Williams et al., 2011].

Geneticists turned to random sample and case-control data because of the relative ease of collecting population data and the computational challenges posed by pedigrees. The tide of computational complexity is now beginning to turn. To handle pedigree data in association testing, statistical geneticists have proposed semiparametric methods such as the generalized linear-mixed model (GLMM) [Amin et al., 2007, Aulchenko et al., 2007] and generalized estimating equations (GEE) [Chen and Yang, 2010, Chen et al., 2011]. Although

*Correspondence to: Eric Sobel, Department of Human Genetics, University of California, Los Angeles, California 90095, United States of America. E-mail: esobel@ucla.edu

such methods work for both quantitative and binary traits, they are compromised by current restrictions that reduce power. The GEE approach requires input of a working correlation structure for each pedigree. The kinship coefficient matrix is a natural candidate. However, current implementations require the same working correlation matrix across all clusters, which implicitly requires all pedigrees to have the same structure [Chen et al., 2011]. This is a dubious and restrictive assumption. In the limited context of case-control studies, recent methods such as M_{QLS} [Thornton and McPeek, 2007], ROADTRIPS [Thornton and McPeek, 2010], and FPCA [Zhu and Xiong, 2012] correct for pedigree and ethnically induced correlations by exploiting dense marker data. Other authors attack the same issues more broadly from the GLMM perspective [Kang et al., 2010, Zhang et al., 2010, Lippert et al., 2011]. Korte et al. [2012] generalizes GLMM to multivariate traits. Models based on the transmission-disequilibrium test (TDT) [Spielman and Ewens, 1998] and its generalization, the family-based association test (FBAT) [Laird et al., 2000, Lange and Laird, 2002, Van Steen and Lange, 2005, Won et al., 2009a, 2009b], are promising but ignore covariates and polygenic background. See Van Steen [2011] for a recent overview of FBAT methods for GWAS. We treat all of these extensions in a unified framework consistent with exceptionally fast computing.

The present paper reexamines the computational bottlenecks encountered in association mapping with pedigree data. It turns out that the previous objections to pedigree GWAS can be overcome. Kinship coefficients can be based on explicitly provided pedigree structure or estimated from dense markers when genealogies are missing or dubious. Frequentist hypothesis testing usually operates by comparing maximum likelihoods under the null and alternative hypotheses. Maximization of the alternative likelihood must be conducted for each and every marker. Score tests constitute a more efficient strategy than likelihood ratio tests. This is the point of departure taken by Chen and Abecasis [2007], but they use approximations that we avoid. The GLOGS program [Stanhope and Abney, 2012] makes similar approximations in the case-control setting. Here, we consider arbitrary pedigrees and multivariate quantitative traits. Score tests require no additional iteration under the alternative model. All that is needed is evaluation of a quadratic form combining the score vector and the expected information matrix at the maximum likelihood estimates under the null model. Although it takes work to assemble these quantities, a careful analysis of the algorithm shows that fast testing is perfectly feasible.

In our implementation of score testing, the few SNPs with the most significant score-test *P*-values are automatically reanalyzed by the slightly more powerful, but much slower, likelihood ratio test (LRT). Our fixed effects (mean component) model assumes Gaussian variation of the trait; the two alleles of a SNP shift trait means. There is no confounding of association and linkage. This framework carries with it several advantages. First, it applies to random sample data, pedigree data, or a mix of both. Second, it enables covariate adjustment, including correction for population stratification. Third, it accommodates additive, dominant, and recessive SNP models. Fourth, it also accommodates both univariate and multivariate traits. And fifth, as just mentioned, it fosters both likelihood ratio tests and score tests. The mean component model is now implemented in our software package MENDEL for easy use by the genetics community. In addition, MENDEL provides a complete suite of tools for pedigree analysis, including GWAS data preparation and manipulation, pedigree genotype simulation (gene dropping), trait simulation, genotype imputation, local and global kinship coefficient estimation, and pedigree-based GWAS (ped-GWAS) [Lange et al., 2005, 2013].

Table 1. Comparison of features in MENDEL, FAST-LMM, and GEMMA for GWAS of QTLs

	MENDEL	FAST-LMM	GEMMA
Multithreaded operation	Yes	Yes	No
Can estimate kinships via SNPs	Yes	Yes	Yes
Imports and exports kinship estimates	Yes	Yes	Yes
Allows retained covariates	Yes	Yes	Yes
Allows linear constraints on covariates	Yes	No	No
Can use either LRT or score test	Yes	No	Yes*
Allows multivariate analysis	Yes	No	Yes
Can perform multiple univariate analyses	Yes	No	No
Allows > 2 variance components	Yes	No	No
Analyzes X-linked loci	Yes	No	No
Automatic SNP filtering on MAF	Yes	No	Yes
Allows nonadditive SNP models	Yes	No	No
Detects outlier pedigrees	Yes	No	No
Detects outlier individuals	Yes	No	No
Can simulate genotype/phenotype data	Yes	No	No
Reads in fractional genotype values	No	Yes	Yes

*GEMMA can use the likelihood ratio, score, or Wald test

The competing software packages EMMAX [Kang et al., 2008], MMM [Pirinen et al., 2013], FAST-LMM [Lippert et al., 2011, Listgarten et al., 2012], GEMMA [Zhou and Stephens, 2012, 2014], and GWAF [Chen and Yang, 2010] already implement variance component models for quantitative trait locus (QTL) analysis. Exhaustive comparison of MENDEL to each of these programs is beyond the scope of the current paper. We limit our comparisons of MENDEL to the state-of-art packages FAST-LMM and GEMMA, arguably the fastest and most sophisticated of the competition. Table 1 summarizes some of the qualitative features of these packages. Our numerical examples also demonstrate an order of magnitude advantage in speed of MENDEL over FAST-LMM, GEMMA, and GWAF. This advantage stems from our careful formulation of the score test and our exploitation of the multicore processors resident in almost all personal computers and computational clusters.

Methods

QTL Association Mapping With Pedigrees

QTL association mapping typically invokes the multivariate Gaussian distribution to model the trait values $\boldsymbol{y} = (y_i)$ over a pedigree. The observed trait value y_i of person i can be either univariate or multivariate. For simplicity we first assume y_i is univariate and later indicate the necessary changes for multivariate y_i. The standard model [Lange, 2002]

Table 2. Genotype encodings for the major gene models

Genotype	Additive	Dominant	Recessive
1/1	−1	−1	−1
1/2	0	−1	+1
2/2	+1	+1	+1

The additive model is the default choice. In the genotype column, "1" and "2" represent the first and second alleles for each SNP. An effect size estimate reflects the change in trait values due to each positive unit change in the encodings. For example, the default additive model estimates the mean trait difference in moving from a 1/2 genotype to a 2/2 genotype.

collects the corresponding trait means into a vector $\boldsymbol{\nu}$ and the corresponding covariances into a matrix $\boldsymbol{\Omega}$ and represents the loglikelihood of a pedigree as

$$L = -\frac{1}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\nu})^t\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\nu}), \tag{1}$$

where det denotes the determinant function and the covariance matrix is typically parametrized as

$$\boldsymbol{\Omega} = 2\sigma_a^2\boldsymbol{\Phi} + \sigma_d^2\boldsymbol{\Delta}_7 + \sigma_h^2\boldsymbol{H} + \sigma_e^2\boldsymbol{I}. \tag{2}$$

Here the variance component $\boldsymbol{\Phi}$ is the global kinship coefficient matrix capturing additive polygenic effects, and $\boldsymbol{\Delta}_7$ is a condensed identity coefficient matrix capturing dominance genetic effects. When pedigree structure is explicitly given, these genetic identity coefficients are easily calculated [Lange, 2002]. With unknown or dubious genealogies, the global kinship coefficient can be accurately estimated from dense markers [Day-Williams et al., 2011]. The household effect matrix $\boldsymbol{H}$ has entries $h_{ij} = 1$ if individuals i and j belong to the same household and 0 otherwise. Individual environmental contributions and trait measurement errors are incorporated via the identity matrix $\boldsymbol{I}$. MENDEL's implementation of this model can include both the two standard variance classes, additive and environmental, as well as the two extra variances classes, dominance and household. Inclusion of additional variance classes has no significant effect on MENDEL's speed of computation.

