Published in final edited form as:

Genet Epidemiol. 2012 November ; 36(7): 675–685. doi:10.1002/gepi.21662.

A Fast and Noise-Resilient Approach to Detect Rare-Variant Associations With Deep Sequencing Data for Complex Disorders

Yee Him Cheung¹, Gao Wang², Suzanne M. Leal², and Shuang Wang^{1,*}

¹Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York

²Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

Abstract

Next generation sequencing technology has enabled the paradigm shift in genetic association studies from the common disease/common variant to common disease/rare-variant hypothesis. Analyzing individual rare variants is known to be underpowered; therefore association methods have been developed that aggregate variants across a genetic region, which for exome sequencing is usually a gene. The foreseeable widespread use of whole genome sequencing poses new challenges in statistical analysis. It calls for new rare-variant association methods that are statistically powerful, robust against high levels of noise due to inclusion of noncausal variants, and yet computationally efficient. We propose a simple and powerful statistic that combines the disease-associated P -values of individual variants using a weight that is the inverse of the expected standard deviation of the allele frequencies under the null. This approach, dubbed as Sigma-P method, is extremely robust to the inclusion of a high proportion of noncausal variants and is also powerful when both detrimental and protective variants are present within a genetic region. The performance of the Sigma-P method was tested using simulated data based on realistic population demographic and disease models and its power was compared to several previously published methods. The results demonstrate that this method generally outperforms other rare-variant association methods over a wide range of models. Additionally, sequence data on the *ANGPTL* family of genes from the Dallas Heart Study were tested for associations with nine metabolic traits and both known and novel putative associations were uncovered using the Sigma-P method.

Keywords

rare-variant analysis; whole genome sequencing; complex disorders

INTRODUCTION

Despite the success of genome-wide association studies (GWAS) in identifying hundreds of common variants that are associated with complex diseases and traits, identified common variants to date have small genetic effects and only explain a small fraction of heritability.

*Correspondence to: Shuang Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, Room 630, 722 West 168th Street, New York, NY 10032., sw2206@columbia.edu.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

One popular theory is that a large proportion of the “missing heritability” lies in rare variants (minor allele frequency, $MAF < 1\%$). Only recently, with the advance in high-throughput sequencing technology, such as Illumina HiSeq, ABI SOLiD, and Roche 454 as well as target enrichment methods, the extension of the association studies to rare variants has become possible. The analysis of sequence data has led to many studies demonstrating the involvement of rare variants in a variety of complex disorders, such as type-1 diabetes, obesity, hypertension, and schizophrenia [Ahituv et al., 2007; Ji et al., 2008; Nejentsev et al., 2009; Nyers et al., 2011; Walsh et al., 2008].

The foreseeable widespread use of full genome sequencing data, despite opening up unprecedented opportunities, poses new challenges in statistical analysis. This data necessitates the development of new methods to detect rare-variant associations that are statistically powerful, robust against a high level of noise introduced by the large number of noncausal variants, and yet computationally efficient [Bansal et al., 2010]. Because of the lack of statistical power to detect phenotypic association with individual rare variants, due to their low frequency in the population and the need to correct for multiple testing on a large number of variants, current methods boost the association signal by aggregating the effects of rare variants within a genomic region.

One of the earliest methods that collapses variants across a region but was not specifically developed for rare variants is the cohort allelic sums test proposed by Morgenthaler and Thilly [2007]. The method simply compares the frequencies of variant carriers in cases and controls using standard chi-square or Fisher’s exact test. The combined multivariate and collapsing (CMC) method proposed by Li and Leal [2008], developed specifically to analyze rare-variants, collapses variant sites within subgroups formed by criteria such as MAF and then applies a multivariate regression analysis. Madsen and Browning proposed a scheme that weights variants according to their estimated frequencies in controls, so that rarer variants are up-weighted [Madsen and Browning, 2009]. Price et al. adopted a similar idea of weighting, but incorporated the computational predictions of functional effects into the weights [Price et al., 2010]. Morris and Zeggini [2010] developed tests that accumulate minor alleles across rare variants within a functional unit specifically for quantitative traits. The RareCover method by Bhatia et al. maximizes Pearson’s chi-square statistic over all possible combinations of variants [Bhatia et al., 2010]. The approach is computationally intensive and can impose a large penalty for multiple testing.

To enhance the detection of association in a mixture of protective and detrimental variants, the data-adaptive sum test (aSum) by Han and Pan [2010] flips the coding of a SNP if the maximum likelihood estimate of its marginal regression coefficient is negative. Ionita-Laza et al. proposed a replication-based strategy (RBS) that chooses the maximum of two separate scores: one computed from only the sites with more variants in cases and the other from only the sites with more variants in controls, and then adjusts for multiple testing by permutation [Ionita-Laza et al., 2011].

In Liu and Leal’s kernel-based adaptive cluster (KBAC) method, weights are assigned to multisite genotypes rather than to individual variant sites [Liu and Leal, 2010]. Neale et al.’s C-Alpha method, designed to be sensitive for detecting associations in a mixture of

detrimental and protective variants, contrasts the observed variance with the expected variance assuming a binomial distribution [Neale et al., 2011]. Two more recent methods, Wu et al.'s sequence kernel association test (SKAT) extends the C-Alpha method to allow for weighting by variant frequency and also allows for analysis within a regression framework [Wu et al., 2011]. Additionally, Lin and Tang's developed a score test that maximizes over several different models under a regression framework [Lin and Tang, 2011].

We propose a simple, fast, and effective Sigma-P method that amplifies the signals of association for individual rare-variant sites, as well as guards against noise caused by the inclusion of noncausal sites, which usually surpass the number of causal sites, particularly in deep sequencing data and large genomic regions under study. The method combines the significance of disease-association per variant site weighted by the inverse of the expected standard deviation of its minor allele frequency under the null hypothesis [Fisher, 1925]. A segregation strategy is also incorporated to make the test powerful in situations where both risk and protective variants are present. Another feature of the Sigma-P method that makes it especially attractive is its general framework. The method can be used for both quantitative and qualitative traits by adopting regression techniques in computing the per-site P -values, as well as for situations when adjusting for covariates is necessary. The straightforward approach makes it computationally efficient for studying the effect of a huge number of variant sites. The performance of the Sigma-P method was evaluated using simulated data as well as sequence data from the Dallas Heart Study (DHS). The results suggest that our proposed approach generally outperforms existing mainstream methods over a wide range of disease models. Furthermore, the method is able to detect both known and novel associations within the DHS between various metabolism traits and the *ANGPTL* family of genes.

