

LEVERAGING LOCAL IDENTITY-BY-DESCENT INCREASES THE POWER OF CASE/CONTROL GWAS WITH RELATED INDIVIDUALS

BY JOSHUA N SAMPSON* BILL WHEELER† PENG LI* JIANXIN SHI*

Division of Cancer Epidemiology and Genetics, National Cancer Institute
** and Information Management Services †*

Large case/control genome-wide association studies (GWAS) often include groups of related individuals with known relationships. When testing for associations at a given locus, current methods incorporate only the familial relationships between individuals. Here, we introduce the chromosome-based Quasi Likelihood Score (cQLS) statistic that incorporates local Identity-By-Descent (IBD) to increase the power to detect associations. In studies robust to population stratification, such as those with case/control sibling pairs, simulations show that the study power can be increased by over 50%. In our example, a GWAS examining late-onset Alzheimers disease, the p-values among the most strongly associated SNPs in the APOE gene tend to decrease, with the smallest p-value decreasing from 1.23×10^{-8} to 7.70×10^{-9} . Furthermore, as a part of our simulations, we reevaluate our expectations about the use of families in GWAS. We show that, although adding only half as many unique chromosomes, genotyping affected siblings is more efficient than genotyping randomly ascertained cases. We also show that genotyping cases with a family history of disease will be less beneficial when searching for SNPs with smaller effect sizes.

1. Introduction. Genome-Wide Association Studies (GWAS) of binary traits can include related individuals from known pedigrees (Barrett et al., 2008; Willer et al., 2008). In case-control studies, GWAS generally increase power by collecting cases with affected relatives, and may therefore find it cost-effective to genotype these related cases as well. Furthermore, GWAS may target families because the appropriate association tests are robust to population stratification (Ewens, Li and Spielman, 2008), the families have previously been collected for linkage analyses, the genetic variants can be called more accurately (Wang et al., 2007), or the effects of parental imprinting can be evaluated (Wilkinson, Davies and Isles, 2007).

There are two general approaches for evaluating GWAS with related individuals from known pedigrees (Ott, Kamatani and Lathrop, 2011; Manichaikul

Keywords and phrases: cQLS, GWAS, related individuals, case-control

et al., 2012). The first approach combines two independent tests, a family based test (Sham et al., 2002; Lange et al., 2003; Laird, Horvath and Xu, 2000), such as the transmission-disequilibrium test, and a population-based test of association. Methods exist for combining these two tests so the resulting analysis is robust to population stratification (Zheng et al., 2010; Won et al., 2012; Mirea et al., 2012; Manichaikul et al., 2012). The second approach, which potentially sacrifices robustness for improved power, calculates a single test statistic of association that accounts for the correlated genotypes among relatives (Bourgain et al., 2003; Thornton and McPeck, 2007; Slager and Schaid, 2001). Our focus is on the second approach and we derive a single test statistic which can use principal components (Price et al., 2006) to correct for population stratification when necessary.

Until now, standard GWAS analyses have focused on the individual. For each individual, a study records their genotype and disease status. If all individuals were unrelated, the appropriate test statistic would simply be the correlation between an individual's genotype and disease status (Hirschhorn and Daly, 2005). However, when some individuals are related, a test statistic should correct for the dependence among genotypes and allow for the minor allele frequency (MAF) at a causal SNP to be higher in controls with affected relatives, as compared to randomly ascertained controls (Thornton and McPeck, 2007; Bourgain et al., 2003; Zhu and Xiong, 2012). The more-powerful Quasi-Likelihood Score (mQLS) statistic (Thornton and McPeck, 2007), an extension of the QLS (Bourgain et al., 2003), accomplishes both objectives. However, because these statistics consider only the overall Identity By Descent (IBD) status between two individuals, they cannot allow for the MAF at a causal SNP to be higher in controls that share both alleles (IBD=2) with an affected sibling, as compared to controls that share no alleles (IBD=0) with an affected sibling.

We introduce a GWAS analysis that takes a different perspective and focuses on the founder chromosomes within each family. For each founder chromosome, we effectively identify its allele (at a given SNP) and the proportion of individuals carrying the SNP from that chromosome who are affected. This step is made possible by recent advances in IBD mapping and haplotyping (Browning and Browning, 2010; Peters et al., 2012; He, 2013). We suggest a chromosome-based Quasi Likelihood Score (cQLS) statistic that, at its simplest, measures the correlation between a binary indicator for the minor allele and the proportion of individuals affected. Formally, this test statistic is a partial score statistic from the retrospective likelihood that includes local IBD status among the observed data. The cQLS effectively leverages local IBD to improve power in GWAS with a large number

of family-based controls. Furthermore, the derivation of the cQLS as a score statistic shows how to appropriately handle families of arbitrary pedigrees, include phenotype data from ungenotyped relatives, accommodate covariates, and allow for arbitrary models (e.g. logistic, liability threshold) linking disease risk and genotype status. In addition to these useful features, the cQLS also permits permutation-based measures of statistical significance. Accounting for relatedness in permutation methods of subject-centric approaches has proven exceptionally difficult (Wang, 2011).

In the next section, we define cQLS, describe the simulated datasets for testing its performance, and introduce a GWAS of Late-Onset Alzheimer’s Disease (LOAD) conducted by the National Institute of Aging and the National Cell Repository for Alzheimer’s Disease (NIA-LOAD/NCRAD) (Lee et al., 2008; Wijsman et al., 2011). In the third section, we evaluate the performance of the cQLS in simulated datasets and the NIA-LOAD/NCRAD dataset. In addition to showing the potential benefit of cQLS, simulations evaluate the power gained by genotyping affected siblings, as compared to randomly ascertained cases, and demonstrate the diminished benefit of recruiting cases with a family history of disease when searching for SNPs with small effects (Ionita-Laza and Ottman, 2011; Hattersley and McCarthy, 8; Teng and Risch, 1999). In the final section, we conclude with a brief discussion.

2. Methods.

2.1. *cQLS: Definition.* We consider a case-control study that contains N_{Fam} families, labeled $j = 1, \dots, N_{Fam}$. Within family j , all genotyped chromosomes are assumed to arise from a family-specific set of n_j founder chromosomes, labeled $k = 1, \dots, n_j$. We denote the total number of chromosomes by N_T

$$(2.1) \quad N_T = \sum_j n_j$$

For a given SNP, we let $Y_{jk} = 1$ if the unique founder chromosome k in family j has a minor allele, and $Y_{jk} = 0$ otherwise. Note that the subscript “ j_1k ” and “ j_2k ” refer to different founders chromosomes from different families. For our discussion here, we will assume that Y_{jk} is uniquely identifiable given the observed genetic data and that we can identify, with certainty, those individuals in family j who inherited the SNP from founder chromosome k . Both assumptions will be relaxed in the appendix. Furthermore, we will let $T_{jik'k} = 1$ if the k'^{th} copy ($k' \in \{1, 2\}$) of the specified SNP in individual i

is descended from founder chromosome k , and $T_{jik'k} = 0$ otherwise. We let $A_{ji} = 1$ if individual i is affected, and $A_{ji} = 0$ otherwise.

For each founder SNP, we have Y_{jk} , indicating the presence or absence of a minor allele, and $C_{jk} = \sum_{i,k'} T_{jik'k} A_{ji}$, a count of the number of affected individuals with that SNP. If we expect that the minor allele increases disease risk, then the presence of a minor allele should be associated with a larger number of affected individuals. Therefore, the squared correlation between Y_{jk} and C_{jk} should be high. The proposed test statistic is a variation of this squared correlation.

Instead of using the count, C_{jk} , we use a version that normalizes each individual's affection status to their expected affection status under the null hypothesis of no association

$$(2.2) \quad Z_{jk} = \sum_i \sum_{k'} T_{jik'k} \left(A_{ji} - \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} (1 - A_{ji}) \right)$$

where $\hat{\pi}_0$ is our estimate of the prevalence of the disease in the population. We presume there are established estimates of $\hat{\pi}_0$ available in the literature. In the appendix, we show how we can modify the expected affection status when other characteristics are known. We further show how we can make specific individuals, such as those with a family history of disease, carry more weight in the analysis by effectively increasing their contribution to Z_{jk} . We denote the average of all observed values of Z_{jk} by

$$(2.3) \quad \bar{Z} \equiv \frac{\sum_{j,k} Z_{jk}}{N_T}$$

Finally, we define a normalized version of the allele for chromosome k ,

$$(2.4) \quad Y_{jk}^\dagger = Y_{jk} - \hat{\phi}$$

where

$$(2.5) \quad \hat{\phi} \equiv \frac{\sum_{j,k} Y_{jk}}{N_T}$$

and is our estimate of ϕ , the minor allele frequency in the population, under the null hypothesis. As promised, our chromosome-based Quasi-Likelihood Score (cQLS) statistic is then proportional to the squared correlation between Y^\dagger and Z

$$(2.6) \quad cQLS \equiv \frac{(\sum_{j,k} (Z_{jk} - \bar{Z}) Y_{jk}^\dagger)^2}{\sum_{j,k} (Z_{jk} - \bar{Z})^2 \hat{\phi} (1 - \hat{\phi})}$$

Because $\text{var}(Y_{jk}^\dagger)$ is calculated as $\hat{\phi}(1-\hat{\phi})$, equation 2.6 requires the assumption of Hardy-Weinberg Equilibrium (HWE). Therefore, the version of cQLS defined by equation 2.6 is appropriate in an ideal scenario where the genotypes and phenotypes of all individuals are known, Y_{jk} can be identified for all families and chromosomes, and HWE holds. In the appendix, we define a more robust cQLS that allows for violations in all three assumptions.

