Genetic Epidemiology 30: 508–518 (2006)

Contrasting Multi-Site Genotypic Distributions Among Discordant Quantitative Phenotypes: The *APOA1/C3/A4/A5* Gene Cluster and Cardiovascular Disease Risk Factors

Bret A. Payseur,[1]* Andrew G. Clark,[1] James Hixson,[2] Eric Boerwinkle[2], and Charles F. Sing[3]

[1]*Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York*
[2]*Human Genetics Center, University of Texas Health Sciences Center, Houston, Texas*
[3]*Department of Human Genetics, University of Michigan, Ann Arbor, Michigan*

Most tests of association between DNA sequence variation and quantitative phenotypes in samples of randomly chosen individuals rely on specification of genotypic strata followed by comparison of phenotypes across these strata. This strategy often succeeds when phenotypic differences are caused by one or two single nucleotide polymorphisms (SNPs) among the surveyed markers. However, when multiple-SNP haplotypes account for observed phenotypic variation, identification of the best partitioning requires examination of an inordinate number of SNP combinations. An alternative approach is to rank individuals by their phenotypic measures and ask whether attributes of the genotypic variation show a non-random distribution along this phenotypic ranking. One simple version of this strategy selects the top and bottom tails of the distribution, and then tests whether genotypes from these two samples are drawn from a single population. This framework does not require the recovery of phased haplotypes and allows contrasts between large numbers of sites at once. We use a method based on this approach to identify associations between plasma triglyceride level, a risk factor for cardiovascular disease, and multi-site genotypes located in the *APOA1/C3/A4/A5* cluster of apolipoprotein genes in unrelated individuals (1,071 African-American females, 780 African-American males, 1,036 European-American females, and 930 European-American males) sampled from four US cities as part of the Coronary Artery Risk Development in Young Adults (CARDIA) study. Method performance is investigated using simulations that model genealogical variation and different genetic architectures. Results indicate that this multi-site test can identify genotype-phenotype associations with reasonable power, including those generated by some simple epistatic models. *Genet. Epidemiol.* 30:508–518, 2006. © 2006 Wiley-Liss, Inc.

Key words: haplotype; phenotypic extremes; population differentiation; population genetics

Contract grant sponsor: NIH; Contract grant number: GM65509; HL072904; Contract grant sponsor: University of Alabama at Birmingham, Coordinating Center; Contract grant number: N01-HC-95095; Contract grant sponsor: University of Alabama at Birmingham, Field Center; Contract grant number: N01-HC-48047; Contract grant sponsor: University of Minnesota, Field Center; Contract grant number: N01-HC-48048; Contract grant sponsor: Northwestern University, Field Center; Contract grant number: N01-HC-48049; Contract grant sponsor: Kaiser Foundation Research Institute; Contract grant number: N01-HC-48050; Contract grant sponsor: University of California, Irvine, Echocardiography Reading Center; Contract grant number: N01-HC-45134; Contract grant sponsor: Harbor-UCLA Research Education Institute, Computed Tomography Reading Center; Contract grant number: N01-HC-05187; Contract grant sponsor: National Heart, Lung and Blood Institute.
*Correspondence to: Bret A. Payseur, Laboratory of Genetics, Genetics/Biotechnology 2428, University of Wisconsin, Madison, WI 53706. E-mail: payseur@wisc.edu
Received 20 November 2005; Revised 13 March 2006; Accepted 28 April 2006
Published online 23 June 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20163

INTRODUCTION

A complete understanding of genetic variation underlying risk of disease requires identification of the functional DNA polymorphisms. With increasing availability of molecular markers, genomic scanning for association between phenotypic variation and randomly chosen polymorphisms, conducted with or without pedigree information, provides promise as a method for identifying disease susceptibility genes. When clear candidate genes exist, surveys of densely spaced polymorphisms within these genes in large samples of individuals can also provide crucial insights into variation at the gene level and its statistical association with variation in disease risk.

A common approach to candidate gene analysis involves the collection of phenotypic measure-

ments and genotypes for a sample of unrelated individuals and comparison of phenotypes among classes defined by single nucleotide polymorphism (SNP) genotypes using analysis of variance or related methods. This ''genotype-first'' paradigm is powerful if the SNPs that contribute to risk are in the marker set, or if they are in linkage disequilibrium with the unobserved causal SNPs. Nevertheless, two arguments suggest that alternative approaches may be worthwhile in cases involving quantitative traits. First, much of the power for detecting associations between polymorphism in candidate genes and complex trait variation derives from individuals with extreme phenotypes [Schork et al., 2000]. This idea has been used to identify maximally informative markers in experimental crosses designed to map quantitative traits [''bulked segregant analysis''; Michelmore et al., 1991].

A second reason for considering alternatives to the ''genotype-first'' framework for candidate gene studies is the inherent complexity that arises when multiple polymorphic sites are analyzed simultaneously. The rapid increase in the number of multi-site genotypes that accompanies the addition of SNPs shrinks the sample size for each genotypic class, challenging the accurate estimation of phenotypic means and variances associated with specific genotypic classes and reducing power. Additionally, the genotype-first approach will often fail in cases where combinations of SNPs that exert no individual marginal effects explain the observed phenotypic variation. Tests of association that benefit from the inclusion of more neighboring SNPs are desirable. Methods that evaluate the relationship between phenotype and genotype for a large number of site combinations have been developed and applied with some success [Nelson et al., 2001; Culverhouse et al., 2004; Jannot et al., 2003], but extension of these approaches to many sites is challenging. Another strategy assesses the correlation of phenotypes and haplotypes in a cladistic context [Templeton et al., 1987; Seltman et al., 2003], but these procedures currently require knowledge of haplotype phase and resolution of the genealogy, and propagation of the uncertainty associated with these inferences can be prohibitively cumbersome.