In general, a mixed model for QTL association mapping captures polygenic and other random effects through $\boldsymbol{\Omega}$ and captures QTL fixed effects through $\boldsymbol{\nu}$. Let $\boldsymbol{\beta}$ denote the full vector of regression coefficients parameterizing $\boldsymbol{\nu}$. In a linear model one postulates that $\boldsymbol{\nu} = \boldsymbol{A}\boldsymbol{\beta}$ for some predictor matrix $\boldsymbol{A}$ incorporating relevant covariates such as age, gender, and diet. In testing association against a given SNP, $\boldsymbol{A}$ is augmented by an extra column whose entries encode genotypes according to one of the models (additive, dominant, and recessive) shown in Table 2. To accommodate imprecise imputation in an additive model, these encodings can be made fractional. The corresponding component of $\boldsymbol{\beta}$, β_{SNP}, is the SNP effect size. In likelihood ratio association testing one contrasts the null hypothesis $\beta_{\text{SNP}} = 0$ with the alternative hypothesis $\beta_{\text{SNP}} \neq 0$. In testing a univariate trait, the likelihood ratio statistic asymptotically follows a χ_1^2 distribution. In testing a multivariate trait with $T > 1$ components, each row of $\boldsymbol{A}$ must be replicated T times. The likelihood ratio statistic then asymptotically follows a χ_T^2 distribution. To implement likelihood ratio testing, iterative maximum likelihood estimation must be undertaken for each and every SNP under the alternative hypothesis. This unfortunate requirement is the major stumbling block retarding pedigree analysis.

Score tests serve as convenient substitutes for likelihood ratio tests. The current paper describes how to implement ultra-fast score tests for screening SNPs. Only SNPs with the most significant score-test P-values are further subjected to the more accurate likelihood ratio test. An advantage of the likelihood ratio method is that it estimates effect sizes. In contrast, the score test only requires parameter estimates under the null hypothesis and involves no iteration beyond fitting the null model. The score vector is the gradient $\nabla L(\boldsymbol{\theta})$ of the loglikelihood $L(\boldsymbol{\theta})$, where the full parameter vector $\boldsymbol{\theta}$ includes variance components such as the additive genetic variance in addition to the regression coefficient vector $\boldsymbol{\beta}$. The transpose $dL(\boldsymbol{\theta})$ of the score is a row vector called the first differential of $L(\boldsymbol{\theta})$. The expected information $J(\boldsymbol{\theta})$ is the covariance matrix of the score vector. It is well known that the expected value of the observed information matrix (negative second differential) $-d^2L(\boldsymbol{\theta})$ coincides with $J(\boldsymbol{\theta})$ [Rao, 2009]. The score statistic

$$S(\boldsymbol{\theta}) = dL(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}\nabla L(\boldsymbol{\theta}) \approx dL(\boldsymbol{\theta})[-d^2L(\boldsymbol{\theta})]^{-1}\nabla L(\boldsymbol{\theta})$$

is evaluated at the maximum likelihood estimates under the null hypothesis with the parameter β_{SNP} of the alternative hypothesis set to 0.

Fast Score Test for Individual SNPs

Under the multivariate model, the expected information matrix $J(\boldsymbol{\theta})$ for a single pedigree can be written in the block diagonal form

$$J(\boldsymbol{\theta}) = \begin{pmatrix} \mathrm{E}\left[-d_{\boldsymbol{\beta}}^2 L(\boldsymbol{\theta})\right] & 0 \\ 0 & \mathrm{E}[-d_{\boldsymbol{\sigma}}^2 L(\boldsymbol{\theta})] \end{pmatrix}, \tag{3}$$

where $\boldsymbol{\sigma}$ denotes the vector of variance parameters [Lange, 2002]. For independent pedigrees, the loglikelihoods (1) and corresponding score vectors and expected information matrices add. Hence, the block diagonal form of $J(\boldsymbol{\theta})$ is preserved. Because the inverse of a block diagonal matrix is block diagonal, the score statistic splits into a piece contributed by the variance components plus a piece contributed by the mean components. The maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}})$ under the null model is a stationary point of the loglikelihood. Thus, the variance components segment $\nabla_{\boldsymbol{\sigma}} L(\hat{\boldsymbol{\theta}})$ of the score vector vanishes. We therefore focus on the mean components segment of the score vector.

If the pedigrees are labeled $1, \ldots, n$, then the pertinent quantities for implementing the score test are

$$\sum_{i=1}^{n} \nabla_{\boldsymbol{\beta}} L_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{A}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i$$

$$\sum_{i=1}^{n} \mathrm{E}\left[-d_{\boldsymbol{\beta}}^2 L_i(\boldsymbol{\theta})\right] = \sum_{i=1}^{n} \boldsymbol{A}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{A}_i,$$

where $\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{A}_i\hat{\boldsymbol{\beta}}$ is the residual for pedigree i and the covariance matrix $\boldsymbol{\Omega}_i$ for pedigree i is determined by

Equation (2). See Chapter 8 of Lange [2002] for a detailed derivation of the score and expected information. Since the score statistic is calculated from estimated parameters under the null model, residuals do not change when we expand the null model to the alternative model keeping $\beta_{\text{SNP}} = 0$. Calculation of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ under the null is accomplished by a quasi-Newton algorithm whose initial step reduces to Fisher scoring [Lange et al., 1976, Lange, 2002].

For pedigree i under the alternative hypothesis, the design matrix $\boldsymbol{A}_i$ can be written as $(\boldsymbol{a}_i, \boldsymbol{N}_i)$, where $\boldsymbol{N}_i$ is the design matrix under the null hypothesis and $\boldsymbol{a}_i$ conveys the genotypes at the current SNP. In testing a univariate trait, the entries of $\boldsymbol{a}_i$ are taken from Table 2. If allele counts are imputed under the additive model, then the entries of $\boldsymbol{a}_i$ may be fractional numbers drawn from the interval $[-1, 1]$. In testing a multivariate trait with $T > 1$ components, each row of $\boldsymbol{A}_i = (\boldsymbol{a}_i, \boldsymbol{N}_i)$ must be replicated T times. The only exceptions to this rule occur for people missing some but not all component traits; otherwise, the covariance matrix $\boldsymbol{\Omega}_i$ for pedigree i decomposes into a sum of Kronecker products [Lange, 2002]. Regardless of whether the trait is univariate or multivariate, one must compute the quantities

$$\sum_{i=1}^{n} \nabla_\beta L_i(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i \\ \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i \end{pmatrix}$$

$$\sum_{i=1}^{n} \mathrm{E}[-d_\beta^2 L_i(\boldsymbol{\theta})] = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i & \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i \\ \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i & \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i \end{pmatrix}.$$

At the maximum likelihood estimates under the null model, the partial score vector $\sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i$ vanishes. Hence, the score statistic for testing a SNP can be expressed as

$$S = \boldsymbol{R}^t \left[\boldsymbol{Q} - \boldsymbol{W}^t \left(\sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i \right)^{-1} \boldsymbol{W} \right]^{-1} \boldsymbol{R},$$

where

$$\boldsymbol{Q} = \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i, \quad \boldsymbol{R} = \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i,$$

$$\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i.$$

In the score statistic S, the covariance matrices $\boldsymbol{\Omega}_i^{-1}$ and residual vectors $\boldsymbol{r}_i$ are evaluated at the maximum likelihood estimates under the null model. Large sample theory says that S asymptotically follows a χ_T^2 distribution.

These formulas suggest that we precompute and store the quantities $\boldsymbol{\Omega}_i^{-1}$, $\boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i$, and $\boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i$ for each pedigree i and the overall sum $\sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i$ at the maximum likelihood estimates under the null hypothesis. From these parts, the basic elements of the score statistic can be quickly assembled. The most onerous quantity that must be computed on the fly as each new SNP is encountered is $\sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i$. If there are p_i people in pedigree i, then computation of the quadratic form $\boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i$ requires $O(p_i^2)$ arithmetic operations. This looks worse than it is in practice since the entries of $\boldsymbol{a}_i$ are integers (–1, 0, and 1) in the absence of fractional imputation. This simplification allows one to avoid a fair amount of arithmetic. Assembling the remaining parts of the score statistic requires $O(p_i)$ arithmetic operations.

Individuals missing univariate trait values are omitted from analysis. Individuals missing some but not all components of a multivariate trait are retained in analysis. The proper adjustments for missing data are made automatically in the score statistic because sections of Gaussian random vectors are Gaussian.

SNPs with minor allele counts below a user-designated threshold are also omitted from analysis. Note that if the minor allele count across a study is 0, then the given SNP is mono-allelic and worthless in association testing. MENDEL's default threshold of three is motivated by the rule of thumb in contingency table testing that all cells have an expected count of at least three. For a multivariate trait, a SNP may fall below the threshold for some component traits but not for others. This situation can occur when each trait displays a different pattern of missing data across individuals. MENDEL retains such anomalous SNPs only for those component traits with a sufficient number of minor alleles. Again, proper adjustments are made automatically within the score-test statistic to account for partial data.

MENDEL's analysis yields a score-test P-value for each SNP. For the user-designated most significant SNPs, MENDEL's subsequent likelihood ratio test outputs an estimated SNP effect size, a standard error of that estimate, and the fraction of the total variance explained by that SNP. For a multivariate trait, MENDEL outputs a SNP effect size and associated standard error for each component trait. In the initial analysis under the null model with no SNPs, MENDEL provides estimates with standard errors of all mean and variance components included in the model. Finally, an estimate of heritability with standard error is also provided.

