

A UNIFIED FRAMEWORK FOR VARIANCE COMPONENT ESTIMATION WITH SUMMARY STATISTICS IN GENOME-WIDE ASSOCIATION STUDIES¹

BY XIANG ZHOU

University of Michigan

Linear mixed models (LMMs) are among the most commonly used tools for genetic association studies. However, the standard method for estimating variance components in LMMs—the restricted maximum likelihood estimation method (REML)—suffers from several important drawbacks: REML requires individual-level genotypes and phenotypes from all samples in the study, is computationally slow, and produces downward-biased estimates in case control studies. To remedy these drawbacks, we present an alternative framework for variance component estimation, which we refer to as MQS. MQS is based on the method of moments (MoM) and the minimal norm quadratic unbiased estimation (MINQUE) criterion, and brings two seemingly unrelated methods—the renowned Haseman–Elston (HE) regression and the recent LD score regression (LDSC)—into the same unified statistical framework. With this new framework, we provide an alternative but mathematically equivalent form of HE that allows for the use of summary statistics. We provide an exact estimation form of LDSC to yield unbiased and statistically more efficient estimates. A key feature of our method is its ability to pair marginal z -scores computed using all samples with SNP correlation information computed using a small random subset of individuals (or individuals from a proper reference panel), while capable of producing estimates that can be almost as accurate as if both quantities are computed using the full data. As a result, our method produces unbiased and statistically efficient estimates, and makes use of summary statistics, while it is computationally efficient for large data sets. Using simulations and applications to 37 phenotypes from 8 real data sets, we illustrate the benefits of our method for estimating and partitioning SNP heritability in population studies as well as for heritability estimation in family studies. Our method is implemented in the GEMMA software package, freely available at www.xzlab.org/software.html.

1. Introduction. Linear mixed models (LMMs), sometimes referred to as variance component models, have been widely applied in many areas of genetics. For example, they have been used for linkage analysis and heritability estimation in family studies [Amos (1994); Almasy and Blangero (1998); Abecasis,

Received November 2016; revised March 2017.

¹Supported by NIH Grants R01HG009124, R01GM126553, and NSF Grant DMS-17-12933.

Key words and phrases. Genome-wide association studies, summary statistics, variance component, linear mixed model, MINQUE, method of moments.

Cardon and Cookson (2000); Diao and Lin (2005); Visscher, Hill and Wray (2008)], for association analysis to control for individual relatedness and population stratification [Yu et al. (2006); Kang et al. (2008, 2010); Zhang et al. (2010); Lippert et al. (2011); Zhou and Stephens (2012, 2014); Pirinen, Donnelly and Spencer (2013); Yang et al. (2014); Loh et al. (2015a)], for genomic selection and risk prediction by jointly modeling genome-wide SNPs [Robinson (1991); Hofer (1998); Whittaker, Thompson and Denham (2000); Hayes, Visscher and Goddard (2009); Makowsky et al. (2011); Zhou, Carbonetto and Stephens (2013); Wray et al. (2013)], and for rare variant association tests by grouping individually weak effects to improve power [Wu et al. (2009)]. More recently, with growing interest, LMMs have been applied to estimate the proportion of phenotypic variance explained by available SNPs [Yang et al. (2010); Speed et al. (2012); Zhou, Carbonetto and Stephens (2013); Wray et al. (2013); de Los Campos, Sorensen and Gianola (2015)]—a quantity often referred to as SNP heritability—and to partition the SNP heritability by different chromosome segments or by different functional genomic annotations [Yang et al. (2011a); Kostem and Eskin (2013); Gusev et al. (2014); Finucane et al. (2015); Loh et al. (2015b)]. These applications all require accurate estimation of variance components in LMMs. Here, we will describe a new method for variance component estimation, with main applications for SNP heritability estimation and partition in population studies as well as side applications for heritability estimation in family studies.

The standard method for variance component estimation is the restricted maximum likelihood estimation (REML) method. REML method is a form of maximum likelihood estimation that obtains the variance component estimates by maximizing the restricted likelihood function, a function that is derived from the likelihood function by removing the effects of nuisance parameters. REML is statistically efficient. However, REML suffers from several important statistical and computational drawbacks. Perhaps the most important drawback of REML is that it requires individual-level genotypes and phenotypes from all samples in the study. Because of consent and privacy concerns, as well as logistic limitations (e.g., large-scale data transfer and storage often require high-end computing infrastructure), it is becoming increasingly difficult to access complete individual-level data from large-scale association studies. Indeed, sharing summary statistics (e.g., marginal z -scores) across multiple studies, performing meta-analysis, and releasing results in terms of summary statistics has become a standard practice in most consortium studies [Allen et al. (2010); Speliotes et al. (1974); Teslovich et al. (2010); Manning et al. (2012); Jostins et al. (2012)]. Requiring complete individual-level data thus restricts the use of REML and limits the benefits of LMMs in many large-scale studies. In addition to its use of individual-level data, REML is also computationally slow. Despite many recent computational innovations [Thompson and Shaw (1990); Kang et al. (2008); Lippert et al. (2011); Zhou and Stephens (2012);

Pirinen, Donnelly and Spencer (2013); Loh et al. (2015a, 2015b)], it still can be challenging to apply REML to large data sets. For example, it can take weeks to analyze tens of thousands of individuals and tens of millions of SNPs with the commonly used GCTA software [Yang et al. (2011a)]. Finally, REML relies on the normality assumption of residual errors and is not robust to model misspecification. In particular, in ascertained case control studies or studies with an extreme sample design, REML underestimates SNP heritability [Chen (2014); Golan, Lander and Rosseta (2014)].

To remedy these drawbacks of REML, we present a new, alternative method for variance component estimation with summary statistics. Our method is based on a long existing alternative to REML for variance component estimation, the minimal norm quadratic unbiased estimation (MINQUE) method, a method of moments (MoM) [Rao (1970, 1971)]. The MoM estimation method has been used in animal breeding programs [Zhu and Weir (1996)], but its popularity has faded away since the development of the statistically more efficient REML method. However, as we will show here, with modifications, MoM can be used for variance component estimation with summary statistics and is computationally much more efficient than REML. To realize the benefits of MoM and adapt MINQUE to association studies with summary statistics, we rely on a set of simple second moment matching equations of MINQUE and develop two additional approximations. Our first approximation allows us to make use of summary statistics in terms of marginal z -scores and pair them with the individual relatedness matrices (or SNP correlation matrices) to obtain unbiased estimates at the price of a reduction of statistical efficiency (with respect to REML). Our second approximation allows us to use only a small random subset of individuals to compute the genetic relatedness matrices for this subset of individuals that can be further used to estimate the average relatedness among individuals (or, equivalently, to estimate the average SNP correlation across genome-wide SNPs). This subsampling strategy greatly improves computational efficiency, but does not reduce much further the statistical efficiency (with respect to the MoM estimates obtained using full data). Importantly, our framework unifies two seemingly unrelated methods—the renowned Haseman–Elston (HE) regression [Haseman and Elston (1972); Drigalenko (1998); Elston et al. (2000); Sham and Purcell (2001); Sham et al. (2002); Chen, Broman and Liang (2004); Chen (2014)] and the recent LD score regression (LDSC) [Bulik-Sullivan et al. (2015a, 2015b); Finucane et al. (2015)]—into the same umbrella. We refer to our method as MQS (MinQue for Summary statistics), and we illustrate its benefits in SNP heritability and heritability estimation with simulations and real data applications.

We organize the paper as follows. We provide a brief description of the method in Section 2, with methodological details provided in the Supplementary Material [Zhou (2017)]. In Section 3, we present comparisons between our method and several other variance component estimation methods with simulations. In Section 4,

we describe applications of MQS to 37 phenotypes from 8 real GWAS data, including both individual-level data and data with only summary statistics. We conclude the paper with a summary and discussion in Section 5.

