Published online 19 January 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 8* ***e60***
doi:10.1093/nar/gks024

A powerful test for multiple rare variants association studies that incorporates sequencing qualities

Z. John Daye[1], Hongzhe Li[1] and Zhi Wei[2,*]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA and [2]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA

Received October 28, 2011; Revised December 20, 2011; Accepted January 4, 2012

ABSTRACT

Next-generation sequencing data will soon become routinely available for association studies between complex traits and rare variants. Sequencing data, however, are characterized by the presence of sequencing errors at each individual genotype. This makes it especially challenging to perform association studies of rare variants, which, due to their low minor allele frequencies, can be easily perturbed by genotype errors. In this article, we develop the quality-weighted multivariate score association test (qMSAT), a new procedure that allows powerful association tests between complex traits and multiple rare variants under the presence of sequencing errors. Simulation results based on quality scores from real data show that the qMSAT often dominates over current methods, that do not utilize quality information. In particular, the qMSAT can dramatically increase power over existing methods under moderate sample sizes and relatively low coverage. Moreover, in an obesity data study, we identified using the qMSAT two functional regions (MGLL promoter and MGLL 3′-untranslated region) where rare variants are associated with extreme obesity. Due to the high cost of sequencing data, the qMSAT is especially valuable for large-scale studies involving rare variants, as it can potentially increase power without additional experimental cost. qMSAT is freely available at http://qmsat.sourceforge.net/.

INTRODUCTION

Analysis of common variants through genome-wide association (GWA) studies has enjoyed much success over the last few years. More than a thousand genetic loci associated with more than 200 complex traits have been identified (1). However, these variants typically explain only a small fraction of the inheritable variability for common diseases (2). This may, in part, be explained by the genetic theory that implicates natural selection in pruning out disease-causing variants efficiently before they may become common among the population (3). On the other hand, rare variants, having minor allele frequencies (MAFs) between 0.1 and 1%, are often evolved from more recent mutations and less subjected to natural selection. Thus, association studies of rare variants may hold the potential for identifying genetic components that are functionally relevant and causal for a larger proportion of inheritable variability (4–6). Several studies have already demonstrated the relevance of rare variants on complex traits (7–13). The availability of emerging sequencing technologies will soon allow association studies to be routinely performed between complex traits and rare variants, making the development of robust and powerful procedures for rare variants association especially relevant.

Recent studies reveal several important characteristics of rare variants association. The rare variant hypothesis states that a large proportion of common diseases may be due to the aggregate effects of multiple rare variants, acting independently and with detectable effects on disease risks (4,6,14). Due to low allele frequencies, rare variants often do not exhibit strong linkage disequilibrium (LD) with either rare or common SNPs (6,15). Hence, GWA studies through indirect LD-mapping with tagged SNPs are usually ineffective in detecting rare variants. Rare variants identified in recent literatures usually have strong effects with a mean odds ratio (OR) of 3.74 and most variants having ORs greater than 2 (4). This is quite different from the effects of identified common variants, which usually have ORs between 1.1 and 1.4. Recent studies also present evidences of multiple rare variants acting collectively on disease risks (8,11,16).

*To whom correspondence should be addressed. Tel: +1 973 642 4497; Fax: +1 973 596 5777; Email: zhiwei04@gmail.com

Impelled by growing interests, several association tests have been proposed for rare variants. These tests often employ the idea of pooling or collapsing multiple rare variants. This is partly due to the need to aggregate effects of low-frequency variants in order to increase power and further due to recent evidences suggesting that multiple rare variants often act collectively upon disease risks (8,11,16,17). For example, the combined multivariate and collapsing (CMC) test (18) extends the cohort allelic sum test (CAST) (17) by collapsing variants in subgroups according to allele frequencies and combining these subgroups using Hotelling's T^2 test. The step-up procedure determines whether each variant is protective, deleterious or non-causal and aggregates them accordingly (19). The C-alpha test is a homogeneity test for case-control data via binomial proportions (20). The sequence kernel association test (SKAT) extends kernel machine-based tests for rare variants with more accurate asymptotic approximations in the tail distribution (21). Moreover, several popular multivariate tests for GWA studies have been evaluated for rare variants association, including the mininum of univariate *P*-values (UminP) (22), sum score (23), sum of squared score (SSU) (24) and weighted sum of squared score (SSUw) tests (25).

These existing methods can sometimes be limited for rare variants association as they do not incorporate sequencing quality information. Associating rare variants to complex traits can be challenging due to the presence of sequencing errors. Under low MAFs, a few minor alleles mistaken as major or a few major alleles mistaken as minor can both result in significant loss of power in rare variants association. The quality of individual genotype estimate can vary significantly across both variants and samples, even when the mean sequencing coverage is high. Nevertheless, accurate measures of sequencing qualities can be readily obtained for each genotype using variant calling toolkits (26,27). Genotype quality scores (GQ) are often provided in Phred scale with $GQ = -10\ \log_{10}(1 - Pr)$, where Pr is the probability of a genotype being correctly called (28). Thus, a GQ value of 10 indicates a 10% sequencing error, whereas a value of 30 indicates 0.1% error. Incorporating sequencing qualities in association tests may allow us to increase power and to fully explore the potential of emerging sequencing technologies for identifying causal rare variants.

Some existing methods allow the incorporation of prior weights on each variant (21,24,29), which can potentially incorporate average quality scores across variants. However, sequencing qualities also vary significantly across samples, in addition to variants. For example, a variant may consist of some proportion of high-quality genotype samples but others that are sequenced at low qualities, with the extreme case of having missing values. Utilizing these genotypes without considering varying sequencing qualities across samples can significantly decrease power in rare variants association studies due to large perturbations from low-quality samples. Thus, a new procedure is needed that can address the incorporation of sequencing qualities at each individual genotype.

To achieve a robust and powerful test for rare variants association studies, we propose a quality-weighted multivariate score association test (qMSAT) that directly incorporates sequencing qualities in association tests with multiple rare variants. The qMSAT utilizes a quality-weighted multivariate regression model to incorporate sequencing qualities at each individual genotype. It is robust towards the inclusion of noncausal variants and variants having effects with different magnitudes and directions. Furthermore, it can coherently account for missing genotypes and conjoin in a principled way individuals and variants sequenced at varying coverage depths. When no quality information is utilized, the qMSAT is equivalent to the C-alpha, SSU and SKAT using the linear kernel. The qMSAT is therefore expected to perform equivalently or better than established methods when quality information is utilized.

We present extensive simulation studies based upon sequencing qualities from real data, obtained from the 1000 Genomes Project (30). To guide investigators, we examine the performances of the qMSAT with current methods across a spectrum of MAFs, odds ratios, coverage depths, sample sizes and LD structures. We also provide scenarios of having non-causal variants and variants having effects with different magnitudes and directions. These results suggest that the qMSAT can often dominate over current methods, that do not utilize quality information. In particular, the qMSAT often demonstrates significant improvements in power under moderate sample sizes and relatively low coverage for rare variants association. Furthermore, we apply the qMSAT in an application to the UCSD obesity data study, where we identified two functional regions, the MGLL promoter and MGLL 3′-untranslated region (3′-UTR), that exhibit significant multiple rare variants association with extreme obesity. Most importantly, our results suggest the qMSAT to be a potentially valuable tool for increasing power without incurring additional experimental costs, such as by sequencing at higher coverage depths or larger sample sizes (31).