The extension of the score test to the multivariate t-distribution is straightforward [Lange et al., 1989]. Suppose η equals the degrees of freedom of the t-distribution and m_i equals the number of observed person-trait combinations for pedigree i. The sections of the score and expected information pertinent to the mean components for the pedigree reduce to

$$\nabla_\beta L_i(\boldsymbol{\theta}) = \frac{\eta + m_i}{\eta + s_i} \boldsymbol{A}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i$$

$$\mathrm{E}\left[-d_\beta^2 L_i(\boldsymbol{\theta}) \right] = \frac{\eta + m_i}{\eta + m_i + 2} \boldsymbol{A}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{A}_i,$$

where r_i is the residual and $s_i = \boldsymbol{r}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i$ is the associated Mahalanobis distance. A sensible choice for η is its estimate under the null model.

Kinship Estimation From SNPs

MENDEL can either calculate the global kinship coefficient matrix $\mathbf{\Phi}$ from the provided pedigree structures or estimate it from dense genotypes. In global kinship estimation, MENDEL's default uses an evenly spaced 20% of the available SNPs, and only compares pairs of individuals within defined pedigrees. Hence, $\mathbf{\Phi}$ is block diagonal. Users can trivially elect to exploit a larger fraction of the available SNPs or estimate kinship for *all* pairs of individuals. Given S selected SNPs, MENDEL estimates the global kinship coefficient of individuals i and j based on either the genetic relation matrix (GRM) method

$$\hat{\Phi}_{ij} = \frac{1}{2S}\sum_{k=1}^{S}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$$

or the method of moments (MoM) [Day-Williams et al., 2011, Lange et al., 2014]

$$\hat{\Phi}_{ij} = \frac{e_{ij}-\sum_{k=1}^{S}\left[p_k^2+(1-p_k)^2\right]}{S-\sum_{k=1}^{S}\left[p_k^2+(1-p_k)^2\right]},$$

where p_k is the minor allele frequency at SNP k, x_{ik} is the number of minor alleles in i's genotype at SNP k, and

$$e_{ij} = \frac{1}{4}\sum_{k=1}^{S}\left[x_{ik}x_{jk}+(2-x_{ik})(2-x_{jk})\right]$$

is the observed fraction of alleles identical-by-state (IBS) between i and j. The GRM method is MENDEL's default. In general, one can think of the GRM method centering and scaling each genotype, while the MoM method uses the raw genotypes and then centers and scales the final result.

Other Utilities for Handling Pedigree Data

To encourage thorough testing of new statistical methods, such as the current Ped-GWAS score test, we have implemented both genotype and trait simulation in our genetic analysis program MENDEL [Lange et al., 2013]. MENDEL does genotype simulation (gene dropping) subject to prescribed allele frequencies, a given genetic map, and Hardy–Weinberg and linkage equilibrium. If one fixes founder haplotypes and simulates conditional on these, then the unrealistic assumption of linkage equilibrium can be relaxed. Missing data patterns are respected or imposed by the user. It is also possible to set the rate for randomly deleting data and to simulate genotypes for people of mixed ethnicity by defining different ancestral populations, each with its own allele frequencies. If this feature is invoked, then each pedigree founder should be assigned to a population.

Trait simulation can be layered on top of genotype simulation. MENDEL simulates either univariate traits determined by generalized linear models or multivariate Gaussian traits determined by variance component models. The biggest limitations are the restriction to a single major locus and the generalized linear model assumption that trait correlations are driven solely by this locus. Variance component models enable inclusion of environmental effects and more complicated correlations among relatives. In the variance component setting, univariate as well as multivariate Gaussian traits can be simulated. Most variance component models are built on Gaussian distributions, but MENDEL allows one to replace these by multivariate t-distributions. Thus, users can investigate robust statistics less prone to distortion by outliers. More theoretical and implementation details appear in the MENDEL documentation [Lange et al., 2013].

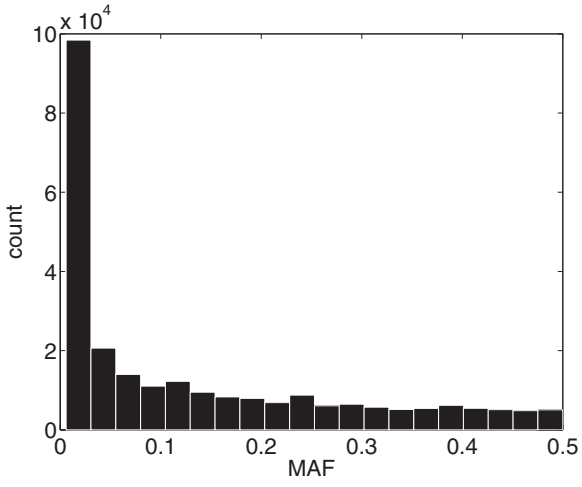

Figure 1. Histogram of minor allele frequencies (MAF) of 253,141 SNPs on chromosome 19 in 85 individuals.

Results

Simulated Data Examples

We performed a variety of simulations to evaluate the score test's computational efficiency, type I error, power, and treatment of multivariate traits. Run times in this section were recorded on a standard laptop computer with a 2.6 GHz Intel i7 CPU.

SNP Data Preparation

To simulate data with realistic linkage disequilibrium (LD) structure, we took advantage of phased sequence data from chromosome 19 on 85 individuals of northern and western European ancestry (originally from the CEPH sample) made publicly available in the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2010]. After we used the VCFTOOLS software [Danecek et al., 2011] to remove markers that were mono-allelic in this set of individuals, 253,141 SNPs remained. Figure 1 displays the histogram of the minor allele frequencies (MAF) in these individuals. Almost half of the SNPs have MAFs below 5%. The haplotype pairs attributed to the 85 CEPH members were reassigned to the 85 founders of 27 pedigree structures selected from the Framingham Heart Study (FHS). The selected Framingham pedigrees were chosen to reflect the kind of pedigrees commonly collected in family-based genetic studies. The 27 pedigrees encompass

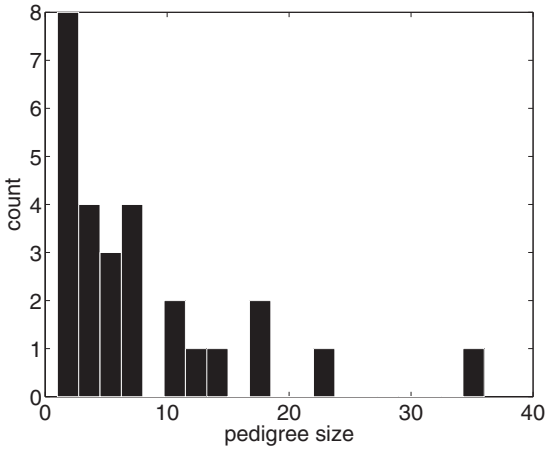

Figure 2. Histogram of pedigree sizes.

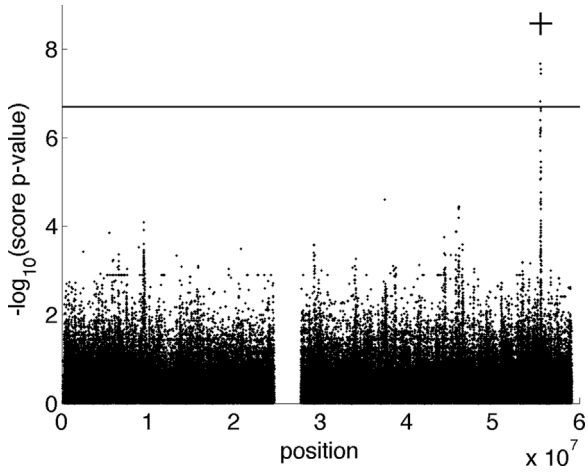

Figure 3. Manhattan plot of the score-test *P*-values for 253,141 SNPs on chromosome 19. Trait values were simulated based on a major locus at SNP `rs10412915` (position 55,494,740) in the `NLRP2` gene. The $-\log_{10}$(score *P*-value) at this SNP is marked with a plus sign. The horizontal line represents the significance threshold for this dataset. See the text for the detailed simulation model.

212 people, range in size from 1 to 36 people and from 1 to 5 generations, and contain sibships of 1–5 children. Figure 2 shows the histogram of the pedigree sizes. The genotypes of nonfounders were simulated conditional on the haplotypes imposed on the founders and recorded as unordered for subsequent analysis purposes.