METHODS

There are four key challenges for analyzing the aggregate effect of rare variants on a phenotype. First, the algorithm should be robust to neutral variants, particularly when the genomic region under study consists of a large proportion of neutral variants. This can occur when testing a biological pathway that involves many exonic regions of multiple genes. Second, the method should preferably be simple and efficient for analyzing a huge number of variant sites, which is particularly important as the volume of data increases substantially for whole genome sequencing. Third, we need a scheme that takes into account the relative significance of each site, as opposed to assuming uniform weights. Fourth, the algorithm should be effective in detecting the association of a mixture of detrimental and protective variants. To address these issues, we propose the Sigma-P method with the following test statistic:

$$S_{\sigma-P}^+ = - \sum_{i=1}^K \xi_i w_i \log p_i$$

where K is the number of variant sites indexed by i , w_i is the weight for site i , and $\log p_i$ is the natural logarithm of the P -value to assess the significance of the disease association for

site i ; ξ_i is an indicator variable that is equal to 1 if site i has an excess of variants in cases than in controls and zero otherwise when testing for risk effect only. On the other hand, rare variants can also be protective. To test for protective effect, ξ_i can be set equal to 1 when site i has an excess of variants in controls than in cases and zero otherwise. The test statistic is then denoted as $S_{\sigma-p}^-$. If the goal is to test for any association regardless of the direction of the effect, the statistic $S_{\sigma-p} = \max(S_{\sigma-p}^+, S_{\sigma-p}^-)$ is used. The P -value of the overall effect of a group of variant sites is then obtained by permutations, which simply shuffle the case/control statuses of the samples and then recompute the test scores for the permuted samples. The permuted P -value is defined as $(k_0 + 1)/(k + 1)$, where k_0 is the number of the k permutations with test scores greater than the one observed.

This approach of combining P -values can effectively guard against statistical noise due to the inclusion of neutral variants, since the strength of association is evaluated for each variant site before aggregation. The succinct method is also computationally efficient and easy to comprehend by clinical scientists. We also note the possible extension of our Sigma-P approach for handling quantitative traits or the adjustment of important covariates and population stratification through the use of generalized linear models (GLMs) in obtaining the P -value p_i for the i th variant site. A more detailed description is given in the Discussion section. Without loss of generality, we focus on the analysis of dichotomous traits in this work and apply the Fisher's exact test based on the hypergeometric distribution to obtain variant level P -values that assess per site significance of the disease association

$$p_i = P(x \geq m_i^A) = \sum_{x=m_i^A}^{m_i} \binom{m_i}{x} \binom{n_i - m_i}{n_i^A - x} / \binom{n_i}{n_i^A}, \quad (1)$$

where x is a random variable of the number rare-variant counts in cases. For the additive model, m_i^A is the variant count in cases, m_i is the total variant count in both cases and controls, n_i^A is the number of alleles in cases, and n_i is the total number of alleles in both cases and controls. If alternatively the dominant/recessive model is assumed, then counts for the heterozygous/homozygous-minor genotypes are substituted for the variant counts.

The discreteness of the hypergeometric distribution, however, leads to conservative P -values, with an expected mean always greater than 0.5. The problem is particularly prominent for rare variants, which have a small number of observations. An effective remedy is to use mid- P -values in which the probability of the observed outcome is multiplied by 0.5, i.e.

$p_i = 0.5P(x = m_i^A) + P(x > m_i^A)$ [Armitage et al., 2002]. The use of mid- P -values has been recommended as the appropriate measure when combining results from several studies [Armitage et al., 2002; Barnard, 1989]. Under the null hypothesis, the average mid- P -value is equal to 0.5. This desirable property makes mid- P -value a good measure for combining P -values from multiple variant sites to assess the overall association. Therefore, two versions

of the proposed statistics are implemented, one using direct P -values from the Fisher's exact test, denoted by $S_{\sigma-P}$, and the other using mid- P -values, denoted by $S_{\sigma-Mid P}$.

Following a similar argument that rarer variants have higher probability to be causal due to purifying selection, we incorporate the same weights proposed by Madsen and Browning [2009], where higher weights are applied to rarer variants

$$w_i = 1/\sqrt{n_i \cdot q_i(1 - q_i)}$$

where $q_i = \frac{m_i^u + 1}{n_i^u + 2}$ is the variant frequency in controls. By adding 1 to the numerator and 2 to

the denominator, it avoids estimates of zero that can cause errors in subsequent calculations. m_i^u and n_i^u denote, respectively, the number of variant alleles and the number of all alleles in controls; n_i , as previously defined, is the number of all alleles in both cases and controls, and can vary due to genotyping errors or missing values. Here, we assume additive inheritance. However, dependent on the assumption of the mode of inheritance, the allele counts can be changed accordingly to counts of heterozygous or homozygous-minor genotypes. Note that other computational predictions of functional effects of each gene can be further incorporated into this weight as multiplicative factors.

SIMULATION STUDIES

SIMULATION MODEL

Sequence data from genes of different lengths were generated based on population demography and mutation/multi-locus selection effects with a random distribution of fitness [Boyko et al., 2008; Eyre-Walker et al., 2006; Kryukov et al., 2009; Williamson et al., 2005]. The generation of sequence data was performed using SimPoP [Peng and Kimmel, 2005] that implements the *srv* [Peng and Liu, 2010] module. Assuming the parameters described by Boyko et al. [2008], a MAF distribution in the coding region was generated by the forward-time simulation of a European population that consists of 100,000 haplotypes. We defined rare variants as having $MAF < 1\%$ in the population and nonsynonymous variants as having selection coefficient $S > 10^{-4}$. The MAF distribution of nonsynonymous variants is shown in Figure S1. On average, there are about 50 non-synonymous variant sites in a 3 kb coding region within the haplotype pool.

We randomly classified nonsynonymous rare variants in a coding region as detrimental, protective, or neutral according to a predefined proportion. The impact of increasing the ratio of detrimental-to-neutral variants from 1:1 to 1:4 was examined. For each ratio, the number of detrimental variants was kept approximately constant (~25) while enlarging the coding region from 3 kb to 7.5 kb, thus increasing the number of neutral variants from ~25 to ~100. We then studied the impact of doubling the coding region length while maintaining the same ratios of detrimental-to-neutral variants as before. Lastly, we considered the scenarios with an equal average number of detrimental and protective variants but an increasing number of neutral variants.

