European Journal of Human Genetics (2011) 19, 164–172
© 2011 Macmillan Publishers Limited All rights reserved 1018-4813/11
www.nature.com/ejhg

ARTICLE

Gene-based interaction analysis by incorporating external linkage disequilibrium information

Jing He[1], Kai Wang[2], Andrew C Edmondson[3], Daniel J Rader[3], Chun Li[4,5] and Mingyao Li*,[1]

Gene–gene interactions have an important role in complex human diseases. Detection of gene–gene interactions has long been a challenge due to their complexity. The standard method aiming at detecting SNP–SNP interactions may be inadequate as it does not model linkage disequilibrium (LD) among SNPs in each gene and may lose power due to a large number of comparisons. To improve power, we propose a principal component (PC)-based framework for gene-based interaction analysis. We analytically derive the optimal weight for both quantitative and binary traits based on pairwise LD information. We then use PCs to summarize the information in each gene and test for interactions between the PCs. We further extend this gene-based interaction analysis procedure to allow the use of imputation dosage scores obtained from a popular imputation software package, MACH, which incorporates multilocus LD information. To evaluate the performance of the gene-based interaction tests, we conducted extensive simulations under various settings. We demonstrate that gene-based interaction tests are more powerful than SNP-based tests when more than two variants interact with each other; moreover, tests that incorporate external LD information are generally more powerful than those that use genotyped markers only. We also apply the proposed gene-based interaction tests to a candidate gene study on high-density lipoprotein. As our method operates at the gene level, it can be applied to a genome-wide association setting and used as a screening tool to detect gene–gene interactions.

European Journal of Human Genetics (2011) **19,** 164–172; doi:10.1038/ejhg.2010.164; published online 6 October 2010

Keywords: gene–gene interaction; linkage disequilibrium; imputation

INTRODUCTION

With continued decreasing cost of high throughput genotyping technology, genome-wide association studies (GWAS) are becoming increasingly popular for gene mapping of complex human diseases. Most of the published GWAS papers report results from single-marker-based analysis in which each SNP is analyzed individually. Although this simple approach has led to the discovery of disease susceptibility genes for many diseases, the identified SNPs often only explain a small fraction of the phenotypic variation, suggesting a large number of disease variants are yet to be discovered. There is growing evidence that gene–gene interactions are important contributors to genetic variation in complex human diseases.[1–6] However, detecting gene–gene interactions remains a challenge due to the lack of powerful statistical methods. The most commonly used statistical approach for studying gene–gene interactions is to use a regression framework in which a pair of markers and their interaction terms are included as predictors. When a large number of markers are available, one might consider doing a stepwise regression[7] or a two-stage analysis.[8] Although such methods have been proven useful in simulation studies, they may lose power when multiple interacting variants exist in each gene.

One potential solution to the aforementioned problem is to perform interaction analysis at the gene level. There is increasing recognition for the importance of gene-based analysis.[9] Several methods have been developed to test whether a gene is associated with the trait of interest.[10–13] The central idea of these methods is to summarize marker genotypes into a few components so that the overall degrees of freedom are reduced while most information in the data is retained. Extensive simulations demonstrate that gene-based association analysis can increase the power of detecting genetic association compared with single-marker-based analysis. It is therefore reasonable to expect that gene-based interaction analysis may outperform SNP-based interaction analysis and lead to identification of novel disease susceptibility genes.

Recently, Li *et al*[13] proposed a novel gene-based association test – ATOM, by combining optimally weighted markers within a gene. For each marker in the gene, either genotyped or untyped, an optimally weighted score is derived based on observed genotypes and linkage disequilibrium (LD) information in a reference data set such as the HapMap.[14,15] To reduce the dimensionality of the data, ATOM tests for association using selected principal components (PCs) of these derived scores. Simulations and analysis of real data showed improved power of ATOM over methods that do not incorporate external LD information, especially when the disease loci are not directly genotyped.