Univariate Trait QTL Mapping

We simulated a univariate quantitative trait with a major locus at SNP `rs10412915` (MAF = 0.259; position 55,494,740 on chromosome 19) using the trait simulation option of MENDEL. The mean effects included the intercept $\mu = 40$ and the regression coefficients $\beta_{\text{snp}} = 2$ and $\beta_{\text{sex}} = 6$; the variance components included $\sigma_a^2 = 5$, $\sigma_e^2 = 1$, and $\sigma_h^2 = \sigma_d^2 = 0$. (See Eq. (2) and the subsequent description of the model for the definition of these parameters.) Power under other effect sizes is explored in a later experiment. Figure 3 displays a Manhattan plot of the *P*-values generated by the score tests. The signal emanating from the major locus is clearly discernible and is the only significant finding. MENDEL took about 6.5 sec for initialization, which includes reading the data, checking for gross errors, performing standard quality control (QC) procedures such as filtering of SNPs and individuals with low genotyping rates, and computing summary statistics. Using all 27 pedigrees, MENDEL then required 5.9 sec to compute the score-test *P*-values at all 253,141 SNPs. Total run time was less than 13 sec.

Score Test vs. LRT

MENDEL allows users to specify how many of the most significant score-test SNPs are reanalyzed using a likelihood ratio test (LRT). In the current example we told MENDEL to calculate the LRTs on the 50 most significant SNPs flagged by the score test. It took MENDEL an additional second to perform these LRTs. This translates into a total run time for data input, QC, and analysis of less than 14 sec. When we told MENDEL to perform LRTs on all SNPs, it took 53 min and 37 sec. The almost 500-fold speedup of the score test over the LRT demonstrates the dramatic gains in computational efficiency possible. In large-scale sequencing studies, we expect an order of magnitude increase in both study individuals and typed SNPs. In later sections we discuss more fully efficiency and power for various models and datasets.

To alleviate concerns about the loss of power in substituting the score test for the LRT, we plot in Figure 4 the top 50 score test and LRT *P*-values. The two top-50 SNP sets coincide. The scatter plot (left panel) shows extremely high correlation ($r = 0.9999$). That all points lie above the 45-degree line indicates that the LRT has uniformly more power (smaller *P*-values) than the score test. The ranking of SNPs is of interest in many pilot studies. The Q–Q plot (right panel) shows that these two tests produce virtually identical rankings. Kendall's τ correlation is 0.9983, and Spearman's correlation is 0.9998.

Discarding vs. Estimating Pedigree Information

We performed two experiments to evaluate the impact of discarding pedigree information in association testing. In the first, we treated all 212 individuals as unrelated and tested all SNPs by linear regression with sex as a covariate. This is the same mean effects model employed in the previous example. It took MENDEL about 6.5 sec for initialization and 5.3 sec for analysis. In the second experiment, we discarded the nonfounders and carried out the same association testing on just the 85 founders. This took MENDEL 4.3 sec for initialization and 4.5 sec for analysis. The top two panels of Figure 5 display the Manhattan plots of the two experiments discarding pedigree information. As expected, ignoring relationships, which means ignoring the reason for some of the genetic similarity, leads to many false positives (as seen in the first experiment). Also as expected, reducing the size of the dataset to include

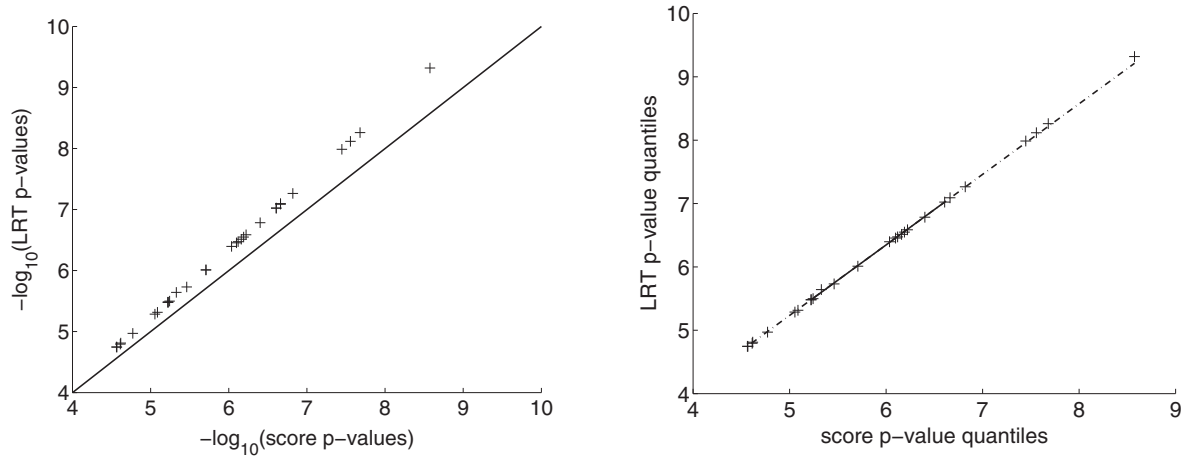

Figure 4. Comparisons of the score and LRT *P*-values. Left: A scatter plot of the top 50 score and LRT *P*-values demonstrates extremely high correlation (r = 0.9999) between the two sets of *P*-values and a uniformly higher power for the LRT. Right: A Q–Q plot of the top 50 score and LRT *P*-values shows that the two tests produce virtually identical rankings. The simulation model is the same as in Figure 3.

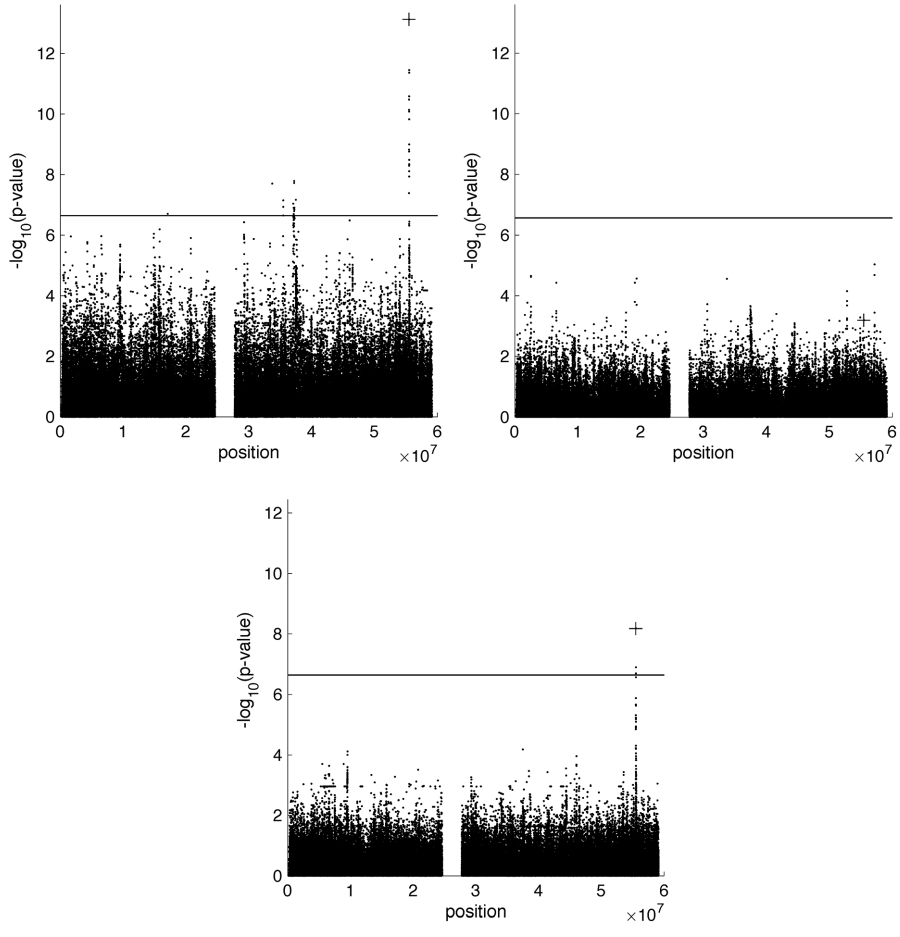

Figure 5. GWAS results suffer when pedigree structure is ignored. Upper left: Manhattan plot of GWAS that treats all 212 individuals as unrelated shows the expected many false positives. Upper right: Manhattan plot of GWAS that includes only the 85 founders shows the expected loss of power. Lower: Manhattan plot of GWAS using all 212 individuals and estimating all kinship information from only the SNP data. A plus sign marks the $-\log_{10}(P\text{-value})$ at the SNP used to simulate the trait. The horizontal line represents the significance threshold for this dataset. The simulation model is the same as in Figure 3.