In simulating the phenotype data we drew upon Risch, which described the modeling of the penetrance for a multilocus genotype [Risch, 1990], and Bodmer and Bonilla who discussed complex traits with rare-variant effect sizes [2008]. This haplotype-based disease model is a realistic approach for simulating the cases of complex disorders in which the risk increases with the number of causal variants carried by an individual. We specified a group population attributable risk (PAR) for the rare variants within a genomic region of interest. The marginal PAR of a variant α was the group PAR divided by the number of causal rare-variant sites in the gene. The odds ratio (OR) of a detrimental heterozygous genotype is approximated by $\gamma_d = \frac{\alpha}{(1-\alpha)q} + 1$, where q is the genotype frequency, and of a protective heterozygous genotype is $\gamma_p = \frac{1}{\gamma_d}$ [Lachin, 2000]. The joint OR Ω , i.e. the product of the independent ORs of all sites, was calculated given the mode of inheritance (additive, multiplicative, dominant, or recessive). Penetrance of the genomic region was defined as $f = \frac{\Gamma}{1+\Gamma}$, where $\Gamma = \frac{f_0}{1-f_0}$ was the odds of disease given the genotype [Risch, 1990]. The calculated penetrance f was then used to determine whether a pair of randomly sampled haplotypes should be assigned the case/control status. In the context of rare-variant analysis, it can be shown that f_0 , the base-line penetrance, is approximated by the disease population prevalence; hence we used this approximation to avoid otherwise complicated exact solutions via numerical approaches. Here, we considered diseases with a population prevalence of 1%, thus f_0 is set as 0.01.

In our simulations, we assumed the additive mode of inheritance. Sample genotypes were generated using the MAF distribution obtained by forward-time simulation as described above. On average, we needed to generate 100 samples from the population in order to obtain one case. Only nonsynonymous variants with $\text{MAF} < 5\%$ in cases and controls combined were included in the analysis and the common variants were discarded. Using the MAF in cases and controls combined rather than the MAF in controls alone in selecting variant sites can avoid increasing type I error due to selection bias [Lemire, 2011; Pearson, 2011]. Each dataset consisted of 500 cases and 500 controls. Power was obtained by finding the percentage of replicates with permuted P -values below the nominal α level, and type I error was simply the power at zero PAR when all variants were neutral.

We further evaluated the performance of the Sigma-P method using a simulation framework that has also been applied by other researchers [Ionita-Laza et al., 2011; Madsen and Browning, 2009]. The framework is based on the Wright's distribution [1949] of MAF and a disease model in which the allele frequencies in cases are calculated independently for each site. Details of the simulation model are included in the supplementary materials.

RESULTS

To evaluate the performance of the Sigma-P method with two versions, σ -P and σ -MidP, we compared Type I error and power with eight other rare-variant association methods: Madsen and Browning's weighted-sum statistic (WSS) [2009], Morris and Zeggini's test of aggregated number of rare-variants (ANRV) [2009], Han and Pan's data-adaptive sum test

(aSum) [2010], Price et al.'s variable threshold (VT) method [2010], Ionita-Laza et al.'s RBS [2011], Neale et al.'s C-Alpha method [2011], Wu et al.'s SKAT [2011], and Lin and Tang's estimated regression coefficient method (EREC) [2011] through extensive simulation studies. Additionally, four methods (WSS, VT, RBS, and C-Alpha) and the Sigma-P method were used to test for rare-variant associations between metabolism traits and the *ANGTPL* family of genes using sequence data from the DHS.

TYPE I ERROR

Type I errors at the nominal α levels of 0.05 and 0.005 were evaluated for each tested rare-variant method by 10,000 permutations and 10,000 replicates. The results are summarized in Table I. In general, the Type I error is well controlled for all methods. The relationship between the observed and expected P -values at $\text{PAR} = 0$ under various scenarios are depicted in the Q-Q plots in Figures S3 and S4, respectively, for the σ -P and σ -MidP methods.

POWER

First the scenarios were considered where all causal variants are detrimental. The objective was to test for any risk effect associated with the variants, hence the statistic $S_{\sigma-P}^+$ was used.

Power at the nominal α level of 0.05 was evaluated by 1,000 permutations and 1,000 replicates. Figure 1 shows the power for the different methods at group PARs from 0% up to 10%, distributed equally among all detrimental variants in a genetic region. Panels (A–F) in Figure 1 correspond to the scenarios with different length of the coding region while maintaining the same ratios of detrimental:neutral variants. The σ -MidP method outperforms the other eight existing methods across all PAR levels for all scenarios evaluated. For the two versions of the Sigma-P method, σ -P and σ -MidP, σ -MidP consistently shows better performance and the gap between the two methods increase with a larger coding region, in which the group PAR is split among a larger number of detrimental variants. This suggests that the mid-p adjustment is especially beneficial for detecting weak effects.

Overall, the σ -MidP method shows the best performance in all simulation settings and the increase in power compared to other methods becomes more pronounced as the number of neutral variants increases. For example, when there are ~50% detrimental variants and ~50% neutral variants, i.e. ~25 detrimental variants and ~25 neutral variants and a locus PAR of 5%, the power of σ -MidP is ~3.5% greater than VT, ~5.6% greater than RBS, ~30% greater than EREC, and ~50% greater than SKAT. If the number of neutral variants are doubled (a ratio of 1:2 detrimental to neutral variants in a 4.5 kb region, i.e. ~25 detrimental variants and ~50 neutral variants), then the power of σ -MidP is 9.2% higher than for both VT and RBS, and almost 40% higher than EREC and 60% higher than SKAT. If the number of neutral variants is quadrupled (ratio of 1:4 of detrimental to neutral variants in a 7.5 kb region, i.e. ~25 detrimental variants and ~100 neutral variants), the power of σ -MidP compared to VT is 16.8% greater and compared to RBS the power is 12.9% greater. This demonstrates that the σ -MidP method can effectively extract true signals from noise due to the inclusion of noncausal variants. When the number of detrimental variants is doubled, so that there are on average ~50 detrimental variants in each region and examine the power for

ratios of 1:1, 1:2, and 1:4 detrimental to neutral variants, the power of all methods decreases. This is due to a weaker marginal PAR per detrimental variant and a higher noise level due to more noncausal variants in each gene. It is thus of particular importance to differentiate the true signals from noise through applying a proper weighting scheme, such as σ -MidP, σ -P, and WSS, or by removing the noncausal variants usually of higher MAFs as in the VT method.

Next, we considered the scenarios with a mixture of detrimental and protective variants and the statistic $S_{\sigma-P} = \max(S_{\sigma-P}^+, S_{\sigma-P}^-)$ was used. The protective sites were assumed to have the same PAR as the detrimental sites but the direction of the effect was reversed, i.e. more mutant alleles in controls than in cases. To see how different noise levels might affect the powers of the methods, we started with a 1:1 ratio of detrimental to protective variants, and no neutral variant in a 3 kb coding region, i.e. ~25 detrimental variants and ~25 protective variants. Then, while keeping the ratio of detrimental to protective variants 1:1, we gradually increased the proportion of neutral variants from 1:1:2 in a 6 kb region to 1:1:8 in a 15 kb region, i.e. ~25 detrimental variants; ~25 protective variants; and ~25, ~50, ~100, neutral variants. The results are shown in Figure 2. In general, both σ -MidP and σ -P have the best performance and the power advantage increases as the noise level rises. The RBS method, designed for detecting association of variants with mixed effect directions, has better performance than the rest of the methods. The approach of separately evaluating the scores for risk and protective effects, choosing the maximum of the two scores and then computing P -values with adjustment for multiple testing is an effective strategy to detect associations in variants with mixed effects. It is also worth noting that the aSum and SKAT methods are more powerful than the WSS and C-Alpha methods when applied on variant sites with opposite effects.