MATERIALS AND METHODS

Quality-Weighted Multivariate Score Association Test (qMSAT)

In this article, we consider case-control association studies having d multiple rare variants and n individuals. For the i-th individual, y_i denotes a dichotomous trait with $y_i = 1$ for cases and $y_i = 0$ for controls and $\mathbf{G}_i = (G_{i1},\ldots, G_{id})$ are variant genotypes. We assume, in this article, that genotypes are represented by the additive genetic model where $G_{ij} = 0, 1$ or 2 are the numbers of minor alleles at a locus. However, our procedure can be used with other genetic models, such as the dominant or recessive model, and our conclusions regarding the inclusion of quality information will remain the same using another genotypic assignment.

Consider the logistic regression model,

$$\text{logit}\, P(y_i = 1) = \alpha_0 + \sum_{j=1}^{d} \beta_j G_{ij} \quad (1)$$

where α_0 is an intercept term and β_j is the effect of the j-th variant on trait. The rare variants hypothesis suggests that rare variants often act independently on disease risks. Further, having low MAFs, rare variants usually do not exhibit strong LD, even at close proximity (6,11,15). Hence, we consider a quality-weighted log-likelihood,

$$l(\alpha_0, \alpha, \beta) = \sum_{i=1}^{n}\sum_{j=1}^{d} q_{ij}\{y_i \log p_{ij} + (1-y_i)\log(1-p_{ij})\} \quad (2)$$

where

$$p_{ij} = \text{logit}^{-1}(\alpha_0 + \beta_j G_{ij})$$

and q_{ij} are quality weights on each pair (i, j) of individuals and variants. The quality-weighted log-likelihood (2) utilizes relative independence among rare variants by aggregating over weighted marginal likelihoods. It allows effects of individual genotypes to be evaluated based upon sequencing qualities, down-weighting those having relatively low qualities. The weighted multivariate log-likelihood approach has been utilized successfully in several other statistical contexts (32,33). It is often employed for the incorporation of quality-related constituents (34). The approach enables conjoined analysis of sequencing data over a spectrum of qualities and read depths. Moreover, problems associated with imputing rare variants for traditional association tests can be circumvented by assigning $q_{ij} = 0$ at missing genotypes in Equation (2). The quality-weighted likelihood naturally allows missing genotypes without the need for imputation, which can be quite challenging under low MAFs.

For simplicity, we utilize the genotype quality scores directly as weights q_{ij} in this article. The quality-based weights q_{ij} can also be assigned using the number of reads and monotonically non-decreasing functions of quality scores, such as $q_{ij} = (GQ/t)\mathbb{I}\{GQ < t\} + \mathbb{I}\{GQ \geq t\}$ where $\mathbb{I}\{\cdot\}$ is an indicator function and t is some threshold above which qualities are believed to be equally acceptable. A myriad of quality-based weights can be applied based upon an investigator's preferences and prior knowledge.

To achieve a robust and powerful test for multiple rare variants association, we utilize a quality-weighted Rao's score statistic (35) in testing the multivariate null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_d = 0. \quad (3)$$

Computing for the slope of the log-likelihood Equation (2) with genotype coefficients β_j restricted at the null hypothesis, we obtain the quality-weighted score vector $\mathbf{S} = (S_1, S_2, \ldots, S_d)^T$ where

$$S_j = \sum_{i=1}^{n} q_{ij} G_{ij}\left(y_i - \hat{p}_{ij}^0\right) \quad (4)$$

and

$$\hat{p}_{ij}^0 = \frac{\sum_{i=1}^{n} q_{ij} y_i}{\sum_{i=1}^{n} q_{ij}}$$

are the predicted proportions of cases at the j-th variant.

We define the qMSAT statistic as the norm squared of the quality-weighted score vector

$$qMSAT = \mathbf{S}^T\mathbf{S}. \quad (5)$$

If no quality information is utilized ($q_{ij} = 1$), the qMSAT is reduced to the SSU and linear SKAT and is also equivalent to the C-alpha test. The qMSAT provides a computationally efficient means to estimate expected improvements in model fit according to sequencing qualities when the null hypothesis in Equation (3) may be violated. It admits a simple form as explicit computations of genotype coefficients β_j are not required. Utilizing a quadratic form, the qMSAT is robust towards the inclusion of variants having effects with opposing directions. However, with prior weights imposed upon the responses y_i in Equation (2), the qMSAT statistic does not assume a readily approximable asymptotic distribution. Thus, we utilize a permutation-based strategy, where the distribution of the qMSAT statistic is approximated using statistics computed with randomly re-sampled responses $\tilde{y}_i$ in Equation (4) (36). As asymptotic-based tests are not guaranteed to be reliable under low MAFs and limited sample sizes, permutation-based procedures are often preferred in rare variants association studies (20,29). Having a simple form, the qMSAT can be computed efficiently even when large numbers of permutations are required. Indeed, the qMSAT is equivalent in computational speed to C-alpha and other score-based procedures, that have been successfully employed in several earlier studies (20,25,29).

Simulations

We present extensive simulations based upon sequencing qualities from the Pilot Phase 3 study of the 1000 Genomes project (30). The study provides sequencing data for 90 Caucasian individuals over 2385 rare variant sites, amounting to 214 650 genotype quality scores across a spectrum of read depths. In order to compare performances of various methods under realistic error distributions, we sampled directly from these empirical quality scores in generating simulated data.

We generated simulated data over a spectrum of MAFs, odds ratios, coverage depths, sample sizes and LD structures. We also provided scenarios where non-causal variants and variants having effects with different magnitudes and directions are included. As in previous studies (25,37), we first generated d latent variables from the multivariate normal distribution with the autoregressive covariance structure $\mathbf{\Sigma} = \rho^{|i-j|}$, where $\rho = 0$ was used to generate independent variants and ρ close to 1 was used to generate variants in strong LD with its neighbors. The latent variables were then used to produce a haplotype at a given MAF by dichotomizing variables at a specified quantile. Two independent haplotypes thus generated

were combined to obtain the underlying genotypes $\tilde{G}_{ij}$, with which we generated dichotomous phenotypes for a case–control study under the logistic regression model

$$\text{logit } P(y_i = 1) = \alpha_0 + \sum_{j=1}^{d} \beta_j \tilde{G}_{ij} \quad (6)$$

at given effect sizes β_j or odds ratios $\exp(\beta_j)$. The intercept term α_0 was set to attain a disease rate of 5%. As in conventional case–control design, we randomly sampled $n/2$ cases and $n/2$ controls along with their underlying genotypes $\tilde{G}_{ij}$. Next, we generated observed genotypes G_{ij} by perturbing the underlying ones $\tilde{G}_{ij}$ with sequencing qualities sampled from the 1000 Genomes project. At a given mean coverage depth (dp), we simulated the number of reads for each genotype from the gamma distribution with shape and scale parameters at 6.3 and $dp/6.3$, respectively. The gamma distribution with the specified parameters have been found to approximate observed numbers of reads well from mean coverage in previous studies (38,39). The quality score GQ was obtained for each genotype by sampling from quality scores of the 214 650 Pilot Phase 3 study genotypes with an observed number of reads equal to or closest to the number of reads simulated from mean coverage. Finally, we generated genotypes G_{ij} by randomly perturbing $\tilde{G}_{ij}$ using the Bernoulli distribution with error probabilities $10^{(-GQ/10)}$. Each procedure was applied by associating phenotypes y_i with perturbed genotypes G_{ij}. Sampled quality scores at each genotype were utilized as weights q_{ij} in the qMSAT.