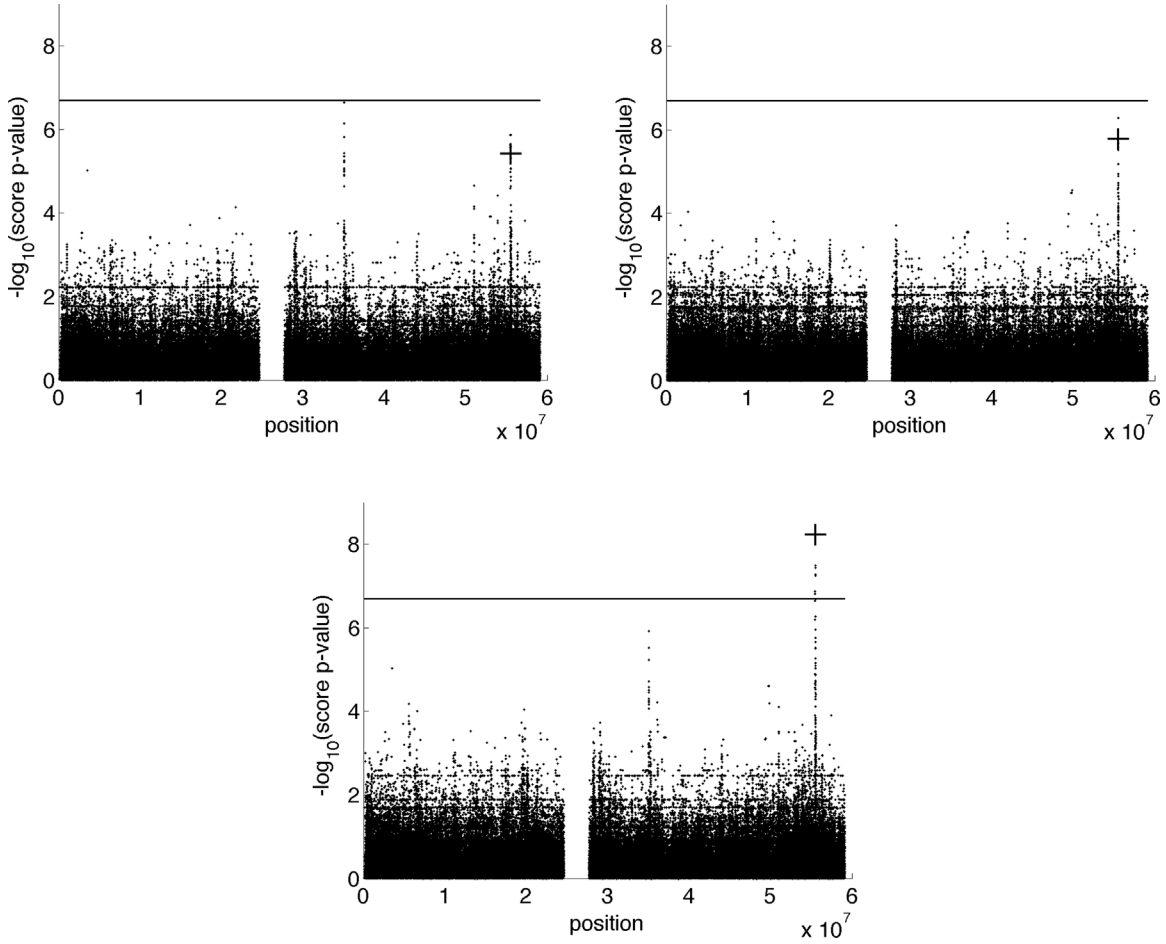

Figure 6. Bivariate QTL mapping. Upper left: Manhattan plot for testing trait 1. Upper right: Manhattan plot for testing trait 2. Lower: Manhattan plot for testing traits 1 and 2 together. Bivariate QTL mapping demonstrates better power than testing each univariate trait separately. The $-\log_{10}$(score *P*-value) at the major locus `rs10412915` (position 55,494,740) is marked with a plus sign. The horizontal line represents the significance threshold for this dataset. See the text for the simulation model.

only unrelateds greatly reduces power (as seen in the second experiment).

Fortunately, when genealogies are missing or dubious, the method of Day-Williams et al [Day-Williams et al., 2011] implemented in MENDEL allows fast and accurate estimation of global kinship coefficients from dense markers. It took MENDEL 18.8 sec to estimate all global kinship coefficients from the 253,141 SNPs. The lower panel in Figure 5 shows the Manhattan plot of the pedigree GWAS based on the estimated kinship coefficients. There is little difference from the results using exact pedigree structures.

Multivariate Trait QTL Mapping

To assess the ability of our ped-GWAS method to detect a pleiotropic effect at the selected major locus `rs10412915`, we simulated two correlated quantitative traits on the previously constructed pedigrees. Trait 1 has mean effects $\mu_1 = 40$, $\beta_{\text{sex},1} = 6$, and $\beta_{\text{snp},1} = 1.5$ and variance components $\sigma^2_{a1} = 5$ and $\sigma^2_{e1} = 1$. Trait 2 has mean effects $\mu_2 = 20$, $\beta_{\text{sex},2} = 4$, and $\beta_{\text{snp},2} = 1.5$ and variance components $\sigma^2_{a2} = 5$ and $\sigma^2_{e2} = 1$. The additive and environmental covariances between the two traits are $\sigma^2_{a1,a2} = 1$ and $\sigma^2_{e1,e2} = 0$. Compared to our earlier univariate trait simulation, SNP effects are reduced for each trait while variance components are held fixed. Figure 6 displays Manhattan plots for testing trait 1 alone, trait 2 alone, and both traits 1 and 2 together. When both traits are tested simultaneously, it takes MENDEL about 6.9 sec for initialization and 9.9 sec for analysis. Despite the reduction in SNP effect sizes, testing both traits simultaneously boosts power significantly. The benefits diminish when the traits are more highly correlated, for example by taking $\sigma^2_{a1,a2} = 3$ and $\sigma^2_{e1,e2} = 0.5$. For the sake of brevity, these further results are not graphed.

Comparison to Current Methods

In this section, we compare the score test to the competing generalized estimating equation (GEE) and variance component model (linear-mixed model, LMM) approaches implemented in the R package GWAF [Chen and Yang, 2010]. Our comparison criteria include computational efficiency, memory usage, type I error, and power. Table 3 shows run

Table 3. Comparison of total run times (in seconds on a standard laptop computer) with GWAF

# SNPs	MENDEL-Score	MENDEL-LRT	GWAF-GEE	GWAF-LMM
100	4.69	5.32	0.71	8.83
1,000	4.75	5.48	7.71	87.06
10,000	5.28	6.05	207.60	894.82
100,000	10.28	11.07	26,486.92	11,703.88

Run times are based on testing the first 100, 1,000, 10,000 and 100,000 SNPs on chromosome 19. The column labeled MENDEL-LRT displays the total run times after adding likelihood ratio tests for the top 50 SNPs identified by the score test. The simulation model is the same as in Figure 3.

Table 4. Empirical power and type I error for various major-locus effect sizes

	MENDEL-Score	MENDEL-LRT	GWAF-GEE	GWAF-LMM
Power				
(β_{snp} = 2.0)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
(β_{snp} = 1.5)	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.02	1.00 ± 0.00
(β_{snp} = 1.2)	0.98 ± 0.01	0.98 ± 0.01	0.89 ± 0.03	0.98 ± 0.01
(β_{snp} = 1.0)	0.92 ± 0.03	0.92 ± 0.03	0.80 ± 0.04	0.92 ± 0.03
(β_{snp} = 0.8)	0.75 ± 0.04	0.75 ± 0.04	0.54 ± 0.05	0.75 ± 0.04
(β_{snp} = 0.5)	0.38 ± 0.05	0.39 ± 0.05	0.29 ± 0.05	0.40 ± 0.05
(β_{snp} = 0.3)	0.14 ± 0.03	0.15 ± 0.04	0.16 ± 0.04	0.15 ± 0.04
Type I Error (β_{snp} = 0.0)	0.04 ± 0.02	0.04 ± 0.02	0.09 ± 0.03	0.04 ± 0.02

The simulation model is the same as in Figure 3. The empirical power is the proportion of replicates with *P*-values less than 0.05 under the alternative model with the listed major-locus effect size. The empirical type I error is the proportion of replicates with *P*-values less than 0.05 under the null model with no major locus. All results represent averages across 100 replicates per model; standard errors appear to the right of each average. The column labeled MENDEL-LRT displays the results after adding likelihood ratio tests for the top 50 SNPs identified by the score test.

times for testing the first 100, 1,000, 10,000, and 100,000 SNPs on chromosome 19. Simulation parameters coincide with those used in Figure 3. MENDEL-LRT lists runs in which the 50 most significant SNPs were further subjected to an LRT. The table lists the total wall clock times for the initialization and analysis phases. In testing 100,000 SNPs, MENDEL shows a roughly 1000-fold speed-up over the GWAF-GEE and GWAF-LMM approaches. This fact validates our initial premise that the score test would offer large gains in speed. When testing 100,000 SNPs, MENDEL never used more than 76 MB of RAM. In contrast, GWAF had a memory footprint larger than 500 MB, a serious concern for testing large-scale GWAS data.