For the different sets of simulations based directly on the Wright's distribution (see Figure S4 and Methods in Supporting Information), similar trends were observed in the power curves, with the lead of the σ -P and σ -MidP methods becoming more prominent when there were more noncausal variants and the marginal PAR of each detrimental variant was reduced (Figures S5–S8).

COMPUTATIONAL SPEED

The speed of σ -P was evaluated by measuring the time of computing a P -value based on 10,000 permutations averaged over 50 replicates. The codes were written in the C programming language and run on a Linux platform with a 3 GHz processor and 2 GB memory. For simulated sequence data of 500 cases and 500 controls in 3 kb coding regions, with approximately 50 nonsynonymous variant sites on average, the time was 2.1 sec for σ -P, compared with 2.4 sec for WSS. When the size of the coding region was doubled to 6 kb, the time became 4.1 sec for σ -P and 4.2 sec for WSS. When the sample size was doubled to 1,000 cases and 1,000 controls, the time increased to 5.3 sec for σ -P and 6.1 sec for WSS. The speed can be further improved by implementing the adaptive permutation procedure, which ends the analysis at a fewer number of permutations if the tentative P -value obtained via permutation is larger than the nominal α level by more than, say, six standard deviations of the mean estimated P -value.

DALLAS HEART STUDY (DHS)

We also evaluate the performances of the methods using the DHS dataset. In addition to sampling DNA from the study subjects, several metabolism traits were characterized in the multiethnic population (African Americans, Hispanics, European Americans, and others) from Dallas County [Victor et al., 2004]. Coding regions of the four genes *ANGPTL3*, 4, 5, and 6 were sequenced in order to investigate their association with nine energy metabolism traits, namely triglyceride (TG), low-density lipoprotein (LDL), very low density lipoprotein (VLDL), high-density lipoprotein (HDL), cholesterol, glucose, body mass index (BMI), systolic (SysBP) and diastolic blood pressure (DiasBP). In the original DHS study, significant associations were identified between *ANGPTL3*, *ANGPTL4*, and TG, and between *ANGPTL6* and cholesterol [Romeo et al., 2007, 2009]. A total of 348 nucleotide sites of sequence variations were discovered in all four genes, with majority of them being rare [Romeo et al., 2007].

In the analysis, a MAF cutoff of 5% was imposed. Individuals with diabetes mellitus, heavy alcohol use, or who were taking lipids lowering drugs were excluded from the analysis. We applied the σ -MidP, σ -P, WSS, VT, RBS, and C-Alpha methods within each ethnic group (1,382 African Americans, 871 European Americans, and 500 Hispanics). Each ethnic group was analyzed separately to avoid spurious associations due to population stratification. For each of the nine quantitative traits, to mimic the case-control design individuals with quantitative trait values within the lower 30th percentile and upper 70th percentile were classified as controls and cases, respectively. Statistic $S_{\sigma-P} = \max(S_{\sigma-P}^+, S_{\sigma-P}^-)$ was applied to test for any association due to either protective or detrimental variants and an additive model was assumed.

Tables II and III summarize the *P*-values of the association analyses within the European American and African American groups and the results for the Hispanic group are listed in Table S2. It should be noted that the main objective here is to compare the *P*-values obtained by different methods, rather than making claims of the discovery of novel associations. Due to the limited sample size, correction for multiple testing was not applied and a nominal significance level of 0.05 was used. The previously reported associations between *ANGPTL4* and TG and between *ANGPTL4* and VLDL [Liu and Leal, 2010] were confirmed in the European group using all rare-variant association methods. In addition, a novel putative association was identified in European Americans between *ANGPTL3* and SysBP (by σ -P, RBS, and WSS).

For the African American group, both σ -P and σ -MidP confirmed the previous association between *ANGPTL5* and BMI [Liu and Leal, 2010]. Additionally, σ -MidP method was able to identify three novel putative associations between *ANGPTL3* and DiasBP, *ANGPTL5* and HDL, and *ANGPTL6* and Glucose. These same results were also observed with at least one of the four other existing methods that were used to perform data analysis.

For the Hispanic group, collectively six putative associations were detected using the four comparison methods or σ -P/ σ -MidP where all of these associations were detected by at least two different methods. More specifically, the σ -MidP method was able to identify five

associations between *ANGPTL3* and DiasBP, *ANGPTL5* and SysBP, *ANGPTL6* and TG, *ANGPTL3* and VLDL, and *ANGPTL4* and VLDL, while the association between *ANGPTL6* and VLDL, was only detected using VT and C-Alpha.

DISCUSSION

Complex disorders are characterized by the involvement of a large number of genes of small effect and incomplete penetrance, and diverse locus and allelic heterogeneity. To uncover these associations, the aggregate effect of variants might have to be analyzed over large genomic regions in one or even multiple underlying biological pathways. This will inevitably lead to the inclusion of a larger proportion of noncausal variants, and also causal variants with effects in different directions. Furthermore, with the availability of data from whole genome sequencing, we need to expand the scope of association analysis to regions outside the exome with nonprotein coding genes and other regulatory sequences. The proposed Sigma-P method is specifically targeted at detecting the collective association of rare variants in large genomic regions. The evaluation of the P -value per site before their group-wise aggregation makes our statistic robust against high level of data noise introduced by the presence of many neutral variants. The strategy of maximizing over separately calculated risk and protectiveness scores makes it also effective for handling mixed effect directions. Our method takes into account both the evidence of association for each variant site i given by $-\log(p_i)$, and its effect size, whose general trend is estimated by the weight $w_i = 1/\sqrt{n_i \cdot q_i(1 - q_i)}$, irrespective of whether the minor allele confers risk or protection [Park et al., 2011]. All these factors contribute to the gain in statistical power of our method. Our straightforward approach is computationally efficient for studying the effect of a large number of variant sites.

In the DHS analysis, although Sigma-P showed satisfactory performance, its P -values for the putative associations were not always the lowest among the tested methods. As we take a closer look at the case-control datasets derived from the top and bottom 30% of each quantitative trait in the European population of the study, we found that the numbers of distinct nonsynonymous sites for genes *ANGPTL3*, 4, 5, and 6 were only 7.4, 10.5, 6.7, and 6.7, respectively, after removing sites that were either monomorphic or with MAF > 5%. Since Sigma-P is especially effective in analyzing large numbers of single nucleotide variant (SNV) with many neutral variants, its edge over other methods was not displayed as prominently when applied on the DHS data, compared with the simulation results. There is, however, no general consensus on which associations are true positive or false positive in the DHS data.

