Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data

Bingshan Li,[1] and Suzanne M. Leal[1,*]

Although whole-genome association studies using tagSNPs are a powerful approach for detecting common variants, they are underpowered for detecting associations with rare variants. Recent studies have demonstrated that common diseases can be due to functional variants with a wide spectrum of allele frequencies, ranging from rare to common. An effective way to identify rare variants is through direct sequencing. The development of cost-effective sequencing technologies enables association studies to use sequence data from candidate genes and, in the future, from the entire genome. Although methods used for analysis of common variants are applicable to sequence data, their performance might not be optimal. In this study, it is shown that the collapsing method, which involves collapsing genotypes across variants and applying a univariate test, is powerful for analyzing rare variants, whereas multivariate analysis is robust against inclusion of noncausal variants. Both methods are superior to analyzing each variant individually with univariate tests. In order to unify the advantages of both collapsing and multiple-marker tests, we developed the Combined Multivariate and Collapsing (CMC) method and demonstrated that the CMC method is both powerful and robust. The CMC method can be applied to either candidate-gene or whole-genome sequence data.

Introduction

For the mapping of common disease susceptibility genes, hundreds of thousands of SNPs are genotyped to facilitate genome-wide association studies in either family- or population-based data. In order for this study design to be successful, the common disease common variant (CDCV) hypothesis must hold true. The CDCV hypothesis asserts that common diseases are caused by common variants with small to modest effects.[1–4] This is currently the most popular theory underlying complex-disease etiology. A well-known example supporting this hypothesis is the APOE gene, in which a single common allele ($\varepsilon 4$) confers high risk of Alzheimer disease and heart disease.[5]

The HapMap project and advances in large-scale SNP genotyping facilitate the identification of disease-susceptibility genes through indirect linkage disequilibrium (LD) mapping. The nonrandom association (i.e., LD) of SNPs is appealing for disease-gene mapping, because a subset of SNPs (tagSNPs) can capture the information of correlated SNPs that are not genotyped, thus vastly reducing the number of SNPs that need to be genotyped for an association study when the CDCV hypothesis holds.[1,6,7] An alternative theory is the common disease rare variant (CDRV) hypothesis, which states that for complex traits there is extreme allelic heterogeneity and that disease etiology is caused collectively by multiple rare variants with moderate to high penetrances.[2,4] Studies based on evolution theories have demonstrated that for complex diseases, allelic heterogeneity might be extensive, with multiple susceptibility alleles of independent origin.[8,9] Analysis based on HapMap data has illustrated that rare variants are more likely to be disease predisposing than are common variants.[10] There is also empirical evidence supporting this hypothesis; e.g., multiple rare variants have also been recently identified to be associated with low plasma levels of HDL cholesterol,[11–15] obesity (MIM 601665),[16] colorectal adenomas (MIM 608456),[17] and schizophrenia (MIM 181500).[18] Although there is substantial evidence that both the CDRV and the CDCV hypotheses are valid, probably a more realistic model for complex traits is that functional variants have a wide spectrum of allele frequencies, which range from rare to common even within the same susceptibility gene.[2]

Recent association studies have been successful for a number of traits, such as age-related macular degeneration (AMD [MIM 603075])[19,20] and Crohn disease (MIM 266600).[21] However, critical assumptions for the efficient detection of associations through LD mapping are that for a specific susceptibility locus there is only low-level allelic heterogeneity and that the variants are common.[6,22] In the presence of allelic heterogeneity, although the power of linkage analysis is not influenced, association studies based on LD mapping will inevitably be low-powered.[10,23] Low frequencies of functional variants result in low r^2 values, with tagSNPs of $\geq$ 5% frequency, and therefore, the power of the indirect LD-mapping approach is low. Alternative approaches are necessary to efficiently identify loci with extreme allelic heterogeneity, i.e., multiple rare variants. Directly sequencing candidate genes—or, in the future, entire genomes—instead of genotyping tagSNPs is an optimal approach for the identification of rare variants associated with disease susceptibility.[13] Recently, candidate-gene resequencing was employed to discover variants in the population for the association of complex traits.[11–17] A major sequencing effort is currently

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
*Correspondence: sleal@bcm.edu
DOI 10.1016/j.ajhg.2008.06.024.

being carried out by an international consortium to sequence at least 1000 genomes, in order to produce the most detailed map of human genetic variations for the support of disease studies (1000 Genomes Project, International Consortium).

Statistical methods for the detection of associations of common variants have been extensively developed and successively applied to numerous studies of complex traits. However, methods for statistical analysis of rare variants are limited. Some methods used for analysis of common variants are readily applicable to rare variants, but their performance may not be optimal. In the next few years, sequencing technology (e.g., 454 and Solexa) will enable the production of large quantities of sequence data on large numbers of individuals and allow for the cost-effective identification of rare variants. This data will enable researchers to investigate the role that rare variants play in disease etiology. In addition to uncovering functional variants, sequence data will also reveal many variants that are not functional. Bioinformatics tools[24] can be used to classify variants as functional or nonfunctional or to quantify the functionality of the variants.

In this article, new methods for the analysis of sequencing data, which are robust and powerful in the presence of allelic heterogeneity and low allele frequencies, are developed, and their performance is evaluated. Although understanding the effect of individual rare variants is ultimately important, an effective first approach is to identify the genes that are involved in the disease etiology. One approach is the single-marker test, whereby individual variant sites within a gene are tested for an association with the disease outcome, with standard univariate statistical tests used (e.g., χ^2 test, Fisher's exact test, or Cochran Armitage test for trend) and with the family-wise error rate (FWER) controlled by a multiple-comparison correction (e.g., Bonferroni, permutation). Another approach is to perform a multiple-marker test, which tests multiple variant sites simultaneously with the use of multivariate methods, such as the Fisher product method,[25] Hotelling's T^2 test,[26,27] or logistic regression. Both single-marker and multiple-marker tests involve multiplicity (i.e., multiple-testing correction or multiple degrees of freedom), which will reduce power. On the other hand, collapsing methods, which combine information across multiple variant sites, could enrich the association signals and at the same time reduce the number of the test's degrees of freedom. However, collapsing nonfunctional variants together with functional variants could adversely affect power. In this article, the performance of single-marker tests, multiple-marker tests, and collapsing methods are investigated analytically and empirically. Additionally, the effects of misclassification on power are evaluated. Misclassification can occur when noncausal variants are included in the analysis or when functional variants are excluded from the analysis because the region has not been sequenced or the variants are falsely deemed nonfunctional through bioinformatics tools. It is demonstrated that collapsing methods are potentially more powerful than are single-marker and multiple-marker tests; however, collapsing methods are not always robust to misclassification of nonfunctional variants, and power loss can be substantial. Although they are less powerful than collapsing methods, multivariate tests are more robust in the presence of misclassification of nonfunctional variants. In order to unify the advantages of both collapsing and multiple-marker tests, the Combined Multivariate and Collapsing (CMC) method is developed. This CMC method is shown to be both powerful and robust against misclassification.