Next we compared the type I error and power of the four methods. In the alternative model, we simulated trait values according to the settings pertinent to Figure 3 with the major locus `rs10412915` retained but with varying effect sizes β_{snp}. In the null model, we discarded the major locus effect and kept the other simulation parameters. All results represent averages across 100 replicates per model. Table 4 tallies the empirical type I error (proportion of replicates with *P*-values less than 0.05 under the null model) and power (proportion of replicates with *P*-values less than 0.05 under the alternative model), along with their standard errors. We observe inflated type I error and lowest power in the GEE results, especially at medium to large effect sizes. This is possibly due to the imposition in the current implementation of GWAF-GEE of a uniform working correlation structure across all pedigrees. Although standard semiparametric theory states that main effects can be consistently estimated even under misspecification of the correlation structure, the sample sizes in real genetic studies are rarely sufficient for such asymptotics to hold. Table 4 suggests that Mendel and GWAF-LMM possess similar operating characteristics. Unfortunately, the extremely low computational efficiency of GWAF-LMM makes it an unattractive choice for GWAS. Modern genetic studies such as those in Framingham and San Antonio often involve at least an order of magnitude more people and (imputed) SNPs than we have simulated here.

Real Data Examples

The San Antonio Family Heart Study

We analyzed a real dataset collected by the San Antonio Family Heart Study (SAFHS) [Mitchell et al., 1996]. The data consist of 3,637 individuals in 211 Mexican American families. High-density lipoprotein (HDL) levels were measured at up to three time points for each of the 1,429 phenotyped individuals. These traits are denoted HDL_1, HDL_2, and HDL_3, measured at corresponding ages AGE_1, AGE_2, and AGE_3. Some of the phenotyped individuals have HDL measurements at only one or two of the time points. Of the 1,429 phenotyped individuals, 1,413 were genotyped at 944,427 genome-wide SNPs. The genotyping success rate exceeded 98% in 1,388 of these individuals over 124 pedigrees. The largest family contains 247 individuals; five others also contain more than 90 individuals. The smallest pedigree was a singleton. Genotyping success rates were above 98% for 935,392 SNPs.

Comparison With FAST-LMM *and* GEMMA

For fair comparisons, we directed MENDEL to estimate SNP-based global kinship coefficients for all pairs of individuals ignoring the input pedigrees. This is the default in FAST-LMM and GEMMA. In addition, we ran MENDEL's default in which the coefficients are estimated only for pairs of individuals within the same input pedigree. We also slightly adjusted some of the default quality control thresholds so the programs would be analyzing roughly the same set of SNPs and individuals. For example, by default MENDEL filters SNPs with fewer than three occurrences of the minor allele in the data; in contrast, FAST-LMM only filters SNPs with zero occurrences of the minor allele, and GEMMA filters SNPs with minor allele frequency (MAF) < 0.01. All other defaults were observed throughout. Users can easily adjust the MENDEL analysis parameters via its control file and the FAST-LMM and GEMMA analysis parameters via their command line.

We first carried out three univariate QTL analyses of HDL_1, HDL_2, and HDL_3, using SEX and AGE_1, AGE_2, or AGE_3 as covariates. We then ran a multivariate QTL analysis of HDL_1, HDL_2, and HDL_3 jointly, which we refer to as HDL_{Joint}. For the multivariate analysis, the most appropriate configuration is to constrain the effects of the SEX and AGE covariates to be the same on all three HDL measurements. Such linear

Table 5. All SNPs with minor allele frequency (MAF) above 0.001 that reach genome-wide significance in any of the analyses of the HDL traits from the San Antonio Family Heart Study (SAFHS)

Trait	SNP	Chr.	Base pair position	MAF	$-\log_{10}(P-\text{val})$ MENDEL default	$-\log_{10}(P-\text{val})$ MENDEL all-pairs	$-\log_{10}(P-\text{val})$ FAST-LMM	$-\log_{10}(P-\text{val})$ GEMMA
HDL_1	rs7303112	12	97,596,023	0.00455	10.21	10.71	7.63	7.24
	rs8040647	15	32,304,988	0.00147*	7.44	7.56	7.35	7.45
	rs9972594	15	32,421,102	0.00147*	7.44	7.56	7.37	7.46
	rs7167103	15	32,830,477	0.00147*	7.44	7.56	7.35	7.44
HDL_2	rs7100957	10	28,207,332	0.00183*	8.84	8.95	8.88	8.82
HDL_3	rs17060933	8	22,510,029	0.00382	8.23	8.28	8.61	8.59
HDL_{Joint} with constrained covariates	rs7303112	12	97,596,023	0.00644	9.89	9.94	Not Available	Not Available
	rs16925210	10	25,308,103	0.00217	8.15	8.33		
	rs7091416	10	25,318,381	0.00217	8.15	8.33		
	rs10075658	5	148,911,957	0.00144*	8.16	8.21		
	rs7733139	5	145,977,990	0.00217	7.36	7.34		
	rs7100957	10	28,207,332	0.00870	7.20	7.30		
HDL_{Joint} without constrained covariates	rs7303112	12	97,596,023	0.00644	9.82	9.88	Not Available	11.08
	rs16925210	10	25,308,103	0.00217	8.04	8.23		3.53
	rs7091416	10	25,318,381	0.00217	8.04	8.23		3.52
	rs10075658	5	148,911,957	0.00144*	8.12	8.17		3.47
	rs7733139	5	145,977,990	0.00217	7.41	7.40		3.47
	rs7100957	10	28,207,332	0.00870	7.19	7.30		4.48
	rs10083226	13	104,434,452	0.00219	7.10	7.31		2.14

All default parameters were used except for minor changes to the quality control thresholds (see text). Also, MENDEL was run in both default and all-pairs modes. MENDEL's default mode estimates nonzero global kinship coefficients only for pairs of individuals within the same input pedigree; MENDEL in all-pairs mode, FAST-LMM, and GEMMA estimate coefficients for all pairs of individuals. Genome-wide significance was declared for P-values $< 5 \times 10^{-8} \Rightarrow -\log_{10}(P-\text{value}) > 7.3$. The SAFHS has 1,413 genotyped and phenotyped individuals in 124 pedigrees. The genotypes include roughly 1 million SNPs. The phenotypes include the subjects' high-density lipoprotein (HDL) level and age at three time points. The HDL_{Joint} runs are multivariate analyses of HDL_1, HDL_2, and HDL_3 jointly; all other runs are univariate analyses. See the text for a list of the covariates used in each analysis. Note that in the multivariate analysis, MENDEL is able to use roughly twice as many individuals as GEMMA (see text and Table 6), which may explain the less significant findings for GEMMA. Each MAF is based on the pedigree founders, except where marked by an asterisk (*). In these cases the minor allele did not appear in the genotyped founders, and its frequency was estimated from all genotyped individuals.

constraints are imposed in MENDEL via a few simple lines in its control file. FAST-LMM and GEMMA do not allow constraints on covariates. Therefore, we also ran a multivariate analysis with only the SEX covariate and no constraints. With no constraints, SEX will have a slightly different effect on each component phenotype in the multivariate analysis. For example, MENDEL's default run estimated a female effect of 2.5 ± 0.3 on HDL_1, 2.1 ± 0.4 on HDL_2, and 2.7 ± 0.4 on HDL_3. FAST-LMM cannot do multivariate analyses.

Table 5 reports all SNPs with MAF > 0.001 that achieve genome-wide significance (P-values less than 5×10^{-8}) as reported by at least one software package. For the univariate analyses, each software package found the same set of significant SNPs, except that one of GEMMA's P-values was slightly short of the significance threshold. Figure 7 shows a Manhattan plot and a Q–Q plot from the HDL_1 analysis by MENDEL given kinship estimates for all pairs of individuals. The results for the other analyses, both univariate and multivariate, were similar. Each MENDEL all-pairs univariate analysis had genomic control λ in the range 1.002 to 1.006; in default mode, λ was in the range 0.992 to 1.022. The various Q–Q plots and associated λ values show there is no systematic biases in the data or analysis. In the all-pairs MENDEL HDL_1 analysis, the grand mean (intercept) was 49.0 ± 0.8. The SEX covariate was significant in all null models. For example, in the all-pairs MENDEL HDL_{Joint} analysis with constrained covariates, the SEX effect was 2.4 ± 0.3 for females and, by design, the opposite for males. The AGE covariate was not significant in any run. For example, again in the all-pairs HDL_{Joint} analysis with parameter constraints, the AGE effect was 0.04 ± 0.02. In the null model for the all-pairs MENDEL HDL_1 analysis, the additive variance was estimated as 78.8 ± 9.9, and the environmental variance was estimated as 78.1 ± 7.2. This gives an overall heritability estimate for HDL_1 of 0.50 ± 0.04. Similar variance estimates were seen in other null models.