2.2. Assigning Chromosomes. To calculate cQLS, we need to determine $T_{jik'k}$, or the identity of the two chromosomes in individual i for all i, j, k' , and k at each SNP. We perform this calculation in three steps. First, we phase subjects using BEAGLE (Browning and Browning, 2011), a software package for analysis of large-scale genetic data sets. Second, we detect shared segments within a family using GERMLINE (Gusev et al., 2009), a software package for discovering long shared segments of Identity-By-Descent (IBD). In the third step, we convert the IBD status to $T_{jik'k}$. The algorithm for this final step is described in Appendix B.4.

2.3. Simulations. Our three aims are to (1) assess the benefit of chromosome-based association tests (2) assess the value of genotyping an affected sibling and (3) assess the benefit of genotyping cases known to have affected siblings.

For our simulations, we assume that the liability threshold model accurately describes disease risk. In the liability threshold model, a complex disease results from an underlying, normally distributed, phenotype, or liability. When an individual's liability exceeds a specific threshold, the individual is affected. In our scenario, the liability, \mathbf{L}_j , for individuals in nuclear family j can be described by a linear function of their genotypes \mathbf{G}_j , where G_{ji} is the number of minor alleles for individual i , $\mathbf{G}_j = [G_{j1} \dots G_{jN_j}]^t$, and the superscript t indicates transpose. Here, \mathbf{E}_j accounts for environmental factors, while \mathbf{F}_j accounts for background genetic correlation.

$$(2.7) \quad \mathbf{L}_j = \beta_G(\mathbf{G}_j - 2\phi) + \mathbf{E}_j + \beta_F\mathbf{F}_j$$

where

$$(2.8) \quad \begin{bmatrix} E_{j1} \\ E_{j2} \\ \vdots \\ E_{jN_j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \right)$$

and

$$(2.9) \quad \begin{bmatrix} F_{j1} \\ F_{j2} \\ F_{j3} \\ \vdots \\ F_{jN_j} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0.5 & \cdots & 0.5 \\ 0 & 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 0.5 & 1 & \cdots & 0.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & 0.5 & \cdots & 1 \end{bmatrix} \right)$$

Here, N_j is the number of individuals in family j and individuals $j1$ and $j2$ are the parents. The liability threshold model will assign all individuals with L_{ji} exceeding the population's 95th percentile to have the disease. We chose β_G so that the corresponding OR for each additional minor allele was between 1.02 and 1.36. We chose β_F so that the sibling recurrence risk ratio, λ_S , was 1.5, 2 or 5 when $\beta_G = 0$. Note that the magnitude of the sibling recurrence risk ratio, as defined here, is independent of the strength of the tested SNP. To calculate β_F , we numerically solved the following equation

$$(2.10) \quad \int_{F_2^{-1}(0.95)}^{\infty} \left[\int_{F_2^{-1}(0.95)}^{\infty} f_1(x|y) \partial x \right] f_2(y) \partial y = 0.05 \lambda_S$$

where F_2 and f_2 are the cumulative distribution and density of $N(0, 1 + \beta_F^2)$, and $f_1(x|y)$ is the density of $N(\beta_F^2 y / (2(1 + \beta_F^2)), (1 + 2\beta_F^2 + 1.75\beta_F^4) / (1 + \beta_F^2))$. The MAF in the population was fixed at 0.1, and genotypes were simulated under HWE.

Our simulated studies, summarized in Table 1, collect families and/or unrelated cases and controls. The first five studies all start by identifying and genotyping 10000 unrelated, randomly ascertained controls (R.A. controls). By randomly ascertained, we mean that we have no knowledge of their family history of disease. Study 1 further genotypes 5000 unrelated, randomly ascertained cases (R.A. cases). Study 2 genotypes 5000 unrelated cases (F.H. cases) with a family history of disease, specifically with an affected sibling. Study 3 genotypes both the 5000 R.A. cases and the 5000 F.H. cases. Study 4 genotypes 10000 R.A. cases and 5000 F.H. cases. Studies 5 and 6 genotype 1 sibling of each proband. Study 5 genotypes the 5000 R.A cases, 5000 F.H. cases and their affected siblings. Study 6 includes 10000 unrelated cases with an unaffected sibling, and their unaffected siblings. Study 7 genotypes multiple relatives of each proband. Specifically, study 7 genotypes the 10000 F.H. cases, their affected siblings, and two unaffected siblings. Studies 6 and 7 highlight the potential power gain for cQLS in scenarios where only family-based controls are available.

For various combinations of β_G , λ_S , and study design, we simulated 10000 datasets and calculated the FBAT test statistic (Laird, Horvath and Xu,

2000), mQLS(Bourgain et al., 2003; Thornton and McPeck, 2007), and cQLS for each dataset. The power was defined as the proportion of p-values that were below the 10^{-7} threshold, a common threshold for GWAS. By using 10000 datasets, the standard errors for our estimates of power are bounded by 0.01. A short description of FBAT and mQLS are provided in Appendix A. Previous comparisons (Manichaikul et al., 2012) of the available test statistics have demonstrated that, among currently available options, mQLS is consistently a top performer, and the appropriate reference for comparison. For all simulations, the cQLS statistic was calculated using the true IBD of all chromosomes within each family. Moreover, the tested SNP was assumed to be the causal SNP. We also evaluated the effect of mistakes in IBD assignment on the power of the cQLS statistic. For a proportion of sibling sets, we randomly assigned the IBD status for each individual in that set, ensuring only that the IBD status was consistent with the observed genotypes. We then recalculated the power for Studies 6 and 7 with $\lambda_S = 1.5$. We examined the scenarios where 2%, 5%, and 8% of sibling sets were allowed to have errors.

To test the accuracy of the p-value, we simulated 10^7 datasets under the null distribution for study designs 5, 6, and 7, where the mQLS and cQLS statistics differ, assuming $\lambda_S = 5$. For computational efficiency, we included only 1000 subjects and only examined larger thresholds of 10^{-3} and 10^{-4} .

2.4. *NIA-LOAD*. The details of the NIA-LOAD/NCRAD GWAS of Late Onset Alzheimer’s Disease(LOAD) have been described elsewhere (Wijsman et al., 2011). Briefly, the study recruited families with multiple affected individuals. Specifically, probands were required to have a diagnosis of Alzheimer’s Disease after the age of 60, have a sibling with a similar diagnosis, and an additional biologically-related family member available for genotyping. In these families, additional relatives over the age of 50 were recruited regardless of cognitive status. Study participants were then genotyped using Illumina’s Human610Quadv1B BeadChips (Illumina, San Diego, CA, USA). We augmented the genotypes by imputing SNPs in the APOE region that have been previously associated with LOAD (Bertram et al., 2007) using IMPUTE2 software version 2.2.2 (Howie, Donnelly and Marchini, 2009), with prephasing by SHAPEIT software version 1 (Delaneau, Zagury and Marchini, 2013), and version 3 of the 1000 Genomes Project data as the reference set. At each imputed SNP, we assigned the most likely genotypes that were consistent with IBD status. In order to create a study guaranteed to be robust to population stratification, we focused on 115 sets of siblings and pruned each set so that it had an equal number of cases and

controls.

3. Results.

3.1. *cQLS vs mQLS*. Chromosome-based test statistics had higher power, as compared to mQLS and FBAT, to detect associations when the study included cases and controls from the same family (Figure 1). We first consider simulations where the disease had a relatively low sibling relative risk of $\lambda_S = 1.5$. If the GWAS included pairs of siblings (Study 6), one affected and one unaffected, then a SNP that would be detected by cQLS in 75% of such studies would only be detected by QLS in 51% (Figure 1a). In studies that included sets of four siblings (Study 7), with each set including two affected and two unaffected individuals, when cQLS provided a power of 0.75, QLS provided a power of 0.36 (Figure 1c). When we simulated a disease with a high sibling relative risk ($\lambda_S = 5$), the power gained from using cQLS decreased. For study designs 6 and 7, mQLS achieved a power of 0.59 and 0.63 for SNPs where cQLS achieved a power of 0.75 (Figure 1b,d).

Errors in IBD assignment decreased the power for association tests using cQLS. With an error rate of 2%, tests based on cQLS still had higher power for Study 6 and Study 7. However, with an error rate of 5%, cQLS performed no better than the other two test statistics in study 6 ($\lambda_S = 1.5$). Specifically, when cQLS had a power of 0.75 (SE = 0.01), QLS provided a similar power of 0.72 (SE = 0.01). When the error rate reached 8%, the three test statistics performed similarly in study 7 ($\lambda_S = 1.5$), with QLS achieving a power of 0.74 (0.01) when cQLS had a power of 0.75 (0.01).

Simulations suggest that the type-I error for the cQLS statistic matched the chosen α threshold when simulating data from the null distribution (Table 2).