2. Method overview. Our method applies to the following LMMs that can be used to partition SNP heritability into k different nonoverlapping categories:

$$(2.1) \quad \mathbf{y} = \sum_{i=1}^k \mathbf{g}_i + \boldsymbol{\varepsilon}, \quad \mathbf{g}_i \sim \text{MVN}(0, \sigma_i^2 \mathbf{K}_i), \quad \boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_{k+1}^2 \mathbf{M}),$$

where \mathbf{y} is an n -vector of phenotypes for n individuals; \mathbf{g}_i is an n -vector of random effects representing the combined genetic effects of SNPs in the i th category; $\mathbf{K}_i = \mathbf{X}_i \mathbf{X}_i^T / p_i$ is an n by n genetic relatedness matrix computed from the n by p_i genotype matrix for p_i SNPs in the i th category; $\sigma^2 = (\sigma_1^2, \dots, \sigma_k^2)$ are the corresponding variance components; $\boldsymbol{\varepsilon}$ is an n -vector of residual errors; σ_{k+1}^2 is the residual error variance; $\mathbf{M} = \mathbf{I} - \mathbf{1}_n \mathbf{1}_n^T / n$ is a projection matrix; and MVN denotes a multivariate normal distribution. Both \mathbf{y} and every column of \mathbf{X} have been centered to have mean zero, allowing us to ignore the intercept and use \mathbf{M} instead of the usual identity matrix \mathbf{I} to constrain the errors to have mean zero. Note that the number of categories, k , is known *a priori* and depends on the particular application. We denote the phenotype variance $s_y^2 = \mathbf{y}^T \mathbf{y} / (n - 1)$. We also define the scaled version of the variance components as $\mathbf{h}^2 = (h_1^2, \dots, h_k^2) = (\sigma_1^2 / s_y^2, \dots, \sigma_k^2 / s_y^2) = \sigma^2 / s_y^2$ and $h_{k+1}^2 = \sigma_{k+1}^2 / s_y^2$. \mathbf{h}^2 represents the proportion of phenotypic variance explained by all SNPs in each category, the estimation of which requires variance component estimates $\hat{\boldsymbol{\sigma}}^2$.

Our method for variance component estimation, which we refer to as MINQUE for summary statistics (MQS), is described in detail in the Supplementary Material. Briefly, MQS is based on a set of second moment matching equations determined by the MINQUE criterion [equation (3)], and has a simple, closed-form solution for estimating the k variance components: $\hat{\boldsymbol{\sigma}}^2 = \mathbf{S}^{-1} \mathbf{q}$, with a k -vector \mathbf{q} and a k by k matrix \mathbf{S} [equation (11)]. Intuitively, the i th element of \mathbf{q} measures the proportion of variance in phenotypes explained (PVE) by SNPs in the i th category when SNPs are independent, while the ij th element of \mathbf{S} accounts for the linkage disequilibrium (LD) between SNPs in the i th and j th categories (and the ii th element of \mathbf{S} accounts for LD within the i th category). \mathbf{q} is computed with the marginal z -scores obtained using all individuals. \mathbf{S} can be computed using the p by p SNP correlation matrices. However, because the relatedness among individuals contains the same amount of information for computing \mathbf{S} as the SNP correlations across genome-wide SNPs, we use the usual n by n genetic relatedness matrices for all individuals to compute \mathbf{S} efficiently. Importantly, computation of \mathbf{S} only requires knowing the average relatedness in the data instead of the detailed pairwise relatedness values. Therefore, as we will show below, \mathbf{S} can also be estimated using

the genetic relatedness matrices for a subset of individuals or for individuals from a reference panel. In addition to \mathbf{q} and \mathbf{S} , MQS requires a set of prespecified SNP weights that are used for both \mathbf{q} and \mathbf{S} . This set of prespecified SNP weights is used in MQS to approximate the optimal MINQUE estimating equations. Different choices of weights represent different ways of approximation and can lead to unbiased estimates with different levels of statistical efficiency.

MQS is a unified framework because different SNP weighting options lead to different estimation methods. We consider two particular weighting options here. The first option is equal SNP weights [equation (9)]. We refer to the variation of MQS under this weighting option MQS-HEW. MQS-HEW is mathematically equivalent to the renowned Haseman–Elston (HE) cross-product regression [Haseman and Elston (1972); Drigalenko (1998); Elston et al. (2000); Sham and Purcell (2001); Sham et al. (2002); Chen, Broman and Liang (2004); Chen (2014)]. However, our particular MQS formulation allows us to both make use of summary statistics and develop an asymptotic form to compute the standard errors. The second weighting option assigns SNP weights as a function of both *a priori* set of variance components and LD scores [equation (10)], where the LD score of a variant is defined as the summation of the r-squared between itself and all SNPs genome-wide. We refer to the variation of MQS under this weighting option MQS-LDW. For $k = 1$, MQS-LDW is equivalent to a special form of LDSC [Bulik-Sullivan et al. (2015a, 2015)] that sets the intercept to be exactly one and that effectively computes LD scores using all SNPs genome-wide. However, our MQS formulation not only provides an asymptotic form to compute the standard errors, but also is capable of measuring SNP correlations among all genome-wide SNPs in a computationally efficient fashion, thus alleviating much of the estimation bias encountered in LDSC (see sections below).

One key feature of MQS is that we can use a random subset of individuals to obtain an estimate of \mathbf{S} without reducing much of the statistical efficiency of the variance component estimates σ^2 with respect to MQS estimation using the full data. The subsampling strategy can be used for variance component estimation with MQS because \mathbf{S} measures only the average SNP correlation or, equivalently, the average individual relatedness in the data, both of which can be estimated by a random subset of samples. In addition, because \mathbf{S} is much easier to be estimated accurately than \mathbf{q} , using estimated \mathbf{S} often leads to only a small reduction in estimation accuracy. Specifically, with subsampling, the variance of the MQS estimates, $\hat{\sigma}^2 = \hat{\mathbf{S}}^{-1}\mathbf{q}$, can be decomposed into two parts: one that is due to $V(\mathbf{q})$ [as a result of $V(\mathbf{y})$] and the other that is due to $V(\hat{\mathbf{S}})$ because of subsampling. Because \mathbf{S} is a quantity computed by averaging across all SNP pairs while \mathbf{q} is a quantity obtained by averaging only across all SNPs, \mathbf{S} is estimated effectively with a sample size that is p' times larger than that for \mathbf{q} , where p' is the effective number of independent SNPs. Therefore, the extra variance due to estimating \mathbf{S} via subsampling— $V(\hat{\mathbf{S}})$ —can be much smaller compared with $V(\mathbf{q})$. As a result,

we can use a much smaller number of individuals m to estimate \mathbf{S} without losing much statistical efficiency. Besides this intuitive explanation, we provide more formal arguments for the subsampling strategy in the Supplementary Material and Figures S1–S3. Certainly, our computation of standard errors still explicitly accounts for the uncertainty introduced by using a much smaller set of individuals to estimate \mathbf{S} instead of computing it. In addition, because the effectiveness of the subsampling strategy depends on the effective number of independent SNPs, the strategy is expected to work better in population studies than in family studies with related individuals and an extended linkage disequilibrium pattern.

As a by-product of the subsampling strategy, estimating \mathbf{S} with a smaller subsample instead of computing \mathbf{S} with the full data also allows us to apply MQS to data from many consortium studies: there, we can pair \mathbf{q} computed from the available marginal z -scores in the consortium study with $\hat{\mathbf{S}}$ estimated from a random subsample of the study using either the relatedness matrices for the subsample or individual-level genotype data for them. When such a random subsample of the study is not available, we can also use individual-level genotype data from a publicly available reference panel, such as the 1000 genome project [The 1000 Genomes Project Consortium (2012)], which also contains a much smaller number of samples compared with the study, to estimate \mathbf{S} , as long as individuals in the reference panel can be viewed as a subsample of the study (e.g., of the same ethnic origin).

Finally, for testing the null hypothesis $H_0 : \sigma^2 = 0$, one could use an exact p -value computation method based on a mixture of chi-square distributions [e.g., based on equation (13)]. Indeed, as we have shown in a separate study that focuses on a different variance component model application, using exact p -value computation methods is the only viable solution to obtain calibrated p -values for genome-wide applications with millions of tests [Crawford et al. (2017)]. However, as our main focus is on SNP heritability estimation or heritability estimation, both of which often involve only one or a few tests [Yang et al. (2010, 2011b); Zhou, Carbonetto and Stephens (2013)], we consider here a simple normal test based on the point estimate and its standard error. We will show in the results section below that the simple normal test provides acceptable type I error control at the usual significance levels (e.g., $\alpha = 0.05$). For convenience, MQS provides two forms to compute the standard errors. The first form is based on asymptotics, and, besides \mathbf{q} and \mathbf{S} , requires either the genetic relatedness matrices from the full data [equations (15), (18), and (20)] or additional summary statistics besides the marginal z -scores [equation (25)]. The second form is based on the blockwise jackknife resampling procedure proposed in LDSC [Bulik-Sullivan et al. (2015a)], and requires only marginal z -scores besides \mathbf{S} . However, the jackknife option assumes blockwise SNP independence and may not yield calibrated standard errors when the LD pattern is complicated. Examples where the jackknife may not apply include ascertained case control studies [Hayes et al. (2005); Zaykin, Meng and

Ehm (2006)], admixture populations [Price et al. (2008)], and related individuals, which are defined loosely as individuals who are not far away in time from their most recent common ancestor [Speed and Balding (2015)].