For the multivariate analysis without parameter constraints, MENDEL is able to include almost twice as many individuals in the analysis as GEMMA (see Table 6). GEMMA only includes individuals phenotyped at all component traits and covariates. This probably explains why MENDEL finds several more SNPs with significant P-values than GEMMA.

Table 6 tallies the run times and memory footprints from each analysis on a typical personal computer with adequate RAM to accommodate FAST-LMM (six CPU cores at 2.67 GHz, with 48 GB total RAM). Even when estimating the global kinship coefficients for all pairs of individuals, each univariate QTL run took MENDEL less than 8 min to read, quality check, and analyze the data for kinship estimates and association tests, roughly 10% of the time required for FAST-LMM and 5% of the time required by GEMMA. (For GEMMA, the kinship estimation and association tests are run separately. The run times reported here are their total.)

The three programs use different association test strategies: MENDEL performs score tests for all SNPs and LRTs for the top

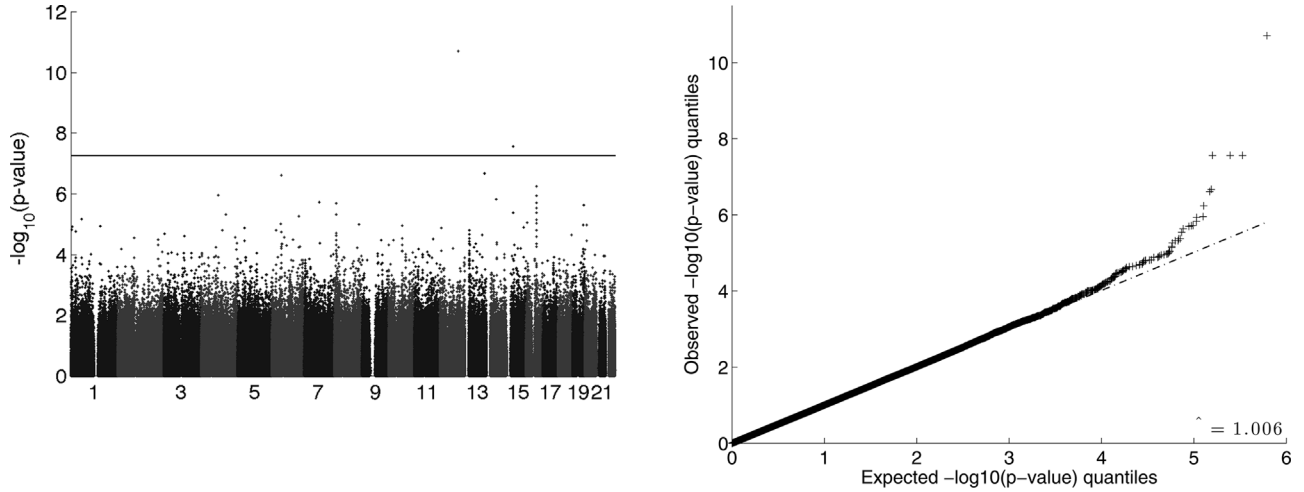

Figure 7. The results of MENDEL's HDL_1 univariate analysis in the SAFHS dataset with global kinship coefficients estimated for all pairs of individuals. Left: The Manhattan plot graphs roughly one million SNPs against their $-\log_{10}(P-\text{value})$. The horizontal line is the genome-wide significance threshold, $7.3 = -\log_{10}(5 \times 10^{-8})$. Right: The Q–Q plot graphs the observed $-\log_{10}(P-\text{value})$ quantiles vs. their expectations. The genomic control value of $\hat{\lambda} = 1.006$ derived from this comparison suggests no systematic biases in the data or analysis.

Table 6. Comparison of run times and memory (RAM) usage on a typical computer but with adequate RAM to accommodate FAST-LMM (six CPU cores at 2.67 GHz, with 48 GB total RAM)

Program	Trait	Analyzed samples	Analyzed SNPs	RunTime (min:sec)	RAM (GB)
MENDEL default		1,357	935,392	1:51	1.2
MENDEL all-pairs	HDL_1	1,357	935,392	7:49	1.2
FAST-LMM		1,397	941,546	76:11	30.0
GEMMA		1,397	919,050	206:54	0.4
MENDEL default		818	935,392	1:33	1.1
MENDEL all-pairs	HDL_2	818	935,392	3:25	1.1
FAST-LMM		840	934,216	49:44	18.0
GEMMA		840	914,051	180:21	0.3
MENDEL default		914	935,392	1:38	1.1
MENDEL all-pairs	HDL_3	914	935,392	3:54	1.1
FAST-LMM		939	937,208	54:58	20.0
GEMMA		939	918,626	182:26	0.3
MENDEL default	HDL_{Joint}	1,388	935,392	4:08	1.2
MENDEL all-pairs	with	1,388	935,392	83:24	1.2
FAST-LMM	constrained		Not available		
GEMMA	covariates		Not available		
MENDEL default	HDL_{Joint}	1,388	935,392	3:49	1.2
MENDEL all-pairs	without	1,388	935,392	80:04	1.2
FAST-LMM	constrained		Not available		
GEMMA	covariates	712	912,318	630:37	0.6

The listed run times include reading the dataset, performing quality checks, estimating the kinship coefficients, and calculating the association test *P*-values. All default parameters were used except for minor changes to the quality control thresholds (see text). Also, MENDEL was run in both default and all-pairs modes. MENDEL's default mode estimates nonzero global kinship coefficients only for pairs of individuals within the same input pedigree; MENDEL in all-pairs mode, FAST-LMM, and GEMMA estimate coefficients for all pairs of individuals. For the multivariate analysis, MENDEL includes roughly twice as many individuals as GEMMA because GEMMA only analyzes individuals phenotyped at all component traits and covariates. MENDEL performs score tests for all SNPs and LRTs for the top SNPs; FAST-LMM performs LRTs; and GEMMA by default performs Wald tests, but the user can change this to LRTs or score tests. Using score tests in GEMMA would make it faster (see text).

SNPs; FAST-LMM performs LRTs; and GEMMA by default performs Wald tests, but the user can change this to LRTs or score tests. For the univariate analyses on a six-core computer, excluding estimation of kinship coefficients, GEMMA's run times under the Wald test and LRT options were roughly similar to FAST-LMM's; GEMMA's run time under the score test option was roughly double MENDEL's in all-pairs mode. This is impressive given GEMMA's lack of multithreading. It is kinship estimation, which in practice can be done once per dataset, that is substantially slower in GEMMA (running roughly 135 minutes) than in FAST-LMM or MENDEL (less than 1 min).

Each trivariate QTL run took MENDEL less than 90 min. MENDEL required roughly one-eighth the time of GEMMA while analyzing almost twice as many individuals. MENDEL is also memory efficient. The univariate and multivariate runs each required less than 1.5 GB of memory, which is well below the amount of RAM in a typical computer. FAST-LMM's memory usage is more than 15 times larger than MENDEL's. GEMMA uses even less memory than MENDEL but is considerably slower.

Discussion

We have implemented an ultra-fast algorithm for QTL analysis of pedigree data or mix of population and pedigree data. In our opinion MENDEL's comprehensive environment for genetic data analysis is a decided advantage. In addition to its exceptional speed and memory efficiency, MENDEL can handle multivariate quantitative traits and detect outlier trait values and pedigrees. Most competing programs ignore multivariate traits and outliers altogether.

A recent review of univariate QTL analysis packages for family data [Eu-ahsunthornwattana et al., 2014] shows that all the explored packages obtain similar results, leaving speed, features, and ease of use as the important factors in choosing between them. Once the current version of MENDEL came out, the authors of the review were kind enough to add a comment (http://www.plosgenetics.org/annotation/listThread.action?root=81847) to their article observing that MENDEL was now the fastest and one of the easiest to use packages they reviewed.

In the SAFHS example dataset we used with HDL phenotypes, all the significant SNPs we found had MAF < 0.01. Due to these low MAFs, we do not claim these SNPs are strong candidates for further study. However, the key point here is that all four methods found the same SNPs, at least for the univariate analyses. We also note that the *P*-values are quite similar regardless of whether one uses kinship estimates between all individuals (MENDEL's all-pairs mode) or only between individuals within the same input pedigrees (MENDEL's default mode). This suggests that the input pedigree structures for this dataset are substantially correct and complete, with few mistaken or hidden relationships. Obviously, this may not be true for other datasets. By supplying good kinship estimates ignoring pedigree structures, the currently reviewed packages make the hard fieldwork of relationship discovery superfluous.