These considerations motivate an approach to association testing that uses the phenotypic distribution to define classes of genotypes for comparison. Associations between genotype and phenotype can be identified as genotypic differentiation between groups of individuals drawn from opposite ends of the phenotypic spectrum. For example, Cohen et al. [2004] discovered that groups of individuals with high vs. low levels of high-density lipoprotein cholesterol differ in the frequencies of non-synonymous SNPs found in candidate genes. The power of this ''phenotype-first'' method has been investigated considering one site at a time [Schork et al., 2000; Tenesa et al., 2003], and the rationale has been discussed in the context of selective genotyping and DNA pooling [Darvasi and Soller, 1992; Sham et al., 2002; Carlson et al., 2004]. Here, we extend this framework to compare phenotypically extreme groups at many sites simultaneously, an important advantage when multiple sites determine the functional haplotype.

Analysis of cardiovascular disease seems particularly amenable to a candidate gene approach. Loss-of-function mutations in several candidate genes co-segregate with familial disorders, and a connection between quantitative variation in risk factors and candidate gene polymorphism has been established. Elevated triglyceride (TG) level is one such risk factor [Austin et al., 1998] that has received considerable attention. Substantial genetic variation for TG level has been detected [Hunt et al., 1989; Brenn, 1994], with biochemical and association studies providing clear candidate genes for this variation.

Genes encoding the apolipoproteins, lipid-binding proteins involved in the transport of lipids in the plasma, are particularly suitable targets for candidate gene surveys of TG-related traits. One 48 kb region of chromosome 11q23–q24 contains four apolipoprotein genes, *APOA1*, *APOC3*, *APOA4*, and *APOA5* [Karathanasis, 1985; Pennacchio et al., 2001]. Several population-based associations between TG level and polymorphisms located in the *APOA1/C3/A4/A5* cluster have been reported [Groenendijk et al., 2001; Pennacchio et al., 2002; Talmud et al., 2002; Austin et al., 2004; Evans et al., 2003; Kao et al., 2003; Lai et al., 2003; Eichenbaum-Voline et al., 2004; Klos et al., 2005].

Here, we compare variation in SNPs densely spaced across the *APOA1/C3/A4/A5* region between subgroups of healthy, young individuals with extreme differences in TG level. We identify genotype-phenotype associations in the form of multi-site genotypic differentiation between groups of individuals sampled from the tails of the phenotypic distribution. Performance of this multi-site method is evaluated using computer simulations.

METHODS

DATA COLLECTION

The complete dataset from which subsets of phenotypes and genotypes were drawn for analysis was collected from individuals surveyed at four field centers (Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA) as part of the longitudinal Coronary Artery Risk Development in Young Adults (CARDIA) study [Friedman et al., 1988; Liu et al., 1989]. In the CARDIA study, subjects were unrelated African-American females ($n = 1071$), African-American males ($n = 780$), European-American females ($n = 1036$), and European-American males ($n = 930$), ranging in age from 18 to 30 years. Individuals were selected without reference to health. Venous blood was drawn after a 12-h fast and plasma TG levels were determined using standard enzymatic methods [Warnick, 1986].

The candidate gene region, the *APOA1/C3/A4/A5* gene cluster, is located on chromosome 11q23–q24 and is approximately 48 kb in size. A resequencing study of this cluster surveyed 17.7 kb, including the intergenic region between *APOA1* and *APOC3*, the exons, intervening introns, and approximately 1000 bp upstream of the *APOA1, APOC3, APOA4*, and *APOA5* genes in 24 unrelated individuals from each of three populations: African-Americans from Jackson, MS; Europeans from North Karelia, Finland; and European-Americans from Rochester, MN [Fullerton et al., 2004]. Average levels of linkage disequilibrium between all pairs of polymorphic sites from across the region (measured by R^2) ranged between 0.087 and 0.122 [depending on the population sampled; Fullerton et al., 2004], suggesting that the simultaneous use of multiple sites in association testing may be worthwhile. The present study focused on genotypes for 80 SNPs (Fig. 1) discovered in this survey. Each SNP was genotyped in the 3,817 individuals (described above) by PCR amplification of genomic DNA, a short extension reaction across the polymorphic site, and detection of allele-specific mass differences of the extension product by mass spectrometry. Allele detection and genotype calling was performed using a MassARRAY System from Sequenom® (San Diego, CA). Blind duplicates were included for assessing the quality of the SNP genotyping.

STATISTICAL ANALYSIS

To reduce skewness and account for important covariates, plasma TG values were adjusted prior to analysis by logarithmic transformation, fitting an ethnic/gender/field center specific linear regression model containing age, age^2, age^3 and BMI, and adding the residuals to the ethnic/gender-specific grand mean. Because the sexes differ in cardiovascular disease risk [Barrett-Connor, 1997], males and females were analyzed separately. African-Americans and European-Americans were also treated separately due to variation in allele frequencies and linkage disequilibrium in the surveyed SNPs [Fullerton et al., 2004] and the expectation of differentiation at other, unlinked regions that may interact with polymorphisms in the *APOA1/C3/A4/A5* cluster.

Individuals with extreme phenotypes were selected from the tails of the TG distributions of each population-sex group in the CARDIA study separately. We conducted three sets of analyses, sampling from the upper and lower 5%, 10%, and 15% of the distributions. These quantiles were chosen to balance the increase in power derived from contrasting more extreme phenotypes with the eventual decrease in power caused by reducing sample size [Schork et al., 2000; Tenesa et al., 2003]. The genotypic state of each SNP in each selected individual was recoded by comparison to sequence from one common chimpanzee into three categories (both alleles identical to chimpanzee, both alleles different from chimpanzee, or one allele different from chimpanzee (heterozygote); Fullerton et al., 2004). The unphased multi-site genotype of each individual was represented by a string of these recoded genotypes.