We compared the qMSAT with 10 other methods that have been proposed recently for rare variants association. The SSU, SSUw, Score, SKAT with linear and quadratic kernels, and UminP are score-based procedures similar to the qMSAT, whereas the Sum and CMC tests are derived from estimates in a logistic regression with collapsed variants. The step-up procedure also performs variable selection to determine causality and directions of effects in addition to collapsing variants together. The C-alpha procedure is a homogeneity test based upon binomial proportions. In a case–control study, the C-alpha is equivalent to the SSU and linear SKAT (21,40). Moreover, the C-alpha, SSU and linear SKAT are, in turn, equivalent to the qMSAT if no quality information is utilized ($q_{ij} = 1$). We implemented the qMSAT in R, available online at http://qmsat.sourceforge.net/. The SSU, SSUw and Score tests were computed using R codes from an earlier study (24), whereas the SKAT was computed using the R package from the original paper (21). The SKAT package also includes options for incorporating MAF-based weights on each variant. Nonetheless, in order to clearly demonstrate the effects of MAF on association tests, our simulations were performed at a spectrum of values at fixed MAFs. Incorporating MAF-based weight on each variant is only beneficial if both rare and common variants are included. Thus, we do not incorporate MAF-based weights for the SKAT in this study. The `thgenetics` R package was used for the step-up procedure (19). All other procedures were computed using available codes in a recent study (40). P-values for the qMSAT, C-alpha, and step-up procedures were computed using 500 permutations. Simulated data for type I errors were generated using the restricted logistic regression model instead of (6),

$$\text{logit } P(y_i = 1) = \alpha_0,$$

where ORs were set to 1 for each variant. Both empirical powers and type I errors for each procedure were estimated using 1000 repetitions as the proportion of P-values < 0.05.

Application to UCSD obesity data set

We analyzed a targeted re-sequencing data of candidate genes at extreme obesity levels from the UCSD obesity study (41,42). The data set consisted of 289 individuals of European ancestry; among which, 73 men and 70 women having body mass index (BMI) $\geq$40 kg/m^2 were selected as cases, and 74 men and 72 women with BMI $\leq$ 30 kg/m^2 were assigned as controls. Two intervals were sequenced over 188 kb that encompassed regions encoding the endocannabinoid metabolic enzymes, fatty-acid amide hydrolase (FAAH) and monoglyceride lipase (MGLL). Earlier studies had found genes in the endocannabinoid system that regulates obesity levels (43,44), making the FAAH and MGLL prime candidates for association analysis of variants to extreme obesity. We mapped short reads using the Burrows-Wheeler Alignment (BWA) algorithm (45). We employed the most recent state-of-the-art variant calling toolkit GATK (26) to perform variant calls and extract corresponding quality scores at each genotype. An earlier algorithm MAQ (46) was used in the original study (42) for variant calling. We obtained 977 variants with a mean coverage of 101×. These variants were then annotated using SeattleSeq (http://gvs.gs.washington.edu/SeattleSeqAnnotation/). Given the annotations, we focused on the 455 rare variants that were not cataloged in the dbSNP database (47); these variants were then divided based on their genomic positions into groups of promoters, 5′-UTRs, exon regions, intron regions and 3′-UTRs, for performing association tests. For each gene, we combined all exon variants together in a group and considered only non-synonymous variants. The FAAH promoter was sequenced over the 11.5 kb upstream region, whereas the MGLL promoter encompassed the 12.5 kb upstream region. Short functional regions with less than two variants were not included in this study. Missing genotypes were imputed by sampling randomly from non-missing ones in the same variant, and their associated quality scores were set to be 0. Number of permutations in statistical tests was increased to 10 000 for increased precision in P-values.

RESULTS

Presence of sequencing errors

Figure 1 presents sequencing qualities of the data from the Pilot Phase 3 study of the 1000 Genomes project at each genotype in terms of genotype quality scores (GQ) and

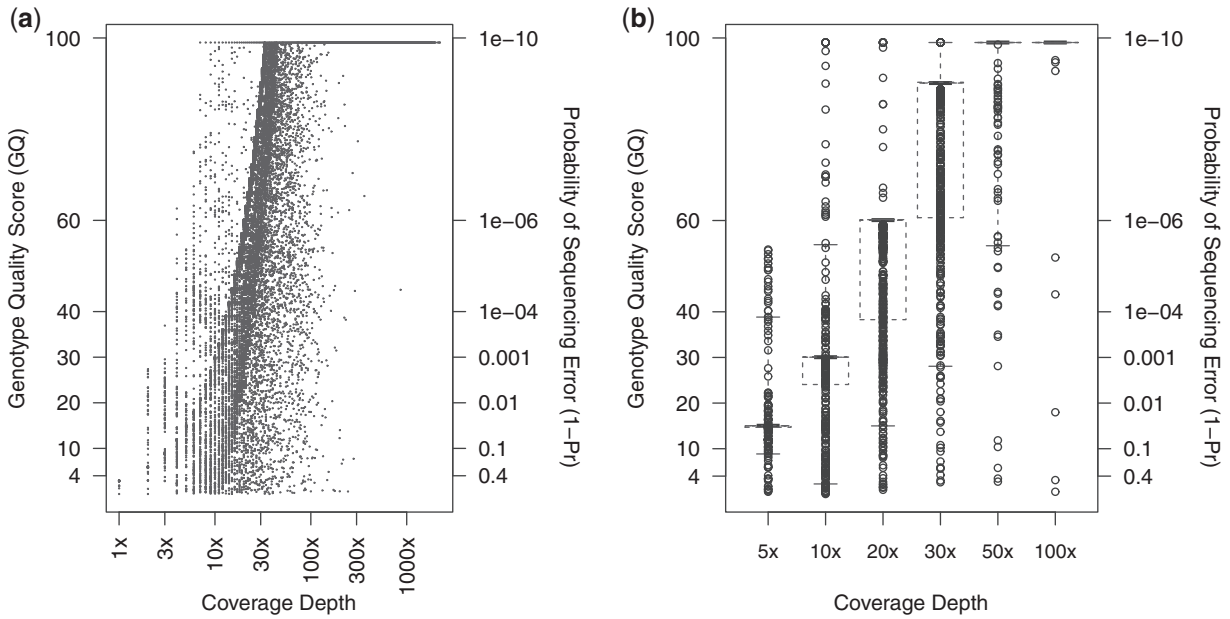

Figure 1. Genotype sequencing qualities generated in Pilot Study 3 of the 1000 Genomes Project. Sequencing qualities in terms of genotype quality scores (GQ) and probability of sequencing errors ($1 - Pr$) are displayed, where $1 - Pr = 10^{-GQ/10}$. A total of 2385 rare variants are called, amounting to 214 650 genotypes at various coverage depths. (**a**) Plot of sequencing qualities at all genotypes, with coverage depths spaced logarithmically. (**b**) Modified box plot of genotypes at coverage depths of 5×, 10×, 20×, 30×, 50× and 100×. Short, bold lines are drawn at the median. Each whisker ends at the 1 and 99% quantiles, and each box encloses the 5–95% interval.

probability of sequencing errors ($1 - Pr$) over varying coverage depths. The quantities are related through the conversion formula $Pr = 1 - 10^{-GQ/10}$. We observe that sequencing qualities tend to increase with increasing coverage depths or numbers of reads at a genotype. Nonetheless, even at relatively high coverage depths, a significant proportion of genotypes still assumes low sequencing qualities. For example, the 99% interval includes genotypes with $GQ = 15$ (or $1 - Pr = 0.03$) at a coverage depth of 20×. In association analysis of rare variants where MAFs are low, even a small proportion of unaccounted sequencing errors can influence results unexpectedly. In the following studies, we evaluate the effects of having sequencing errors under realistic settings by sampling sequencing qualities directly from the Pilot Phase 3 study as presented in Figure 1 (see 'Material and Methods' section).

Evaluation of the effects of coverage depth and sample size

We evaluated empirical powers and type I errors over mean coverage depths of 5×, 10×, 20×, 30×, 50× and 100× and moderate sample sizes of 200, 400, 600, 800 and 1000, each having balanced numbers of cases and controls. We considered 20 rare variants ($d = 20$) consisting of five causal ones with odds ratios of 2, 3, 4, 0.5 and 0.5 for power computations. No LD was assumed among the variants. In each case, we compared the qMSAT with 10 other methods that do not utilize quality information (see 'Material and Methods' section).