Material and Methods

In this article, both analytical and empirical results are presented. Simulations were used for empirical evaluation of type I error and the effect of LD on power; all other power calculations were carried out analytically. Although approximations of prevalence and wild-type penetrance are described here for easier interpretation, only exact analytical calculations were implemented.

Genetic Model

Assume that within a locus there are M variants that can independently cause disease susceptibility. The term "locus" refers to the unit in which the variants will be collectively analyzed. The variants can reside within a gene or a single genomic region. Usually, rare mutations occur on different haplotypes within a locus[8,9] and, therefore, correlation between variants is low. For the analytical calculations, it is assumed that variants are independent. Each of the variants has two alleles, denoted as A_i and a_i, $i = 1,2,\ldots,M$, in which A_i is the rare and high-risk allele and has an allele frequency of p_i. The total frequency of the rare variants in a locus is $p = \sum_{i=1}^{M} p_i$. Let G_k, $k = 0,1,2$ denote the genotypes aa, Aa, and AA, respectively. The genotype frequencies under Hardy-Weinberg Equilibrium (HWE) at the i^{th} variant site are $p_i(G_0) = (1-p_i)^2$, $p_i(G_1) = 2p_i(1-p_i)$ and $p_i(G_2) = p_i^2$. Let the penetrances of genotypes at the i^{th} variant site be represented by f_{ki} for genotypes G_k, $k = 0,1,2$. The locus wild-type penetrance, denoted by f_0, is the probability of an individual being affected if the genotypes across all variant sites are wild-type aa. The overall and individual wild-type penetrances satisfy $f_0 = 1 - \prod_{i=1}^{M}(1-f_{0i})$. For low wild-type penetrances at individual variant sites, the higher-order product terms can be ignored, and the relationship can be approximated by $f_0 = \sum_{i=1}^{M} f_{0i}$. If the assumption is made that wild-type genotypes at different sites have the same penetrance, the relationship can be simplified to $f_0 = Mf_{0i}$. The locus relative risk (RR) at the i^{th} variant site is defined as $\gamma_{1i} = f_{1i}/f_0$, $\gamma_{2i} = f_{2i}/f_0$. For the additive model, $\gamma_{2i} = 2\gamma_{1i} - 1$; for the multiplicative model, $\gamma_{2i} = \gamma_{1i}^2$; for the dominant model, $\gamma_{2i} = \gamma_{1i}$; for the recessive model, $\gamma_{1i} = 1$. The prevalence of the disease caused by each individual variant is calculated as

$$K_i = \sum_{k=0}^{2} p_i(G_k) f_{ki}.$$

Under the heterogeneity model, the prevalence caused by the entire locus is given by

$$K = 1 - \prod_{i=1}^{M} (1 - K_i).$$

If individual prevalences due to a single variant are low, the higher-order product terms can be ignored and the total

prevalence can be approximated by the sum of the individual prevalences: $K = \sum_{i=1}^{M} K_i$.

As a result of allelic heterogeneity, affected individuals can have the same phenotype due to different causal variants. The proportion of individuals affected as a result of the i^{th} variant in the ascertained cases is given by

$$\pi_i = \frac{K_i}{\sum_{j=1}^{M} K_j}.$$

Individuals with diseases due to the i^{th} variant are members of the i^{th} "group," with a total of M groups in the ascertained cases, and the relative sample size of the i^{th} group is π_i. For the i^{th} group, the expected genotype frequency at the i^{th} variant site in cases is

$$p_i(G_k \mid g_i) = \frac{p_i(G_k) f_{ki}}{K_i},\ k = 0,1,2.$$

The expected frequency of genotype G_k at the i^{th} variant site across all M groups in cases is given by

$$p_i^D(G_k) = \pi_i p(G_k \mid g_i) + (1 - \pi_i) p(G_k),\ k = 0,1,2.$$

The controls are disease-free, and the expected genotype frequencies at the i^{th} variant site in controls is given by

$$p_i^N(G_k) = \frac{p_i(G_k)(1 - f_{ki})}{1 - K_i},\ k = 0,1,2.$$

Information on the expected genotype frequencies at each variant site in the sample can be used for various methods to analytically calculate the power to detect an association. The focus in this article is the omnibus test, which provides an association test of the entire locus and is not focused on any specific variant within the locus.

Single-Marker Test

One approach of association studies is to test each variant site individually with the use of a univariate test and assess the significance of the omnibus test after correction for multiple comparisons. For univariate tests, a 2×3 contingency table can be constructed to compare genotype frequencies at each variant site in cases and controls. Because an observation of individuals that are homozygous for the high-risk rare allele is extremely rare, *AA* genotypes are collapsed with *Aa* genotypes, and a 2×2 table is constructed. For an equal number of cases and controls, $N_A = N_{\bar{A}} = N$, the classical Pearson χ^2 statistic[28] for testing equal genotype frequencies in cases and controls is given by

$$X_i^2 = N\left\{\frac{\left[\widehat{p}_i^D(aa) - \widehat{p}_i^N(aa)\right]^2}{\widehat{p}_i^D(aa) + \widehat{p}_i^N(aa)} + \frac{\left[\widehat{p}_i^D(Aa) + \widehat{p}_i^D(AA) - \widehat{p}_i^N(Aa) - \widehat{p}_i^N(AA)\right]^2}{\widehat{p}_i^D(Aa) + \widehat{p}_i^D(AA) + \widehat{p}_i^N(Aa) + \widehat{p}_i^N(AA)}\right\}$$

in which each $\widehat{p}_i$ is the observed genotype frequency at the i^{th} variant site in cases and controls. The power of the test is dependent on the noncentrality parameter (NCP), denoted as ν_i, of a noncentral χ_1^2 distribution, and the NCP is given by

$$\nu_i = N\left\{\frac{\left[p_i^D(aa) - p_i^N(aa)\right]^2}{p_i^D(aa) + p_i^N(aa)} + \frac{\left[p_i^D(Aa) + p_i^D(AA) - p_i^N(Aa) - p_i^N(AA)\right]^2}{p_i^D(Aa) + p_i^D(AA) + p_i^N(Aa) + p_i^N(AA)}\right\}.$$