We modified the S_{NN} statistic and program [Hudson, 2000; http://home.uchicago.edu/~rhudson1/source/permtest.html], which was devised to detect multi-site haplotypic differentiation among populations, to compare collections of unphased genotypes sampled from the upper and lower tails of the phenotypic distribution. Following Hudson [2000], we compared individual i to each other individual. In our comparison between diploid genotypes, alternative homozygotes were assigned a value of 1, heterozygote-homozygote contrasts were assigned a value of 0.5, and identical genotypes (homozygote-homozygote and heterozygote-heterozygote) were assigned a value of 0, at each site. These distances were summed across sites. The number of individuals that showed the smallest number of differences with i was determined. Then, the fraction of these "nearest-neighbor" individuals that were found in the same phenotypic group as i (X_i) was

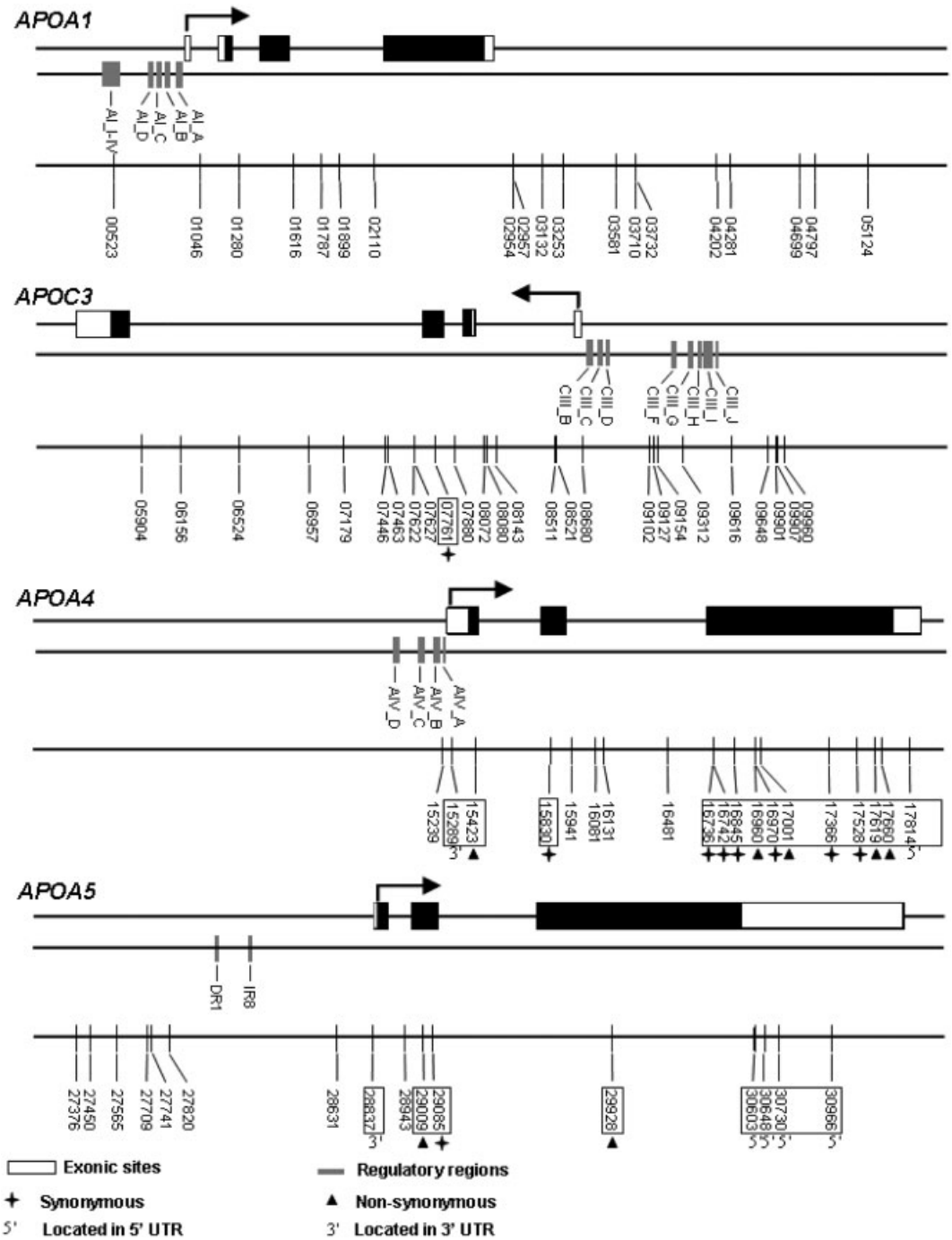

Fig. 1. Locations of 80 SNPs from the *APOA1/C3/A4/A5* gene cluster used in this study. Exon structure, direction of transcription, and regulatory regions are also shown.

calculated. S_{NN} was estimated as

$$S_{NN} = \frac{\sum_{j=1}^{n} X_j}{n}.$$

S_{NN} measures the tendency of individuals with similar sequences to be found in the same group. When phenotypic variation is associated with multi-site genotypes, we expect individuals with the same phenotype to exhibit more similar genotypes. Statistical significance was assessed by permutating sequences between phenotypic groups 10,000 times. This approach uses sequence similarity without requiring the reconstruction of genealogies or phase information, and allows

the incorporation of large numbers of polymorphic sites.