Figure 2 presents type I errors at an MAF of 0.5%. Type I errors are generally well controlled. However, score-based procedures using asymptotic distributions (SSU, SSUw, Score, SKAT and UminP) tend to be conservative at moderate or high coverage, especially when sample sizes are small. On the other hand, permutation-based methods (qMSAT, C-alpha and step-up) are usually well controlled around the 0.05 error rate. This may suggest that permutation methods can sometimes be advantageous under finite samples. Supplementary Figures S1 and S3 further depict type I errors at MAFs of 0.1% and 1%, respectively.

Empirical powers are provided in Figure 3 at an MAF of 0.5%. We observe that the qMSAT often dominates over other procedures, that do not utilize quality information. Improvements in power are most significant at relatively low coverage. Consider the permutation-based C-alpha, which is related to the qMSAT when no quality information is utilized (see 'Material and Methods' section). At $n = 600$, the qMSAT outperforms C-alpha with 78.0, 158, 61.2, 27.4, 21.0 and 14.8% increases in power at mean coverage depths of 5×, 10×, 20×, 30×, 50×, and 100×, respectively (Figure 3). Moreover, the qMSAT often achieves fairly high powers even at relatively low coverage. At $n = 1000$, the qMSAT attains powers of 0.530, 0.739 and 0.773 at mean coverage depths of 10×, 20× and 30×, respectively, whereas a similar level of power is attained at 0.780 with a much higher mean coverage depth of 100× (Figure 3). This suggests the qMSAT as a valuable tool for achieving powerful results without incurring additional costs, such as by sequencing at higher coverage depths. The SSU usually tends to have higher powers than SSUw and score tests. Previous studies have suggested that the SSUw and score tests that depend upon accurate

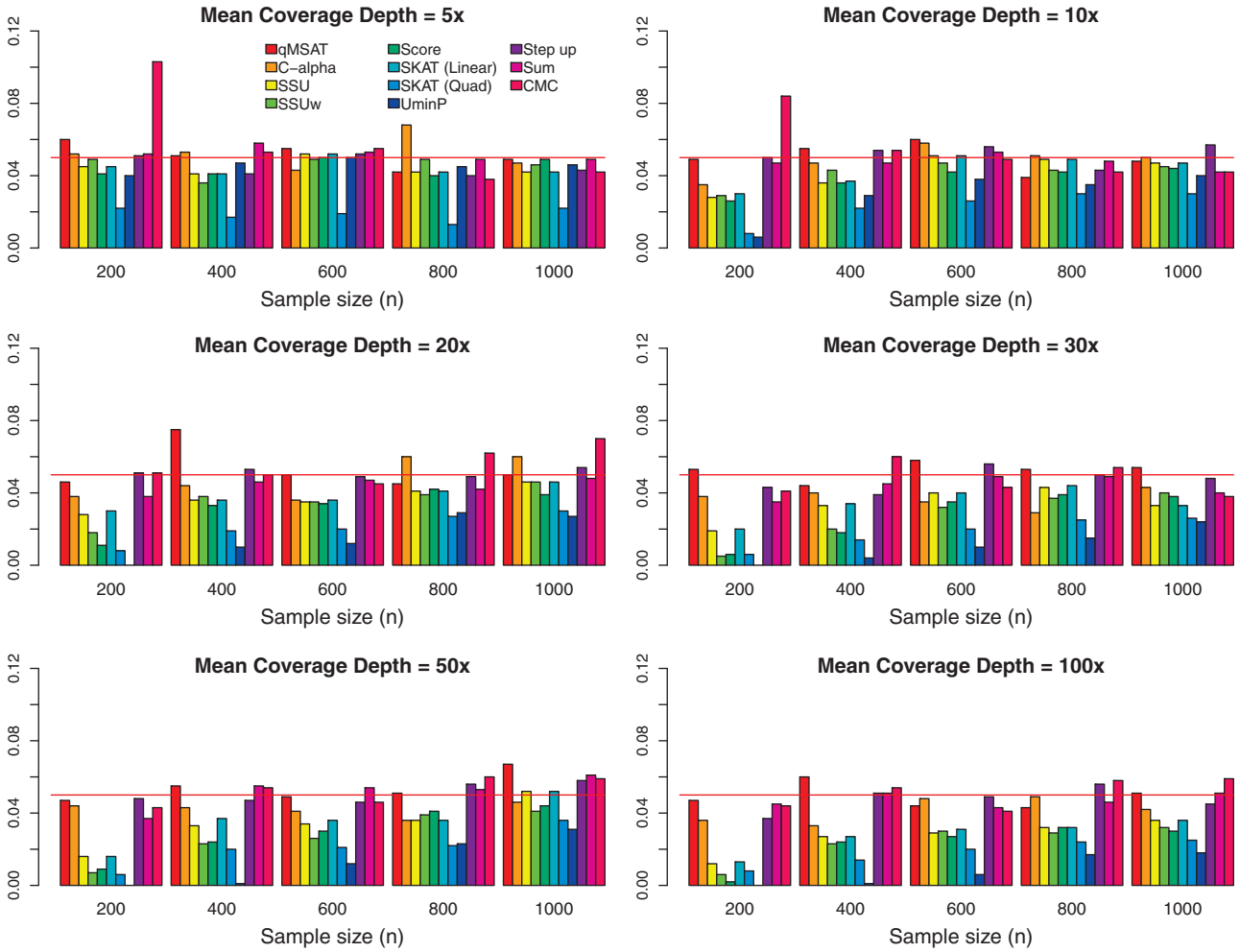

Figure 2. Type I errors over varying coverage depths and sample sizes at MAF = 0.5%. There are $n/2$ numbers of cases and controls each. Twenty non-causal rare variants ($d = 20$) are considered with odds ratios at 1 for all variants. No LD is assumed among the variants. Reference lines (in red) are drawn at the 0.05 error rate.

estimations of variances and the covariance matrix, respectively, may be limited for multivariate analysis when the dimension d is large (24,48,49). The SKAT using linear kernels often has higher powers than the SKAT using quadratic kernels as epistatic effects were not included in this analysis. Due to the inclusion of causal variants having opposing directions, the sum test often has limited powers. Supplementary Figures S2 and S4 further provide power evaluations at MAFs of 0.1 and 1%, respectively.

Evaluation of the effects of MAF and odds ratio

We compared empirical powers and type I errors over varying MAFs of 0.1, 0.5, 1, 5 and 10% and ORs of 1–6 for each of the five variants among the 20 ($d = 20$) considered, in which the remaining are non-causal with ORs of 1. Identified rare variants in recent literatures have a mean OR of 3.74, and most have ORs greater than 2 (4). To focus on evaluating effects of MAF and OR, we considered $n = 500$ observations with balanced numbers of cases and controls and assumed no LD among the variants.

Figure 4 presents empirical powers and type I errors (ORs = 1) at a mean coverage depth of 30×. Power increases for all methods with increasing MAF and OR. The qMSAT often outperforms other procedures at low MAFs for rare variants, whereas all procedures do well at high MAFs for common variants. For low MAFs, power for the qMSAT improves dramatically for increasing ORs. For example, at an MAF of 0.1%, the qMSAT outperforms C-alpha with 66.0, 84.4, 113, 124 and 127% increases in power at causal variants' ORs of 2, 3, 4, 5 and 6, respectively (Figure 4). This may suggest the qMSAT as a potentially useful procedure for identifying rare variants (with MAFs between 0.1 and 1%) under moderate sample sizes and, possibly, for identifying novel variants (with MAFs <0.1%) having strong effect sizes in large-scale studies. The sum test performs competitively with related methods in this example, as all causal variants have the same direction. Supplementary Figures S5–S9 present empirical powers and type I errors at mean coverage depths of 5×, 10×, 20×, 50× and 100×, respectively. Improvements of the qMSAT tend to be more pronounced at lower coverage depths in this example.