We summarize the key features of MQS along with several other variance component estimation methods in Table 1. Because of the subsampling strategy, MQS is efficient in terms of computation and memory usage compared with a range of commonly used methods (Figure 1).

3. Simulations. We perform simulations to compare the performance of several different methods on variance component estimation. We use two real genotype data sets for simulation: an Australian data with $n = 3925$ individuals and $p = 4,352,968$ imputed SNPs [Yang et al. (2010)], and a Finnish data with $n = 5123$ individuals and $p = 319,148$ genotyped SNPs [Sabatti et al. (2008)]. We choose these two data sets not only because both consist of white individuals of European ancestry, but also because the two differ in LD pattern: the Finland data displays longer LD than the Australia data (Figure S4). The Finland data displays a long LD pattern presumably because individuals from the Finland data are more closely related to each other than individuals from the Australia data; however, the Finland study is not a family study. The long LD pattern in the Finland data makes it easy to validate some of our expectations. For each data set, the real genotypes are used to compute genetic relatedness matrices, with which we simulate phenotypes based on LMMs (details in Supplementary Material). Environment effects are simulated from independent normal distributions that do not depend on genetic relatedness; thus no population stratification is present in the simulations.

We compare six different methods: (1) REML that uses individual-level phenotypes and genotypes; (2) HE regression that uses individual-level phenotypes and genotypes; (3) LDSC that uses z -scores computed from the full data and LD scores estimated based on the LDSC-recommended 1 MB window size from individual-level genotypes of the full data; (4) LDSC that uses LD scores estimated based on a 10 MB window size instead; (5) MQS-HEW that uses z -scores computed from the full data and \hat{S} estimated from individual-level genotypes of $m = 400$ randomly selected individuals; and (6) MQS-LDW that uses z -scores computed from the full data, LD scores estimated based on 1 MB window from genotypes of $m = 400$ randomly selected individuals, and \hat{S} estimated from genotypes of the same subsample. We evaluate and compare different methods based on their statistical properties including unbiasedness and statistical efficiency.

We first simulate phenotypes under LMMs with $k = 1$ (Supplementary Material). We check three scenarios: $h^2 = 0, 0.25, \text{ or } 0.5$; note that most quantitative traits have an SNP heritability below 0.5 [Furlotte, Heckerman and Lippert (2014)]. For each scenario, we perform 1000 replicates. For MQS-HEW and MQS-LDW, a different set of $m = 400$ individuals are used in each simulation replicate.

TABLE 1

Computational complexity and memory usage of different methods for variance component estimation. The computational complexity includes time to compute the genetic relatedness matrices. REML relies on maximizing the restricted likelihood while MoM is based on moments matching. AI: average information method; NR: Newton Raphson's algorithm; MC: Monte Carlo algorithm; HE: Haseman-Elston regression; LDSC: LD score regression. n is the number of individuals; m is the number of randomly selected subset of individuals ($m < n$); k is the number of variance components; p is the number of genetic markers; w is the average number of variants used to estimate the LD scores; c is the number of jackknife samples for estimating the standard error; t is the number of iterations used in estimation: $t = 1$ for MQS-HEW, $t = 2$ for MQS-LDW, and $t > 2$ for LDSC and REML; (a): asymptotic; (j): jackknife

Type	Methods	Computational complexity (O)		Memory usage (O)	Summary statistics	Example software	Selected references
		Point estimate	Standard error				
REML	NR-AI	$pn^2 + t(n^3 + k^2n^2)$	(a): k^2n^2	kn^2	No	GCTA, GEMMA	Gilmour, Thompson and Cullis (1995) Yang et al. (2011c) Zhou and Stephens (2012)
	MC-AI	$tpn^{1.5}$	(a): k^2n^2	pn	No	BOLT-REML	García-Cortés et al. (1992) Matilainen et al. (2012) Loh et al. (2015b)
MoM	HE	$pn^2 + k^2n^2$	(j): ck^2n^2	kn^2	No	PCGC	Haseman and Elston (1972) Golan, Lander and Rosseta (2014)
	HE	$pn^2 + k^2n^2$	(a): k^2n^2	kn^2	No	GEMMA	this work
	LDSC	$wpn + tkp$	(j): ckp	pn	Yes	LDSC	Bulik-Sullivan et al. (2015a) Bulik-Sullivan et al. (2015b) Finucane et al. (2015)
	MQS	$pn + t(pm^2 + k^2m^2)$	(a): kpn ; (j): cp	km^2	Yes	GEMMA	this work

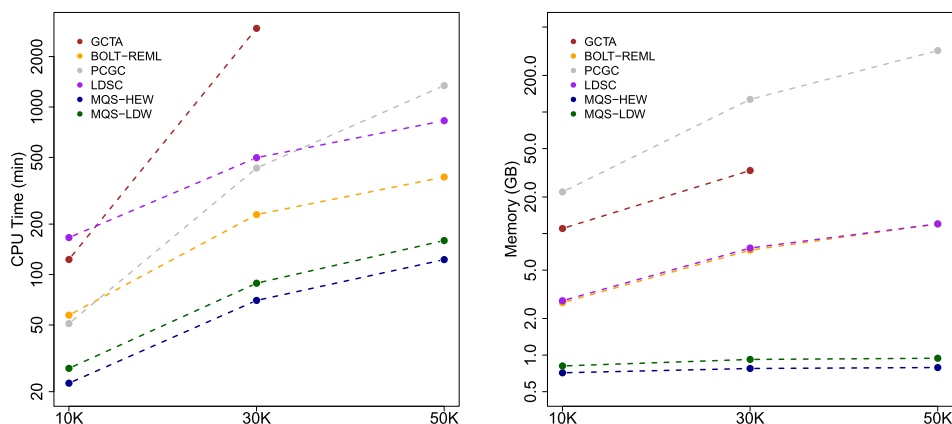


FIG. 1. Comparison of CPU time (left) and memory usage (right) for commonly used variance component estimation software for data with 10,000, 30,000, or 50,000 individuals and 1 million SNPs. Computation is performed on a single core of an Intel Xeon CPU E5-2683. Software compared include GCTA (brown), BOLT-REML (orange), PCGC (grey), LDSC (purple), MQS-HEW (blue), and MQS-LDW (green). PCGC uses 100 jackknife samples to compute the standard errors; its computing time is linear with the increase in number of jackknife samples. LDSC uses a neighbor window of 1000 SNPs to compute LD scores. The standard errors in MQS-HEW and MQS-LDW are computed based on the asymptotic form; using the blockwise jackknife will make MQS-HEW and MQS-LDW two times faster. MQS-HEW and MQS-LDW use $m = 400, 1200, 2000$ for the three sample sizes in the data. The x-axis shows the number of samples in the data. The y-axis is on a log scale.

We obtain the variance component estimates from different methods. The left two panels of Figure 2 show boxplots of these estimates. The right two panels of Figures 2 show the inverse of the estimated relative statistical efficiency of the estimates, which is estimated by contrasting the mean squared error (MSE) of the estimates from one method to the MSE of the REML estimates; a higher relative MSE indicates lower statistical efficiency compared with REML.

The results fit our expectations:

First, because MQS-HEW with \mathbf{S} is identical to HE and because $\hat{\mathbf{S}}$ is an accurate estimate of \mathbf{S} , estimates from MQS-HEW with $\hat{\mathbf{S}}$ are similar to those from HE (Figure S5). In fact, the statistical efficiency loss of using MQS-HEW instead of HE (i.e., using $m = 400$ individuals to estimate \mathbf{S} instead of computing it with the full data) is estimated to be only 0.75%/1.4%/2.5% in the Australia data and 0.31%/2.2%/4.0% in the Finland data for $h^2 = 0/0.25/0.5$ scenarios, respectively. In addition, because the variance of the MQS estimates depends on a product of $V(\hat{\mathbf{S}})$ and the heritability parameter h^2 , MQS-HEW estimates are closer to HE estimates with smaller h^2 .