So far we have only discussed the application of Sigma-P on the study of dichotomous traits. The approach can be extended to handle quantitative traits, as well as situations where adjustment for continuous or categorical covariates is necessary. This can be achieved by adopting GLMs, instead of Fisher's exact test, in computing the per-site P -values. To calculate the weight w_i for quantitative-trait analysis, where samples are not labeled directly as cases or controls, we may estimate the population allele frequency q_i by counting only alleles in samples whose trait values fall within the range to which the majority of the population belongs. To account for any possible correlations between the genotypes and the

covariates, conditional permutations [Strobl et al., 2008] can be used in obtaining the group wise P -values. However, for covariates unlikely to be correlated to the genotypes, such as age and gender, we may simply permute the genotypes, while keeping the covariates with the trait values of the samples. The implementation and evaluation of the generalized Sigma-P approach should be further explored in the future.

Another possible solution for the control of covariates is by assigning cases and controls to matched strata based on a stratification score that estimates the probability of disease conditional on potential confounders, such as components of ancestry [Epstein et al., 2012]. Stratified rare-variant analysis can then be performed by applying our Sigma-P statistic per stratum, followed by a global test that combines the results across. This approach can in general be applied to any existing rare-variant methods for handling covariates in a study.

The operating principles behind the method are best explained through simple examples. Listed below in Table IV are five types of SNVs categorized by their MAF and effect size with some hypothetical data among 500 cases and 500 controls. In computing the marginal OR, a continuity correction of 0.5 is added to all cells to avoid problems arising from zero cell counts. In general, if two SNVs have the same P -value of the strength of association, the one with a lower MAF must have a stronger phenotypic effect size, as illustrated by SNV types A and C. In other words, rare variants, even exhibiting strong effects, have P -values constrained by the small number of observed variants. By assigning larger weights to variants of lower MAF, such inherent lack of statistical power in rare variants can be duly mitigated.

The WSS method [Madsen and Browning, 2009] is similar in concept to the CMC method [Li and Leal, 2008], in which the statistic contrasts the aggregate number of mutations in cases against that in controls. In WSS, larger weights are assigned to variant sites of lower frequencies. The approach works well for homogeneous datasets with mostly detrimental variants. However, the performance declines as the datasets become more heterogeneous and have a larger proportion of neutral variants. Consider a SNV pair consisting of one strongly predisposing site of Type A and a neutral site with 30 variants, respectively, in cases and controls. The aggregate numbers of variants in cases and controls are then 33 vs. 30, with a $-\log(p)$ value of 1.045, which is much lower than the value of 2.774 with SNV A alone. Obviously, the contrast will be further weakened by SNVs of opposing effects along the same line of reasoning.

The VT statistic [Price et al., 2010] is based on measuring the aggregate difference in the mutation counts between cases and controls. It then maximizes the scores over different MAF thresholds, with the option of incorporating the computational predictions of functional effects (e.g. incorporating weights from Polyphen-2) in the weighting scheme. Similarly, this method suffers from the loss of power when there are mixed effect directions, but the VT approach is effective in filtering out the noise components in the data if most causal sites do have MAFs that are below a certain threshold.

The one-sided RBS statistic [Ionita-Laza et al., 2011] sums uniformly the $-\log(p)$ value of all sites that have a greater number of variants in cases than in controls. The P -values are

approximated by assuming that the number of variants at each site follows the Poisson distribution. By independently evaluating the statistical significance of each site before aggregating their effects, the impact of neutral variants can be effectively reduced. The strategy of computing two one-sided scores separately for risk and protectiveness, choosing the larger one of them, then adjusting for performing two tests through permutation is also proved to be effective for dealing with variants with effects in opposite directions. The method can also remove some of the non-causal sites in the one-sided risk score if these sites have more variants in controls than in cases, and vice versa for the protectiveness score. However, without a proper scheme for weighting each site, the influence of rare causal variants is inherently under-measured due to their limited numbers, as discussed earlier in this section.

The C-Alpha method [Neale et al., 2011] tests for a deviation in the distribution of variants through a statistic that contrasts the variance of each observed count of variants in cases with the expected variance, assuming a binomial distribution under the null hypothesis of no association. Since a variant observed only once provides no direct information on over-dispersion, singletons are pooled into a single binomial count. The method is powerful for variants with very strong effect. This can be readily explained by the fact that the variant count of each site contributes quadratically to the normalized Z statistic, which asymptotically approximates the normal distribution. However, the method performs unsatisfactorily when there are a large number of neutral sites as in our simulation analysis, because those sites may contribute negatively to the one-sided statistic that supports evidence of association only at the positive tail.

The Sigma-P method, on the other hand, takes into account both the effect size and strength of statistical evidence for each SNV, and incorporates the strategy of maximizing over two one-sided score tests, respectively, for detrimental and protective variants. The simple form of the statistic makes it easy to understand and interpret by researchers from different disciplines. It is more powerful than several existing rare-variant association analysis methods for a wide range of simulation settings and has detected novel putative associations in the DHS data that needs to be confirmed by future studies. Our proposed method is effective for handling large datasets with rare variants of mixed effects and is robust against noise introduced by neutral variants. These desirable properties make the Sigma-P method a promising tool for identifying complex disease rare-variant associations where the pathological mechanisms may involve scores of rare variants sporadically distributed in wide stretches of genomic regions with a large proportion of noncausal variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors want to thank the Yale High Performance Computing center for providing computing facility (supported by NIH grant RR19895). This research is supported by National Institutes of Health grants 1RC4MD005964 and 1RC2HL102926 (S.M.L.). No conflict of interest for this work is declared.