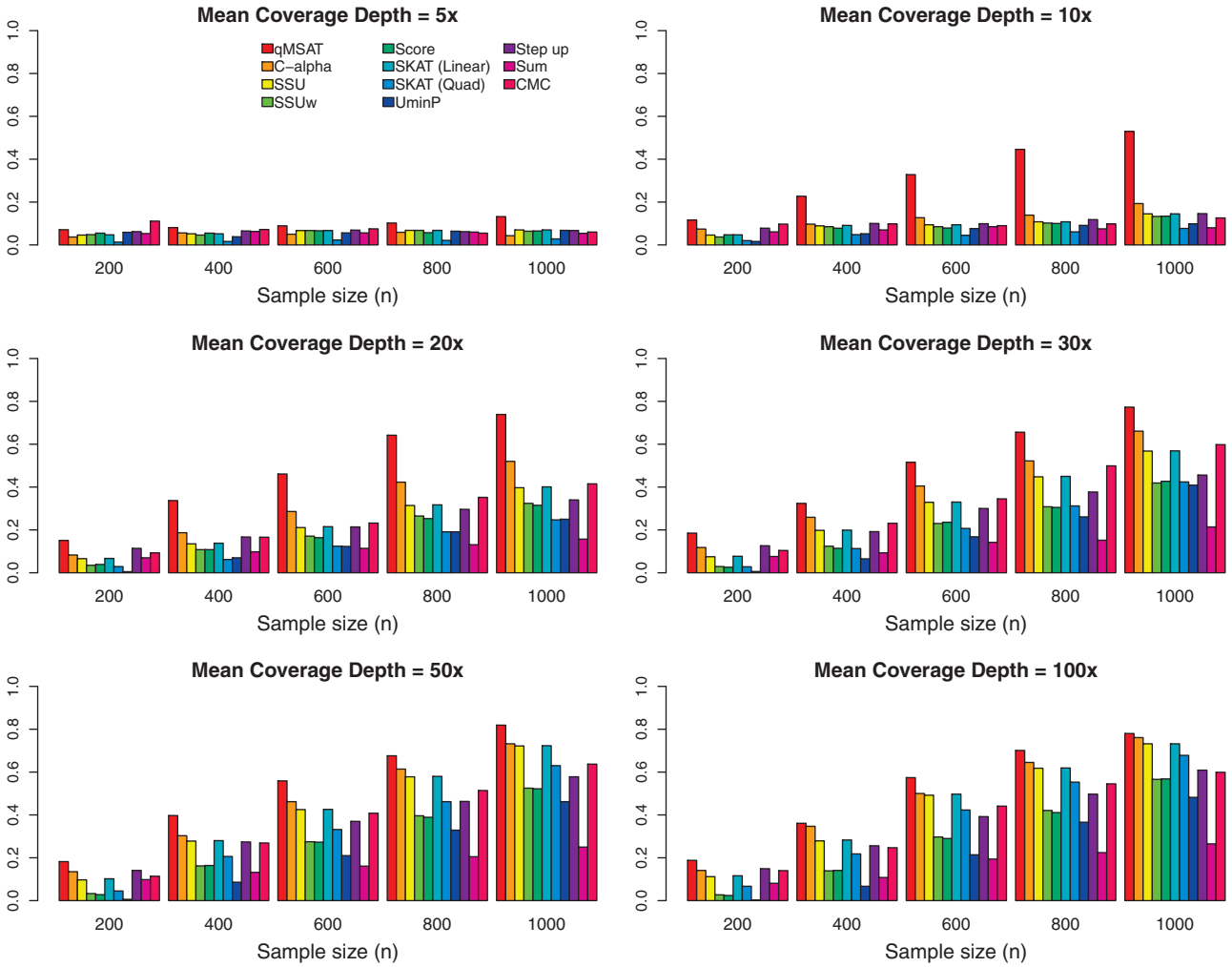

Figure 3. Power evaluation of the effects of coverage depths and sample sizes at MAF = 0.5%. There are $n/2$ numbers of cases and controls each. Twenty rare variants ($d = 20$) are considered, of which five are causal with odds ratios at 2, 3, 4, 0.5 and 0.5. No LD is assumed among the variants.

Evaluation of the effects of LD

We evaluated the effects of LD at a mean coverage depth of 30× and sample sizes of $n = 500$, with equal numbers of cases and controls. A spectrum of LD structures was evaluated with latent variables having autoregressive covariance matrices $\mathbf{\Sigma} = \rho^{|i-j|}$ with ρ at 0, 0.25, 0.5, 0.75 and 0.9 (Figure 5). We employed 20 variants ($d = 20$) with ORs of (3, 2, 1, 1, 4, 1, 1, 1, 0.5, 0.5, 1,. . .,1). Variants generated with $\rho = 0$ are independent, whereas those generated with large values of ρ are in LD with its neighbors.

Figure 5 presents empirical powers for evaluating LD effects. Power increases for all methods with increasing ρ or degree of LD among neighboring variants. This is largely because variants in LD tend to have effect sizes augmented by aggregation. The qMSAT tends to dominate other methods at relatively low MAFs, even at high LD. Rare variants usually do not exhibit strong LD (6,11,15). However, these results suggest that the qMSAT can also be effective when LD effects are strong. Supplementary Figure S10 further provides type I errors.

Application to UCSD obesity data set

We analyzed the UCSD obesity data from a previous study that suggested multiple rare variants to be associated with the FAAH promoter, the MGLL promoter, and a region of MGLL introns (42). We performed multiple rare variants association tests using all statistical approaches as above. As in the original study (42), none of these methods found significant rare variants in the FAAH and MGLL exon regions (Supplementary Table S1); several methods confirmed associations of multiple rare variants in the FAAH promoter and MGLL promoter regions (Table 1). Specifically, the qMSAT attained *P*-values of 0.061 and 0.037 for the FAAH promoter and MGLL promoter regions, respectively. No intron regions were found to be significant by any of these methods (Supplementary Table S1). Interestingly, several methods, including the qMSAT (Table 1), identified the MGLL 3′-UTR to be significant, which was missed in the previous study (42). Promoters are well known to be gene transcription regulators that specify binding of transcription factors to DNA