Second, LDSC uses LD scores to quantify SNP correlations (unlike MQS, which uses \mathbf{S} to quantify correlations). However, LDSC uses only neighborhood SNPs instead of all genome-wide SNPs to estimate LD scores through a sliding

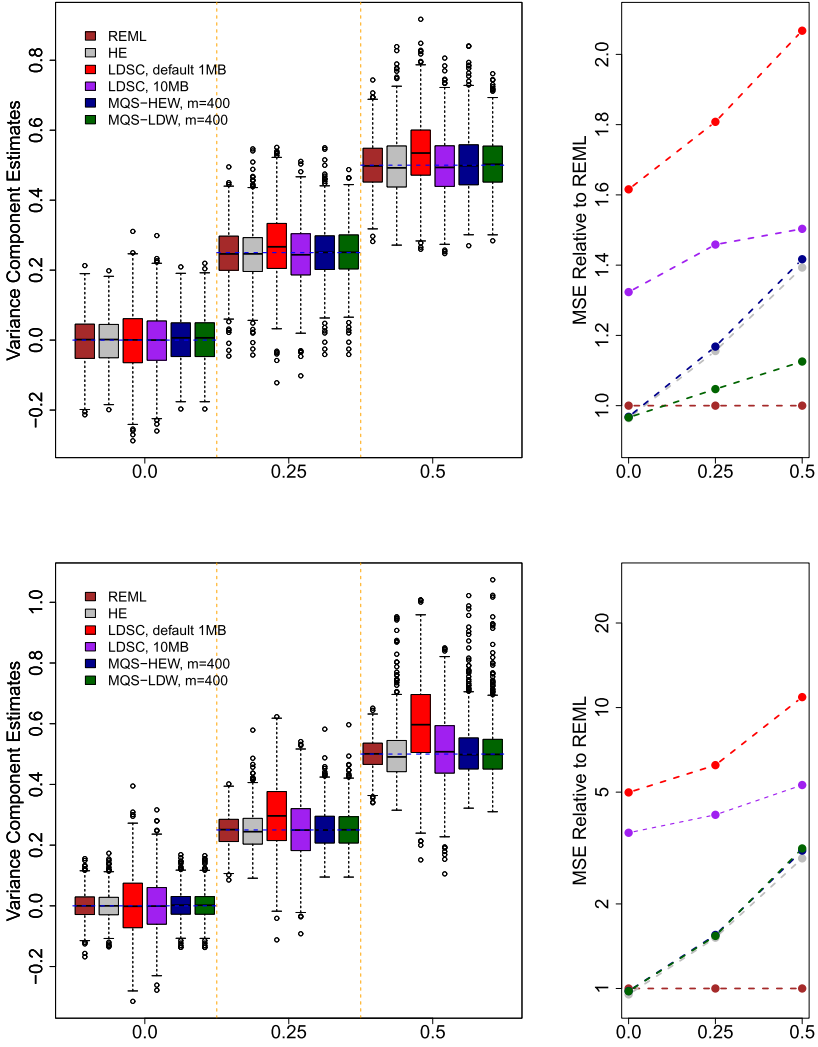


FIG. 2. Comparison of variance component estimates from REML (brown), HE (grey), LDSC (red and purple), MQS-HEW (blue), and MQS-LDW (green) for $k = 1$ simulations based on the Australian data (top panels) or the Finland data (bottom panels). LDSC estimates are obtained using either the default 1 MB window (red) or an extended 10 MB window (purple). The left two panels show boxplots. The true variance components (0, 0.25, and 0.5) are shown as blue horizontal lines. The right two panels show the mean squared error (MSE) relative to REML. MSE relative to REML measures the statistical efficiency of REML with respect to other methods; a higher relative MSE thus indicates low statistical efficiency. The x-axis shows the true variance components. The y-axis in the bottom right panel is on a log scale.

window-based approach. As a consequence, the SNP correlations are underestimated and the variance component estimates from LDSC are upward biased: they are on average 7.5% higher than the truth in the Australia data, and are 20.0% higher in the Finland data with longer LD. Such upward bias can be mitigated in the two data by using a larger sliding window size. However, for any given data, it is unclear how large a window size one should choose *a priori* to mitigate the bias in LDSC. Therefore, our MQS-LDW formulation represents a much attractive alternative to LDSC, as it uses \mathbf{S} to quantify genome-wide SNP correlations. In addition to the bias, we find that LDSC is noticeably much less statistically efficient than the other methods, even when its bias is largely mitigated by the larger sliding window size. The statistical inefficiency of LDSC presumably stems from the fact that LDSC does not set its intercept to be exactly one as should be under the LMM assumption.

Third, MQS-HEW/HE and MQS-LDW approximate MINQUE in different ways, but both approximations are only accurate when the variance component is small and/or when individuals are unrelated. Thus both MQS-HEW/HE and MQS-LDW are statistically more efficient for a small h^2 than for a large h^2 , and more efficient in the Australia data than in the Finland data. In addition, in the case of $h^2 = 0$, both MQS-HEW/HE and MQS-LDW are statistically more efficient than REML because they effectively assume $h^2 = 0$ *a priori*. In other cases, REML is statistically the most efficient method since it is based on maximizing (the restricted) likelihood.

Fourth, presumably because MQS-LDW uses estimates from MQS-HEW as initial values, and presumably because MQS-LDW uses the extra information of LD scores to compute the SNP weights, MQS-LDW is statistically more efficient than MQS-HEW in the Australia data. However, because long LD reduces the accuracy of LD score estimates, MQS-LDW is as efficient as MQS-HEW in the Finland data.

Besides comparing point estimates, we also examine the type I error control by different methods for testing the null hypothesis $H_0 : \sigma^2 = 0$. While we caution that exact p -value computation should be performed based on a mixture of chi-square distributions, we find that a simple normal test based on the point estimate and its standard error computed from the asymptotic form do provide acceptable, though slightly inflated, type I error control at the significance levels of 0.01–0.10 (Table S2). The standard error computed from the jackknife procedure is reasonable in the Australia data but causes inflated type I error in the Finland data where a longer LD pattern is observed.

Next, we perform simulations using LMMs with $k = 6$ (details in Supplementary Material). The results are shown in Figures S6–S8 and are similar to the case of $k = 1$ as described above. We also examine the robustness of MQS with respect to model misspecifications. The equivalence between HE and MQS-HEW guarantees the MQS-HEW estimates to be unbiased for case control studies [Chen

(2014); Golan, Lander and Rosseta (2014)]. To validate this and check if MQS-LDW is equally unbiased in case control studies, we perform a set of ascertained case control simulations. We simulate 2500 cases, 2500 controls and 10,000 independent causal SNPs from a liability threshold model with $h^2 = 0.5$ and a disease prevalence of 0.1%, following exactly the early study [Golan, Lander and Rosseta (2014)]. We either use the 10,000 causal SNPs directly to form a nonsparse simulation scenario or pair them with 90,000 noncausal SNPs to form a sparse simulation scenario. We then estimate heritability on the observed scale with different methods and transform the estimates back to the liability scale [Lee et al. (2011); Zhou, Carbonetto and Stephens (2013); Golan, Lander and Rosseta (2014)]. Because of the small number of simulated SNPs, we use genome-wide SNPs to compute the LD scores. The comparison results are shown in Figure S9. HE, MQS-HEW, and MQS-LDW all produce unbiased and statistically efficient estimates. In contrast, REML is downward biased, while LDSC displays huge variance. Finally, we also explore sparse simulation scenarios where only a small proportion of SNPs are causal for continuous trait simulations. To do so, we randomly select 100 or 1000 SNPs from the Finland data to be causal and we simulate their effects to explain a fixed $h^2 = 0.5$ (details in Supplementary Material). The comparison results are shown in Figure S10. Consistent with the above continuous trait simulations as well as previous studies that have validated the robustness of LMM methods in various genetic architectures [Speed et al. (2012); Zhou, Carbonetto and Stephens (2013)], REML, HE, MQS-HEW, and MQS-LDW are all statistically efficient and robust with respect to model misspecifications. LDSC again displays large estimation variance.

While we have mainly focused on SNP heritability estimation in population studies, we also explore the applicability of our methods for heritability estimation in family studies. To do so, we obtain genotype data from the Framingham heart study (FHS) with $n = 6850$ related individuals and $p = 394,174$ SNPs [Splansky et al., (2007)]. As above, we compute the relatedness matrix, simulate phenotypes under LMMs with $k = 1$, and examine three scenarios with $h^2 = 0, 0.25, \text{ or } 0.5$ (details in Supplementary Material). Point estimates and relative MSE are shown in Figure S11. The results are largely consistent with the $k = 1$ simulations described above. Both MQS-HEW and MQS-LDW estimates are approximately unbiased. The LDSC estimates display upward bias with large variance, and such bias cannot be mitigated by using the large 10 MB window size. Different from the population-based simulation studies, however, due to strong individual relatedness and subsequently strong LD in the family data, the subsampling strategy becomes less effective compared with the Australia- and Finland-based simulations (as we explained earlier in the Method Overview). Indeed, both MQS-HEW and MQS-LDW estimates here are considerably less efficient than the HE estimates, especially for large h^2 . However, both MQS estimates are still substantially more accurate than LDSC. We also present the type I error results of various methods from the FHS-based simulations in Table S2. The results there again are consistent

with early simulations: while the asymptotic form provides calibrated type I error control, the jackknife procedure leads to inflated type I error due to strong LD in the data.