We applied the multi-site test to several combinations of SNPs from the *APOA1/C3/A4/A5* region. First, we asked whether individuals with high vs. low TG levels differed in genotypes comprised of surveyed SNPs from across the gene cluster. This was a single global test of association between variation in the *APOA1/C3/A4/A5* candidate region and TG level. Next we looked for genotypic differentiation between the phenotypic tails by grouping SNPs from the same gene (one test each for *APOA1*, *APOC3*, *APOA4*, and *APOA5*). This set of contrasts was motivated by the notion that functional polymorphisms within genes may interact. Finally, because nonsynonymous variation is likely to affect protein function, we also evaluated evidence for differentiation between genotypes formed by combining amino acid variants from across the entire region. Due to the relatively small number of tests, we did not adjust statistical significance criteria for the performance of multiple tests.

Statistical power was evaluated by computer simulation. To mimic variation in an individual gene from the *APOA1/C3/A4/A5* region, 20-site genotypes sampled from 4 kb of sequence were simulated for 1,000 diploid individuals and phenotypes were generated under three genetic models (Appendix). All simulations compared collections of genotypes from the upper and lower 10% (100 individuals from each tail) of the phenotypic distribution. Power was estimated as the fraction of 1,000 simulation replicates that yielded $P < 0.05$. The false-positive rate (also using a significance criterion of $P < 0.05$), estimated by simulating population histories with all sites exhibiting no phenotypic effect, was 0.053.

RESULTS

MULTI-SITE ASSOCIATION TESTING ACROSS THE *APOA1/C3/A4/A5* REGION

Populations sampled from opposite ends of the TG distribution harbored different collections of multi-site genotypes in several contrasts (Table I).

TABLE I. Results of tests comparing multi-site genotypes from the *APOA1/C3/A4/A5* gene cluster between the tails of the TG distribution

		5% Tails[a]			10% Tails			15% Tails		
Subset	Stratum[b]	Sites[c]	S_{NN}	P	Sites	S_{NN}	P	Sites	S_{NN}	P
All sites	EM	64	0.5358	0.2252	65	**0.5697**[d]	**0.0294**	65	0.5193	0.2428
All sites	EF	60	0.5159	0.3407	61	0.5089	0.3752	67	0.5139	0.2933
All sites	AM	69	0.4892	0.5195	74	0.5567	0.0886	77	0.4737	0.7488
All sites	AF	75	0.5204	0.2750	76	0.4864	0.6354	76	0.4983	0.5031
APOA1	EM	17	**0.5582**	**0.0186**	18	**0.5695**	**0.0003**	18	**0.5249**	**0.0320**
APOA1	EF	16	0.4739	0.7402	16	0.4748	0.8816	17	0.4946	0.5966
APOA1	AM	18	0.4731	0.6997	19	0.4911	0.5739	19	0.4682	0.9483
APOA1	AF	18	0.4950	0.4835	18	0.5138	0.1873	18	0.5084	0.2241
APOC3	EM	18	0.5577	0.0627	18	**0.5679**	**0.0033**	18	0.5089	0.2499
APOC3	EF	16	0.5134	0.3223	16	0.5182	0.1991	19	0.5058	0.3255
APOC3	AM	24	0.4696	0.6695	25	0.5444	0.0962	25	0.4926	0.5681
APOC3	AF	25	0.5486	0.1262	25	0.4851	0.6591	25	0.4843	0.7338
APOA4	EM	14	0.5239	0.2089	14	**0.5426**	**0.0231**	14	**0.5304**	**0.0286**
APOA4	EF	14	0.5065	0.3375	15	0.5165	0.1374	16	0.5155	0.1150
APOA4	AM	12	0.4522	0.8496	15	0.5319	0.0744	17	**0.5382**	**0.0145**
APOA4	AF	15	0.5061	0.3221	15	0.5112	0.1681	15	**0.5172**	**0.0361**
APOA5	EM	15	**0.5483**	**0.0476**	15	**0.5773**	**0.0001**	15	**0.5526**	**0.0003**
APOA5	EF	14	0.5163	0.2261	14	**0.5250**	**0.0406**	15	**0.5278**	**0.0054**
APOA5	AM	15	0.4911	0.5121	15	0.4957	0.5135	16	0.4989	0.4621
APOA5	AF	15	0.4992	0.4680	15	0.4745	0.8118	15	0.4922	0.6335
Non-synonymous	EM	7	0.5107	0.2484	7	**0.5527**	**0.0013**	7	**0.5202**	**0.0221**
Non-synonymous	EF	7	**0.5287**	**0.0444**	7	0.5124	0.0916	7	0.5105	0.0701
Non-synonymous	AM	5	0.4868	0.5652	5	0.5073	0.2042	5	0.5076	0.1298
Non-synonymous	AF	7	0.5149	0.1461	7	**0.5147**	**0.0382**	7	0.5044	0.1548

[a]Percentage of individuals drawn from the upper and lower tails of the TG distribution.
[b]EM = European-American males; EF = European-American females; AM = African-American males; AF = African-American females.
[c]Number of variable sites in the combined sample from the tails of the TG distribution.
[d]Values in bold were statistically significant at the $P < 0.05$ level, without correcting for the performance of multiple tests.

European-American males showed the strongest evidence of association, with genotypic differentiation between the phenotypic tails observed in separate tests of each gene. Association tests at *APOA1* and *APOA5* showed statistical significance ($P < 0.05$) across all thresholds (5%, 10%, or 15%) used to define the phenotypic extremes. Seven-site genotypes composed of all surveyed non-synonymous variants from across the gene cluster (five sites in *APOA4* and two sites in *APOA5*) also yielded evidence of association. Comparison of individuals from the upper and lower 10% of the TG distribution combining sites from across the *APOA1/C3/A4/A5* region again identified genotypic differentiation, although this result was not replicated at other thresholds.