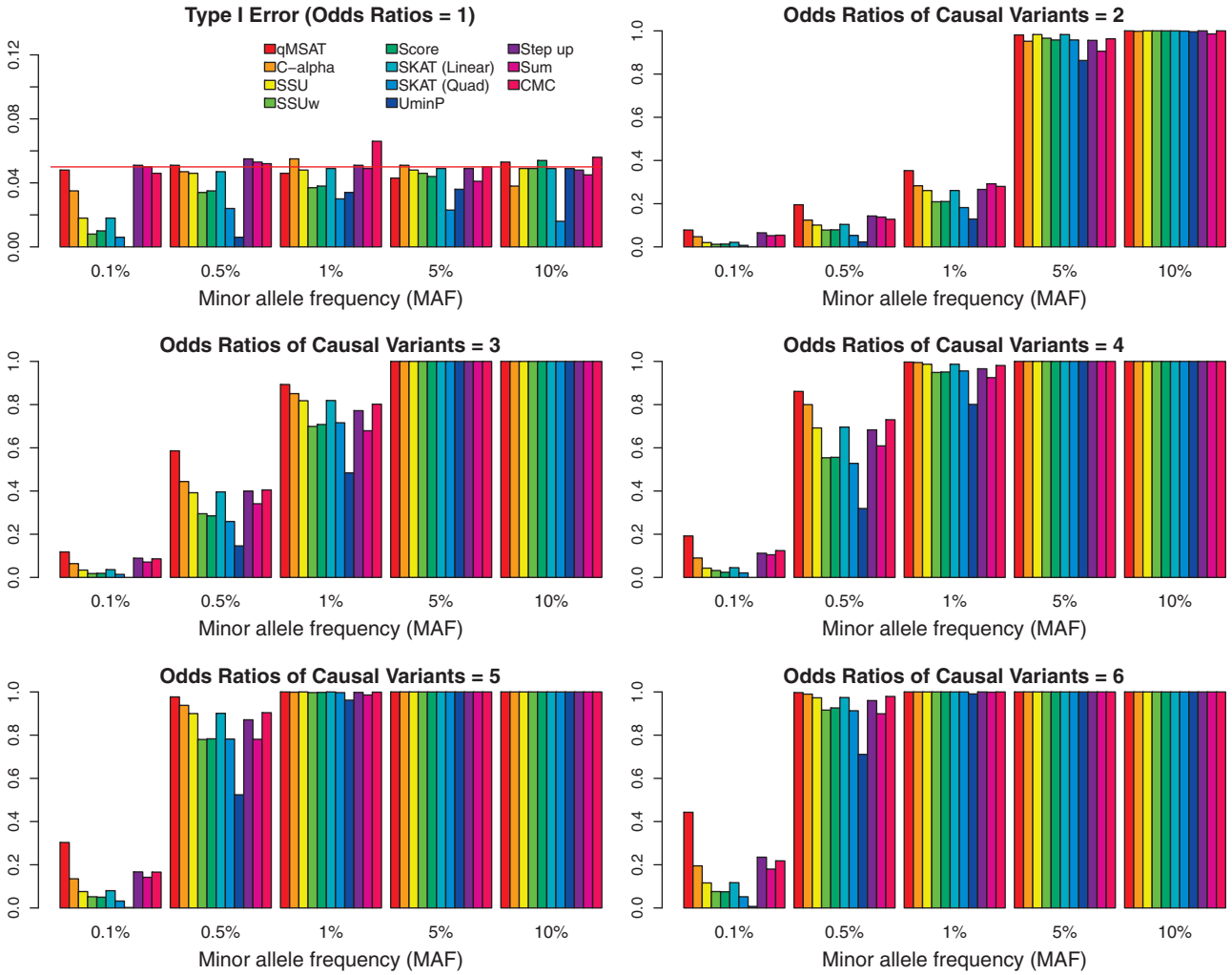

Figure 4. Effects of MAFs and odds ratios at a mean coverage depth of 30×. Empirical powers and type I errors are presented at a mean coverage depth of 30× with varying MAFs and odds ratios for five variants amongst twenty ($d = 20$) considered altogether. We use $n = 500$ observations, where there are $n/2 = 250$ numbers of cases and controls each. No LD is assumed among the variants. Reference lines (in red) are drawn at the 0.05 error rate.

sequences. Moreover, protein binding to 3′-UTR is known to function in some situations by protecting mRNA from nuclease degradation (50). These results suggest that recent mutations in the FAAH promoter, MGLL promoter, and MGLL 3′-UTR may have led to extreme obesity levels in some individuals by altering gene transcription mechanisms. Further studies, such as by associating rare variants to gene expression levels, may help to verify these results. On the other hand, recent mutations in MGLL introns might not have played as significant a role as previous study (42) suggested in extreme obesity. Differences between our results from those of the previous study might be due to using more recent variant calling algorithm in our analysis (see 'Material and Methods' section). Among the procedures evaluated, the qMSAT has the smallest *P*-values for association of rare variants in the MGLL promoter and MGLL 3′-UTR with extreme obesity levels (Table 1). The C-alpha, SSU and linear SKAT are equivalent to the qMSAT when no quality information is utilized. Both C-alpha and qMSAT utilize a permutation-based estimation of *P*-values in order to better control type I errors under moderate samples. In all three regions where at least one method indicates significant association, the qMSAT has significantly smaller *P*-values than the C-alpha (Table 1). These results confirm our simulation studies suggesting that the incorporation of quality information can often improve power in rare variants association analysis. We observe that the SKAT using linear kernels has smaller *P*-values than the SKAT using quadratic kernels (Table 1). This may suggest that rare variants, having low MAFs, might not be able to exhibit strong interaction and higher-order effects. In addition, we performed marginal association analysis of rare variants in the FAAH promoter, MGLL promoter and MGLL 3′-UTR (Supplementary Table S2). The MGLL promoter contains two variants with marginal associations significant at the 0.05 *P*-value level but with opposing effects on extreme obesity (Supplementary Table S2). In this region at the MGLL promoter, the multivariate sum test has the largest *P*-value of 0.578 (Table 1).

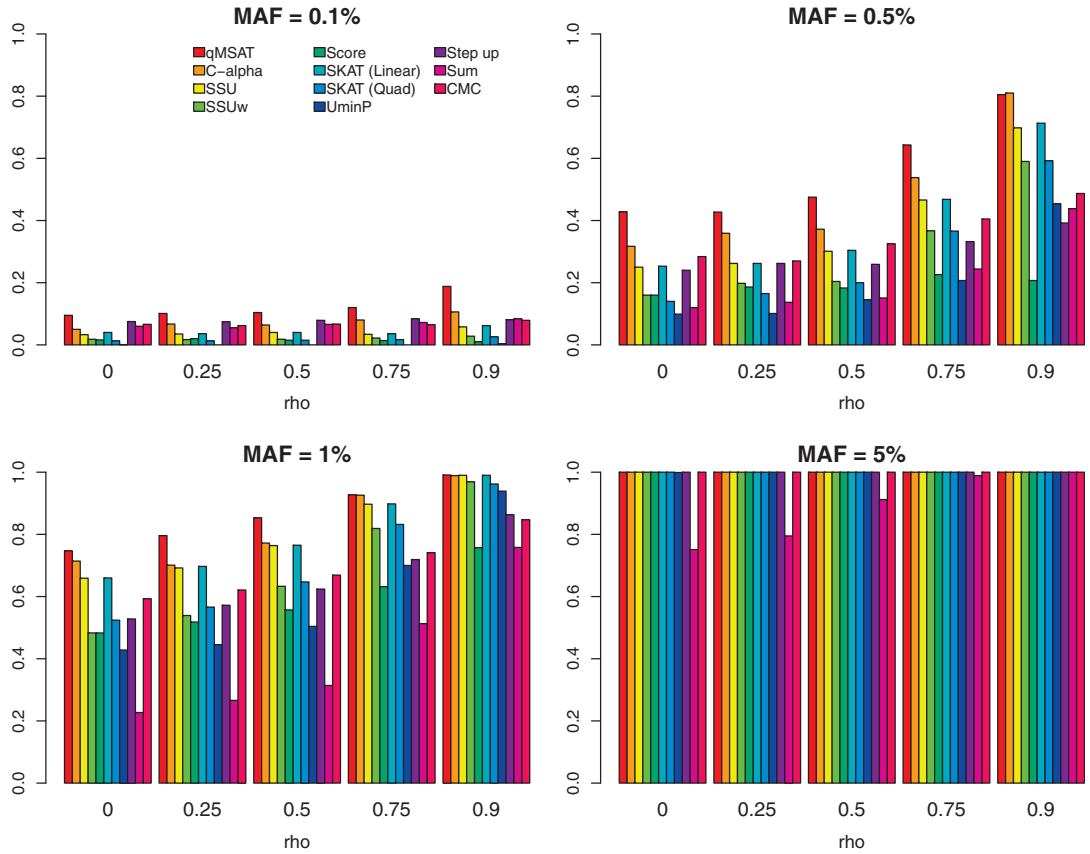

Figure 5. Power evaluation of the effects of linkage disequilibrium. Empirical powers are presented at a mean coverage depth of 30× and sample sizes $n = 500$ with equal numbers of cases and controls for a spectrum of LD structures generated from latent variables having autoregressive covariance matrices $\mathbf{\Sigma} = \rho^{|i-j|}$. Variants generated with $\rho = 0$ have no LD, whereas variants generated with large values of ρ are in high LD. Twenty variants ($d = 20$) are considered having odds ratios of (3, 2, 1, 1, 4, 1, 1, 1, 0.5, 0.5, 1,. . .,1).