3.2. *Genotyping F.H. Cases vs R.A. Cases*. Genotyping cases with a family history of disease provides a study with significantly higher power than genotyping randomly ascertained cases. For a disease with a low sibling relative risk ($\lambda_S = 1.5$), when the effect size for a SNP was large enough so that the study with all F.H. cases (Study 2) had a power of 0.75, a study with all R.A. cases (Study 1) had a power of only 0.20 (Figure 2a). However, as we increase the total heritability, shrinking the proportion of heritability attributable to the tested SNP, the power gained from using F.H. cases is decreased. When $\lambda_S = 5$, a SNP with a power of 0.75 in Study 2 would have had a power of 0.55 (Figure 2b) in Study 1. For interpretation, recall that the sibling relative risk (λ_S) reflects the heritability from genetic variants other than the tested SNP. Therefore, as Figures 2a and 2b show, the power

of Study 1, which collects R.A. cases, does not depend on λ_S , but only on the relative risk of the tested SNP.

3.3. Genotyping Affected Sibling. Genotyping the affected siblings of F.H. cases increases the power. The additional genotyping of 5000 siblings, or moving from Study 3 and a simple score test to Study 5 and cQLS, will increase the power to detect a SNP with OR=1.15 from 0.57 to 0.75 (Figure 3a). Each sibling, on average, only offers one new chromosome, but siblings are also F.H. cases, and as such, are enriched for the causal allele. Therefore, genotyping the 5000 siblings promises higher power than genotyping an additional 5000, unrelated, but randomly ascertained cases. Moving from study 3 to study 4, which includes 10000 R.A. cases and 5000 F.H. cases only increases the power from 0.57 to 0.63 (Figure 3a). In studies, with a mixture of F.H. cases and R.A. cases, such as Study 3, the power of a standard association test can be improved by appropriately upweighting F.H. cases through use of either mQLS or cQLS. When moving from Study 3 and cQLS to Study 5 and cQLS, the power gain is less impressive, increasing from 0.60 to 0.75 (Figure 3c). When the total heritability is high and $\lambda_S = 5$, mQLS and cQLS overweight the F.H. cases, and including information about ungenotyped individuals can potentially lower power (Figure 3d).

3.4. LOAD. In addition to the simulations, we tested 11 SNPs in the APOE region of chromosome 19 for an association with LOAD in participants of the NIA-LOAD/NCRAD GWAS. Table 3 shows that eight of these 11 SNPs were associated with LOAD at a p-value below 0.05. In 7 of these 8 SNPs, the test based on the cQLS statistic resulted in a lower p-value, as compared to using the mQLS or FBAT statistic. However, all three methods provide similar evidence that these SNPs are associated with Alzheimer’s disease. For the two most strongly associated SNPs, rs429358 and rs4420638, the p-values were reduced from 1.23^{-8} and 5.23^{-7} based on the mQLS statistic or 5.9^{-8} and 1.2^{-6} based on the FBAT statistic to 7.70^{-9} and 1.79^{-7} based on the cQLS statistic.

4. Discussion. Our primary objective was to introduce the chromosome-based Quasi-likelihood Score (cQLS) statistic and demonstrate that it can offer increased power to detect associations in GWAS with related individuals. Specifically, in studies designed to be robust to population stratification, such as those including sibling sets equally divided between cases and controls, statistical power can be increased by over 50%. The new statistic can also be applied to less robust study designs, but like GWAS with unrelated

individuals, would then require adjusting for population-eigenvectors. The derivation of cQLS as a partial likelihood shows how to easily adjust for covariates in both logistic and liability-threshold models.

Although our evaluation has focused on single SNP tests with fixed thresholds for statistical significance (e.g. 10^{-7}), cQLS can offer a key, additional, advantage when testing groups of SNPs in linkage disequilibrium. In GWAS with unrelated individuals, genotypes can be permuted among individuals to obtain permutation-based measures of significance. In GWAS with related individuals, standard methods are not appropriate as individuals are not independent (Wang, 2011). However, the founder chromosomes are independent, and therefore it is straightforward to apply permutation methods with cQLS in GWAS of related individuals.

In addition to illustrating the improved power, we designed our simulations to reevaluate our expectations about the use of families in GWAS. First, affected siblings are often not genotyped because adding, on average, only one unique chromosome to the study is thought not to be worth the cost. However, we show that the power gained from genotyping an affected sibling can actually exceed the power from genotyping a randomly ascertained case. Second, it is well known that studies will have higher power when including cases (F.H. cases) with a family history of disease. F.H. cases should be enriched with disease-causing variants. However, we show that when a disease is highly heritable, the enrichment for any specific disease-causing variant is weaker. Thus, the benefit from genotyping cases with a family history of disease is lower, as demonstrated by our comparison of diseases with $\lambda_S = 1.5$ and $\lambda_S = 5$.

Like mQLS, cQLS offers the ability to use phenotyped, but not genotyped family members. Such an advantage could also be gained by imputing the genotype of such individuals and then performing the GWAS using the entire population, with appropriate adjustment for the uncertainty introduced by imputation. However such methods are not easily available and would still not offer the other benefit of cQLS, identifying local IBD.

Although cQLS requires calculating haplotypes for determining IBD, the test statistic still focuses on finding associations with single SNPs as opposed to haplotypes. A haplotype analysis (Akey, Jin and Xiong, 2001), which looks for associations with a specific haplotype, will decrease power when the causal SNP is directly genotyped, as is likely to be the case when using dense arrays or sequencing. However, having identified the haplotypes, this information can be used to adjust for local ancestry instead of using a more global principal components approach (Wang et al., 2011).

The cQLS has limitations. First, the statistic will lose power in the pres-

ence of IBD error. In our examples, we found an error rate of 5% was large enough to offset the benefit of cQLS in the smaller studies. A second issue is that this new statistic requires a larger computational investment. Specifically, the first three steps of phasing, detecting shared segments of IBD, and calculating $T_{jik'k}$ required a total of 8.8 hours on a single 2.8 GHz Intel X5660 processor for the NIA-LOAD/NCRAD GWAS containing 575003 SNPs. After these initial steps, calculating the cQLS statistic for the NIA-LOAD/NCRAD GWAS required only 3.2 minutes. In comparison, mQLS required 70.4 minutes and FBAT required 11.1 minutes. Both mQLS and FBAT required less time by dividing the genome into 22 regions. For mQLS, we divided the genome into 22 intervals with an equal number of SNPs and for FBAT, we divided the genome into chromosomes. The mQLS program was not designed to handle GWAS, and we would expect that if optimized, performing an mQLS analysis would require less computational time than a cQLS analysis.

The benefit of cQLS depends on study design. For those designs that include mixtures of family-based and randomly ascertained controls, the increased power offered by cQLS will be lowered. Therefore, the additional computational cost would offer less value. Second, we have examined cQLS only in nuclear families and simple three-generation families (data not shown) where IBD can be reconstructed with high accuracy. We need further testing to assess the quality of IBD estimates from more distant relationships. Third, cQLS tests for association, and, unlike FBAT, will have no power when there is linkage but no association. Although sequencing has removed our reliance on tag SNPs, a linkage analysis may still offer advantages in the presence of epistasis.

Finally, we remark that there is no consensus on how to best combine within-family and between-family information in GWAS with related individuals. However, a technical way to address this question would be to examine the full-data likelihood. The derivation of the cQLS (Appendix B) starts by defining this likelihood. The form of the final test statistic is the likelihood ratio test statistic based on the key partial likelihood, suggesting that the cQLS offers a near optimal combination of the two types of information.

Study	Cases			Controls		
	Random	Fam His	Sib 1	Random	Sib 1	Sib 2
1	5000	0	0	10000	0	0
2	0	5000	0	10000	0	0
3	5000	5000	0	10000	0	0
4	10000	5000	0	10000	0	0
Multiple Family Members Genotyped						
5	5000	5000	5000	10000	0	0
6	0	10000	0	0	10000	0
7	0	10000	10000	0	10000	10000

TABLE 1

Each evaluated study design, labeled 1 through 7, genotypes different types of cases and controls. The first set of columns list the number of randomly ascertained cases (Random), identified cases - or probands - with a family history (Fam His), and affected siblings of probands (Sib 1) genotyped for the given study. The second set of columns list the number of randomly ascertained controls (Rand), unaffected "older" siblings (Sib 1), and unaffected "younger" siblings (Sib 2) genotyped for the given study. Studies 1 and 2 are used to assess the benefit of genotyping individuals with a family history of disease. Studies 3, 4, and 5 are used to assess the benefit of genotyping siblings of probands. Studies 6 and 7 are used to assess chromosome-based tests.

Study	cQLS		mQLS		FBAT	
	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$
5	9.81×10^{-4}	9.65×10^{-5}	9.79×10^{-4}	9.03×10^{-5}	9.44×10^{-4}	8.38×10^{-5}
6	9.80×10^{-4}	9.25×10^{-5}	1.00×10^{-3}	1.08×10^{-4}	9.56×10^{-4}	1.04×10^{-4}
7	9.70×10^{-4}	1.00×10^{-4}	9.78×10^{-4}	9.68×10^{-5}	9.62×10^{-4}	9.28×10^{-5}

TABLE 2

The proportions of 10^7 null simulations where the cQLS, mQLS, and FBAT p -values are below the specified α threshold. With 10^7 simulations, the standard errors for our empirical α -levels of 10^{-3} and 10^{-4} are approximately $10^{-5} = \sqrt{10^{-3}/10^7}$ and $3 \times 10^{-6} = \sqrt{10^{-4}/10^7}$.