Although population-sex groups other than European-American males generally showed weaker evidence for genotypic differentiation between the tails of the TG distribution, some associations emerged. European-American females showed evidence of association in two sets of tests: those involving multi-site genotypes from *APOA5* and those involving seven-site non-synonymous genotypes. *APOA4* genotypes were associated with TG levels in both African-American males and African-American females (when comparing the upper and lower 15% of the distribution). Interestingly, African-American females also showed signs of significant genotypic differentiation between the phenotypic tails at non-synonymous sites (using a threshold of 10%), even though none of these sites appeared to be related to the phenotype in single-site tests (data not shown). Some of the surveyed non-synonymous sites have been previously connected with TG level, including the S19W variant in *APOA5* (site 29009), which appears to affect TG in both healthy and hypertriglyceridemic individuals [reviewed in Shoulders et al., 2004; Klos et al., 2005].

POWER

Figure 2 illustrates the effects of heritability, mode of SNP action, and recombination rate on power for the multi-site test. Considering that only one-fifth (upper 10% + lower 10%) of the total number of individuals was used, the multi-site method retained good power to detect associations under the conditions we investigated. Across recombination rates, power was at least 80% for the additive and epistatic cases in which associations explained 5% or more of the phenotypic variance.

Power was affected by the mode of action of the causal site(s). Across the spectrum of heritabilities, the method showed greater power to detect associations caused by additive SNPs than those caused by recessive action (Fig. 2a and b). When genetic variance was generated purely by an interaction between two sites (additive by additive epistasis), the power was similar to that observed for additive models (Fig. 2c). This result was probably caused by the deviation of allele frequencies from 0.5, which can change the additive and dominance variance components [Falconer and Mackay, 1996], thereby creating marginal population associations even when phenotypic differences are caused by epistasis. Power was also higher for the epistatic model than for two-site additive models (data not shown).

We simulated recombination fractions appropriate for a 4 kb region, assuming an effective population size of 10,000. Power decreased with increasing recombination rate, although not dramatically. Comparing the largest ($4Nr = 4.8$, corresponding to 3 cM/Mb) and smallest ($4Nr = 0$) rates simulated, the average reduction in power across genetic architectures was 5%.

DISCUSSION

THE ROLE OF *APOA1/C3/A4/A5* GENOTYPES IN TG LEVELS

We uncovered evidence of multi-site genotypic differentiation between individuals with different TG levels, from across the *APOA1/C3/A4/A5* cluster. Genotypes constructed from sites grouped by (i) gene cluster membership, (ii) location in individual genes, and (iii) non-synonymous nature each displayed heterogeneity among phenotypically defined populations. Although combinations of sites for testing were chosen based on biological (and not statistical) criteria, the identification of the multi-site associations between genotype and phenotype we report here would be difficult to accomplish using genotype-first methods. In this context, a genotype-first association test would require estimation of phenotypic means and variances for each observed haplotype (genotype) formed by combining polymorphism at large numbers of sites. Our approach suggests that polymorphisms in the *APOA1/C3/A4/A5* gene cluster affect TG levels and that this relationship is visible in multi-site, unphased genotypes.

Our analyses revealed stronger evidence for associations between TG level and genotypes in