4. Real data applications. To obtain further insights into the differences between various methods, we apply all five methods to estimate SNP heritability for 18 phenotypes in three human GWAS data sets. The first GWAS data is the Australian data that contains height measurements for Australian. The second GWAS data is the Finland data that contains 10 quantitative traits, including body mass index (BMI), C-reactive protein (CRP), glucose, insulin, high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides (TG), total cholesterol (TC), systolic blood pressure (SysBP), and diastolic blood pressure (DiaBP). The third GWAS data is the WTCCC data [[The Wellcome Trust Case Control Consortium \(2007\)](#)], which includes about 14,000 cases from 7 common diseases and about 3000 shared controls, all typed on a common set of 458,868 SNPs. The 7 common diseases are bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). We apply the five methods to these data in the same way as described in the simulations. For LDSC, we use a 10 MB sliding window, as we have seen in simulations that longer than default window size is necessary to mitigate the bias of LDSC estimates in population studies.

The SNP heritability estimates are presented in Table 2. For case control studies, we present estimates on the observed scale, which can be easily converted to the liability scale if the disease prevalence in the population is known [[Lee et al. \(2011\)](#); [Zhou, Carbonetto and Stephens \(2013\)](#); [Golan, Lander and Rosseta \(2014\)](#)]. For example, when the disease prevalence is 0.5% in the population and the case proportion is 50% in the case control study, then the scaling factor is 0.47. Because the scaling factor is always smaller than one, the observed scale heritability estimates in case control studies can be bigger than one, allowing the liability scale heritability estimates not to be bounded by the small scaling factor. To ensure unbiasedness, we do not constrain variance components to be positive during estimation [[Price et al. \(2011\)](#)]. Thus a low heritable trait may have its SNP heritability estimated to be below zero. The heritability estimates in the Finland data are largely consistent with a previous study [[Browning and Browning \(2013\)](#)]. The estimates in the WTCCC are also consistent with a previous study [[Golan, Lander and Rosseta \(2014\)](#)]. The heritability estimate for height in the Australia data is slightly smaller than that from previous studies [[Yang et al. \(2010\)](#); [Zhou, Carbonetto and Stephens \(2013\)](#)], a phenomenon observed with imputed data elsewhere [[Gusev et al. \(2013\)](#)]. Table 2 also lists the genomic control factors and the intercept estimates from LDSC. Both values are close to one for all traits, suggesting limited population stratification in the three data sets. To further reduce the influence of population stratification, we remove the top two principal components (PCs) or the

top ten PCs and present the results with PCs removed in Tables S3 and S4. Results are similar with or without PCs removed. Because PCs come from the genetic relatedness matrix, we do caution that over-correcting population stratification by removing too many PCs can reduce SNP heritability—a phenomenon recognized elsewhere [Finucane et al. (2015)].

The real data results are consistent with the simulations. First, MQS-HEW estimates and the standard errors based on the asymptotic form are almost identical to that of HE, consistent with the accuracy of the subsampling strategy. Second, the estimates from LDSC are consistently different from estimates of other methods, and often come with a much larger standard error—again consistent with its statistical inefficiency in the simulations. Third, consistent with its known downward bias in case control studies [Chen (2014); Golan, Lander and Rosseta (2014)], REML estimates are smaller than HE or MQS estimates in the seven diseases. Fourth, also consistent with simulations, the jackknife procedure in MQS often produces overly narrow standard errors when compared with the asymptotic form (Table S5). However, for two disease phenotypes (RA and T1D), the standard errors from jackknife are extremely large, suggesting that the calibration issue with jackknife may not always favor one direction.

Our method requires individual-level genotypes from a random subsample of the study to estimate \mathbf{S} . When such a subset of individuals is not available, we can use a reference panel to estimate \mathbf{S} , as long as individuals in the reference panel can be viewed as a subsample of the study. However, a mismatch between the reference panel and the study sample can cause estimation bias. In addition, using a separate reference panel prevents us from using the asymptotic form to compute the standard errors. Here, we explore the use of genotype data from the 1000 genomes project [The 1000 Genomes Project Consortium (2012)] for SNP heritability estimation in the three GWASs. Specifically, instead of using 400 randomly selected individual from the study sample, we use 503 individuals of European ancestry from the 1000 genomes project to estimate \mathbf{S} . The SNP heritability estimates for all traits from the three data sets are shown in Table S5. For both the Australia and WTCCC data, using the 1000 genomes data as a reference panel produces similar results. However, for the Finland data, the estimates from using the 1000 genomes data are much larger, suggesting a potential overestimation. The results suggest that a match between the reference panel and study sample is critical for accurate estimation. In addition, because we can only use the jackknife to compute the standard errors, the standard errors suffer from the same drawback as detailed in the previous paragraph.

Next, we explore the use of our methods in estimating heritability in family studies using the FHS data set [Splansky et al., (2007)]. We perform analysis on four blood lipid traits that include HDL ($n = 6850$), LDL ($n = 6855$), TC ($n = 3806$), and TG ($n = 3806$). Table 2 shows the heritability estimates from different methods for these traits. The results are largely consistent with both simulations and

TABLE 2

*SNP heritability or heritability estimates from different methods for 15 quantitative traits and 7 binary phenotypes from four GWASs. FHS is a family study, while the rest are population studies. Values in parentheses are standard errors. For MQS-HEW and MQS-LDW, the standard errors are computed by the asymptotic form. The standard errors computed by the blockwise jackknife are available in a supplementary table. The heritability estimates for the WTCCC data are presented at the observed scale. A small scaling factor is required to transform the estimates to a liability scale. To ensure unbiasedness, we do not constrain variance components to be positive during estimation. Thus a low heritable trait may have its SNP heritability estimated to be below zero. λ_{GC} is the genomic control factor, while *icpt* is the intercept estimate from LDSC. Note that both the LDSC intercept estimates and the genomic control factors in the FHS data are larger than one, indicating strong relatedness in this data*

Trait	Methods						λ_{GC}
	REML	HE	LDSC, 10 MB	MQS (<i>m</i> = 400; asymptotic)		<i>icpt</i>	
				HEW	LDW		
Australia, <i>n</i> = 3925, <i>p</i> = 4,352,968							
Height	0.27 (0.072)	0.25 (0.072)	0.21 (0.089)	0.26 (0.072)	0.28 (0.072)	1.008	1.027
Finland, <i>n</i> = 5123, <i>p</i> = 319,148							
BMI	0.20 (0.047)	0.18 (0.053)	0.15 (0.11)	0.19 (0.053)	0.19 (0.054)	1.012	1.026
CRP	0.043 (0.049)	0.034 (0.045)	0.15 (0.094)	0.037 (0.045)	0.037 (0.045)	0.996	0.996
DiaBP	0.071 (0.046)	0.073 (0.049)	0.047 (0.074)	0.075 (0.049)	0.076 (0.049)	1.007	1.012
Glucose	0.17 (0.046)	0.21 (0.069)	0.25 (0.057)	0.21 (0.070)	0.21 (0.068)	1.016	1.038
HDL	0.34 (0.043)	0.36 (0.071)	0.51 (0.12)	0.36 (0.072)	0.35 (0.068)	0.987	1.051
Insulin	-0.063 (0.037)	-0.12 (0.057)	-0.10 (0.048)	-0.082 (0.044)	-0.081 (0.044)	0.992	0.986
LDL	0.38 (0.041)	0.60 (0.13)	0.20 (0.11)	0.61 (0.13)	0.61 (0.13)	1.080	1.110
SysBP	0.19 (0.045)	0.27 (0.087)	0.13 (0.092)	0.27 (0.088)	0.28 (0.090)	1.030	1.057
TC	0.29 (0.043)	0.42 (0.097)	0.15 (0.12)	0.42 (0.098)	0.43 (0.099)	1.055	1.068
TG	0.19 (0.047)	0.17 (0.053)	0.15 (0.12)	0.18 (0.053)	0.17 (0.053)	1.010	1.034

TABLE 2
(Continued)