The success of ATOM motivated us to extend it to the analysis of gene–gene interactions. Here we describe a PC framework for gene-based interaction analysis. We analytically derive the optimal weight for both quantitative and binary traits based on pairwise LD information. We then use PCs to summarize the information in each gene, and test for interactions between the PCs. We further extend this gene-based interaction analysis procedure to allow the

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA; [2]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; [3]Cardiovascular Institute, University of Pennsylvania, Philadelphia, PA, USA; [4]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA; [5]Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, TN, USA
*Correspondence: Dr M Li, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 213 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA. Tel: +1 215 746 3916; Fax: +1 215 573 4865; E-mail: mingyao@mail.med.upenn.edu
Received 10 December 2009; revised 13 July 2010; accepted 26 August 2010; published online 6 October 2010

use of imputation dosage scores obtained from popular imputation software packages MACH[16] or IMPUTE,[17] which incorporates multi-locus LD information. We evaluate the performance of the proposed tests by extensive simulations and the analysis of a candidate gene study on high-density lipoprotein cholesterol (HDL-C).

METHODS

We consider the problem of gene-based interaction analysis between two genes with multiple markers in each gene. We first present the analytical solutions for quantitative and binary traits assuming the interacting trait loci are known. We then extend the method to the more realistic situation in which the interacting trait loci are unknown.

Quantitative trait

Suppose the quantitative trait of interest, *Y*, is influenced by the interaction between two diallelic quantitative trait loci (QTLs) located in two different genes. Let T_j and t_j (with frequencies p_{T_j} and p_{t_j}, respectively) denote the two alleles at QTL j (=1, 2). Assume the mean of the trait value Y given genotypes g_{T_1} and g_{T_2} can be written as

$$E(Y|g_{T_1}, g_{T_2}) = \alpha_T + \beta_{T_1} g_{T_1} + \beta_{T_2} g_{T_2} + \beta_{T_1,T_2} g_{T_1} g_{T_2}, \tag{1}$$

where $g_{T_j} \in \{0, 1, 2\}$ is the number of allele T_j at QTL j. To detect interaction between the two QTLs, we test $H_0\colon \beta_{T_1,T_2} = 0$. However, as g_{T_1} and g_{T_2} may not be directly observed, the test of interaction is often accomplished through examination of interactions between genotyped markers. Assume a diallelic marker j in gene j (with alleles A_j and a_j and allele frequencies p_{A_j} and p_{a_j}, respectively) is in LD with QTL j. We will show that the mean of Y given genotypes g_1 and g_2 at markers 1 and 2 can be written as

$$E(Y|g_1, g_2) = \alpha + \beta_1 g_1 + \beta_2 g_2 + \beta_{1,2} g_1 g_2, \tag{2}$$

Equation (2) allows indirect assessment of interaction between the QTLs by testing H_0: $\beta_{1,2}$=0.

The regression coefficients β_{T_1,T_2} and $\beta_{1,2}$ reflect the magnitude of interaction effects between the QTLs and between the markers, respectively, and their relationship depends on the degree of LD between the QTLs and the markers. To explicitly derive their relationship, we note that $E(Y|g_1, g_2)$ can be written as

$$\begin{aligned} E(Y|g_1, g_2) &= \sum_{g_{T_1}} \sum_{g_{T_2}} E(Y|g_{T_1}, g_{T_2}) P(g_{T_1}|g_1) P(g_{T_2}|g_2) \\ &= \alpha_T + \beta_{T_1} H_1(g_1) + \beta_{T_2} H_2(g_2) + \beta_{T_1,T_2} H_1(g_1) H_2(g_2), \end{aligned} \tag{3}$$

where $H_j(g_j) = \sum_{g_{T_j}} g_{T_j} P(g_{T_j}|g_j)$, $j = 1, 2$. Li *et al*[13] have shown that when the QTLs and the markers are in Hardy–Weinberg equilibrium in the population,

$$H_j(g_j) = 2(p_{T_j} - \Delta_j/p_{a_j}) + [\Delta_j/(p_{A_j} p_{a_j})] \times g_j, \tag{4}$$

where $\Delta_j = p_{A_j T_j} - p_{A_j} p_{T_j}$ is the LD coefficient between QTL j and marker j. Therefore,

$$\begin{aligned} H_1(g_1) H_2(g_2) = &4[p_{T_1} - \Delta_1/p_{a_1}][p_{T_2} - \Delta_2/p_{a_2}] \\ &+2[p_{T_2} - \Delta_2/p_{a_2}][\Delta_1/(p_{A_1} p_{a_1})] \times g_1 \\ &+2[p_{T_1} - \Delta_1/p_{a_1}][\Delta_2/(p_{A_2} p_{a_2})] \times g_2 \\ &+[\Delta_1/(p_{A_1} p_{a_1})][\Delta_2/(p_{A_2} p_{a_2})] \times g_1 g_2. \end{aligned} \tag{5}$$