SNP	Chr	Position	mQLS p-value	FBAT p-value	cQLS p-value
rs4806173	19	36024925	8.5×10^{-1}	7.7×10^{-1}	9.1×10^{-1}
rs12984928	19	36029852	8.5×10^{-1}	7.7×10^{-1}	9.1×10^{-1}
rs6857	19	45392254	6.0×10^{-5}	5.8×10^{-6}	2.0×10^{-5}
rs157582	19	45396219	4.7×10^{-5}	4.8×10^{-6}	1.4×10^{-5}
rs449647	19	45408564	1.9×10^{-1}	2.1×10^{-1}	7.5×10^{-2}
rs440446	19	45409167	1.8×10^{-2}	2.4×10^{-2}	2.1×10^{-2}
rs429358	19	45411941	1.2×10^{-8}	5.9×10^{-8}	7.7×10^{-9}
rs4420638	19	45422946	5.2×10^{-7}	1.1×10^{-6}	1.8×10^{-7}
rs157580	19	50087106	3.9×10^{-2}	3.2×10^{-2}	1.8×10^{-2}
rs2075650	19	50087459	3.0×10^{-4}	7.1×10^{-5}	1.7×10^{-4}
rs405509	19	50100676	4.5×10^{-2}	2.3×10^{-2}	4.2×10^{-2}

TABLE 3

The p-values test from association tests of Late-Onset Alzheimer's Disease and SNPs in the APOE region among the NIA-LOAD/NCRAD population. The first three columns indicate SNP ID, chromosome, and position. The last three columns indicate the p-values calculated from the mQLS, FBAT, and cQLS statistics.

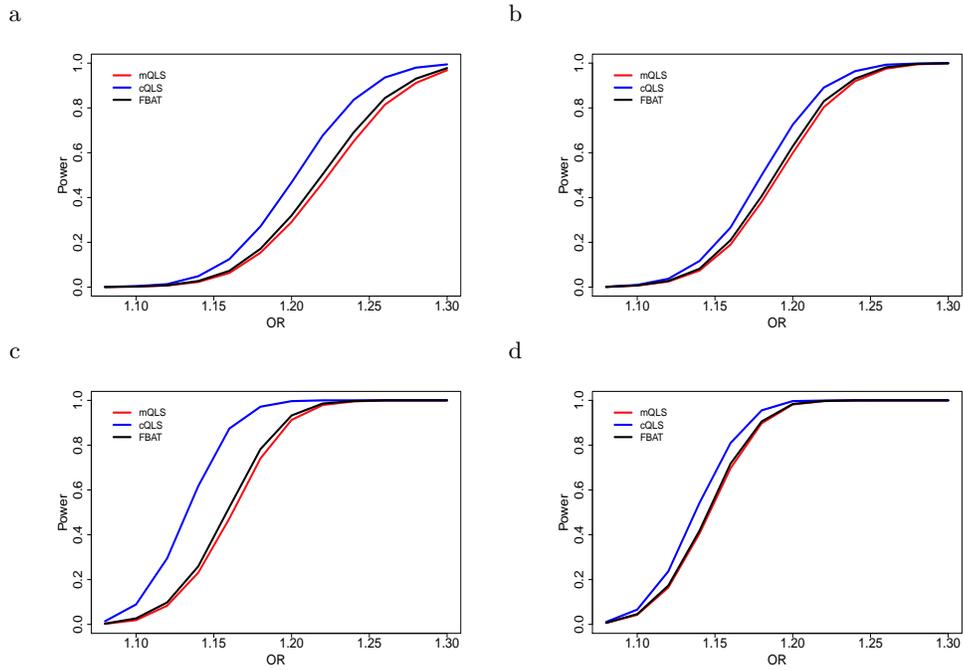
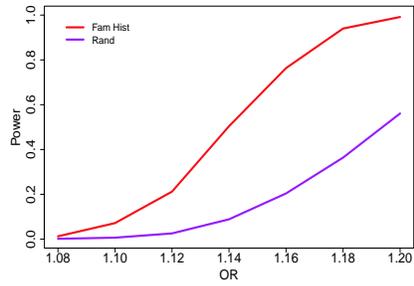


FIG 1. *

The power for cQLS (blue), mQLS (red), and FBAT (black) to detect the association between a SNP and the disease as a function of the OR when (a) $\lambda_S = 1.5$ in Study 6 (b) $\lambda_S = 5$ in Study 6 (c) $\lambda_S = 1.5$ in Study 7 (d) $\lambda_S = 5$ in Study 7. Studies 6 and 7 correspond to genotyping pairs of siblings and groups of four siblings, respectively. λ_S is the sibling relative risk of the disease.

a



b

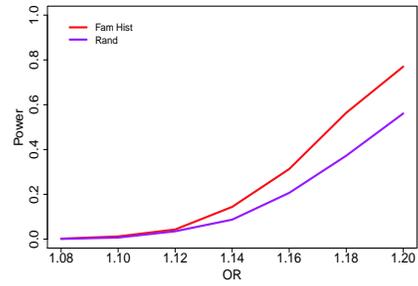


FIG 2. *

a) The power to detect the association between a SNP and the disease, as a function of OR, when the study compares 10000 randomly ascertained controls with either 5000 R.A. cases (red, Study 1) or 5000 F.H. cases (blue, Study 2) when the sibling relative risk, λ_S , is 1.5. b) Same, but with $\lambda_S = 5$. For tests of unrelated individuals, all QLS statistics give identical results.

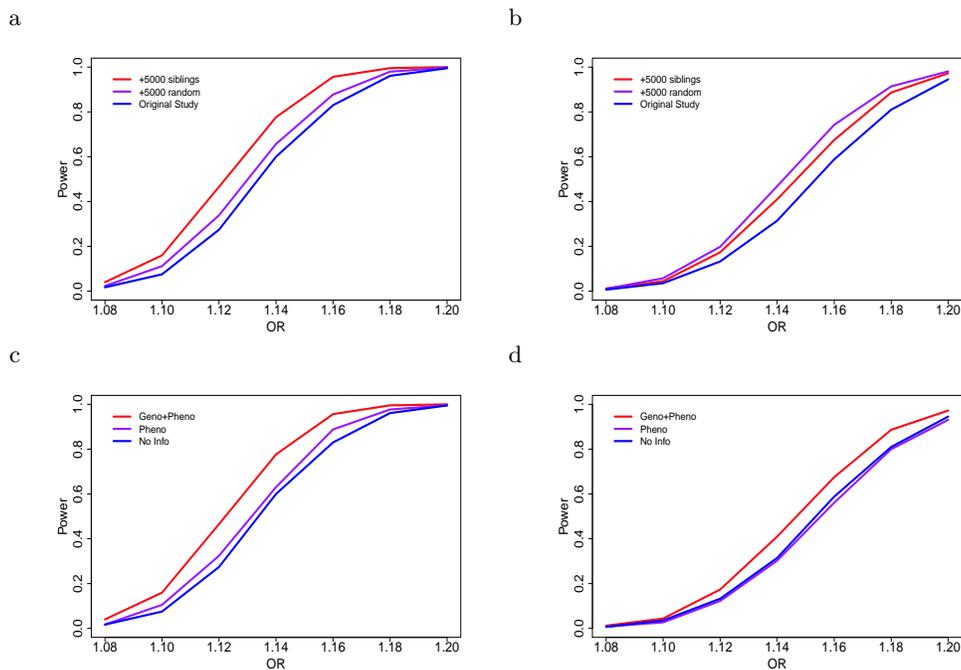


FIG 3. *

a) The power to detect the association between a SNP and the disease, as a function of OR, when the study compares 10000 randomly ascertained controls with either 5000 R.A. cases + 5000 F.H. cases (blue, study 3), 10000 R.A. cases + 5000 F.H. cases (purple, study 4), or 5000 R.A. cases + 5000 pairs of affected siblings (red, study 5) given a sibling relative risk, λ_S , of 1.5 and all tests based on the score statistic. b) Same, but with $\lambda_S = 5$. c) The power when the study compares 10000 R.A. controls and either 5000 R.A. cases + 5000 F.H. cases and the score statistic (blue), 5000 R.A. cases + 5000 F.H. cases and cQLS (purple), or 5000 R.A. cases + 5000 pairs of affected siblings and the cQLS (red) given a sibling relative risk, λ_S , of 1.5. d) Same, but with $\lambda_S = 5$.

APPENDIX A: MQLS AND FBAT

We compared the performance of cQLS test to two standard tests: mQLS (Thornton and McPeck, 2007) and FBAT (Laird, Horvath and Xu, 2000). mQLS is a quasi-likelihood score statistic that presumes an individual's expected genotype increases with the sum, over all affected family members, of their kinship coefficients with that individual. A complete, but terse, definition follows.