A future version of MENDEL will address its failure to read fractional genotype values. This is simply a logistical issue, as all MENDEL's internal genotype computations are already handled as floating point operations. Another imminent feature is a fourth style of kinship coefficient estimation that allows the user to force theoretical kinship coefficients for pairs of individuals within the same pedigree and estimated kinships for all other pairs.

By supplying a comprehensive, fast, and easy to use package for GWAS on quantitative traits in general pedigrees, we hope to encourage exploitation of family-based datasets for gene mapping. A gene mapping study should collect as large a sample as possible consistent with economic constraints and uniform trait phenotyping. If the sample includes pedigrees, all the better. One should *not* let the choice of statistical test determine the data collected; on the contrary, the data should determine the test. Here, we have argued that score tests can efficiently handle unrelated individuals, pedigrees, or a mixture of both. For human studies, where controlled breeding is forbidden, nature has provided pedigrees segregating every genetic trait. Many of these pedigrees are known from earlier linkage era studies and should be treasured as valuable resources.

Let us suggest a few directions for future work. The current method works marker by marker and is ill equipped to perform model selection. Penalized regression, such as lasso, is available to handle model selection for case-control and random sample data [Wu and Lange, 2008, Wu et al., 2009, Zhou et al., 2010, 2011] and can be generalized to variance component models. Although we have generalized the score test to distributions such as the multivariate *t*, extending it to discrete traits may be out of reach. For likelihood-based methods, there simply are no discrete analogues of the Gaussian distribution that lend themselves to graceful evaluation of pedigree likelihoods. Treating case/control data as a 0/1 quantitative variable is a possibility that has been explored by Pirinen et al. [2013]. The GEE method is another fallback option because it does not depend on precise distributional assumptions.

In rare variant mapping, grouping related SNPs in a variance component may be a good alternative to the mean component models used here. Each variant may be too rare to achieve significance in hypothesis testing. Fortunately, aggregating genotype information within biological units such as genes or pathways offer better power than marginal testing of individual SNPs. See Asimit and Zeggini [2010] for a recent review of aggregation strategies. Kwee et al. [2008] have successfully applied a variance component model for association testing of SNP sets in a sample of unrelated subjects. Rönnegård et al. [2008] consider score tests for random effects models in the context of experimental line crosses. Score tests may well be the key to implementing random effect models in pedigrees. However, the computational demands are apt to be more formidable than those encountered here with fixed effects models. In particular, if tests are based simply on local identity-by-descent (IBD) sharing, then the boundaries between pedigrees disappear, and the entire sample collapses to one large pedigree. The required local kinship coefficients can again be well estimated from dense markers, but this demands more computation than the estimation of global kinship coefficients under the mean components model advocated here [Day-Williams et al., 2011]. Since inversion of a pedigree covariance matrix scales as the cube of the number of individuals in the pedigree, treating the entire sample as a single pedigree will put a practical upper limit on sample size. There are other issues in implementing variance component models such as assigning *P*-values and dealing with multivariate traits that are best left to a separate paper.

Acknowledgment

The authors gratefully acknowledge the NIH grants GM053275 (E.M.S. and K.L.), HG006139 (H.Z., E.M.S., and K.L.), MH059490 (J.B., T.D.D., E.M.S., and K.L.), and GM105785 (H.Z.) and NSF grant DMS1310319 (H.Z.) supporting this research. K.K.C. also gratefully acknowledges the fellowship support from the Burroughs Wellcome Fund Inter-school Training Program in Metabolic Diseases.

References

Amin N, van Duijn CM, Aulchenko YS. 2007. A genomic background based method for association analysis in related individuals. *PLoS ONE* 2(12):e1274.

Asimit J, Zeggini E. 2010. Rare variant association analysis methods for complex traits. *Ann Rev Genet* 44(1):293–308.

Aulchenko YS, de Koning D-J, Haley C. 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177(1):577–585.

Chen M-H, Liu X, Wei F, Larson MG, Fox CS, Vasan RS, Yang Q. 2011. A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees. *Genet Epidemiol* 35(7):650–657.

Chen M-H, Yang Q. 2010. GWAF: an R package for genome-wide association analyses with family data. *Bioinformatics* 26(4):580–581.

Chen W-M, Abecasis GR. 2007. Family-based association tests for genome-wide association scans. *Am J Hum Genet* 81(5):913–926.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST and others. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM. 2011. Linkage analysis without defined pedigrees. *Genet Epidemiol* 35(5):360–370.

Eu-ahsunthornwattana J, Miller EN, Fakiola M, Wellcome Trust Case Control Consortium 2, Jeronimo SMB, Blackwell JM, Cordell HJ. 2014. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet* 10(7):e1004445.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 106(23): 9362–9367.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42(4):348–354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709–1723.

Korte A, Vilhjalmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* 44(9):1066–1071.

Ku CS, Loy EY, Pawitan Y, Chia KS. 2010. The pursuit of genome-wide association studies: where are we now? *J Hum Genet* 55(4):195–206.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* 82(2):386–397.

Laird NM, Horvath S, Xu X. 2000. Implementing a unified approach to family-based tests of association. *Genet Epidemiol* 19(Suppl 1):S36–S42.

Lange C, Laird NM. 2002. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genet Epidemiol* 23(2):165–180.

Lange K. 2002. *Mathematical and Statistical Methods for Genetic Analysis. Statistics for Biology and Health* (2nd edn.). New York: Springer-Verlag.

Lange K, Little RJA, Taylor JMG. 1989. Robust statistical modeling using the *t* distribution. *J Am Stat Assoc* 84(408):881–896.

Lange K, Papp JC, Sinsheimer JS, Sobel EM. 2014. Next-generation statistical genetics: modeling, penalization, and optimization in high-dimensional data. *Ann Rev Stat Its Appl* 1(1):279–300.

Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM. 2013. Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics* 29(12):1568–1570.

Lange K, Sinsheimer JS, Sobel E. 2005. Association testing with Mendel. *Genet Epidemiol* 29(1):36–50.

Lange K, Westlake J, Spence MA. 1976. Extensions to pedigree analysis iii. variance components by the scoring method. *Annal Hum Genet* 39(4):485–491.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nat Methods* 8(10):833–835.

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. 2012. Improved linear mixed models for genome-wide association studies. *Nat Methods* 9(6):525–526.

Mitchell BD, Kammerer CM, Blangero J, Mahaney MC, Rainwater DL, Dyke B, Hixson JE, Henkel RD, Sharp RM, Comuzzie AG, VandeBerg JL, Stern MP, MacCluer JW. 1996. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans: the San Antonio Family Heart Study. *Circulation* 94(9):2159–2170.

Ott J, Kamatani Y, Lathrop M. 2011. Family-based designs for genome-wide association studies. *Nat Rev Genet* 12(7):465–474.

Pirinen M, Donnelly P, Spencer CCA. 2013. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Annal Appl Stat* 7(1):369–390.

Rao C. 2009. *Linear Statistical Inference And Its Applications* (2nd edn). New York: Wiley.

Rönnegård L, Besnier F, Carlborg O. 2008. An improved method for quantitative trait loci detection and identification of within-line segregation in f2 intercross designs. *Genetics* 178(4):2315–2326.

Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 62(2):450–458.

Stanhope SA, Abney M. 2012. GLOGS: a fast and powerful method for GWAS of binary traits with risk covariates in related populations. *Bioinformatics* 28(11):1553–1554.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.

Thornton T, McPeek MS. 2007. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet* 81(2):321–337.

Thornton T, McPeek MS. 2010. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet* 86(2):172–184.

Van Steen K. 2011. Perspectives on genome-wide multi-stage family-based association studies. *Stat Med* 30(18):2201–2221.

Van Steen K, Lange C. 2005. PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. *Hum Genomics* 2(1):67–69.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* 90(1):7–24.

Won S, Bertram L, Becker D, Tanzi R, Lange C. 2009a. Maximizing the power of genome-wide association studies: a novel class of powerful family-based association tests. *Stat Biosci* 1(2):125–143.

Won S, Wilk JB, Mathias RA, O'Donnell CJ, Silverman EK, Barnes K, O'Connor GT, Weiss ST, Lange C. 2009b. On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS Genet* 5(11):e1000741.

Wu TT, Chen Y, Hastie T, Sobel EM, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25(6):714–721.

Wu TT, Lange K. 2008. Coordinate descent algorithms for lasso penalized regression. *Annal Appl Stat* 2(1):224–244.

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42(4):355–360.

Zhou H, Alexander D, Sehl M, Sinsheimer JS, Sobel EM, Lange K. 2011. Penalized regression for genome-wide association screening of sequence data. *Pacific Symp Biocomput* 2011: 106–117.

Zhou H, Sehl ME, Sinsheimer JS, Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19):2375–2382.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44: 821–824.

Zhou X, Stephens M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 11(4):407–409.

Zhu Y, Xiong M. 2012. Family-based association studies for next-generation sequencing. *Am J Hum Genet* 90(6):1028–1045.