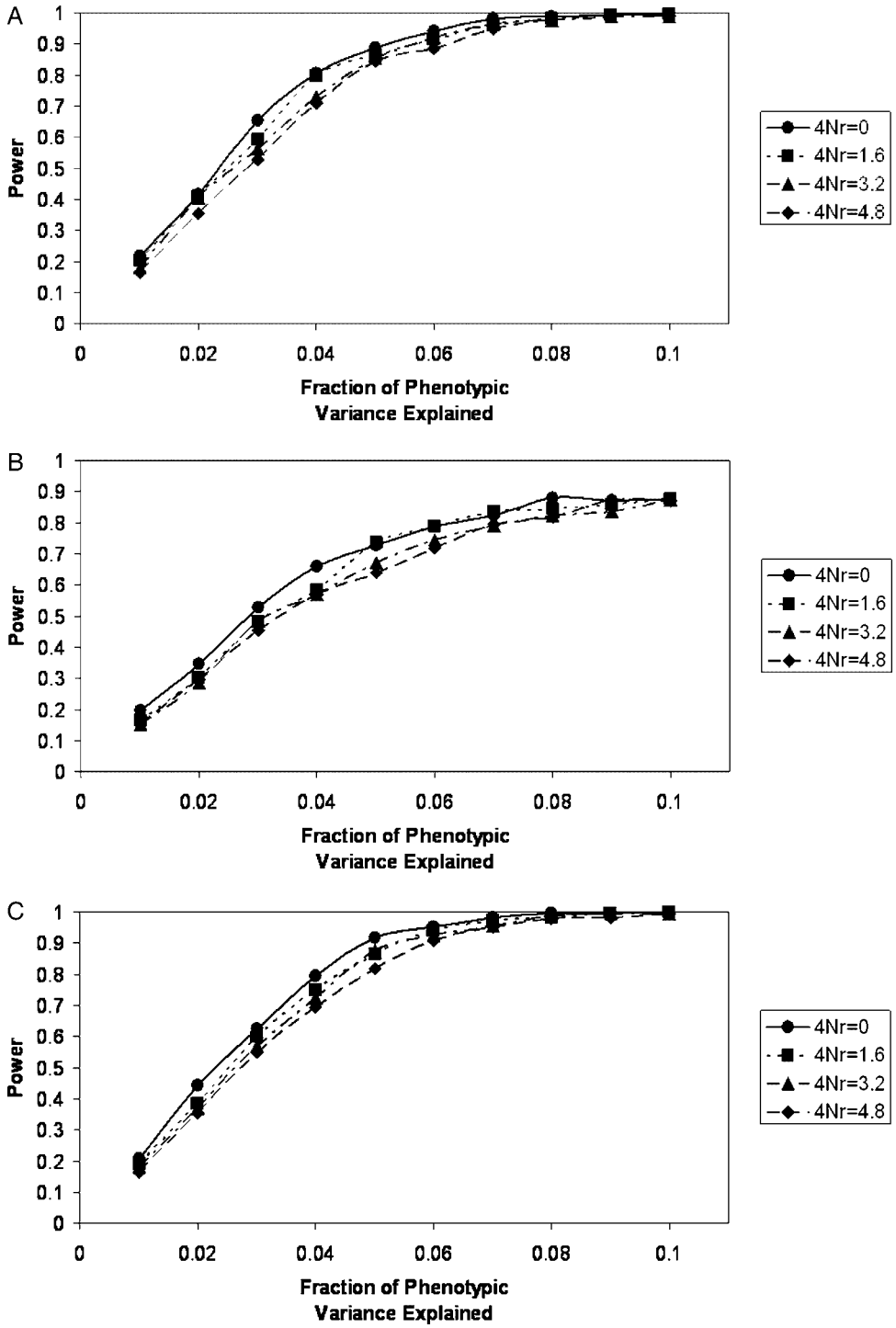

Fig. 2. Power of multi-site approach to detect associations. Genotypes were simulated at 20 sites for 1,000 diploid individuals, and 100 individuals were drawn from each phenotypic tail for comparison. Power is the fraction of 1,000 simulations with $P < 0.05$. (A) One-site additive model. (B) One-site fully recessive model. (C) Two-site (additive by additive) epistatic model. 4Nr is the per-locus population recombination rate.

the *APOA1/C3/A4/A5* gene cluster among European-American males than in other population-sex strata. This discrepancy has several potential causes. First, levels of linkage disequilibrium might be higher in European-American males, providing increased power for detecting associa-

tions. For example, populations with predominantly European ancestry tend to show higher linkage disequilibrium at loci sampled from throughout the genome than do populations derived from Africa [Tishkoff et al., 2000; Reich et al., 2001], an observation that is likely due to the reduced effective population size of European populations. However, the relative paucity of genotype-phenotype associations seen in European-American females in our study argues against the idea that differences in linkage disequilibrium completely explain the results. Second, the relationship between TG level and genotype may differ between European-American males and the other strata.

A third possibility is a combination of increased linkage disequilibrium in European-Americans and a different genetic basis for variation in TG levels among males and females. Biological differences among the sexes in the development of cardiovascular disease, and related risk factors, have been identified [Barrett-Connor, 1997]. Studies of the association between plasma TG levels and variation in genes from the *APOA1/C3/A4/A5* cluster suggest effects of sex in some cases [Lai et al., 2003], but not in others [Pennacchio et al., 2002]. Differences in subject age between studies (our data come from healthy, 18–30-year-olds, while other studies focus on middle-aged, diseased individuals) complicate comparisons between results, however.

ASSOCIATION TESTING BY MULTI-SITE COMPARISON OF PHENOTYPICALLY DEFINED STRATA

Investigators have previously used genotypic comparisons among individuals with extreme phenotypes in linkage mapping [Michelmore et al., 1991; Darvasi and Soller, 1992; Risch and Zhang, 1995], the transmission disequilibrium test [Deng and Li, 2002], and association testing among unrelated individuals [Schork et al., 2000; Tenesa et al., 2003; Cohen et al., 2004]. We have extended this approach to measure multi-site genotypic differentiation between the tails of the phenotypic distribution.

Several caveats accompany our method. First, the best criteria for delimiting the sections of the phenotypic distribution for comparison are unclear. We reported results using three tail sizes: 5%, 10%, and 15%. While some associations between TG level and multi-site genotypes were observed across thresholds (in European-American males), other associations were only discovered using particular criteria. The optimal tail size must balance the benefits of comparing individuals with increasingly divergent phenotypes with the accompanying reduction in sample size. One study of association testing in the context of DNA pooling estimated that for individual sites harboring common alleles with additive effects, sampling 27% from each tail of the phenotypic distribution represents the optimal design [Jawaid et al., 2002], while other authors suggested 5% [Tenesa et al., 2003]; similar calculations have not been reported for multi-site tests. Whether the method may benefit from unequal sampling from the tails of the distribution under certain genetic architectures [Jawaid et al., 2002] also remains to be investigated.

Our simulations focused on relatively simple genetic architectures, with genetic variation attributed to one or two SNPs. Although such a scenario is reasonable for candidate genes underlying variation in complex traits, evaluation of test performance under models with more than two causal sites would provide a more biologically sound assessment. The possible increase in efficiency derived from focusing on the tails of the phenotypic distribution is also partly obviated by the restriction to considering one phenotypic dimension at a time. Furthermore, as is the case for several other available methods, false positives could arise with our approach if the sample is comprised of multiple populations showing both phenotypic and genotypic differences.

Our multi-site method features several important advantages over more traditional approaches to association testing. First, the method directly and simultaneously uses information from many polymorphic sites in a candidate region. This feature facilitates the identification of associations generated by epistasis and reduces the number of tests required to find associations. Second, our test uses unphased genotypes, obviating the need for haplotype reconstruction. Although knowledge of haplotype configurations may aid attempts to correlate genotype with phenotype [Clark, 2004], statistical inference of phase introduces uncertainty that complicates subsequent association testing and reduces power [Morris et al., 2004; Schaid, 2004]. Third, the delineation of groups for comparison based on the phenotype avoids reductions in power that accompany genotype-first approaches. Finally, the tail-based method requires genotypes from fewer individuals, raising the possibility of

selectively genotyping individuals based on their phenotypic values.