For purposes of defining mQLS, we assume there are $n = N_g + M$ total subjects, of which N_g have been genotyped. We let Φ be the $(N_g + M) \times (N_g + M)$ matrix of kinship coefficients and $\Phi_{N,M}$ be the submatrix containing the last M columns of the first N_g rows. The phenotype data are coded as A^\ddagger , a column vector of length $N_g + M$ having i^{th} entry 1 if individual i is affected, 0 if unknown, and $\hat{\pi}_0/(1 - \hat{\pi}_0)$ otherwise. A_N^\ddagger is the vector containing the first N_g elements (e.g. genotyped individuals) and A_M^\ddagger contains the last M elements. Let G be the vector of observed genotypes ($\frac{1}{2}G_i \in \{0, 0.5, 1\}$) for the first N_g individuals. Finally, let $\mathbf{1}$ be a column vector of 1's. Then mQLS is defined as

$$(A.1) \quad mQLS = \hat{\sigma}^{-2} \left(\frac{1}{2}G - \hat{\mu}_0 \right)^T \alpha \Gamma^{-1} \alpha^T \left(\frac{1}{2}G - \hat{\mu}_0 \right)$$

where

$$(A.2) \quad \alpha = A_N^\ddagger + \Phi^{-1} \Phi_{N,M} A_M^\ddagger$$

$$(A.3) \quad \Gamma = \alpha^T (\Phi A_n^\ddagger + \Phi_{N,M} A_M^\ddagger) - (\mathbf{1}^T \alpha)^2 (\mathbf{1}^T \Phi^{-1} \mathbf{1}^T)^{-1}$$

$$(A.4) \quad \hat{\mu}_0 = \hat{p}_{null} \mathbf{1}$$

$$(A.5) \quad \hat{\sigma}_0^{-2} = \left[\frac{1}{2} \hat{p}_{null} (1 - \hat{p}_{null}) \right]^{-1}$$

$$(A.6) \quad \hat{p}_{null} = \frac{1}{2} (\mathbf{1}^T \Phi^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi^{-1} G$$

For implementation, we downloaded the software from <http://galton.uchicago.edu/~mcpeek/software/MQLS/index.html> and, in simulations, set the prevalence of the disease to the “true” value.

FBAT compares the genotypes observed in the cases to their expected value under the null hypothesis of “no linkage and no association” or “no association, in the presence of linkage”, conditioned on the parent's genotypes (or the appropriate sufficient statistic if parental genotypes are unknown). For details, we suggest the user's manual for the software downloadable from <http://www.biostat.harvard.edu/~fbat/default.html>.

APPENDIX B: CQLS: DERIVATION

B.1. Model Assumptions. Without loss of generality, we can assume that all N_{ind} individuals come from the same family, and therefore drop the subscript j from notation. All notation and discussion assume a single SNP under study.

We will assume random mating.

We will assume

$$P[A_i = 1|G_i = 2, \mathbf{X}_i, \Theta] = H_i^2 P[A_i = 1|G_i = 0, \mathbf{X}_i, \Theta]$$

where G_i is defined as the genotype, or the number of minor alleles, for individual i , \mathbf{X}_i is a vector of covariates, Θ is a set of parameters, and

$$(B.1) \quad H_i = \frac{P[A_i = 1|G_i = 1, \mathbf{X}_i]}{P[A_i = 1|G_i = 0, \mathbf{X}_i]}$$

The immediate consequence is that

$$P[A_i = 0|G_i = 2, \mathbf{X}_i, \Theta] = (1 - \pi_i) \left[\frac{1 - H_i \pi_i}{1 - \pi_i} \right]^2 + \frac{\pi_i}{1 - \pi_i} (H_i - 1)^2$$

where π_i is the probability individual i is affected.

For purposes of *deriving* our test statistic, we further assume that $(\pi_i/(1 - \pi_i))(H_i - 1)^2$ is small, or that we can treat the following approximation as an equality without issue

$$P[A_i = 0|G_i = 2, \mathbf{X}_i, \Theta] \approx \left(\frac{1 - H_i \pi_i}{1 - \pi_i} \right)^2 P[A_i = 0|G_i = 0, \mathbf{X}_i]$$

The properties of the test statistic, such as being distributed as a χ_1^2 variable under the null, will not depend on this approximation holding.

These approximations give us two simplifying results

(B.2)

$$P[Y_1, \dots, Y_n | \mathbf{A}, \mathbf{X}, M^*, \Theta] = \prod_k P[Y_k | \mathbf{A}, \mathbf{X}, M^*, \Theta]$$

(B.3)

$$P[Y_k = 1 | \mathbf{A}, \mathbf{X}, M^*, \Theta] = \frac{p \prod_i h_i^{\sum_{k'} T_{ik'k}}}{1 + p \prod_i h_i^{\sum_{k'} T_{ik'k}}}$$

where $h_i = H_i D_i + (1 - H_i \pi_i)(1 - \pi_i)^{-1}(1 - D_i)$, $p = P[Y_i = 1]/P[Y_i = 0] = \phi/(1 - \phi)$ and $M^* = \{T_{ik'k} : i \in 1, \dots, N_{ind}, k' \in \{1, 2\}, k \in \{1, \dots, n\}\}$ is the IBD architecture. We will derive equations (B.2) and (B.3) in a later section of the appendix.

B.2. Probability and Score. We are interested in the distribution of \mathbf{Y} and \mathbf{M}^O given \mathbf{A} , \mathbf{X} , S , and Θ . Here, \mathbf{M}^O are the observed values of \mathbf{M}^* . The variables $T_{ik'/k}$ cannot be identified in family members that are phenotyped, but not genotyped. We have chosen to treat Y as the outcome because we will not need to know the selection procedure used to choose families and we do not need to estimate the nuisance parameter that is the correlation of disease status in the family due to non-genetic similarities. We do note that a full probability would be $P[\mathbf{Y}, \mathbf{M}^O, \mathbf{A}, \mathbf{X}|S, \Theta]$ and we ignore $P[\mathbf{A}, \mathbf{X}|S, \Theta]$ because it carries little information about Θ .

$$P[\mathbf{Y}, \mathbf{M}^O|\mathbf{A}, \mathbf{X}, S, \theta] = P[\mathbf{Y}|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, S, \theta]P[\mathbf{M}^O|\mathbf{A}, \mathbf{X}, S, \theta]$$

However, we consider only the conditional probability, as this half is far more sensitive to Θ . Also, although not mentioned above, we will make the additional assumption that conditional on all other information, the genotypes and the characteristics used to select the individuals are independent. Then we know that S drops out of the desired probability.

(B.4)

$$P[\mathbf{Y}|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, S, \Theta] = \frac{P[\mathbf{Y}, S|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, \Theta]}{P[S|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, \Theta]} = P[\mathbf{Y}|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, \Theta]$$

We will show that the score statistic for $P[\mathbf{Y}|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, \Theta]$ is defined by equation (2.6) when $M^O = M^*$ in an upcoming section.

B.3. Mathematical Detail.

B.3.1. *Detail for section B.1.* We first demonstrate equation B.2.

(B.5)

$$P[Y_1, \dots, Y_{n_1}|\mathbf{A}, \mathbf{X}, M^*, \Theta] =$$

(B.6)

$$\frac{P[\mathbf{A}|Y_1, \dots, Y_{n_1}, \mathbf{X}, \Theta] \prod_k P[Y_k]}{P[\mathbf{A}|\mathbf{X}, M^*, \Theta]} =$$

(B.7)

$$\frac{P[\mathbf{A}|\mathbf{Y} = 0, \mathbf{X}, \Theta] \prod_k \prod_i h_i^{\sum_{k'} G_{ik'} T_{ik'/k}^{M^*}} P[Y_k]}{P[\mathbf{A}|M, \mathbf{X}, \Theta]} =$$

(B.8)

$$\prod_k P[Y_k|\mathbf{A}, M, \mathbf{X}, \Theta]$$

We next demonstrate equation B.3, where we let \mathbf{Y}_{-k} be the vector of all alleles except for k .