Trait	Methods						λ_{GC}
	REML	HE	LDSC, 10 MB	MQS ($m = 400$; asymptotic)		icpt	
				HEW	LDW		
WTCCC, $n = \sim 5000$, $p = 458,868$							
BD	0.83 (0.057)	1.05 (0.095)	0.44 (0.11)	1.08 (0.098)	1.16 (0.10)	1.079	1.115
CAD	0.57 (0.061)	0.58 (0.071)	0.19 (0.11)	0.60 (0.073)	0.63 (0.074)	1.049	1.073
CD	0.71 (0.060)	1.06 (0.15)	0.42 (0.14)	1.08 (0.16)	1.12 (0.17)	1.080	1.113
HT	0.58 (0.060)	0.62 (0.072)	0.13 (0.098)	0.63 (0.072)	0.66 (0.073)	1.061	1.067
RA	0.69 (0.059)	0.81 (0.083)	0.50 (0.34)	0.84 (0.085)	0.83 (0.083)	1.041	1.040
T1D	0.97 (0.052)	1.53 (0.15)	1.15 (0.88)	1.60 (0.15)	1.41 (0.11)	1.031	1.056
T2D	0.61 (0.060)	0.69 (0.082)	0.21 (0.11)	0.71 (0.084)	0.76 (0.084)	1.061	1.082
FHS, $n = 3806-6855$, $p = 372,131$							
HDL	0.43 (0.024)	0.41 (0.041)	0.15 (0.095)	0.40 (0.081)	0.40 (0.082)	1.394	1.423
LDL	0.42 (0.024)	0.50 (0.10)	0.16 (0.11)	0.58 (0.16)	0.58 (0.16)	1.488	1.514
TC	0.42 (0.037)	0.43 (0.068)	0.17 (0.15)	0.54 (0.11)	0.55 (0.11)	1.258	1.288
TG	0.34 (0.036)	0.38 (0.052)	0.17 (0.14)	0.47 (0.084)	0.47 (0.084)	1.225	1.239

the above real data applications. In particular, due to individual relatedness and subsequently extended LD structure, the MQS-HEW and MQS-LDW estimates are almost identical to each other, and both are close to the HE estimates. Consistent with the simulations, the standard errors of MQS-HEW and MQS-LDW are larger than that of HE, suggesting that the subsampling strategy in a family study can introduce extra estimation variance and are less effective than in a population study. In addition, the estimates from LDSC are consistently different from estimates of other methods, and again come with large standard errors. Table S5 also shows the standard errors of MQS estimates from the jackknife as well as estimates and standard errors of MQS from using the 503 individuals of European ancestry from the 1000 genomes project. Consistent with the results described in the previous paragraph, the results in Table S5 again illustrate calibration issues with the jackknife and overestimation phenomenon when the LD pattern (or, equivalently, individual relatedness) in the reference panel is weaker than that in the study.

Finally, we apply MQS-HEW and MQS-LDW methods to analyze 8 phenotypes from four consortium studies. These phenotypes include BMI ($n = 120,569$), height (HT, $n = 129,945$) from the GIANT consortium [Allen et al. (2010); Speliotes et al. (1974)], HDL ($n = 88,754$), LDL ($n = 84,685$), TC ($n = 89,005$) and TG ($n = 85,691$) from the Global Lipids Genetics Consortium [Teslovich et al. (2010)], fasting glucose (FG, $n = 58,074$) from the MAGIC consortium [Manning et al. (2012)], and Crohn's disease (CD, $n = 21,447$) from the International Inflammatory Bowel Disease Genetics Consortium [Jostins et al. (2012)]. The data have been preprocessed by a previous study [Pickrell (2014)]. We further select a common set of $p = 5,014,740$ SNPs among these phenotypes for analysis. We partition SNPs into the same six functional categories (coding, UTR, promoter, DHS, intronic, and else) as before [Gusev et al. (2014)]. Because only z -scores are available for these phenotypes, we use the jackknife to compute the standard errors and use genotypes from 503 individuals of European ancestry in the 1000 genomes project [The 1000 Genomes Project Consortium (2012)] as a reference panel to estimate \mathbf{S} . In addition, following previous approaches [Finucane et al. (2015), Loh et al. (2015b)], to contrast the importance of different categories, we focus on estimating the relative value instead of the absolute value of variance components. Specifically, as in [Finucane et al. (2015)], we construct a fold enrichment parameter, defined as the ratio between the per-SNP variance in one category and the per-SNP variance in all categories, to quantify the relative importance of different functional categories (details in Supplementary Material).

Figure 3 shows the enrichment parameters for six categories in 8 phenotypes estimated by either MQS-HEW or MQS-LDW. The results from both MQS-HEW and MQS-LDW are consistent with what we expect [Finucane et al. (2015)]: for most phenotypes (with the notable exception of BMI), the per-SNP explained variance in the coding region is the largest, followed by the UTR, promoter, and the

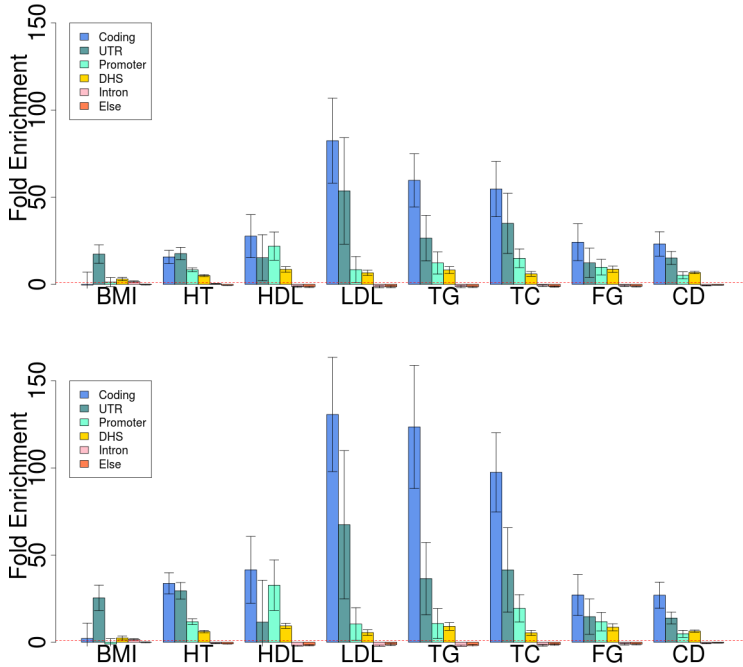


FIG. 3. *MQS-HEW* (top) and *MQS-LDW* (bottom) reveal the importance of six functional categories in 8 phenotypes from four GWAS data sets. The y-axis shows the fold enrichment, computed as a ratio between the average SNP effect size in one category and the average SNP effect size across the whole genome. Both *MQS-HEW* and *MQS-LDW* use marginal z-scores together with genotypes of 503 individuals with European ancestry from the 1000 genomes project. The asymptotic form is used to construct the confidence intervals.

DNS regions. The per-SNP variance for both the intronic and intergenic regions are close to zero. The enrichment estimates between *MQS-HEW* and *MQS-LDW* are similar overall, though the enrichment of the coding region is estimated to be larger in *MQS-LDW* than in *MQS-HEW* for the lipid phenotypes.

5. Discussion. We have presented a statistical method, *MQS*, for variance component estimation with summary statistics. *MQS* produces unbiased estimates and is computationally efficient for large data. *MQS* is also flexible: it can model the effect size dependency on minor allele frequencies [Zhou, Carbonetto and Stephens (2013)]; it can incorporate overlapping SNP functional annotations; and it can control for other covariates such as genotype PCs (details in Supplementary Material). *MQS* can be used in pair with other methods to model uneven LD [Speed et al. (2012), Yang et al. (2015)], and can be extended to model multiple correlated phenotypes [Zhou and Stephens (2014); Bulik-Sullivan et al. (2015b)]. With simulations and applications to 37 phenotypes from 8 GWASs, we have shown the benefits of our method.

In the present study, we have focused on two variations of MQS that are distinguished from each other in using different SNP weighting options. Both variations, MQS-HEW and MQS-LDW, use \mathbf{q} and \mathbf{S} for computation and yield unbiased estimates but with different statistical efficiency. Although we cannot tell in advance which method is statistically more efficient for a particular data set, simulations suggest that MQS-LDW is statistically more efficient than MQS-HEW in unrelated individuals. The superior statistical efficiency of MQS-LDW presumably stems from the fact that MQS-LDW uses in the SNP weights the extra LD score information and relies on *a priori* set of variance component estimates from MQS-HEW. However, because LD scores are necessarily estimated via a sliding window-based approach, the LD scores become inaccurate in related individuals. As a consequence, in related individuals, the weights used in MQS-LDW become almost indistinguishable from the equal SNP weights used in MQS-HEW, leading to comparable performance between the two methods. In addition, MQS-LDW can be computationally inconvenient: because MQS-LDW requires an iterative procedure, computing its standard errors based on the asymptotic form also requires recomputing summary statistics from the study sample. Therefore, we recommend the use of MQS-LDW for unrelated individuals due to its statistical efficiency, and MQS-HEW for related individuals due to its convenience and its comparable performance with MQS-LDW there. Importantly, our framework paves ways for future extensions of the two basic variations. Specifically, our derivation of MQS-LDW suggests that using other SNP weighting matrices, in particular, nondiagonal ones, holds the promise of more statistically efficient estimates. Extending MQS and exploring the use of other weighing matrices is an interesting avenue for future research.