REFERENCES

- Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hébert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA. 2007 Medical sequencing at the extremes of human body mass. *Am J Hum Genet* 80:779–791. [PubMed: 17357083]
- Armitage P, Berry G, Matthews JNS. 2002 *Statistical Methods in Medical Research* (4th ed.) Malden, MA: Blackwell Publishing.
- Bansal V, Libiger O, Torkamani A, Schork NJ. 2010 Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11:773–785. [PubMed: 20940738]
- Barnard GA. 1989 On alleged gains in power from lower P-values. *Stat Med* 8:1469–1477. [PubMed: 2616936]
- Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V. 2010 A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 6, e1000954. [PubMed: 20976246]
- Bodmer W, Bonilla C. 2008 Common and rare variants in multi-factorial susceptibility to common diseases. *Nat Genet* 40:695–701. [PubMed: 18509313]
- Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD. 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4, e1000083. [PubMed: 18516229]
- Epstein MP, Duncan R, Broadaway A, He M, Allen AS, Satten GA. 2012 Stratification-score matching improves correction for confounding by population stratification in case-control association studies. *Genet Epidemiol* 36:195–205. [PubMed: 22714934]
- Eyre-Walker A, Woolfit M, Phelps T. 2006 The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900. [PubMed: 16547091]
- Fisher RA. 1925 *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd.
- Han F, Pan W. 2010 A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70:42–54. [PubMed: 20413981]
- Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. 2011 A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 7, e10001289.
- Ji W, Foo JN, O’Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008 Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 40:592–599. [PubMed: 18391953]
- Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. 2009 Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A* 106:3871–3876. [PubMed: 19202052]
- Lachin JM. 2000 *Biostatistical Methods: The Assessment of Relative Risks*. New York: Wiley & Sons.
- Lemire M. 2011 Defining rare variants by their frequencies in controls may increase type I error. *Nat Genet* 43:391–392. [PubMed: 21522172]
- Li B, Leal SM. 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311–321. [PubMed: 18691683]
- Lin DY, Tang ZZ. 2011 A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89:354–367. [PubMed: 21885029]
- Liu DJ, Leal SM. 2010 A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 6, e10001156.
- Madsen BE, Browning SR. 2009 A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5, e1000384. [PubMed: 19214210]
- Morgenthaler S, Thilly WG. 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(1–2):28–56. [PubMed: 17101154]

- Morris AP, Zeggini E. 2010 An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34:188–193. [PubMed: 19810025]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011 Testing for an unusual distribution of rare variants. *PLoS Genet* 7:e10001322.
- Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009 Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324:387–389. [PubMed: 19264985]
- Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Jr., Chatterjee N. 2011 Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A* 108(44):18026–31. [PubMed: 22003128]
- Pearson RD. 2011 Bias due to selection of rare variants using frequency in controls. *Nat Genet* 43:392–393. [PubMed: 21522173]
- Peng B, Kimmel M. 2005 simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* 21:3686–3687. [PubMed: 16020469]
- Peng B, Liu X. 2010 Simulating sequences of the human genome with rare variants. *Hum Hered* 70:287–291. [PubMed: 21212684]
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010 Pooled association tests for rare variants in exon resequencing studies. *Am J Hum Genet* 86:832–838. [PubMed: 20471002]
- Risch N. 1990 Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 46:222–228. [PubMed: 2301392]
- Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC. 2007 Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* 39:513–516. [PubMed: 17322881]
- Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC. 2009 Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest* 119:70–79. [PubMed: 19075393]
- Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. 2008 Conditional variable importance for random forests. *BMC Bioinformatics* 9:307. [PubMed: 18620558]
- Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA, Staab JM, Hobbs HH, Dallas Heart Study Investigators. 2004 The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am J Cardiol* 93:1473–1480. [PubMed: 15194016]
- Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. 2008 Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320:539–543. [PubMed: 18369103]
- Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci USA* 102:7882–7887. [PubMed: 15905331]
- Wright S. 1949 Adaptation and selection. In: Jepsen GL, Simpson GG, Mayr E (Eds.), *Genetics, Paleontology, and Evolution*. Princeton: Princeton University Press p 365–389.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82–93. [PubMed: 21737059]

Fig. 1.

Power comparisons under the additive model with a decreasing ratio of detrimental to neutral variants in genomic regions with different lengths: (panel A) 1:1 (~25 detrimental variants: ~25 neutral variants) in 3 kb, (panel B) 1:2 (~25 detrimental variants: ~50 neutral variants) in 4.5 kb, (panel C) 1:4 (~25 detrimental variants: ~100 neutral variants) in 7.5 kb, (panel D) 1:1 (~50 detrimental variants: ~50 neutral variants) in 6 kb, (panel E) 1:2 (~50 detrimental variants: ~100 neutral variants) in 9 kb, and (panel F) 1:4 (~50 detrimental variants: ~200 neutral variants) in 15 kb regions. Sample size is 500 cases and 500 controls.

Fig. 2.

Power comparisons under the dominant model with the ratio of detrimental to protective variants fixed at 1:1, and an increasing number of neutral variants: (panel A) 1:1:0 (~25 detrimental variants: ~25 protective variants: 0 neutral variants) in 3 kb, (panel B) 1:1:2 (~25 detrimental variants: ~25 protective variants: ~50 neutral variants) in 6 kb, (panel C) 1:1:4 (~25 detrimental variants: ~25 protective variants: ~100 neutral variants) in 9 kb, and (panel D) 1:1:8 (~25 detrimental variants: ~25 protective variants: ~200 neutral variants) in 15 kb regions. Sample size is 500 cases and 500 controls.

Type I error at 0.05, 0.01, and 0.005 nominal levels, respectively, for varying length of the coding region. There are approximately 50, 75, 100, 125, 150, and 250 neutral variant sites in 3 kb, 4.5 kb, 6 kb, 7.5 kb, 9 kb, and 15 kb coding regions, respectively

TABLE I.

α	Length of coding region	σ -MidP	σ -P	WSS	ANRV	aSum	VT	RBS	C-Alpha	SKAT	EREC
0.05	3 kb	0.0489	0.0508	0.0529	0.0383	0.0357	0.0467	0.0504	0.0451	0.0488	0.0506
	4.5 kb	0.0516	0.0516	0.0478	0.0381	0.0406	0.0505	0.0469	0.0472	0.0531	0.0494
	6 kb	0.0498	0.0502	0.0529	0.0433	0.0417	0.0517	0.0512	0.0473	0.0482	0.0531
	7.5 kb	0.0507	0.0484	0.0504	0.0449	0.0429	0.0504	0.0558	0.0501	0.0514	0.0511
	9 kb	0.0514	0.0505	0.0502	0.0446	0.0459	0.0473	0.0511	0.0507	0.0504	0.047
	15 kb	0.0479	0.0524	0.0494	0.0449	0.0457	0.0499	0.049	0.0502	0.0499	0.0501
0.01	3 kb	0.0105	0.0089	0.0099	0.0071	0.0058	0.009	0.0095	0.0081	0.0096	0.0104
	4.5 kb	0.0101	0.0118	0.0107	0.0077	0.0078	0.0092	0.0102	0.0091	0.0095	0.0102
	6 kb	0.0107	0.0101	0.0104	0.0084	0.0082	0.0117	0.0113	0.0093	0.0102	0.0095
	7.5 kb	0.0112	0.0091	0.0098	0.0105	0.0086	0.008	0.0108	0.0097	0.0109	0.0105
	9 kb	0.0082	0.0088	0.0108	0.0081	0.0099	0.0093	0.0104	0.0097	0.0095	0.0107
	15 kb	0.0105	0.0106	0.0116	0.0098	0.0098	0.0101	0.0108	0.0109	0.0096	0.0111
0.005	3 kb	0.0048	0.0050	0.0050	0.0035	0.0026	0.0047	0.0045	0.0039	0.0058	0.005
	4.5 kb	0.0050	0.0054	0.0057	0.0034	0.004	0.0038	0.0052	0.0044	0.0042	0.0057
	6 kb	0.0054	0.0051	0.0059	0.0041	0.0044	0.0066	0.0056	0.0048	0.0063	0.0043
	7.5 kb	0.0049	0.0040	0.0055	0.0048	0.0036	0.0042	0.0051	0.0049	0.0059	0.0061
	9 kb	0.0038	0.0048	0.0060	0.0041	0.0048	0.0047	0.0058	0.0055	0.0058	0.0053
	15 kb	0.0056	0.0051	0.0060	0.0042	0.0045	0.0051	0.0057	0.0057	0.004	0.0043

TABLE II.