Table 1. Multiple rare variants association analysis of the UCSD obesity data

	FAAH promoter ($d=17$)	MGLL promoter ($d=25$)	MGLL 3′-UTR ($d=6$)
qMSAT	0.061	**0.037**	**0.002**
C-alpha	0.077	**0.049**	**0.044**
SSU	0.080	**0.038**	0.246
SSUw	0.145	0.234	0.308
Score	0.146	0.123	0.303
SKAT (Linear)	0.079	**0.038**	0.246
SKAT (Quad)	0.107	0.050	0.250
UminP	0.631	0.211	0.627
Step up	**0.022**	0.063	**0.015**
Sum	**0.002**	0.578	0.052
CMC	**0.015**	0.109	0.052

P-values are provided for association tests performed over d multiple rare variants. FAAH promoter, MGLL promoter and MGLL 3′-UTR are regions exhibiting significant association of multiple rare variants with extreme BMI at the 0.05 level by one or more methods. *P*-values significant at the 0.05 level are boldfaced.

This may support previous studies asserting the sum test to be limited when association of significant variants in a region are in opposite directions, though it might be powerful when directions are the same (25). Due to low MAFs, marginal association tests for rare variants often have low power and can be subjected to perturbations from genotype errors. Both the FAAH promoter and MGLL 3′-UTR, indicated to be significant by several multivariate tests, do not have rare variants exhibiting marginal associations significant at the 0.05 *P*-value level (Supplementary Table S2). This result further suggests the benefit of pooling multiple variants together in identifying rare variants associated with common diseases.

In this application, the qMSAT *P*-values were computed in 25.039 s for all 12 regions considered using 10 000 permutations with the R programming language. In large-scale rare variants association studies involving hundreds of sequenced genes, it is expected to take only a couple of hours. Computing speed can also be greatly improved by using more efficient programming languages, such as C or Fortran. Among the methods evaluated, the step-up procedure has comparatively much slower computational speed, as it has to ascertain whether each variant is protective, deleterious or non-causal.

DISCUSSION

We have proposed a novel procedure that allows the incorporation of sequencing qualities directly in association tests of complex traits and multiple rare variants.

Extensive simulation studies with sequencing qualities sampled directly from real data have been used to evaluate our method across a spectrum of factors of interest. Results have suggested the qMSAT as a potentially very useful method for improving power, especially under moderate sample sizes and relatively low coverage. Moreover, our analysis of the UCSD obesity data set using the qMSAT has identified two regions (MGLL promoter and MGLL 3′-UTR), where multiple rare variants are significantly associated with extreme obesity.

We have not considered the inclusion of additional covariates in order to provide a concise presentation. It is clear that our conclusions regarding the incorporation of sequencing qualities will remain the same whether additional covariates are included or not. The qMSAT can be easily extended to include additional covariates, such as relevant common variants and environmental effects. For dichotomous traits y_i and $\mathbf{X}_i$ a d_c-vector of additional covariates, the qMSAT statistic $qMSAT = \mathbf{S}^T\ \mathbf{S}$ is defined using scores in Equation (4) with

$$\hat{p}_{ij}^0 = \text{logit}^{-1}(\hat{\alpha}_0^{(j)} + \mathbf{X}_i^T\hat{\alpha}^{(j)})$$

where $\hat{\alpha}_0^{(j)}$ and $\hat{\alpha}^{(j)}$ are estimated under the null hypothesis by maximizing the log-likelihood in Equation (2) with $p_{ij} = \text{logit}^{-1}(\alpha_0 + \mathbf{X}_i^T\alpha + \beta_j G_{ij})$ using only the covariates $\mathbf{X}_i$ and quality-based weights q_{1j}, q_{2j},..., q_{nj} at the j-th variant, i.e. $q_{ij'} = 0$ for all $i = 1,\ldots, n$ and $j' \neq j$.

Moreover, we focused on association tests with the logistic regression for case–control studies. This is partly due to the prevalence of case–control data in biomedical researches and the fact that effects of rare variants, having extremely low MAFs, can be more readily modeled with dichotomous traits than continuous ones. However, with the advent of large-scale rare variants association studies in the near future, quantitative traits may also be considered. For continuous traits y_i, the qMSAT statistic is defined with scores

$$S_j = \sum_{i=1}^{n} q_{ij}G_{ij}(y_i - \hat{\alpha}_0^{(j)} - \mathbf{X}_i^T\hat{\alpha}^{(j)}),$$

for which the quality-weighted log-likelihood under the linear model

$$l(\alpha_0, \alpha, \beta) = -\sum_{i=1}^{n}\sum_{j=1}^{d} q_{ij}(y_i - \alpha_0 - \mathbf{X}_i^T\alpha - \beta_j G_{ij})^2$$

is employed in place of Equation (2). The coefficients $\hat{\alpha}_0^{(j)}$ and $\hat{\alpha}^{(j)}$ are estimated under the null hypothesis by weighted least squares regression of y_i on additional covariates $\mathbf{X}_i$ only with quality-based weights q_{1j}, q_{2j},..., q_{nj} at the j-th variant. The same benefits demonstrated for the qMSAT for case–control data are expected to extend to applications involving quantitative traits.

We have not considered association studies simultaneously involving both rare and common variants. A common strategy proposed in the literature is to re-weight each variant according to its MAF, augmenting effects of rare variants relative to common ones. However, it is not clear which is the best way to re-weight variants objectively. For example, re-weighting by standard deviations, control MAFs (51), and beta transformations of MAFs (21) have been proposed. As rare variants usually act on disease risks independently of common ones (6,15), a natural approach is to first identify common variants that may be causal, such as by routine GWA studies, and then incorporate them as additional covariates $\mathbf{X}_i$ as in the qMSAT.

In this article, we have utilized genotype quality scores as weights in the qMSAT. These scores can account for errors, such as those arising from base calling in sequencer traces and mapping of short sequencing reads, and can quantify genotype uncertainty accurately at most loci. However, variant calling at some loci may suffer due to artifacts, such as from strand bias, that cannot be easily accounted for and may lead to less accurate genotype quality scores. Nonetheless, current variant calling programs usually can filter out the variants at these suspicious loci (26,27,52). In the situation when genotype qualities are less accurate, the qMSAT may be less powerful while type I errors are expected to be controlled due to permutation (36). Nevertheless, our results demonstrate that the incorporation of sequencing qualities can consistently improve power over methods that do not incorporate quality information. Moreover, the development of variant calling methods is an active area (26,27,52,53), and our procedure will continue to benefit from future improvements in variant calling programs.

On the other hand, the qMSAT is quite flexible and can incorporate other forms of weights, in addition to quality information. For example, if there is evidence to believe that some variants are more likely to be functional, such as by prior knowledge or evolutionary models (54), or some individuals are more representative of the effected population, the qMSAT weights q_{ij} can be redefined accordingly in order to incorporate additional information at each variant or individual. Indeed, the qMSAT extends earlier methods with weights on each variant (24,51,55) to allow varying weights on both individuals and variants.