MQS uses a small random subset of individuals to estimate \mathbf{S} . In population studies, using $\hat{\mathbf{S}}$ instead of \mathbf{S} reduces much of the computational cost while yielding estimates that are almost as statistically efficient as if the full data were used. For instance, in both simulations and real data applications on SNP heritability estimation, we have used $\sim 10\%$ of the data to estimate \mathbf{S} . Using $\sim 10\%$ of the data incurs minimal loss of statistical efficiency (a few percent in simulations) but results in an effective ~ 100 -fold speed gain (because computational complexity scales with m^2). Our subsampling approach of using $\hat{\mathbf{S}}$ instead of \mathbf{S} is motivated by recent genetic studies that make use of a reference panel for genotype imputation [Browning (2006); Guan and Stephens (2008); Wen and Stephens (2010); Howie et al. (2012)] and, more recently, for multi-loci analysis [Yang et al. (2012); Bulik-Sullivan et al. (2015a)]: when the full data is not completely observed, these studies rely on a reference panel to impute the missing pieces to construct a complete data. Our approach, however, differs from the previous approaches in two important ways: we actively use a subset of data to estimate certain quantities even when the full data is completely observed [i.e., in line with the idea of stochastic approximation method as in Robbins and Monro (1951)], and we account for

the extra uncertainty introduced by using a smaller subset of data. Importantly, in the present study, we provide an initial set of statistical reasoning to justify our subsampling approach in MQS. However, even with our guidelines, it often remains difficult to choose the right number of subsamples, m , for practical analysis. A large m would not save much computational time, while a small m could be insufficient to produce unbiased estimates or introduce substantial estimation variance. In practice, the optimal choice of m will likely depend on both the number of categories k and the effective number of independent SNPs in each category (which in turn depends on individual relatedness); thus we caution against the use of an m that is too small when k is large and/or when individuals are highly related. Indeed, as we have shown with the family-based simulations and applications, using a small m for related individuals can introduce considerable estimation variance. Therefore, we recommend in practice examining the estimates and the standard errors for a range of m values to choose an m that can balance computation and accuracy. In addition, while we have only explored a simple random subsampling strategy in the present study, we note that other more sophisticated sampling methods can be easily paired with MQS to further improve estimation efficiency. Overall, we believe the subsampling strategy allows MQS to achieve an appealing balance between computational efficiency and statistical efficiency. With increasing data sizes, exploring the benefits of subsampling strategy in other statistical methods for large-scale GWASs—as well as other big data applications—is likely to yield fruitful results in the future.

Acknowledgments. We thank Nick Martin and the Queensland Institute of Medical Research for making the Australia height data available to us. We thank Joseph K. Pickrell for making the summary data from 8 quantitative human traits available to us. We thank the NFBC1966 Study Investigators for making the Finland NFBC1966 data available to us. The NFBC1966 study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the Broad Institute, University of California Los Angeles, University of Oulu, and the National Institute for Health and Welfare in Finland. This manuscript was not prepared in collaboration with investigators of the NFBC1966 study and does not necessarily reflect their views or those of their host institutions. This study also makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113 and 085475. This research was conducted in part using data and resources from the Framingham Heart Study of the NHLBI and Boston University School of Medicine, which was partially supported by the NHLBI Framingham Heart Study (Contract No. N01-HC-25195) and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). We thank all participants and staff from the Framingham Heart Study. We thank Chaolong Wang, William Wen, Ping Zeng, and Xiang Zhu for helpful

comments on a previous version of the manuscript. We thank the Associate Editor and the two anonymous reviewers whose comments have greatly improved the quality of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary Material (DOI: [10.1214/17-AOAS1052SUPP](https://doi.org/10.1214/17-AOAS1052SUPP); .pdf). Supplementary figures, tables and text.

REFERENCES

- ABECASIS, G. R., CARDON, L. R. and COOKSON, W. O. (2000). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66** 279–292.
- ALLEN, H. L., ESTRADA, K., LETTRE, G., BERNDT, S. I., WEEDON, M. N., RIVADENEIRA, F., WILLER, C. J., JACKSON, A. U., VEDANTAM, S. et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467** 832–838.
- ALMASY, L. and BLANGERO, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62** 1198–1211.
- AMOS, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54** 535–543.
- BROWNING, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* **78** 903–913.
- BROWNING, S. R. and BROWNING, B. L. (2013). Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Hum. Genet.* **132** 129–138.
- BULIK-SULLIVAN, B. (2015). Relationship between LD score and Haseman–Elston regression. *BioRxiv* **0** 018283.
- BULIK-SULLIVAN, B. K., LOH, P.-R., FINUCANE, H. K., RIPKE, S., YANG, J., SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM, PATTERSON, N., DALY, M. J., PRICE, A. L. and NEALE, B. M. (2015a). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47** 291–295.
- BULIK-SULLIVAN, B., FINUCANE, H. K., ANTILA, V., GUSEV, A., DAY, F. R., LOH, P.-R., REPROGEN CONSORTIUM, PSYCHIATRIC GENOMICS CONSORTIUM, GENETIC CONSORTIUM FOR ANOREXIA NERVOSA OF THE WELLCOME TRUST CASE CONTROL CONSORTIUM 3 et al. (2015b). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47** 1236–1241.
- CHEN, G.-B. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman–Elston regression. *Front. Genet.* **5** 107.
- CHEN, W.-M., BROMAN, K. W. and LIANG, K.-Y. (2004). Quantitative trait linkage analysis by generalized estimating equations: Unification of variance components and Haseman–Elston regression. *Genet. Epidemiol.* **26** 265–272.
- CRAWFORD, L., ZENG, P., MUKHERJEE, S. and ZHOU, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *BioRxiv*.
- DE LOS CAMPOS, G., SORENSEN, D. and GIANOLA, D. (2015). Genomic heritability: What is it? *PLoS Genet.* **11** e1005048.
- DIAO, G. and LIN, D. Y. (2005). A powerful and robust method for mapping quantitative trait loci in general pedigrees. *Am. J. Hum. Genet.* **77** 97–111.
- DRIGALENKO, E. (1998). How sib-pairs reveal linkage. *Am. J. Hum. Genet.* **63** 1242–1245.
- ELSTON, R. C., BUXBAUM, S., JACOBS, K. B. and OLSON, J. M. (2000). Haseman and Elston revisited. *Genet. Epidemiol.* **19** 1–17.

- FINUCANE, H. K., BULIK-SULLIVAN, B., GUSEV, A., TRYNKA, G., RESHEF, Y., LOH, P.-R., ANTILLA, V., XU, H., ZANG, C. et al. (2015). Partitioning heritability by functional category using GWAS summary statistics. *Nat. Genet.* **47** 1228–1235.
- FURLOTTE, N. A., HECKERMAN, D. and LIPPERT, C. (2014). Quantifying the uncertainty in heritability. *J. Hum. Genet.* **59** 269–275.
- GARCÍA-CORTÉS, L. A., MORENO, C., VARONA, L. and ALTARRIBA, J. (1992). Variance component estimation by resampling. *J. Anim. Breed. Genet.* **109** 358–363.
- GILMOUR, A. R., THOMPSON, R. and CULLIS, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51** 1440–1450.
- GOLAN, D., LANDER, E. S. and ROSSETA, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA* **111** E5272–E5281.
- GUAN, Y. and STEPHENS, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genet.* **4** e1000279.
- GUSEV, A., BHATIA, G., ZAITLEN, N., VILHJALMSSON, B. J., DIOGO, D., STAHL, E. A., GREGERSEN, P. K., WORTHINGTON, J., KLARESKOG, L. et al. (2013). Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9** e1003993.
- GUSEV, A., LEE, S. H., TRYNKA, G., FINUCANE, H., VILHJALMSSON, B. J., XU, H., ZANG, C., RIPKE, S., BULIK-SULLIVAN, B. et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **5** 535–552.
- HASEMAN, J. K. and ELSTON, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2** 3–19.
- HAYES, B. J., VISSCHER, P. M. and GODDARD, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb.)* **91** 47–60.
- HAYES, M. G., DEL BOSQUE-PLATA, L., TSUCHIYA, T., HANIS, C. L., BELL, G. I. and COX, N. J. (2005). Patterns of linkage disequilibrium in the type 2 diabetes gene calpain-10. *Diabetes* **54** 3573–3576.
- HOFER, A. (1998). Variance component estimation in animal breeding: A review. *J. Anim. Breed. Genet.* **115** 247–265.
- HOWIE, B., FUCHSBERGER, C., STEPHENS, M., MARCHINI, J. and ABECASIS, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44** 955–959.
- JOSTINS, L., RIPKE, S., WEERSMA, R. K., DUERR, R. H., MCGOVERN, D. P., HUI, K. Y., LEE, J. C., SCHUMM, L. P., SHARMA, Y. et al. (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491** 119–124.
- KANG, H. M., ZAITLEN, N. A., WADE, C. M., KIRBY, A., HECKERMAN, D., DALY, M. J. and ESKIN, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178** 1709–1723.
- KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S.-Y., FREIMER, N. B., SABATTI, C. and ESKIN, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42** 348–354.
- KOSTEM, E. and ESKIN, E. (2013). Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *Am. J. Hum. Genet.* **92** 558–564.
- LEE, S. H., WRAY, N. R., GODDARD, M. E. and VISSCHER, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88** 294–305.
- LIPPERT, C., LISTGARTEN, J., LIU, Y., KADIE, C. M., DAVIDSON, R. I. and HECKERMAN, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8** 833–835.
- LOH, P.-R., TUCKER, G., BULIK-SULLIVAN, B. K., VILHJALMSSON, B. J., FINUCANE, H. K., CHASMAN, D. I., RIDKER, P. M., NEALE, B. M., BERGER, B. et al. (2015a). Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nat. Genet.* **47** 284–290.