Several extensions to the multi-site contrast between phenotypic categories can be imagined. For example, samples formed by combining individuals with extreme phenotypes may exhibit deeper genealogical splits and therefore harbor more intermediate-frequency variants than samples drawn randomly with respect to phenotype. Furthermore, the phenotype-first approach is not limited in application to the tails of the phenotypic distribution. Comparing patterns of multi-site genotypic variation across multiple segments of the phenotypic distribution may yield additional insights, particularly when sites have small phenotypic effects (so that alternative genotypes may not be clearly concentrated in the tails) or interactions between sites are important (since genetic background may differ across the phenotypic distribution). Because the identification of associations in this context is equivalent to detecting population differentiation, any approach that contrasts patterns of variation between populations based on multiple sites could be used. Strategies that explicitly accommodate variation in the underlying genealogy [Zöllner and Pritchard, 2004] seem especially promising.

ACKNOWLEDGMENTS

We thank Sara Hamon for providing Figure 1. We thank Sara Hamon, Kathy Klos, Aida M. Andrés, Jian Li, Kevin Thornton, Scott Williamson, Patricia Wittkopp, and Kristi Montooth for advice, assistance with analyses, and comments on the manuscript. We appreciate the thorough comments of two anonymous reviewers. We thank the CARDIA review board for its guidance and approval of this manuscript.

REFERENCES

Austin MA, Hokanson JE, Edwards KL. 1998. Hypertriglyceridemia as a cardiovascular risk factor. Am J Cardiol 81:7B–12B.

Austin MA, Talmud PJ, Farin FM, Nickerson DA, Edwards KL, Leonetti D, McNeely MJ, Viernes HM, Humphries SE, Fujimoto WY. 2004. Association of apolipoprotein A5 variants with LDL particle size and triglyceride in Japanese Americans. Biochim Biophys Acta—Mol Basis Disease 1688:1–9.

Barrett-Connor E. 1997. Sex differences in coronary heart disease. Why are women so superior? The 1995 Ancel Keys Lecture. Circulation 95:252–264.

Brenn T. 1994. Genetic and environmental effects on coronary heart disease risk factors in northern Norway. The cardiovascular disease study in Finnmark. Ann Hum Genet 58: 369–379.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. 2004. Mapping complex disease loci in whole-genome association studies. Nature 429:446–452.

Clark AG. 2004. The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321–333.

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma cholesterol levels of HDL cholesterol. Science 305:869–872.

Culverhouse R, Klein T, Shannon W. 2004. Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27:1–12.

Darvasi A, Soller M. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor Appl Genet 85:353–359.

Deng HW, Li J. 2002. The effects of selected sampling on the transmission disequilibrium test of a quantitative trait locus. Genet Res 79:161–174.

Eichenbaum-Voline S, Olivier M, Jones EL, Naoumova RP, Jones B, Gau B, Patel HN, Seed M, Betteridge DJ, Galton DJ, Rubin EM, Scott J, Shoulders CC, Pennacchio LA. 2004. Linkage and association between distinct variants of the APOA1/C3/A4/A5 gene cluster and familial combined hyperlipidemia. Arterioscler Thromb Vasc Biol 24:167–174.

Evans D, Buchwald A, Biel FU. 2003. The single nucleotide polymorphism—1131T>C in the apolipoprotein A5 (APOA5) gene is associated with elevated triglycerides in patients with hyperlipidemia. J Mol Med 81:645–654.

Falconer DS, Mackay TFC. 1996. Introduction to Quantitative Genetics. England: Addison Wesley Longman Limited.

Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, Jacobs DR Jr, Liu K, Savage PJ. 1988. CARDIA: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol 41:1105–1116.

Fullerton SM, Buchanan AV, Sonpar VA, Taylor SL, Smith JD, Carlson CS, Salomaa V, Stengard JH, Boerwinkle E, Clark AG, Nickerson DA, Weiss KM. 2004. The effects of scale: variation in the *APOA1/C3/A4/A5* gene. Hum Genet 115:36–56.

Gavrilets S, de Jong G. 1993. Pleiotropic models of polygenic variation, stabilizing selection, and epistasis. Genetics 134: 609–625.

Groenendijk M, Cantor RM, de Bruin TWA, Dallinga-Thie GM. 2001. The apoAI-CIII-AIV gene cluster. Atherosclerosis 157:1–11.

Hudson RR. 1990. Gene genealogies and the coalescent process. Oxford Surveys Evol Biol 7:1–44.

Hudson RR. 2000. A new statistic for detecting genetic differentiation. Genetics 155:2011–2014.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hunt SC, Hasstedt SJ, Kuida H, Stults BM, Hopkins PN, Williams RR. 1989. Genetic heritability and common environmental components of resting and stressed blood pressure, lipids, and body mass index in Utah pedigrees and twins. Am J Epidemiol 129:625–638.

Jannot AS, Essioux L, Reese MG, Clerget-Darpoux F. 2003. Improved use of SNP information to detect the role of genes. Genet Epidemiol 25:158–167.

Jawaid A, Bader JS, Purcell S, Cherny SS, Sham P. 2002. Optimal selection strategies for QTL mapping using pooled DNA samples. Am J Hum Genet 10:125–132.

Kao JT, Wen HC, Chien KL, Hsu HC, Lin SW. 2003. A novel genetic variant in the apolipoprotein A5 gene is associated with hypertriglyceridemia. Hum Mol Genet 12:2533–2539.

Karathanasis SK. 1985. Apolipoprotein multigene family: tandem organization of human apolipoprotein AI, CIII, and AIV genes. Proc Natl Acad Sci USA 82:6374–6378.