(B.9)

$$P[Y_k = 1 | \mathbf{A}, \mathbf{X}, M^*, \Theta] =$$

(B.10)

$$P[Y_k = 1 | \mathbf{A}, \mathbf{X}, M^*, \Theta, \mathbf{Y}_{-k}] =$$

(B.11)

$$\frac{P[\mathbf{A} | \mathbf{Y}_{-k}, Y_k = 1, \mathbf{X}, M^*, \Theta] P[Y_k = 1]}{P[\mathbf{A} | \mathbf{Y}_{-k}, \mathbf{X}, M^*, \Theta]} =$$

(B.12)

$$\frac{P[\mathbf{A} | \mathbf{Y}_{-k}, Y_k = 1, \mathbf{X}, M^*, \Theta] P[Y_k = 1]}{P[\mathbf{A} | \mathbf{Y}_{-k}, Y_k = 0, \mathbf{X}, M^*, \Theta] P[Y_k = 0]} P[\mathbf{A} | \mathbf{Y}_{-k}, Y_k = 0, \mathbf{X}, M^*, \Theta] P[Y_k = 0]}{P[\mathbf{A} | \mathbf{Y}_{-k}, \mathbf{X}, M^*, \Theta]} =$$

(B.13)

$$\frac{P[\mathbf{A} | \mathbf{Y}_{-k}, Y_k = 1, \mathbf{X}, M^*, \Theta] P[Y_k = 1]}{P[\mathbf{A} | \mathbf{Y}_{-k}, Y_k = 0, \mathbf{X}, M^*, \Theta] P[Y_k = 0]} P[Y_k = 0 | \mathbf{A}, \mathbf{X}, M^*, \Theta] =$$

(B.14)

$$\frac{\prod_i h_i \sum_{k'} T_{ik'k}^{M^*} P[A_i | \mathbf{Y}_{-k}, Y_k = 0, \mathbf{X}, M^*, \Theta] P[Y_k = 1]}{\prod_i P[A_i | \mathbf{Y}_{-k}, Y_k = 0, \mathbf{X}, M^*, \Theta] P[Y_k = 0]} P[Y_k = 0 | \mathbf{A}, \mathbf{X}, M^*, \Theta] =$$

(B.15)

$$\prod_i h_i \sum_{k'} T_{ik'k}^{M^*} \frac{P[Y_k = 1]}{P[Y_k = 0]} P[Y_k = 0 | \mathbf{A}, \mathbf{X}, M^*, \Theta]$$

B.3.2. Score Statistic: Y_{jk} and M^* can be uniquely identified. The overall probability can be written as the product of the probabilities for each chromosome with the assumptions in place.

(B.16)

$$P[\mathbf{Y} | D, p, \Theta] = \prod_k P[Y_k | D, p, \Theta]$$

where D abbreviates the collected data, $D = \{\mathbf{A}, \mathbf{X}, M^*\}$.

Because we must account for the nuisance parameter p , the score statistic for Θ is equation B.17 evaluated under the null hypothesis

(B.17)

$$- \left[\sum_k \frac{d}{d\Theta} \ell_D(Y_k) \quad \sum_k \frac{d}{dp} \ell_D(Y_k) \right] \left[\begin{array}{cc} \sum_k \frac{d}{d^2\Theta} \ell_D(Y_k) & \sum_k \frac{d}{d\Theta dp} \ell_D(Y_k) \\ \sum_k \frac{d}{d\Theta dp} \ell_D(Y_k) & \sum_k \frac{d}{d^2p} \ell_D(Y_k) \end{array} \right]^{-1} \left[\begin{array}{c} \sum_k \frac{d}{d\Theta} \ell_D(Y_k) \\ \sum_k \frac{d}{dp} \ell_D(Y_k) \end{array} \right]$$

where

$$(B.18) \quad \ell_D(Y_k) = \log(P[Y_k | \mathbf{A}, \mathbf{X}, M^*, p, \Theta])$$

It is straight forward to evaluate the needed derivative

$$(B.19) \quad U \equiv \sum_k \frac{d}{d\Theta} |_{H_0} \ell_D(Y_k) = \sum_k Z_k (Y_k - \hat{\phi})$$

$$(B.20) \quad \hat{\phi} = \frac{\sum_k Y_k}{\sum_k 1}$$

where

$$(B.21) \quad Z_k \equiv \sum_i \sum_{k'} T_{ik'k}^{M^*} \frac{\dot{h}_i}{h_i}$$

$$(B.22) \quad \dot{h}_i = \frac{d}{d\Theta} h_i$$

We can rewrite U , so that we can calculate its variance, σ_U^2 , without computing/inverting the matrix in equation (B.17).

$$(B.23) \quad U \equiv \sum_k \frac{d}{d\Theta} |_{H_0} \ell_D(Y_k) = \sum_k (Z_k - \bar{Z})(Y_k - \hat{\phi})$$

$$(B.24) \quad \bar{Z} = \frac{\sum_k Z_k}{\sum_k 1}$$

$$(B.25) \quad \sigma_U^2 = \phi(1 - \phi) \sum_k (Z_k - \bar{Z})^2$$

B.3.3. *Score Statistic: All individuals are not genotyped and M^* cannot be uniquely identified.* When all family members are not genotyped, the probability must be averaged over the B possible IBD states

(B.26)

$$P[\mathbf{Y}|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, p, \Theta] = \sum_{b=1}^B P[\mathbf{Y}|\mathbf{A}, \mathbf{X}, M^* = m_b, p, \Theta] P[m_b|M^O]$$

We can take advantage of the equality

(B.27)

$$\frac{d}{d\Theta} \Big|_{H_0} \log\left(\sum_b c_b P[\mathbf{Y}|D(m_b), p, \Theta]\right) = \frac{d}{d\Theta} \Big|_{H_0} \sum_b c_b \log(P[\mathbf{Y}|D(m_b), p, \Theta])$$

where we use the abbreviation $c_b = P[m_b|M^O]$ and use $D(m_b)$ because M^* is no longer known.

Equation (B.27) shows us that the score for $P[\mathbf{Y}|\mathbf{A}, \mathbf{X}, \mathbf{M}^O, p, \Theta]$ is proportional to the score we would observe had there been N_M families, where all families had the observed \mathbf{Y} and $c_b N_M$ of those families had IBD structure m_b . Therefore, equation (B.19) still holds, so long as we now let

(B.28)

$$Z_k \equiv \sum_i \sum_{k'} P[T_{ik'k} = 1] \frac{\dot{h}_i}{h_i}$$

B.3.4. *Score Statistic: Y_{jk} cannot be uniquely identified.* In some scenarios, Y_k cannot be uniquely identified given the available genetic information. In these scenarios, we must average the two possibilities to obtain the value of cQLS. As an example, this situation occurs in our simulated studies of sibling pairs. When two siblings have IBD=1 and are each heterozygous, we cannot determine whether the shared chromosome has the minor or common allele. We focus on this specific example to explain the needed adjustment.

Let $Q_j = 1$ if the siblings have IBD = 1 and are both heterozygous. Furthermore, in such a family, let $Y_1, Y_2,$ and Y_3 denote the alleles on the chromosome uniquely in the first brother, in both brothers, and uniquely in the second brother respectively. Let $Z_1, Z_2,$ and Z_3 be the disease variable for each of those chromosomes. We know

(B.29) $P(Q = 1 | A, X, M^*, p, \Theta) =$

(B.30) $P(\{Y_1, Y_2, Y_3\} = \{1, 0, 1\} | A, X, M^*, p, \Theta) +$

(B.31) $P(\{Y_1, Y_2, Y_3\} = \{0, 1, 0\} | A, X, M^*, p, \Theta)$

For families with $Q = 1$, we must reevaluate their contribution to the score equations. Under the null hypothesis, we find

$$(B.32) \quad \frac{d \log(P(Q = 1|A, X, M^*, p, \Theta))}{dp} \Big|_{H_0} = (2 - \phi) \frac{-\frac{\phi^2}{p^2}}{1 - \phi} + (1 + \phi) \frac{\phi}{p^2}$$

and

$$(B.33) \quad \frac{d \log(P(Q = 1|A, X, M^*, p, \Theta))}{d\Theta} \Big|_{H_0} = (1 - 2\phi)Z_2$$

Our new contributions to the score equations lead us to the following MLE of ϕ

$$(B.34) \quad \hat{\phi} = \frac{N_1 + N_3}{N_0 + N_1 + N_3}$$

where

$$(B.35) \quad N_0 = \sum_{j,k} 1(Q_j = 0)1(Y_{jk} = 0)$$

$$(B.36) \quad N_1 = \sum_{j,k} 1(Q_j = 0)1(Y_{jk} = 1)$$

$$(B.37) \quad N_3 = \sum_{j,k} 1(Q_j = 1)$$

Furthermore, we can rewrite U as

$$(B.38) \quad U = \sum_{j,k:Q_j=1} (Z_k - \bar{Z})(Y_k - \hat{\phi}) + \sum_{j:Q_j=0} (1 - 2\hat{\phi})(Z_2 - \bar{Z})$$

and the score statistic as

$$(B.39) \quad \frac{U^2}{\hat{\sigma}_U^2}$$

where we let p_u be the unique probabilities of each of the eight possible combinations of $\{Y_1, Y_2, Y_3\}$ when $IBD=1$, Z_{j*} be the possible corresponding contributions from family j , and $\hat{\sigma}_U^2$ be the appropriate estimate of the variance under the null.

$$\begin{aligned}
\text{(B.40)} \quad p_{u1} &= \hat{\phi}^3 \\
\text{(B.41)} \quad p_{u2} &= \hat{\phi}^2(1 - \hat{\phi}) \\
\text{(B.42)} \quad p_{u3} &= \hat{\phi}(1 - \hat{\phi})^2 \\
\text{(B.43)} \quad p_{u4} &= (1 - \hat{\phi})^3
\end{aligned}$$

$$\begin{aligned}
Z_{*j1} &= \left(\sum_{k=1}^3 (Z_k - \bar{Z})(1 - \hat{\phi}) \right)^2 \\
Z_{*j2} &= \left(\sum_{k=1}^3 (Z_k - \bar{Z})(1 - \hat{\phi}) - Z_3 + \bar{Z} \right)^2 + \left((Z_2 - \bar{Z})(1 - \hat{\phi}) \right)^2 + \\
&\quad \left(\sum_{k=1}^3 (Z_k - \bar{Z})(1 - \hat{\phi}) - Z_1 + \bar{Z} \right)^2 \\
Z_{*j3} &= \left(\sum_{k=1}^3 (Z_k - \bar{Z})(1 - \hat{\phi}) - Z_2 - Z_3 + 2\bar{Z} \right)^2 + \left((Z_2 - \bar{Z})(1 - \hat{\phi}) \right)^2 + \\
&\quad \left(\sum_{k=1}^3 (Z_k - \bar{Z})(1 - \hat{\phi}) - Z_1 - Z_2 + 2\bar{Z} \right)^2 \\
Z_{*j4} &= \sum_{k=1}^3 \left((Z_k - \bar{Z})(-\hat{\phi}) \right)^2
\end{aligned}$$

and

$$\text{(B.44)} \hat{\sigma}_U^2 = \hat{\phi}(1 - \hat{\phi}) \sum_{j,k:IBD(j) \neq 1} (Z_k - \bar{Z})^2 + \sum_{j:IBD(j)=1} \sum_{t=1}^4 p_{ut} Z_{*jt}$$

Accurate haplotyping would overcome this difficulty and allow us to uniquely identify Y_{jk} . As we expect haplotyping to become standard practice in the very near future (Peters et al., 2012), we expect that this step will soon be unnecessary.