The power to detect an association at the i^{th} variant site at level α is

$$\eta_i = Pr\left(\chi_1^2(\nu_i) \geq \chi_{1,1-\alpha}^2\right).$$

Because M tests are performed at M variant sites, it is necessary to correct for multiple comparisons in order to control the FWER. Because all rare variants are assumed to be independent, a Bonferroni correction is used, and after controlling for the FWER, the power of the i^{th} test is

$$\eta_i^B = Pr\left(\chi_1^2(\nu_i) \geq \chi_{1,1-\alpha/M}^2\right).$$

The power of the omnibus test for the locus is given by

$$\eta_S = 1 - \prod_{i}^{M} (1 - \eta_i^B).$$

Multiple-Marker Test

Another approach for the study of association is to test all variants simultaneously with the use of a multivariate test; e.g., the Fisher product method, Hotelling's T^2 test, or multiple logistic regression. Hotelling's T^2 test is used as an example of multivariate tests, and the power is calculated analytically for the analysis of rare variants. Following Xiong et al.,[27] an indicator variable is defined for the genotype at the i^{th} variant site for the j^{th} individual in the case population:

$$X_{ji} = \begin{cases} 1 & \text{Genotype is AA} \\ 0 & \text{Genotype is Aa} \\ -1 & \text{Genotype is aa} \end{cases}$$

Similarly, Y_{ji} is defined for the control population. Let $X_j = (X_{j1},\ldots,X_{jM})^T$, $Y_j = (Y_{j1},\ldots, Y_{jM})^T$. Then $\overline{X}_i = 1/N_A \sum_{j=1}^{N_A} X_{ji}$, $\overline{Y}_i = 1/N_{\bar{A}} \sum_{j=1}^{N_{\bar{A}}} Y_{ji}$ and $\overline{X} = (\overline{X}_1,\ldots,\overline{X}_M)^T$, $\overline{Y} = (\overline{Y}_1,\ldots,\overline{Y}_M)^T$. The covariance matrix of the pooled sample for the indicator variables across M variants is given by

$$S = \frac{1}{N_A + N_{\bar{A}} - 2}\left\{\sum_{j=1}^{N_A}(X_j - \overline{X})(X_j - \overline{X})^T + \sum_{j=1}^{N_{\bar{A}}}(Y_j - \overline{Y})(Y_j - \overline{Y})^T\right\}.$$

Hotelling's T^2 statistic is defined as

$$T^2 = \frac{N_A N_{\bar{A}}}{N_A + N_{\bar{A}}}(\overline{X} - \overline{Y})^T S^{-1}(\overline{X} - \overline{Y}).$$

Under the null hypothesis that none of the variants is associated with disease susceptibility, for a large sample size of cases and controls,

$$\frac{N_A + N_{\bar{A}} - M - 1}{M(N_A + N_{\bar{A}} - 2)} T^2$$

is asymptotically distributed as an F distribution, with M and $N_A + N_{\bar{A}} - M - 1$ degrees of freedom. Under the alternative hypothesis that at least one of the variants is associated with the disease, the T^2 statistic is asymptotically distributed as a noncentral χ_M^2 distribution, with M degrees of freedom, and the NCP is given by

$$\nu_H = \mu^T\left(\frac{1}{N_A}\sum\nolimits_A + \frac{1}{N_{\bar{A}}}\sum\nolimits_{\bar{A}}\right)^{-1}\mu,$$

in which μ is the vector of expected difference between cases and controls, $\mu = (\mu_1,\ldots,\mu_M)^T$, and $\mu_i = E[\overline{X}_i] - E[\overline{Y}_i]$. The covariance matrices, Σ_A for cases and $\Sigma_{\bar{A}}$ for controls, can be simplified under the assumption of independence of the rare variants. The i^{th} diagonal

element of the matrix is the variance of the indicator variable at the i^{th} variant site, and off-diagonal elements of the matrix are zero. From the expected genotype frequencies at each variant site, $p_i^D(G_k)$ for cases and $p_i^N(G_k)$ for controls, μ, Σ_A, and $\Sigma_{\bar{A}}$ can be calculated, and the power to detect an association for at least one variant is given by

$$\eta_H = Pr\left(\chi_M^2(\nu_H) \geq \chi_{M,1-\alpha}^2\right).$$

Collapsing Method

Given that single-marker tests involve correcting for multiple comparisons and that multiple-marker tests can have a large number of degrees of freedom, another approach, which collapses the genotypes across variants and results in enriched signals and a reduced number of degrees of freedom, is proposed.

For this method, define an indicator variable X for the j^{th} case individual as

$$X_j = \begin{cases} 1 & \text{rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

Y_j is similarly defined for control individuals. Due to the rarity of variants, the probability of carrying more than one variant for an individual is low, and the method collapses genotypes across all variants, such that an individual is coded as 1 if a rare allele is present at any of the variant sites and as 0 otherwise. The detection of an association of multiple rare variants is transformed into a test of whether the proportions of individuals with rare variants in cases and controls differ. Let ϕ_A and $\phi_{\bar{A}}$ denote the frequencies of individuals carrying rare variants, in cases and controls, respectively. The probability of no variants at all sites in the i^{th} group in cases is given by

$$\lambda_i = p_i(aa\,|\,g_i)\prod_{j\neq i} p_j(aa). \tag{1}$$

Summing over all groups, the proportion of individuals with at least one variant in cases is given by

$$\phi_A = 1 - \sum_{i=1}^{M}(\pi_i\lambda_i).$$

In controls, the probability of carrying no variants at all M sites is $\prod_{i=1}^{M} p_i^N(aa)$, and therefore, the proportion of rare-variant carriers in controls is given by

$$\phi_{\bar{A}} = 1 - \prod_{i=1}^{M} p_i^N(aa). \tag{2}$$

The classic Pearson χ^2 statistic can be used to test the null hypothesis that $\phi_A = \phi_{\bar{A}}$, and the NCP of the noncentral χ_1^2 distribution is

$$\nu_c = N\left[\frac{(\phi_A - \phi_{\bar{A}})^2}{\phi_A + \phi_{\bar{A}}} + \frac{(\phi_A - \phi_{\bar{A}})^2}{2 - \phi_A - \phi_{\bar{A}}}\right].$$

The power of the χ^2 test for the collapsing method is given by

$$\eta_c = Pr\left(\chi_1^2(\nu_c) \geq \chi_{1,1-\alpha}^2\right).$$

CMC Method

The CMC method is a unified approach that combines collapsing and multivariate tests. For the CMC method, markers are divided into subgroups on the basis of predefined criteria (e.g., allele frequencies), and within each group, marker data are collapsed. A multivariate test (e.g., Hotelling's T^2 test) is then applied for analysis of the groups of marker data. Suppose the M markers at the locus are classified into k groups, $\{g_j, j = 1,\ldots,k\}$, and that the number of markers in group g_j is n_j. Within g_j, the n_j markers in the set are collapsed as described in the previous section (Collapsing Method). Collapsing is carried out for each of the groups in the same manner. For those groups in which the number of markers equals 1, no collapsing is necessary. A multivariate test can then be applied to the data, in which within each group the individuals are coded as either 1 (a carrier of one or more variants) or 0 (wild-type). With Hotelling's T^2 test used, the power of the CMC method is calculated on the k dimensional data in the same manner as described in the Multiple-Marker Test section.