Klos KLE, Hamon S, Clark AG, Boerwinkle E, Liu K, Sing CF. 2005. APOA5 polymorphisms influence plasma triglycerides in young, healthy African Americans and whites of the CARDIA Study. J Lipid Res 46:564–570.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution genetic map of the human genome. Nat Genet 31:241–247.

Lai C-Q, Tai E-S, Tan CE, Cutter J, Chew SK, Zhu Y-P, Adiconis X, Ordovas JM. 2003. The *APOA5* locus is a strong determinant of plasma triglyceride concentrations across ethnic groups in Singapore. J Lipid Res 44:2365–2373.

Liu K, Ballew C, Jacobs DR Jr, Sidney S, Savage PJ, Dyer A, Hughes G, Blanton MM. 1989. Ethnic differences in blood pressure, pulse rate, and related characteristics in young adults. The CARDIA study. Hypertension 14:218–226.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720–731.

Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA 88:9828–9832.

Morris RW, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet 74:945–953.

Nelson MR, Kardia SL, Ferrell RE, Sing CF. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11:458–470.

Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM, Rubin EM. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science 294:169–173.

Pennacchio LA, Olivier M, Hubacek JA, Krauss RM, Rubin EM, Cohen JC. 2002. Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. Hum Mol Genet 11:3031–3038.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

Risch N, Zhang H. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584–1589.

Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, Richter D, Lavery T, Kouyoumjian R, Farhadian S, Ward R, Lander E. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Schaid DJ. 2004. Linkage disequilibrium testing when linkage phase is unknown. Genetics 166:505–512.

Schork NJ, Nath SK, Fallin D, Chakravarti A. 2000. Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. Am J Hum Genet 67:1208–1218.

Seltman H, Roeder K, Devlin B. 2003. Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58.

Sham P, Bader JS, Craig I, O'Donovan M, Owen M. 2002. DNA pooling: a tool for large-scaling association studies. Nat Genet 3:862–871.

Shoulders CC, Jones EL, Naoumova RP. 2004. Genetics of familial combined hyperlipidemia and risk of coronary heart disease. Hum Mol Genet 13:R149–R160.

Talmud PJ, Hawe E, Martin S, Olivier M, Miller GJ, Rubin EM, Pennacchio LA, Humphries SE. 2002. Relative contribution of variation withing the *APOC3*/*A4*/*A5* gene cluster in determining plasma triglycerides. Hum Mol Genet 11:3039–3046.

Templeton AR, Boerwinkle E, Sing CF. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics 117: 343–351.

Tenesa A, Knott SA, Carothers AD, Visscher PM. 2003. Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals. Ann Hum Genet 67:557–566.

Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, Destro-Bisol G, Sanjantila A, Lu RB, Deinard AS, Sirugo G, Jenkins T, Kidd KK, Clark AG. 2000. Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins. Am J Hum Genet 67:901–925.

Warnick GR. 1986. Enzymatic methods for quantification of lipoprotein lipids. Methods Enzymol 129:101–123.

Wright S. 1951. The genetical structure of populations. Ann Eugen 15:323–354.

Zöllner S, Pritchard JK. 2004. Coalescent-based association mapping and fine-mapping of complex traits. Genetics 169:1071–1092.

Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association. Nat Rev Genet 5: 89–100.

APPENDIX
COMPUTER SIMULATIONS TO MEASURE STATISTICAL POWER OF THE MULTI-SITE TEST

The coalescent provides a useful tool for investigating the performance of tests that depend on genealogical history [Hudson, 1990], including methods designed to associate genotype with phenotype [Long and Langley, 1999; Zondervan and Cardon, 2004]. Coalescent simulations generate samples of genotypes by (i) creating random genealogies and (ii) adding mutations to these genealogies. To mimic variation in an individual gene from the *APOA1*/*C3*/*A4*/*A5* region, 20-site haplotypes sampled from 4 kb of sequence were simulated, with recombination (see below), for 2,000 chromosomes (1,000 diploid individuals), using the MS program [Hudson, 2002]. Simulations assumed that all variation was neutral and derived from a population at demographic equilibrium. Haplotypes were randomly paired to form genotypes. One or two SNPs (the first in the sequence, or the first and the last, respectively) were used to determine the genotypic effect influencing the phenotype. Only simulation replicates yielding allele frequencies of at least 0.1 at these quantitative trait nucleotides (QTN) were retained. Genetic effects (g_j) were generated for

each individual from these QTN according to the model [Gavrilets and de Jong, 1993].

$$g_j = a(i_m + i_p) + 2di_m i_p + e(i_m + i_p)(i_m + i_p)$$

where i_m and i_p are the two alleles (each with state 0 or 1) that individual j carries at the QTN, a represents an additive effect, d denotes a dominance effect, and e is an epistatic effect (the epistatic case modeled here was additive by additive epistasis). The genetic variance in the population (v_g) was measured as the variance of g. For each individual, a random environmental effect was drawn from a $N(0,v_e)$ distribution, where

$$v_e = v_g\left(\frac{1}{f} - 1\right)$$

and f is the desired fraction of phenotypic variation that is genetic [the broad-sense heritability; Falconer and Mackay, 1996].

Individuals were then selected from the upper and lower 10% of the phenotypic distribution (100 individuals from each tail) and their genotypes were compared using the multi-site test. We evaluated power under varying levels of recombination (the population recombination rate, $4Nr = 0$, 1.6, 3.2, and 4.8, where N is the effective population size and r is the per-locus recombination rate). These values were chosen to sample a reasonable range of recombination rates for the *APOA1/C3/A4/A5* region [0, 1, 2, and 3 cM/Mb; Kong et al., 2002], assuming an effective population size of 10,000. For each recombination rate, we investigated 10 broad-sense heritabilities (ranging from 1% to 10% of the phenotypic variance explained) and three modes of action (one-site additive, one-site fully recessive, and two-site epistatic).