Rare-variant association analyses in European Americans. The *ANGPTL* family of genes were analyzed to test for rare-variant associations with nine metabolic traits in a sample of 871 European Americans from the DHS. In order to mimic a qualitative trait, only those individuals with quantitative trait values in the upper 70th percentile and lower 30th percentile were analyzed. Those *P*-values with a nominal significance level of 0.05 are highlighted in bold with an asterisk.

Phenotype	Gene name	No. of cases/controls	P-Values					C-Alpha	No. of variant sites	No. of variant carriers in cases/controls	No. of variant alleles in cases/controls
			(σ-MidP	σ-P	VT	RBS	WSS				
BMI	ANGPTL3	239/239	0.5874	0.6394	0.3407	0.7413	0.7193	0.4635	31	8/7	8/7
	ANGPTL4	262/267	0.1528	0.1988	0.1728	0.1199	0.3157	0.3846	31	9/14	9/14
	ANGPTL5	259/259	0.7742	0.7792	0.5714	0.5664	0.3297	0.4416	22	19/14	19/14
DiasBP	ANGPTL6	265/268	0.2178	0.1858	0.6823	0.2308	0.3616	0.8362	39	23/18	24/18
	ANGPTL3	239/239	0.1379	0.0979	0.6474	0.1429	0.2697	0.2617	31	6/11	6/11
	ANGPTL4	262/267	0.3467	0.2817	0.6134	0.2577	0.2557	0.4296	31	10/16	10/16
SysBP	ANGPTL5	259/259	0.6813	0.4346	0.2677	0.1728	0.046*	0.0639	22	26/15	26/15
	ANGPTL6	265/268	0.8242	0.7752	0.7043	0.8122	0.6094	0.4376	39	27/25	27/27
	ANGPTL3	239/239	0.0519	0.014*	0.0709	0.023*	0.026*	0.1778	31	5/14	5/14
Cholesterol	ANGPTL4	262/267	0.5005	0.5135	0.2667	0.3147	0.1818	0.3806	31	11/18	11/18
	ANGPTL5	259/259	0.8182	0.8332	0.6014	0.956	0.8971	0.968	22	18/19	18/19
	ANGPTL6	265/268	0.4206	0.3467	0.6364	0.3407	0.1978	0.2907	39	22/32	23/33
LDL	ANGPTL3	239/239	0.7852	0.7373	0.9121	0.5894	0.3776	0.9021	31	10/14	10/14
	ANGPTL4	262/267	0.5744	0.6843	0.7502	0.7582	0.7133	0.5345	31	11/12	11/12
	ANGPTL5	259/259	0.7952	0.4056	0.2118	0.1888	0.1439	0.042*	22	23/13	24/13
HDL	ANGPTL6	265/268	0.4066	0.5315	0.2747	0.5165	0.3457	0.7842	39	28/22	28/24
	ANGPTL3	239/239	0.9121	0.8871	0.9031	0.7303	0.4036	0.8551	31	11/15	11/15
	ANGPTL4	262/267	0.2907	0.3936	0.5345	0.4975	0.7992	0.1069	31	13/11	13/11
HDL	ANGPTL5	259/259	0.9451	0.8182	0.8372	0.5994	0.4535	0.2817	22	24/19	25/19
	ANGPTL6	265/268	0.4845	0.3037	0.7792	0.5574	0.957	0.3666	39	28/28	28/31
	ANGPTL3	239/239	0.1748	0.3087	0.0549	0.3956	0.8731	0.0729	31	11/9	11/9
HDL	ANGPTL4	262/267	0.1229	0.1518	0.2408	0.1059	0.0799	0.3327	31	19/10	19/10
	ANGPTL5	259/259	0.5085	0.4436	0.6254	0.8232	0.5874	0.3996	22	23/17	23/18
	ANGPTL6	265/268	0.7213	0.7253	0.8382	0.7403	0.7552	0.8781	39	33/29	34/29

Phenotype	Gene name	No. of cases/controls	P-Values				C-Alpha	No. of variant sites	No. of variant carriers in cases/controls	No. of variant alleles in cases/controls
			σ -P	VT	RBS	WSS				
TG	ANGPTL3	239/239	0.6883	0.6174	0.2607	0.5614	0.8012	0.4655	6/8	6/8
	ANGPTL4	262/267	0.002*	0.001*	0.003*	0.001*	0.002*	0.005*	4/20	4/20
	ANGPTL5	259/259	0.7103	0.9151	0.6863	0.8931	0.5734	0.3027	19/16	20/16
	ANGPTL6	265/268	0.4216	0.4296	0.6683	0.7063	0.3317	0.9091	27/21	27/24
VLDL	ANGPTL3	239/239	0.7552	0.5954	0.2308	0.5634	0.8092	0.3467	6/9	6/9
	ANGPTL4	262/267	0.006*	0.003*	0.004*	0.001*	0.004*	0.004*	5/22	5/22
	ANGPTL5	259/259	0.7213	0.968	0.6813	0.9081	0.7972	0.5534	19/19	20/19
	ANGPTL6	265/268	0.3187	0.3996	0.7712	0.7103	0.5774	0.6923	27/26	27/29
Glucose	ANGPTL3	239/239	0.0669	0.0709	0.1888	0.0959	0.1319	0.4476	6/12	6/12
	ANGPTL4	262/267	0.5574	0.4885	0.6843	0.3437	0.3017	0.2657	18/11	18/11
	ANGPTL5	259/259	0.6613	0.5045	0.4286	0.2537	0.2897	0.1459	24/15	25/15
	ANGPTL6	265/268	0.962	0.9461	0.7023	0.8731	0.4196	0.8002	30/24	31/26

TABLE III.