B.3.5. Violation of Hardy Weinberg Equilibrium. Our estimate for the variance of U in equation B.25 assumes that the genotypes are in HWE. As an alternative, start by calculating the 16 possible values (one for each genotype) of $U = \sum_k (Z_k - \bar{Z})(Y_k - \hat{\phi})$ for each family j . The second step is to calculate the probability of each of the 16 genotypes. Given these probabilities and the possible values of U , it is straight-forward to calculate the variance of U for family any family under the null hypothesis, conditional

on IBD architecture. Currently, this alternative is only available for families with at most four founding chromosomes (e.g. nuclear families) and therefore the remaining goal is to estimate $p = \{p_{200}, p_{110}, p_{101}, p_{020}, p_{011}, p_{002}\}$ where p_{xyz} is the probability that the founding individuals include x , y , and z individuals with genotypes $G_i = 0$, $G_i = 1$, and $G_i = 2$. We estimate these six probabilities by effectively maximizing $P(Y|M^O, p)$ with the constraints that $p_{xyz} \geq 0$ and that $0.25p_{110} + 0.5p_{020} + 0.5p_{101} + 0.75p_{011} + p_{002} = \hat{\phi}$. Specifically, we minimize the following function

$$(B.45) \quad 2\left(\sum_j 1(n_j = 2)\right)(M_2 V_2 - A_2)^2 + 3\left(\sum_j 1(n_j = 3)\right)(M_3 V_3 - A_3)^2 + 4\left(\sum_j 1(n_j = 4)\right)(M_4 V_4 - A_4)^2$$

where M_4 a 6×6 identity matrix,

$$M_2 = \begin{bmatrix} 1 & 0.5 & 0.25 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 1 & 0.5 & 0 \\ 0 & 0 & 0.25 & 0 & 0.5 & 1 \end{bmatrix} \quad \text{and} \quad M_3 = \begin{bmatrix} 1 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 1 \\ 0 & 0.25 & 0 & 0 & 0.5 & 0 \end{bmatrix}$$

and V_2 is the vector estimating $\{P(\sum_k Y_k = 0|n_j = 2), P(\sum_k Y_k = 1|n_j = 2), P(\sum_k Y_k = 2|n_j = 2)\}$, V_3 estimates $\{P(\sum_k Y_k = 0|n_j = 3), P(\sum_k Y_k = 1, W = 1|n_j = 3), P(\sum_k Y_k = 2, W = 1|n_j = 3), P(\sum_k Y_k = 3|n_j = 3), P(\sum_k Y_k = 2, W = 0|n_j = 3)\}$, V_4 estimates $\{P(\sum_k Y_k = 0|n_j = 4), P(\sum_k Y_k = 1|n_j = 4), P(\sum_k Y_k = 1, G_2 = 1|n_j = 4), P(\sum_k Y_k = 1, G_2 \neq 1|n_j = 4), P(\sum_k Y_k = 3|n_j = 4), P(\sum_k Y_k = 4|n_j = 4)\}$, W is a binary variable indicating whether all alleles are identifiable, and $\{G_1, G_2\}$ are the founder genotypes.

B.4. Algorithm for assigning $T_{ijk'k}$.

We start by arbitrarily assigning numbers to the chromosomes of the founder individuals and trimming the family so that no two individuals have $IBD = 2$. Founder individuals are defined to be the largest group possible such that all pairs of founder individuals have $IBD = 0$. Let A be initialized as the founder individuals.

Find an individual, i_3 , in the compliment of A , that meets the first of the following possible criteria and follow the assignment mechanism. Add individual i_3 to A and then repeat.

a) i_3 has $IBD = 1$ with two individuals in A , say i_1 and i_2 , that are also

IBD = 1 with each other. Count the number of minor alleles, among individuals i_1 , i_2 , and i_3 , at all loci in the shared region. Assume chromosomes in individuals i_1 and i_2 have been labeled as $\{1, 2\}$ and $\{1, 4\}$.

Option a1) If (nearly) all counts are even, the chromosomes in individual i_3 are assigned as $\{2, 4\}$.

Option a2) (Nearly) All counts are not even, and individual i_3 is either IBD = 0 with all other individuals in A or IBD = 1 only with individuals who are IBD = 1 with both i_1 and i_2 . Then the chromosomes in i_3 are labeled as $\{1, X\}$, where X is a new chromosome number.

Option a3) (Nearly) All counts are not even, and individual i_3 is IBD = 1 with at least two more individuals in A , say i_4 and i_5 , that are IBD = 1 with each other, but IBD = 0 with both i_1 and i_2 . Then the chromosomes in i_3 are labeled as $\{1, 5\}$, where we assume the chromosomes in i_4 and i_5 are labeled as $\{5, 6\}$ and $\{5, 8\}$.

Option a4) (Nearly) All counts are not even, and individual i_3 is IBD = 1 with exactly one other individual in A , say i_4 , that is IBD = 0 with both i_1 and i_2 . Then, we label i_3 as $\{1, 5\}$, where the chromosomes in individuals i_4 have been labeled as $\{5, 6\}$.

b) i_3 has IBD = 1 with two individuals in A , say i_2 and i_4 , that share IBD = 0 with each other.

Option b1) Individual i_1 (or i_2) has IBD = 1 with another individual in A , say i_4 , in A . Then, we label i_3 as $\{2, 3\}$, where the chromosomes in individuals i_1 , i_2 , and i_4 have been labeled as $\{1, 2\}$, $\{3, 4\}$, and $\{1, 6\}$.

Option b2) Individual i_1 and i_2 have IBD = 0 with all other individuals in A . Then, we label i_3 as $\{1, 3\}$, where the chromosomes in individuals i_1 and i_2 have been labeled as $\{1, 2\}$ and $\{3, 4\}$.

c) i_3 has IBD = 1 with only one individual in A , say i_1 .

Option c1) Individual i_1 has IBD = 1 with another individual in A , say i_2 . Assign the chromosomes in individual i_3 as $\{2, X\}$, where the chromosomes in individuals i_1 and i_2 have been labeled as $\{1, 2\}$ and $\{1, 4\}$ and X is a new chromosome number.

Option c2) Individual i_1 has IBD = 0 with all other individuals in A . Assign the chromosomes in individual i_3 as $\{2, X\}$, where the chromosomes in individuals i_1 have been labeled as $\{1, 2\}$ and X is a new chromosome number.

B.5. Limitations. The standard method for finding an optimal test statistic starts by defining the parameter of interest and then writing out the likelihood of the observed data given this, and possibly other, parame-

ters. In the GWAS discussed here, such a likelihood would necessarily bridge the within-family and between-family information, and immediately show how the two pieces of information should be combined. Here, we have defined this likelihood and show that the cQLS is derived as the score statistic to a specific partial likelihood. However, as that likelihood shows, we ignore information that can be derived from the observed IBD structure. For example, if all affected siblings are IBD=2 at a SNP, that provides some evidence of an association between SNP and disease. Although that information is minimal, we are currently looking into methods for capturing and including this independent information as well. By using only the partial likelihood, the cQLS is not guaranteed to result in the most powerful test.

ACKNOWLEDGEMENTS

The research of Drs. Joshua Sampson and Jianxin Shi was supported by the intramural program of the National Institute of Cancer. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (<http://biowulf.nih.gov>). The NIA-LOAD and NCRAD data was downloaded from dbGaP.