Misclassification

Two types of misclassifications are considered: inclusion of nonfunctional variants and exclusion of functional variants. First, consider inclusion of W nonfunctional variants in the analysis. For the single-marker test, the power to detect an association at nonfunctional variant sites is equal to α and the total number of tests is $M + W$. For Hotelling's T^2 test, the mean vector μ and covariance matrices Σ_A and $\Sigma_{\bar{A}}$ are modified by appending the zero vector of length W to the μ vector and adding variances of nonfunctional variants to the diagonal entries of the covariance matrices Σ_A and $\Sigma_{\bar{A}}$. For the collapsing method, the genotype frequencies of nonfunctional variants are included in Equation (1), for cases, and in Equation (2), for controls, to calculate ϕ_A and $\phi_{\bar{A}}$. For the CMC method, the modification is made within each collapsing group and then the power of Hotelling's T^2 test is calculated.

For the case in which T functional variants are excluded from the analysis, the power calculations for the single-marker test, Hotelling's T^2 test, and the collapsing method are carried out in the same manner, except that only $M - T$ out of M variants are analyzed. The number of tests for the single-marker test is $M - T$, and the number of degrees of freedom for Hotelling's T^2 test is $M - T$. The collapsing method remains a univariate test in this situation. For the CMC method, the power of Hotelling's T^2 is calculated on the basis of the modified data within each collapsing group.

Effects of Linkage Disequilibrium

Simulation was used to investigate the effect of LD on power for the single-marker test, Hotelling's T^2 test, and the collapsing method. The locus has six variants, with a total allele frequency of 0.05. Four of the variants have an allele frequency of 0.01 and are on different haplotypes. Each of the remaining two variants, with allele frequencies of 0.005, is on one of the haplotypes where a variant with allele frequency of 0.01 resides; there is complete LD between these variants ($r^2 \approx 0.5$). For comparison purposes, a second simulation was carried out, in which all variants were on separate haplotypes. For generating the data, two haplotypes were randomly sampled and assigned to either case or control status on the basis of an additive model with a locus RR of 2.0, assuming that variants on different sites cause the disease independently. The process was repeated until a sample of 250 cases and 250 controls was obtained, and the single-marker test, Hotelling's T^2 test, and the collapsing method were applied to the generated sample. One thousand replicates were generated, and the power was evaluated for an α level of 0.001.

Table 1. Type I Error Rates at the α Level of 0.05 for Data Analyzed with and without Collapsing

Freq.	No. of Variants	Collapsing and Logistic Reg.	Collapsing and Pearson χ^2	Hotelling's T^2	Logistic Reg.
0.05[a]	5	0.054	0.033	0.045	0.074
	10	0.051	0.032	0.028	0.115
	20	0.048	0.029	0.010	0.204
0.01[b]	5	0.051	0.032	0.044	0.084
	10	0.054	0.034	0.020	0.115
	20	0.054	0.032	0.006	0.191

Type I error rates were evaluated, for a total variant frequency ("Freq.") of 0.05 and 0.01 with 5, 10, and 20 variants. Pearson χ^2 test and logistic regression were applied on the collapsed data. Hotelling's T^2 test and logistic regression were applied on data that were not collapsed. Results are based on 5000 replicates.

[a] Sample size of 250 cases and 250 controls.

[b] Sample size of 1000 cases and 1000 controls.

Table 2. Type I Error Rates at the α Level of 0.05 for the CMC Method

		Minor-Allele Frequency of a High-Frequency Variant			
		0.02		0.05	
Freq.	No. of Variants	Hotelling's T^2	Logistic Reg.	Hotelling's T^2	Logistic Reg.
0.05[a]	5	0.050	0.055	0.051	0.054
	10	0.050	0.055	0.050	0.053
	20	0.043	0.048	0.049	0.052
0.01[b]	5	0.052	0.054	0.055	0.057
	10	0.051	0.052	0.051	0.055
	20	0.049	0.053	0.049	0.050

Within the locus, there is one high-frequency variant with an allele frequency of either 0.02 or 0.05 and 5, 10, or 20 rare variants with a total variant frequency ("Freq.") of 0.05 or 0.01. The CMC method was evaluated with both Hotelling's T^2 test and logistic regression. Results are based on 5000 replicates.

[a] Sample size of 250 cases and 250 controls.

[b] Sample size of 1000 cases and 1000 controls.

Evaluation of Type I Error Rate

In order to evaluate the type I error rate for each test, simulation was used to generate data under the null hypothesis of no association between variants and disease status. Genotypes for each of the M variants within a locus were generated on the basis of population allele frequencies. This sequence of M genotypes was randomly assigned either case or control status. This process was repeated until the desired sample sizes for cases (N_A) and controls ($N_{\bar{A}}$) were obtained for each replicate, and the tests of interest were performed on the data set. This process was repeated for 5000 replicates. It was then evaluated whether or not each replicate had a p value ≤ 0.05. The type I error rate was estimated by the proportion of replicates with a p value ≤ 0.05. A type I error rate > 0.05 signifies a higher false-positive rate, and conversely, a type I error rate < 0.05 indicates a conservative test.

Parameters

In order to evaluate power and type I error rate, total sample sizes of 500 and 2000 were used, with an equal number of cases and controls. For the analysis, total locus variant frequencies of 0.05 and 0.01 were utilized, with each locus composed of 5–20 rare variants with equal or unequal frequencies. The power at the α level of 0.001 was evaluated at the locus RRs of 1.5, 2.0, and 3.0 for the additive model, in which the locus wild-type penetrance $f_0 = 0.01$. For comparison purposes, the power was also calculated at the locus RR of 2.0 for the multiplicative, dominant, and recessive models. Unless otherwise stated, the results are given for a sample size of 250 cases and 250 controls, for a total locus variant frequency of 0.05, with ten variants of equal frequency and a locus RR of 2.0 under the additive model.