Rare-variant association analyses in African Americans. The *ANGPTL* family of genes are analyzed to test for rare-variant associations with nine metabolic traits in a sample of 1,382 African Americans from the DHS. In order to mimic a qualitative trait, only those individuals with quantitative trait values in the upper 70th percentile and lower 30th percentile were analyzed. Those *P*-values with a nominal significance level of 0.05 are highlighted in bold with an asterisk.

Phenotype	Gene name	No. of cases/controls	<i>P</i> -Values					C-Alpha	No. of variant sites	No. of variant carriers in cases/controls	No. of variant alleles in cases/controls
			σ -MidP	σ -P	VT	RBS	WSS				
BMI	<i>ANGPTL3</i>	346/346	0.7862	0.8252	0.3387	0.7722	0.2258	0.7023	30	4/6	4/6
	<i>ANGPTL4</i>	414/425	0.6863	0.6903	1	0.6573	0.4635	0.8002	30	7/10	7/10
	<i>ANGPTL5</i>	407/402	0.018*	0.02*	0.2238	0.1918	0.3097	0.5015	22	45/41	45/41
DiasBP	<i>ANGPTL6</i>	423/419	0.964	0.962	0.6434	0.9441	0.9431	0.982	39	10/9	10/9
	<i>ANGPTL3</i>	346/346	0.025*	0.043*	0.015*	0.0579	0.7532	0.1608	31	44/48	45/49
	<i>ANGPTL4</i>	414/425	0.7722	0.6104	0.5914	0.4885	0.4176	0.1778	30	11/7	11/7
SysBP	<i>ANGPTL5</i>	407/402	0.4106	0.4685	0.1149	0.4655	0.5135	0.045*	22	48/57	48/57
	<i>ANGPTL6</i>	423/419	0.6364	0.4426	0.6603	0.4356	0.4336	0.4126	39	12/8	12/8
	<i>ANGPTL3</i>	346/346	0.5485	0.6693	0.7393	0.5455	0.6414	0.3407	30	6/8	6/8
Cholesterol	<i>ANGPTL4</i>	414/425	0.9101	0.9191	1	0.952	0.9471	0.6474	30	9/10	9/10
	<i>ANGPTL5</i>	407/402	0.8192	0.7443	0.4625	0.5544	0.4575	0.6444	22	63/54	64/54
	<i>ANGPTL6</i>	423/419	0.5405	0.6494	0.7802	0.5774	0.7592	0.4456	39	9/10	9/10
LDL	<i>ANGPTL3</i>	346/346	0.962	0.965	0.9121	0.999	0.5814	0.3247	31	33/36	33/38
	<i>ANGPTL4</i>	414/425	0.3776	0.5335	0.1279	0.4995	0.4496	0.6264	30	10/13	10/13
	<i>ANGPTL5</i>	407/402	0.3816	0.3277	0.8871	0.5045	1	0.5584	22	50/51	51/51
HDL	<i>ANGPTL6</i>	423/419	0.8621	0.9311	0.3237	0.983	0.951	0.7932	39	9/9	9/9
	<i>ANGPTL3</i>	346/346	0.7003	0.6264	0.7602	0.5814	0.6973	0.8741	31	41/36	41/36
	<i>ANGPTL4</i>	414/425	0.8122	0.7992	0.5155	0.6953	0.3916	0.4715	30	10/14	10/14
HDL	<i>ANGPTL5</i>	407/402	0.5325	0.5754	0.7862	0.5784	0.6813	0.2917	22	50/57	50/57
	<i>ANGPTL6</i>	423/419	0.4496	0.5864	0.2018	0.6723	0.5694	0.4166	39	11/9	11/9
	<i>ANGPTL3</i>	346/346	0.6963	0.7473	0.7622	0.5874	0.7353	0.6553	31	38/43	38/44
HDL	<i>ANGPTL4</i>	414/425	0.958	0.98	0.952	0.983	1	0.9281	30	10/10	10/10
	<i>ANGPTL5</i>	407/402	0.026*	0.031*	0.033*	0.0819	0.2517	0.2298	22	60/47	60/47
	<i>ANGPTL6</i>	423/419	0.4186	0.6124	0.4675	0.7023	0.991	0.1329	39	9/8	9/8

Phenotype	Gene name	No. of cases/controls	P-Values					C-Alpha	No. of variant sites	No. of variant carriers in cases/controls	No. of variant alleles in cases/controls
			σ-MidP	σ-P	VT	RBS	WSS				
TG	ANGPTL3	346/346	0.4496	0.5684	0.3796	0.1968	0.1039	0.0859	31	32/47	32/48
	ANGPTL4	414/425	0.7113	0.8312	0.3267	0.9081	0.987	0.7892	30	10/10	10/10
	ANGPTL5	407/402	0.1179	0.1558	0.2198	0.1598	0.1878	0.8162	22	45/58	46/58
	ANGPTL6	423/419	0.3886	0.2338	0.6244	0.2258	0.2657	0.2737	39	7/13	7/13
VLDL	ANGPTL3	346/346	0.4066	0.5065	0.3387	0.1958	0.1059	0.1239	31	33/45	33/46
	ANGPTL4	414/425	0.8332	0.8811	0.4935	0.8422	0.6234	0.5604	30	11/8	11/8
	ANGPTL5	407/402	0.2118	0.2318	0.2168	0.1169	0.2038	0.7023	22	47/69	48/69
	ANGPTL6	423/419	0.2567	0.1499	0.4905	0.1479	0.2318	0.2218	39	7/13	7/13
Glucose	ANGPTL3	346/346	0.8531	0.8362	0.8521	0.9181	0.9451	0.9161	31	44/48	46/49
	ANGPTL4	414/425	0.98	0.987	1	0.981	0.9351	0.975	30	8/9	8/9
	ANGPTL5	407/402	0.2278	0.3457	0.0859	0.4456	0.8252	0.9421	22	58/66	58/66
	ANGPTL6	423/419	0.02*	0.0679	0.007*	0.0529	0.044*	0.017*	39	15/7	15/7

TABLE IV.

Simple examples of different SNV types and their basic statistics

SNV	No. of variants in cases	No. of variants in controls	MAF ^a in controls (<i>q</i>)	Odds ratio	Inverse-sigma weight (<i>w</i>)	$-\log(p)$	$-\log(p)$ with mid-p adjustment
Very rare, strong effect	3	0	0.001	7.021	31.670	2.081	2.774
Very rare, weak effect	4	2	0.003	1.804	18.303	1.069	1.486
Rare, weak effect	16	10	0.011	1.581	9.597	1.821	2.100
Uncommon, weak effect	42	30	0.031	1.411	5.775	2.373	2.576
Uncommon, very weak/neutral effect	32	30	0.031	1.068	5.775	0.801	0.919

^aMAF = $(m + 1)/(n + 2)$, where *m* is the number of variants and *n* is the number of all alleles in controls.