Statistical analysis of sequencing data is ultimately limited by inadequacies in the underlying technology employed. With the qMSAT, we approached this challenge by incorporating sequencing quality, a technology-derived constituent, directly in a statistical procedure. The qMSAT also offers a coherent strategy for integrating missing genotypes and individual genotypes sequenced at varying coverage depths and quality scores. Improvements in performances that we observed for the qMSAT in both empirical and real-data studies may suggest the importance of considering the limitations of the technology at hand in order to improve statistical analysis of genomic data sets.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables S1–S2 and Supplementary Figures S1–S10.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous referees for their very generous comments and improvements made in the article based on their suggestions.

FUNDING

National Institutes of Health (grant numbers CA127334, ES009911). Funding for open access charge: Waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Hindorff,L.A., Junkins,H.A., Hall,P.N., Mehta,J.P. and Manolio,T.A. (2011) *A catalog of published genome-wide association studies*., Available at: www.genome.gov/gwastudies. Accessed July 15, 2011.
2. Maher,B. (2008) Personal genomes: the case of the missing heritability. *Nature*, **456**, 18–21.
3. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
4. Bodmer,W. and Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.
5. Kryukov,G.V., Pennacchio,L.A. and Sunyaev,S.R. (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, **80**, 727–739.
6. Pritchard,J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.
7. Ahituv,N., Kavaslar,N., Schackwitz,W., Ustaszewska,A., Martin,J., Hebert,S., Doelle,H., Ersoy,B., Kryukov,G., Schmidt,S. *et al.* (2007) Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.*, **80**, 779–791.
8. Cohen,J.C., Kiss,R.S., Pertsemlidis,A., Marcel,Y.L., McPherson,R. and Hobbs,H.H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.
9. Cohen,J.C., Boerwinkle,E. Jr, T.H.M. and Hobbs,H.H. (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.*, **354**, 1264–1272.
10. Ji,W., Foo,J.N., O'Roak,B.J., Zhao,H., Larson,M.G., Simon,D.B., Newton-Cheh,C., State,M.W., Levy,D. and Lifton,R.P. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.
11. Nejentsev,S., Walker,N., Riches,D., Egholm,M. and Todd,J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.
12. Romeo,S., Pennacchio,L.A., Fu,Y., Boerwinkle,E., Tybjaerg-Hansen,A., Hobbs,H.H. and Cohen,J.C. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513–516.
13. Romeo,S., Yin,W., Kozlitina,J., Pennacchio,L.A., Boerwinkle,E., Hobbs,H.H. and Cohen,J.C. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.*, **119**, 70–79.
14. Gorlov,I.P., Gorlova,O.Y., Sunyaev,S.R., Spitz,M.R. and Amos,C.I. (2008) Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **82**, 100–112.
15. Pritchard,J.K. and Cox,N.J. (2002) The allelic architecture of human disease genes: common disease-common variant… or not? *Hum. Mol. Genet.*, **11**, 2417–2423.
16. Fearnhead,N.S., Wilding,J.L., Winney,B., Tonks,S., Bartlett,S., Bicknell,D.C., Tomlinson,I.P., Mortensen,N.J. and Bodmer,W.F. (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl Acad. Sci. USA*, **101**, 15992–15997.
17. Morgenthaler,S. and Thilly,W.G. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, **615**, 28–56.
18. Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.
19. Hoffmann,T.J., Marini,N.J. and Witte,J.S. (2010) Comprehensive approach to analyzing rare genetic variants. *PLoS One*, **5**, e1001289.
20. Neale,B.M., Rivas,M.A., Voight,B.F., Altshuler,D., Devlin,B., Orho-Melander,M., Kathiresan,S., Purcell,S.M., Roeder,K. and Daly,M.J. (2011) Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e1001322.
21. Wu,M.C., Lee,S., Cai,T., Li,Y., Boehnke,M. and Lin,X. (2011) Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am. J. Hum. Genet.*, **89**, 82–93.
22. Conneely,K.N. and Boehnke,M. (2007) So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *Am. J. Hum. Genet.*, **81**, 1158–1168.
23. Chapman,J. and Whittaker,J. (2008) Analysis of multiple SNPs in a candidate gene or region. *Genet. Epidemiol.*, **32**, 560–566.
24. Pan,W. (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.*, **33**, 497–507.
25. Basu,S. and Pan,W. (2011) Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.*, **35**, 606–619.
26. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S. *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
27. Wei,Z., Wang,W., Hu,P., Lyon,G.J. and Hakonarson,H. (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.*, **39**, e132.
28. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
29. Lin,D.Y. and Tang,Z.Z. (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.*, **89**, 354–367.
30. , The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
31. Li,Y., Sidore,C., Kang,H.M., Boehnke,M. and Abecasis,G.R. (2011) Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.
32. Ruppert,D. and Wand,M.P. (1994) Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**, 1346–1370.
33. Zellner,A. (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.*, **57**, 348–368.
34. Carroll,R.J. and Ruppert,D. (1988) *Transformation and Weighting in Regression*. Chapman and Hall/CRC, New York.
35. Rao,C.R. (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Camb. Phil. Soc.*, **44**, 50–57.
36. Churchill,G.A. and Doerge,R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.
37. Wang,T. and Elston,R.C. (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.*, **80**, 353–360.
38. Prabhu,S. and Pe'er,I. (2009) Overlapping pools for high-throughput targeted resequencing. *Genome Res.*, **19**, 1254–1261.

39. Sarin,S., Prabhu,S., O'Meara,M.M., Pe'er,I. and Hobert,O. (2008) Caenorhabditis elegans mutant allele identification by whole-genome sequencing. *Nat. Methods*, **5**, 865–867.
40. Pan,W. (2011) Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.*, **35**, 211–216.
41. Bhatia,G., Bansal,O., Harismendy,O., Schork,N.J., Topol,E.J., Frazer,K. and Bafna,V. (2010) A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput. Biol.*, **6**, e1000954.
42. Harismendy,O., Bansal,V., Bhatia,G., Nakano,M., Scott,M., Wang,X., Dib,C., Turlotte,E., Sipe,J.C., Murray,S.S. *et al.* (2010) Population sequencing of two endocannabinoid metabolic genes identifies rare and common regulatory variants associated with extreme obesity and metabolite level. *Genome Biol.*, **11**, R118.
43. Rodriguez de Fonseca,F., Del Arco,I., Bermudez-Silva,F.J., Bilbao,A., Cippitelli,A. and Navarro,M. (2005) The endocannabinoid system: physiology and parmacology. *Alcohol Alcohol*, **40**, 2–14.
44. Walker,J.M., Krey,J.F., Chu,C.J. and Huang,S.M. (2002) Endocannabinoids and related fatty acid derivatives in pain modulation. *Chem. Phys. Lipids*, **121**, 159–172.
45. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595, Mar.
46. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
47. Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **8**, 677–679.
48. Bai,Z. and Saranadasa,H. (1996) Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, **6**, 311–329.
49. Chen,S.X. and Qin,Y.L. (2010) A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808–835.
50. Cooper,G.M. and Hausman,R.E. (2007) *The Cell A Molecular Approach*, 4th edn. Sinauer Associates Inc., Washington, DC, p. 326.
51. Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.
52. Meacham,F., Boffelli,D., Dhahbi,J., Martin,D.I.K., Singer,M. and Pachter,L. (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**.
53. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
54. King,C.R., Rathouz,P.J. and Nicolae,D.L. (2010) An evolutionary framework for association testing in resequencing studies. *PLoS Genet.*, **6**, e1001202.
55. Kwee,L.C., Liu,D., Lin,X., Ghosh,D. and Epstein,M.P. (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**, 386–397.