Results

Evaluation of Type I Error

The type I error rate is well controlled and slightly conservative for Hotelling's T^2 test and the collapsing method (Table 1). This is not the case when logistic regression is used for the multiple-marker test and the likelihood-ratio test is performed on the basis of an asymptotic χ^2 distribution. Logistic regression is anticonservative, and type I error is inflated. This inflation increases with decreasing allele frequencies (Table 1). For the CMC method, when either the multivariate Hotelling's T^2 test or logistic regression is used for analysis of the data, the type I error is well controlled (Table 2).

Analysis of Functional Variants

For a total locus variant frequency of 0.01 and a locus RR of 2.0, the power is the lowest for the single-marker test, with an increase in power for the multiple-marker test (Hotelling's T^2) and the greatest power observed for the collapsing method (analysis of the collapsed genotypes with the use of the Pearson χ^2 test statistic). When there are ten variants within the locus, the power is 0.05, 0.39, and 0.83 for the single-marker test, Hotelling's T^2 test, and the collapsing method, respectively (Table 3). As the number of variants within the locus is increased from 5 to 20, the power for both the single-marker test and the multiple-marker test decreases but, conversely, the power for the collapsing method increases (Table 3; Figure S1). For example, when the total locus variant frequency is 0.05 and the number of variants is increased from 5 to 20, the power for the single-marker test decreases from 0.14 to 0.02, the power for Hotelling's T^2 test decreases from 0.52 to 0.25, and the power for the collapsing method increases from 0.81 to 0.88. This effect holds when the total locus variant frequency is decreased to 0.01, when one variant's frequency is half of the total variant frequency and the other variant frequencies are equal, and when half of the variants have a locus RR of 3.0 and the remaining variants have a lower locus RR (e.g., 2.0 or 1.5) (Table 3). For these situations, the power of the single-marker test is always the smallest of the three tests. Increasing the frequency of one of the variants to half of the total variant frequency increases the power of the single-marker test, whereas the power for the other tests remains approximately the same (Table 3).

Table 3. The Power of the Single-Marker Test, Hotelling's T^2 Test, and the Collapsing Method

			5			10			20		
Freq.	RR	Model	S	H	C	S	H	C	S	H	C
0.05[a]	2	equal freq.	0.14	0.52	0.81	0.05	0.40	0.86	0.02	0.25	0.88
		unequal freq.[c]	0.24	0.52	0.81	0.20	0.40	0.85	0.17	0.25	0.87
		unequal RR[d]	0.33	0.81	0.96	0.11	0.66	0.97	0.04	0.47	0.97
	1.5	equal freq.	0.06	0.23	0.50	0.03	0.17	0.58	0.01	0.10	0.62
		unequal freq.	0.10	0.23	0.50	0.08	0.17	0.57	0.06	0.10	0.61
		unequal RR	0.32	0.75	0.94	0.10	0.56	0.93	0.03	0.38	0.94
0.01[b]	2	equal freq.	0.13	0.50	0.78	0.05	0.39	0.83	0.02	0.25	0.85
		unequal freq.	0.23	0.50	0.78	0.20	0.39	0.83	0.16	0.25	0.85
		unequal RR	0.34	0.82	0.96	0.11	0.67	0.96	0.04	0.49	0.97
	1.5	equal freq.	0.05	0.20	0.45	0.02	0.15	0.53	0.01	0.08	0.57
		unequal freq.	0.09	0.20	0.45	0.07	0.15	0.53	0.06	0.08	0.56
		unequal RR	0.32	0.75	0.92	0.10	0.56	0.91	0.03	0.38	0.93

The power of the single-marker test (S), Hotelling's T^2 test (H), and the collapsing method (C) when there are 5, 10, or 20 causal variants within each locus that have a total variant frequency of either 0.05 or 0.01. The analytical power is displayed for equal and unequal allele frequencies and for variants having equal and unequal locus RR.

[a] Sample size of 250 cases and 250 controls.

[b] Sample size of 1000 cases and 1000 controls.

[c] One variant was assigned half of the total allele frequency, and the remaining variants have equal frequencies.

[d] Half of the variants were assigned an RR of 3.0, and the remaining variants have an RR of either 2.0 or 1.5.

Misclassification: Excluding Functional Variants

In the situation during which functional variants are excluded from the analysis, the power of the single-marker test remains consistently low, whereas Hotelling's T^2 test and the collapsing method decrease in power with the increasing number of causal variants that are excluded (Table 4, Figure S2). The collapsing method has much greater power than does Hotelling's T^2 test when there are no causal variants missing, but as the proportion of variants excluded from the analysis increases, the power also decreases more dramatically. For a total locus variant frequency of 0.05, consisting of ten causal variants of equal frequency and a locus RR of 2.0 when there are no variants excluded, the power is 0.86 and 0.40 for the collapsing method and Hotelling's T^2 test, respectively. When 20% of the causal variants are excluded, the power falls to 0.72 and 0.31 for the collapsing method and Hotelling's T^2 test, respectively. Even when 60% of the causal variants are excluded, the collapsing method still has greater power than does Hotelling's T^2 test (0.28 versus 0.12).

When high-frequency causal variants (e.g., those with a frequency of 0.02 or 0.05) are excluded from the analysis, the drop in power is most dramatic for the single-marker test and Hotelling's T^2 test. For the single-marker test and Hotelling's T^2 test, the power drops from 0.46 and 0.75 to 0.04 and 0.26, respectively, when a causal variant with a frequency of 0.05 is excluded from the analysis. Although the initial power is greater and the reduction in power is not as large for the collapsing method, the decrease in power is not inconsequential. For example, the power for the collapsing method falls from 0.95 to 0.81 when a functional variant with an allele frequency of 0.02 is excluded from the analysis. The reduction in power is even more dramatic when an allele with a frequency of 0.05 is excluded from the analysis, with the power decreasing from 0.99 to 0.73.

Misclassification: Inclusion of Nonfunctional Variants

When nonfunctional rare variants with the same allele frequencies as those of functional variants are included in the analysis, power decreases for all three tests. The power for the single-marker test is consistently low (Table 4, Figure S2). The power decreases more slowly for Hotelling's T^2 test than for the collapsing method (Table 4, Figure S2). As a result of the higher initial power of the collapsing method, even when 20 nonfunctional rare variants with frequencies of 0.005 are included in the analysis, the power for the collapsing method (0.33) is still greater than the power for Hotelling's T^2 test (0.16) (Table 4, Figure S2).