REFERENCES

- AKEY, J., JIN, L. and XIONG, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* **9** 291–300.
- BARRETT, J. C., HANSOUL, S., NICOLAE, D. L., CHO, J. H., DUERR, R. H., RIPOUX, J. D., BRANT, S. R., SILVERBERG, M. S., TAYLOR, K. D., BARMADA, M. M., BITTON, A., DASSOPOULOS, T., DATTA, L. W., GREEN, T., GRIFFITHS, A. M., KISTNER, E. O., MURTHA, M. T., REGUEIRO, M. D., ROTTER, J. I., SCHUMM, L. P., STEINHART, A. H., TARGAN, S. R., XAVIER, R. J., LIBIOULLE, C., SANDOR, C., LATHROP, M., BELAICHE, J., DEWIT, O., GUT, I., HEATH, S., LAUKENS, D., MNI, M., RUTGEERTS, P., VAN GOSSUM, A., ZELENKA, D., FRANCHIMONT, D., HUGOT, J.-P., DE VOS, M., VERMEIRE, S., LOUIS, E., CARDON, L. R., ANDERSON, C. A., DRUMMOND, H., NIMMO, E., AHMAD, T., PRESCOTT, N. J., ONNIE, C. M., FISHER, S. A., MARCHINI, J., GHORI, J., BUMPSTEAD, S., GWILLIAM, R., TREMELLING, M., DELOUKAS, P., MANSFIELD, J., JEWELL, D., SATSANGI, J., MATHEW, C. G., PARKES, M., GEORGES, M. and DALY, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40** 955–962.
- BERTRAM, L., MCQUEEN, M. B., MULLIN, K., BLACKER, D. and TANZI, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet* **39** 17–23.
- BOURGAIN, C., HOFFJAN, S., NICOLAE, R., NEWMAN, D., STEINER, L., WALKER, K., REYNOLDS, R., OBER, C. and MCPEEK, M. S. (2003). Novel Case-Control Test in a Founder Population Identifies P-Selectin as an Atopy-Susceptibility Locus. *The American Journal of Human Genetics* **73** 612 - 626.
- BROWNING, S. R. and BROWNING, B. L. (2010). High-Resolution Detection of Identity by Descent in Unrelated Individuals. *The American Journal of Human Genetics* **86** 526–539.

- BROWNING, B. L. and BROWNING, S. R. (2011). A Fast, Powerful Method for Detecting Identity by Descent. *Am J Hum Genet* **88** 173–182.
- DELANEAU, O., ZAGURY, J.-F. and MARCHINI, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat Meth* **10** 5–6.
- EWENS, W. J., LI, M. and SPIELMAN, R. S. (2008). A Review of Family-Based Tests for Linkage Disequilibrium between a Quantitative Trait and a Genetic Marker. *PLoS Genet* **4** e1000180.
- GUSEV, A., LOWE, J. K., STOFFEL, M., DALY, M. J., ALTSHULER, D., BRESLOW, J. L., FRIEDMAN, J. M. and PE'ER, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19** 318–326.
- HATTERSLEY, A. T. and MCCARTHY, M. I. (8). What makes a good genetic association study? *The Lancet* **366** 1315–1323.
- HE, D. (2013). IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. *Bioinformatics* **29** i162-i170.
- HIRSCHHORN, J. N. and DALY, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6** 95–108.
- HOWIE, B. N., DONNELLY, P. and MARCHINI, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5** e1000529–.
- IONITA-LAZA, I. and OTTMAN, R. (2011). Study Designs for Identification of Rare Disease Variants in Complex Diseases: The Utility of Family-Based Designs. *Genetics* **189** 1061–1068.
- LAIRD, N. M., HORVATH, S. and XU, X. (2000). Implementing a unified approach to family-based tests of association. *Genetic Epidemiology* **19** S36-S42.
- LANGE, C., SILVERMAN, E. K., XU, X., WEISS, S. T. and LAIRD, N. M. (2003). A multivariate family based association test using generalized estimating equations: FBAT GEE. *Biostatistics* **4** 195–206.
- LEE, J., CHENG, R., GRAFF-RADFORD, N., FOROUD, T. and R, M. (2008). Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: Implication of additional loci. *Archives of Neurology* **65** 1518–1526.
- MANICHAIKUL, A., CHEN, W.-M., WILLIAMS, K., WONG, Q., SALE, M., PANKOW, J., TSAI, M., ROTTER, J., RICH, S. and MYCHALECKYJ, J. (2012). Analysis of family- and population-based samples in cohort genome-wide association studies. *Human Genetics* **131** 275–287.
- MIREA, L., INFANTE-RIVARD, C., SUN, L. and BULL, S. B. (2012). Strategies for Genetic Association Analyses Combining Unrelated Case-Control Individuals and Family Trios. *American Journal of Epidemiology* **176** 70–79.
- OTT, J., KAMATANI, Y. and LATHROP, M. (2011). Family-based designs for genome-wide association studies. *Nat Rev Genet* **12** 465–474.
- PETERS, B. A., KERMANI, B. G., SPARKS, A. B., ALFEROV, O., HONG, P., ALEXEEV, A., JIANG, Y., DAHL, F., TANG, Y. T., HAAS, J., ROBASKY, K., ZARANEK, A. W., LEE, J.-H., BALL, M. P., PETERSON, J. E., PERAZICH, H., YEUNG, G., LIU, J., CHEN, L., KENNEMER, M. I., POTHURAJU, K., KONVICKA, K., TSOUPOKO-SITNIKOV, M., PANT, K. P., EBERT, J. C., NILSEN, G. B., BACCASH, J., HALPERN, A. L., CHURCH, G. M. and DRMANAC, R. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487** 190–195.
- PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38** 904–909.

- SHAM, P. C., PURCELL, S., CHERNY, S. S. and ABECASIS, G. R. (2002). Powerful Regression-Based Quantitative-Trait Linkage Analysis of General Pedigrees. *The American Journal of Human Genetics* **71** 238 - 253.
- SLAGER, S. L. and SCHAID, D. J. (2001). Evaluation of Candidate Genes in Case-Control Studies: A Statistical Method to Account for Related Subjects. *The American Journal of Human Genetics* **68** 1457 - 1462.
- TENG, J. and RISCH, N. (1999). The Relative Power of Family-Based and Case-Control Designs for Linkage Disequilibrium Studies of Complex Human Diseases. II. Individual Genotyping. *Genome Research* **9** 234-241.
- THORNTON, T. and MCPEEK, M. S. (2007). Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test. *The American Journal of Human Genetics* **81** 321 - 337.
- WANG, Z. (2011). Direct assessment of multiple testing correction in case-control association studies with related individuals. *Genet. Epidemiol.* **35** 70–79.
- WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F. A., HAKONARSON, H. and BUCAN, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17** 1665–1674.
- WANG, X., ZHU, X., QIN, H., COOPER, R. S., EWENS, W. J., LI, C. and LI, M. (2011). Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* **27** 670-677.
- WIJSMAN, E. M., PANKRATZ, N. D., CHOI, Y., ROTHSTEIN, J. H., FABER, K. M., CHENG, R., LEE, J. H., BIRD, T. D., BENNETT, D. A., DIAZ-ARRASTIA, R., GOATE, A. M., FARLOW, M., GHETTI, B., SWEET, R. A., FOROUD, T. M., MAYEUX, R. and GROUP, T. N.-L. F. S. (2011). Genome-Wide Association of Familial Late-Onset Alzheimer's Disease Replicates BIN1 and CLU and Nominates CUGBP2 in Interaction with APOE. *PLoS Genet* **7** e1001308.
- WILKINSON, L. S., DAVIES, W. and ISLES, A. R. (2007). Genomic imprinting effects on brain development and function. *Nat Rev Neurosci* **8** 832–843.
- WILLER, C. J., SANNA, S., JACKSON, A. U., SCUTERI, A., BONNYCASTLE, L. L., CLARKE, R., HEATH, S. C., TIMPSON, N. J., NAJJAR, S. S., STRINGHAM, H. M., STRAIT, J., DUREN, W. L., MASCHIO, A., BUSONERO, F., MULAS, A., ALBAL, G., SWIFT, A. J., MORKEN, M. A., NARISU, N., BENNETT, D., PARISH, S., SHEN, H., GALAN, P., MENETON, P., HERCBERG, S., ZELENKA, D., CHEN, W.-M., LI, Y., SCOTT, L. J., SCHEET, P. A., SUNDVALL, J., WATANABE, R. M., NAGARAJA, R., EBRAHIM, S., LAWLOR, D. A., BEN-SHLOMO, Y., DAVEY-SMITH, G., SHULDINER, A. R., COLLINS, R., BERGMAN, R. N., UDA, M., TUOMILEHTO, J., CAO, A., COLLINS, F. S., LAKATTA, E., LATHROP, G. M., BOEHNKE, M., SCHLESSINGER, D., MOHLKE, K. L. and ABECASIS, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40** 161–169.
- WON, S., LU, Q., BERTRAM, L., TANZI, R. E. and LANGE, C. (2012). On the Meta-Analysis of Genome-Wide Association Studies: A Robust and Efficient Approach to Combine Population and Family-Based Studies. *Hum Hered* **73** 35–46.
- ZHENG, Y., HEAGERTY, P. J., HSU, L. and NEWCOMB, P. A. (2010). On Combining Family-Based and Population-Based Case Control Data in Association Studies. *Biometrics* **66** 1024–1033.
- ZHU, Y. and XIONG, M. (2012). Family-Based Association Studies for Next-Generation Sequencing. *The American Journal of Human Genetics* **90** 1028 - 1045.

NCI
9609 MEDICAL CENTER DRIVE
ROCKVILLE, MD 20850
E-MAIL: joshua.sampson@nih.gov
peng.li@nih.gov
jianxin.shi@nih.gov

IMS
ROCKVILLE MD
wheelerb@imsweb.com