- LOH, P.-R., BHATIA, G., GUSEV, A., FINUCANE, H. K., BULIK-SULLIVAN, B. K., POLLACK, S. J., SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM, DE CANDIA, T. R., LEE, S. H. et al. (2015b). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47** 1385–1392.
- MAKOWSKY, R., PAJEWSKI, N. M., KLIMENTIDIS, Y. C., VAZQUEZ, A. I., DUARTE, C. W., ALLISON, D. B. and DE LOS CAMPOS, G. (2011). Beyond missing heritability: Prediction of complex traits. *PLoS Genet.* **7** e1002051.
- MANNING, A. K., HIVERT, M.-F., SCOTT, R. A., GRIMSBY, J. L., BOUATIA-NAJI, N., CHEN, H., RYBIN, D., LIU, C.-T., BIELAK, L. F. et al. (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44** 659–669.
- MATILAINEN, K., MÄNTYSAARI, E. A., LIDAUER, M. H., STRANDÉN, I. and THOMPSON, R. (2012). Employing a Monte Carlo algorithm in expectation maximization restricted maximum likelihood estimation of the linear mixed model. *J. Anim. Breed. Genet.* **129** 457–468.
- PICKRELL, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94** 559–573.
- PIRINEN, M., DONNELLY, P. and SPENCER, C. C. A. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7** 369–390. [MR3086423](#)
- PRICE, A. L., WEALE, M. E., PATTERSON, N., MYERS, S. R., NEED, A. C., SHIANNA, K. V., GE, D., ROTTER, J. I., TORRES, E. et al. (2008). Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **1** 132–135.
- PRICE, A. L., HELGASON, A., THORLEIFSSON, G., MCCARROLL, S. A., KONG, A. and STEFANSSON, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7** e1001317.
- RAO, C. R. (1970). Estimation of heteroscedastic variances in linear models. *J. Amer. Statist. Assoc.* **65** 161–172. [MR0286221](#)
- RAO, C. R. (1971). Estimation of variance and covariance components—MINQUE theory. *J. Multivariate Anal.* **1** 257–275. [MR0301869](#)
- ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. [MR0042668](#)
- ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6** 15–51. [MR1108815](#)
- SABATTI, C., SERVICE, S. K., HARTIKAINEN, A.-L., POUTA, A., RIPATTI, S., BRODSKY, J., JONES, C. G., ZAITLEN, N. A., VARILO, T. et al. (2008). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41** 35–46.
- SHAM, P. C. and PURCELL, S. (2001). Equivalence between Haseman–Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.* **68** 1527–1532.
- SHAM, P. C., PURCELL, S., CHERNY, S. S. and ABECASIS, G. R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am. J. Hum. Genet.* **71** 238–253.
- SPEED, D. and BALDING, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nat. Rev. Genet.* **16** 33–33.
- SPEED, D., HEMANI, G., JOHNSON, M. R. and BALDING, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91** 1011–1021.
- SPELIOTES, E. K., WILLER, C. J., BERNDT, S. I., MONDA, K. L., THORLEIFSSON, G., JACKSON, A. U., ALLEN, H. L., LINDGREN, C. M., LUAN, J. et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42** 937–948.
- SPLANSKY, G. L., COREY, D., YANG, Q., ATWOOD, L. D., CUPPLES, L. A., BENJAMIN, E. J., D’AGOSTINO, R. B., FOX, C. S., LARSON, M. G. et al. (2007). The third generation cohort of

- the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, recruitment, and initial examination. *Am. J. Epidemiol.* **165** 1328–1335.
- TESLOVICH, T. M., MUSUNURU, K., SMITH, A. V., EDMONDSON, A. C., STYLIANOU, I. M., KOSEKI, M., PIRRUCCELLO, J. P., RIPATTI, S., CHASMAN, D. I. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466** 707–713.
- THE 1000 GENOMES PROJECT CONSORTIUM (2012). An integrated map of genetic variation from 1092 human genomes. *Nature* **491** 56–65.
- THE WELLCOME TRUST CASE CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447** 661–678.
- THOMPSON, E. A. and SHAW, R. G. (1990). Pedigree analysis for quantitative traits: Variance components without matrix inversion. *Biometrics* **46** 399–413.
- VISSCHER, P. M., HILL, W. G. and WRAY, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9** 255–266.
- WEN, X. and STEPHENS, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.* **4** 1158–1182. [MR2751337](#)
- WHITTAKER, J. C., THOMPSON, R. and DENHAM, M. (2000). Marker-assisted selection using ridge regression. *Genet. Res.* **75** 249–252.
- WRAY, N. R., YANG, J., HAYES, B. J., PRICE, A. L., GODDARD, M. E. and VISSCHER, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14** 507–515.
- WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.
- YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.
- YANG, J., MANOLIO, T. A., PASQUALE, L. R., BOERWINKLE, E., CAPORASO, N., CUNNINGHAM, J. M., DE ANDRADE, M., FEENSTRA, B., FEINGOLD, E. et al. (2011a). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43** 519–525.
- YANG, J., WEEDON, M. N., PURCELL, S., LETTRE, G., ESTRADA, K., WILLER, C. J., SMITH, A. V., INGELSSON, E., O'CONNELL, J. R. et al. (2011b). Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19** 807–812.
- YANG, J., LEE, S. H., GODDARD, M. E. and VISSCHER, P. M. (2011c). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88** 76–82.
- YANG, J., FERREIRA, T., MORRIS, A. P., MEDLAND, S. E., GENETIC INVESTIGATION OF ANTHROPOMETRIC TRAITS (GIANT) CONSORTIUM, DIABETES GENETICS REPLICATION AND META-ANALYSIS (DIAGRAM) CONSORTIUM, MADDEN, P. A. F., HEATH, A. C., MARTIN, N. G. et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44** 369–375.
- YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. and PRICE, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46** 100–106.
- YANG, J., BAKSHI, A., ZHU, Z., HEMANI, G., VINKHUYZEN, A. A. E., LEE, S. H., ROBINSON, M. R., PERRY, J. R. B., NOLTE, I. M. et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47** 1114–1120.
- YU, J., PRESSOIR, G., BRIGGS, W. H., BI, I. V., YAMASAKI, M., DOEBLEY, J. F., McMULLEN, M. D., GAUT, B. S., NIELSEN, D. M. et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38** 203–208.
- ZAYKIN, D. V., MENG, Z. and EHM, M. G. (2006). Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am. J. Hum. Genet.* **78** 737–746.

- ZHANG, Z., ERSOZ, E., LAI, C.-Q., TODHUNTER, R. J., TIWARI, H. K., GORE, M. A., BRADBURY, P. J., YU, J., ARNETT, D. K. et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42** 355–360.
- ZHOU, X. (2017). Supplement to “A unified framework for variance component estimation with summary statistics in genome-wide association studies.” DOI:[10.1214/17-AOAS1052SUPP](https://doi.org/10.1214/17-AOAS1052SUPP).
- ZHOU, X., CARBONETTO, P. and STEPHENS, M. (2013). Polygenic modelling with Bayesian sparse linear mixed models. *PLoS Genet.* **9** e1003264.
- ZHOU, X. and STEPHENS, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44** 821–824.
- ZHOU, X. and STEPHENS, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11** 407–409.
- ZHU, J. and WEIR, B. (1996). Mixed model approaches for diallele analysis based on a bio-model. *Genet. Res.* **68** 233–240.

DEPARTMENT OF BIOSTATISTICS
CENTER FOR STATISTICAL GENETICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: xzhousph@umich.edu