When one or more high-frequency noncausal variants (e.g., those with a frequency of 0.02 or 0.05) are included in the analysis, the power of the single-marker test remains lower than that of both Hotelling's T^2 test and the collapsing method. For Hotelling's T^2 test, although there is a slight drop in power for each additional noncausal variant included in the analysis, the allele frequency of the noncausal variant does not affect the power of the test. For example, the power of Hotelling's T^2 test is 0.4 when all variants are causal; when a nonfunctional variant is included in the analysis, regardless of its allele frequency, the power drops to 0.38, and the power falls slightly more to 0.36 when two nonfunctional variants are included. This is not the case for collapsing method; the power decreases with the increasing allele frequency of the nonfunctional variant, and the decrease in power is even more drastic when two high-frequency noncausal variants are included in the analysis (Table 5, Figure S3). For the collapsing method, the power decreases from 0.86 to 0.73 when

Table 4. The Power of the Single-Marker Test, Hotelling's T^2 Test, and the Collapsing Method when Noncausal Rare Variants Are Included and Causal Rare Variants Are Excluded

	Total Variant Frequency					
	0.05[a]			0.01[b]		
No. Included	**S**	**H**	**C**	**S**	**H**	**C**
0	0.05	0.40	0.86	0.05	0.39	0.83
5	0.04	0.30	0.70	0.04	0.29	0.68
10	0.03	0.23	0.55	0.03	0.23	0.55
20	0.03	0.16	0.33	0.02	0.15	0.36
No. Excluded	**S**	**H**	**C**	**S**	**H**	**C**
2	0.05	0.31	0.72	0.05	0.30	0.69
4	0.05	0.21	0.52	0.04	0.20	0.50
6	0.04	0.12	0.28	0.04	0.12	0.27
8	0.03	0.05	0.08	0.03	0.05	0.08

The effect of including noncausal variants and excluding causal variants on the power of the single-marker test (S), Hotelling's T^2 test (H), and the collapsing method (C) when there are ten rare causal variants in a gene, with a total variant frequency of either 0.05 or 0.01. In the upper section of the table, also included in the analysis are 5, 10, and 20 rare noncausal variants, with the same allele frequencies as the causal variants. In the lower section of the table, 2, 4, 6, and 8 causal variants are excluded from the analysis.

[a] Sample size of 250 cases and 250 controls.

[b] Sample size of 1000 cases and 1000 controls.

Table 5. The Power of the Single-Marker Test, Hotelling's T^2 Test, the Collapsing Method, and the CMC Method when High-Frequency Causal and Noncausal Variants Are Included in the Analysis

No. of Variants	Freq.	S	H	C	CMC
High-Frequency Functional Variants Included					
0	NA	0.05	0.40	0.86	NA
1	0.02	0.13	0.57	0.95	0.91
1	0.05	0.46	0.75	0.99	0.97
2	0.02	0.16	0.69	0.98	0.93
2	0.05	0.51	0.86	0.99	0.98
High-Frequency Nonfunctional Variants Included					
1	0.02	0.05	0.38	0.73	0.80
1	0.05	0.05	0.38	0.54	0.80
2	0.02	0.05	0.36	0.60	0.74
2	0.05	0.05	0.36	0.32	0.74

The power of the single-marker test (S), Hotelling's T^2 test (H), the collapsing method (C), and the CMC method when there are ten rare causal variants, with a total variant frequency of 0.05, for a sample size of 250 cases and 250 controls. Also included in the analysis are one or two high-frequency causal or noncausal variants with an allele frequency of 0.02 or 0.05.

one noncausal variant with an allele frequency of 0.02 is included in the analysis. The power decreases further, to 0.54, when the noncausal variant's allele frequency is increased to 0.05, and the power reduces further, to 0.32, when two noncausal variants with allele frequencies of 0.05 are included in the analysis.

Power of the CMC Method

Variants that have an allele frequency ≤ 0.01 are collapsed, whereas variants with a frequency of > 0.01 are not collapsed. There is a large increase in power if the CMC method is used when there is misclassification, as compared to the collapsing method, particularly when the allele frequency of the noncausal variant is high. For example, when one noncausal variant with an allele frequency of 0.05 is included in the analysis, the power for the collapsing method, Hotelling's T^2 test, and the CMC method is 0.54, 0.38, and 0.80, respectively (Table 5, Figure S4). Although for the CMC method the allele frequency of the noncausal allele does not affect the power, the power is reduced as additional noncausal variants are included in the analysis. However, the CMC method is still more powerful than both the collapsing method and Hotelling's T^2 test (Table 5, Figure S4). When two high-frequency noncausal variants with allele frequencies of 0.05 are included in the analysis, the power is 0.74 for the CMC method, 0.36 for Hotelling's T^2 test, and 0.32 for the collapsing method (Table 5, Figure S4).

Also evaluated was how much power is lost when the CMC method is used to analyze data in which high-frequency variants included in the analysis are truly functional. It is observed that for the CMC method, when two functional variants are included in the analysis, there is only a slight loss in power as compared to the collapsing method (Table 5, Figure S5). For example, when two causal variants with allele frequencies of 0.05 are included in the analysis, the power for the collapsing method is 0.99. The power drops to 0.98 when the CMC method is used to analyze the data (Table 5, Figure S5).

Effect of Linkage Disequilibrium

In the presence of LD, the power for the single-marker test, Hotelling's T^2 test, and the collapsing method is 0.075, 0.63, and 0.85, respectively. For the example in which the data were generated with each variant on a separate haplotype, the corresponding powers are 0.011, 0.451, and 0.737, respectively.

Discussion

Before statistical analysis of sequence data can be carried out, the first step is quantifying which variants are potentially functional or neutral. Bioinformatics tools[24] such as Polyphen,[29] SIFT,[30] and Evolutionary Trace[31] can be used to classify variants as potentially functional or neutral or to quantify the certainty of the functionality. The results obtained from bioinformatics tools can be used to determine which variants should be included in the analysis. In an ideal situation, all variants that are included in the analysis are functional and no functional variants are excluded.

When there is no misclassification of variants, the single-marker test has the lowest power. Not only does this

test pay a penalty for multiple testing, but it is also affected by the low allele frequency at each variant, where the power for each individual χ^2 test is low. It should be noted that Fisher's exact test should be used instead of the χ^2 test when the expected cell counts are low, in order to avoid inflation of type I error. Because Fisher's exact test is more conservative than the χ^2 test, the power can be even lower than that shown for the χ^2 test. The power for Hotelling's T^2 test is superior to that for the single-marker test but is less powerful than that for the collapsing method. The improvement of power for the collapsing method is due to an enrichment of signals across variants and the single univariate test performed.

Although the highest power is obtained when all variants are correctly classified, it is unrealistic to assume, even when bioinformatics tools are used for classification of functional status, that errors will not occur. Misclassification of rare variants does not have a dramatic effect on power unless the functional status is incorrectly assigned for a substantial number of variants. Retention of power is observed when either rare functional variants are incorrectly removed from the analysis or nonfunctional variants are included in the analysis. The exclusion of rare functional variants has a more striking effect on the reduction of power than does the inclusion of rare nonfunctional variants.

When analyzing rare variants, high allele frequency is not a sufficient basis for excluding variants from the analysis. The allelic spectrum for complex disease is usually unknown; however, a number of studies have demonstrated that alleles with a wide range of frequencies are involved in disease etiology.[11–18,32] For example, for HDL cholesterol it was recently shown that both common and rare variants were responsible for modifying HDL cholesterol levels.[32] If high-frequency functional variants are removed from the analysis, the effect on power can be extremely detrimental, and if high-frequency nonfunctional variants are included in the analysis and the collapsing method is used, the power is also severely weakened. However, with the use of CMC method, which applies a multivariate test (e.g., Hotelling's T^2 test) on the collapsed rare variants and the uncollapsed high-frequency variants, the high power is retained even if the high-frequency variants are nonfunctional. If the high-frequency variant is causal, there is only a slight decrease in power with the use of the CMC method as compared to the collapsing method. Although the allele frequency of 0.01 was used for classification of rare and high-frequency variants, the cutoff is subjective and dependent on the spectrum of the variant frequency within a locus. This cutoff criterion might be too high if the total allele frequency for the functional variants is low (e.g., ≤ 0.01). If a wide spectrum of allele frequencies is observed, several cutoffs can be used for the classification of variants into multiple groups. Variants that have very different allele frequencies should not be collapsed into the same group, in order to avoid a substantial loss of power when misclassification is present.

If within a locus there are both rare and common functional variants, use of the CMC method can increase power, as compared to separate analysis of either the rare variants or the common variants. Although in some circumstances there might be sufficient power to detect an association when a single common causal variant is analyzed, even for a functional variant with allele frequency of ≥ 0.05 the power to detect an association might be low if the genotypic RR is small (e.g., $1.0 \leq \text{RR} \leq 1.2$). In the presence of common variants, it can be advantageous to analyze both common and rare variants simultaneously with the CMC method; including rare variants in the analysis can greatly increase power if the rare variants have high genotypic RRs and are either numerous or not extremely rare. The amount of increase in power with the CMC method will be dependent upon the total minor-allele frequency of the rare variants, the strength of the rare variants' genotypic RRs, and the underlying genetic model.

In this article, it is shown how the CMC method can be used to analyze data on the basis of allele frequencies; e.g., on the basis of high-frequency or rare variants. The CMC method can also be used when classification is made on the basis of certainty of functionality. For example, scores from Polyphen, Evolutionary Trace, or SIFT can be used to group variants into multiple classes depending on user-defined cutoffs that reflect their potential functional role in disease etiology. Even when classification is made on the basis of confidence in functionality, it is still inadvisable to collapse rare and high-frequency variants because, as previously discussed, if functionality classification is incorrect, then a large penalty in power can be incurred.

There is a caveat when collapsing rare variants across multiple markers. When all of the functional variants confer high risk or are protective, collapsing will enrich the signal. However, the signal will be weakened if some variants are protective whereas others increase disease risk. Although this situation is probably uncommon, when prior information is available on high-risk and protective variants it should be taken into account when deciding how to collapse variants, in order to obtain optimal power. The CMC method can be applied when protective and high-risk variants are collapsed separately.

Due to low allele frequencies of rare variants, the probability of individuals who are homozygous for the minor allele being ascertained is extremely low. Therefore, even though the locus RR for the multiplicative model (i.e., $\gamma_{2i} = \gamma_{1i}^2$) is greater than the locus RR for the additive model (i.e., $\gamma_{2i} = 2\gamma_{1i} - 1$), for all of the tests there is little difference in power between these two models for rare variants, with the power for the multiplicative model being slightly higher than the additive model. Similarly, there is only a slight increase in power for the additive model compared to the dominant model (i.e., $\gamma_{2i} = \gamma_{1i}$) (data not shown). The situation is quite different for the recessive model, in which the locus RR $\gamma_{1i} = 1$ and $\gamma_{2i} > 1$.

Due to the rarity of homozygous genotypes for the minor allele, very large sample sizes are necessary for sufficient power under the recessive model. For example, for $\gamma_{2i} = 2.0$, a total locus allele frequency of 0.05 with ten causal variants and an α level of 0.001, a sample size of $>$ 20,000 cases is necessary to obtain a power of 0.8 with the collapsing method.

For rare variants, it is reasonable to assume that within a locus they reside on different haplotypes.[8,9] Under this assumption, the frequency of haplotype $h_{A_1A_2}$ is zero and the LD between two rare variants is $D = -p_1p_2 \approx 0$ and $r^2 = D^2/p_1(1-p_1)p_2(1-p_2) \approx p_1p_2 \approx 0$, in which $h_{A_1A_2}$ is the haplotype of the two variants. Therefore, it is usually reasonable to assume that the variants within a locus are usually independent for power calculation. If this assumption is violated and two functional variants are on the same haplotype, the power is increased, because there is a higher probability of carrying more than one functional variant that increases the probability of an individual being a case. The application and the validity of the single-marker test, Hotelling's T^2 test, the collapsing method, and the CMC method are not altered by the presence of LD. In the absence or presence of LD between rare variants, the collapsing and CMC methods are more powerful than Hotelling's T^2 test and the single-marker test.

A drawback of the described analysis methods is that covariates that could be potential confounders are not easily controlled for in the analysis. It has been demonstrated for association studies that it is important to control for potential confounders, including population stratification.[33] For both the collapsing and CMC methods, this problem can be overcome by implementing logistic regression, in which covariates can be included in the analysis.

For all of the methods that were evaluated, type I error was well controlled, except when logistic regression was implemented to analyze uncollapsed rare variants (Table 1). It is a well-known phenomenon that low cell counts or empty cells can cause numerical instability of the maximum-likelihood estimation.[34] When logistic-regression analysis was applied to collapsed variants or to the CMC method, type I error was well controlled; however, this might not be the case if after collapsing the total allele frequency is still very low. This problem can be circumvented by estimation of empirical p values via permutation or use of exact logistic regression.[35,36]

The collapsing method used in this article was based on whether or not an individual had at least one copy of a rare variant. There are other collapsing methods, involving haplotype reconstruction, that can be used. One method involves testing a 2×3 table, in which individuals are classified as homozygous wild-type, having one or more variants on the same haplotype and the other haplotype containing only wild-type alleles, or having at least two variants on different haplotypes. Another approach is to test a 2×2 table, in which individual haplotypes are classified into having at least one variant or no variants; in this situation, the sample size is 2N. Both of these methods had power similar to that of the collapsing method described in this article (data not shown). It should be noted that for the methods involving haplotype reconstruction, it was assumed that the haplotypes were known. However, in reality, haplotypes are not known with 100% accuracy, and these errors in classification will reduce power.

Although it is not necessary to correct for testing multiple variants within a locus when the described methods are used, if multiple regions are being tested, the FWER should be controlled. The α value that should be used is dependent on the number of tests that will be performed and whether or not these tests are independent. Currently, for whole genome association studies, a p value of 5×10^{-7} or smaller is used for genome-wide significance, and this criterion takes into consideration the correlation of the common SNPs.[37] For genome-wide association studies that use sequence data, a more stringent criterion is necessary because rare variants are not highly correlated. The α level that should be used to sufficiently control type I error for whole-genome sequence data is currently unknown; however, it will be dependent not only on the number of variants that are analyzed but also on how the data is analyzed. For example, a more stringent criterion would be necessary if every variant were analyzed separately, compared to if variants across a locus were analyzed simultaneously. The examples in this article are given for a single locus, and an α level of 0.001 was used. However, if more than one locus is being analyzed, a more stringent α value would have to be used in order to control the FWER.

In this study, the focus is on a locus with multiple variants and the main interest is the association of the locus with the disease phenotype. In addition to allelic heterogeneity, locus heterogeneity will also be involved in the etiology of complex traits. The methods described here are able to detect multiple loci in the case of locus heterogeneity by analyzing individual loci separately. However, the methods are not designed to detect gene × gene interactions. The CMC method is a powerful and robust tool for elucidating the main effects of susceptibility genes that are involved in complex traits, for which the CDRV hypothesis holds true. This method can be implemented with the use of standard statistical software packages and readily applied to candidate-gene sequence data or extended for analysis of whole-genome sequence data.

Supplemental Data

Supplemental Data include five figures and are available with this article online at http://www.ajhg.org/.

Acknowledgments

The work was funded by National Institutes of Health grants R01-DC03594 and R01-NS049130. The authors would like to thank Andrew DeWan and Michael Nothnagel for their useful comments and suggestions.

Received: April 7, 2008
Revised: May 29, 2008
Accepted: June 29, 2008
Published online: August 7, 2008

Web Resources

The URLs for data presented herein are as follows:

The 1000 Genomes Project, International Consortium, http://www.1000genomes.org/

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

References

1. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.
2. Iyengar, S.K., and Elston, R.C. (2007). The genetic basis of complex traits: Rare variants or "common gene, common disease"? Methods Mol. Biol. *376*, 71–84.
3. Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502–510.
4. Smith, D.J., and Lusis, A.J. (2002). The allelic structure of common disease. Hum. Mol. Genet. *11*, 2455–2461.
5. Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., and Pericak-Vance, M.A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science *261*, 921–923.
6. Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. *22*, 139–144.
7. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.
8. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.
9. 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.
10. 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.
11. Brunham, L.R., Singaraja, R.R., and Hayden, M.R. (2006). Variations on a gene: Rare and common variants in ABCA1 and their impact on HDL cholesterol levels and atherosclerosis. Annu. Rev. Nutr. *26*, 105–129.
12. 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.
13. Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc. Natl. Acad. Sci. USA *103*, 1810–1815.
14. 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.
15. Slatter, T.L., Jones, G.T., Williams, M.J., van Rij, A.M., and McCormick, S.P. (2008). Novel rare mutations and promoter haplotypes in ABCA1 contribute to low-HDL-C levels. Clin. Genet. *73*, 179–184.
16. 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.
17. Azzopardi, D., Dallosso, A.R., Eliason, K., Hendrickson, B.C., Jones, N., Rawstorne, E., Colley, J., Moskvina, V., Frye, C., Sampson, J.R., et al. (2008). Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res. *68*, 358–363.
18. Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science *320*, 539–543.
19. Dewan, A., Liu, M., Hartman, S., Zhang, S.S., Liu, D.T., Zhao, C., Tam, P.O., Chan, W.M., Lam, D.S., Snyder, M., et al. (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration. Science *314*, 989–992.
20. Yang, Z., Camp, N.J., Sun, H., Tong, Z., Gibbs, D., Cameron, D.J., Chen, H., Zhao, Y., Pearson, E., Li, X., et al. (2006). A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. Science *314*, 992–993.
21. Mathew, C.G. (2008). New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nat. Rev. Genet. *9*, 9–14.
22. Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.
23. Slager, S.L., Huang, J., and Vieland, V.J. (2000). Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genet. Epidemiol. *18*, 143–156.
24. Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief. Bioinform. *6*, 44–56.
25. Fisher, R.A. (1932). Statistical Methods for Research Workers (London: Oliver and Boyd).
26. Hotelling, H. (1931). The generalization of student's ratio. Ann. Math. Stat. *2*, 360–378.
27. Xiong, M., Zhao, J., and Boerwinkle, E. (2002). Generalized T2 test for genome association studies. Am. J. Hum. Genet. *70*, 1257–1268.
28. Chapman, N.H., and Wijsman, E.M. (1998). Genome screens using linkage disequilibrium tests: Optimal marker characteristics and feasibility. Am. J. Hum. Genet. *63*, 1872–1885.
29. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: Server and survey. Nucleic Acids Res. *30*, 3894–3900.
30. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.
31. Lichtarge, O., Bourne, H.R., and Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. *257*, 342–358.
32. Spirin, V., Schmidt, S., Pertsemlidis, A., Cooper, R.S., Cohen, J.C., and Sunyaev, S.R. (2007). Common Single-Nucleotide Polymorphisms Act in Concert to Affect Plasma Levels of

High-Density Lipoprotein Cholesterol. Am. J. Hum. Genet. *81*, 1298–1303.
33. Cardon, L.R., and Palmer, L.J. (2003). Population stratification and spurious allelic association. Lancet *361*, 598–604.
34. Heinze, G., and Schemper, M. (2002). A solution to the problem of separation in logistic regression. Stat. Med. *21*, 2409–2419.
35. Agresti, A. (2002). Categorical Data Analysis (New Jersey: Wiley).
36. Mehta, C.R., and Patel, N.R. (1995). Exact logistic regression: Theory and examples. Stat. Med. *14*, 2143–2160.
37. Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol. *32*, 227–234.