If we replace the items in equation (3) accordingly by those in equations (4) and (5), it becomes apparent that equation (2) holds, with

$$\beta_{1,2} = [\Delta_1/(p_{A_1} p_{a_1})][\Delta_2/(p_{A_2} p_{a_2})] \times \beta_{T_1,T_2}. \tag{6}$$

Equation (6) indicates that the two interaction coefficients, β_{T_1,T_2} and $\beta_{1,2}$, differ only by a factor $[\Delta_1/(p_{A_1} p_{a_1})][\Delta_2/(p_{A_2} p_{a_2})]$, which is a function of the marker allele frequencies and the LD coefficients between the QTLs and the markers. The above derivation can be readily extended to binary traits such as disease affection status (see Supplementary Materials).

From the above derivation, we can see that if we define a weighted genotype score for marker j, $g_j^* = [\Delta_j/(p_{A_j} p_{a_j})] g_j$, then the corresponding mean model for Y given the weighted genotype scores becomes

$$\begin{aligned} E(Y|g_1^*, g_2^*) = &\alpha^* + [\beta_{T_1} + 2\beta_{T_1,T_2}(p_{T_2} - \Delta_2/p_{a_2})] g_1^* \\ &+ [\beta_{T_2} + 2\beta_{T_1,T_2}(p_{T_1} - \Delta_1/p_{a_1})] g_2^* + \beta_{T_1,T_2} g_1^* g_2^* \\ = &\alpha^* + \beta_1^* g_1^* + \beta_2^* g_2^* + \beta_{T_1,T_2} g_1^* g_2^*, \end{aligned} \tag{7}$$

where the interaction coefficient becomes the same as that in equation (1). This indicates that for any pair of markers with one from each of the two genes, using weighted genotype scores will result in models that share the same interaction coefficient β_{T_1,T_2}, a fact that we will use below to combine multiple markers within a gene.

Suppose m_j diallelic markers in gene j are genotyped, with alleles $1_{l_j}^{(j)}$ and $0_{l_j}^{(j)}$ for marker l_j $(1 \le l_j \le m_j)$ and allele frequencies $p_{l_j}^{(j)}$ and $q_{l_j}^{(j)}$, respectively. The above derivations suggest that for individual i $(1 \le i \le n)$ and marker l_j $(1 \le l_j \le m_j)$ in gene j, we may consider the weighted genotype score $g_{i,l_j}^{(j)*} = [\Delta_{l_j}^{(j)}/(p_{l_j}^{(j)} q_{l_j}^{(j)})] \times g_{i,l_j}^{(j)}$, where $\Delta_{l_j}^{(j)}$ is the LD coefficient between QTL j and marker l_j in gene j, and $g_{i,l_j}^{(j)}$ denotes the number of allele $1_{l_j}^{(j)}$ carried by individual i. Then the mean of the trait value Y_i given weighted genotype scores at marker l_1 in gene 1 and marker l_2 in gene 2 will be $E(Y_i|g_{i,l_1}^{(1)*}, g_{i,l_2}^{(2)*}) = \alpha_{l_1,l_2}^* + \beta_{l_1}^* g_{i,l_1}^{(1)*} + \beta_{l_2}^* g_{i,l_2}^{(2)*} + \beta_{T_1,T_2} g_{i,l_1}^{(1)*} g_{i,l_2}^{(2)*}$. When all possible marker combinations in the two genes are considered, then all $m_1 \times m_2$ interaction terms share a common interaction coefficient β_{T_1,T_2}. This suggests that for individual i, we can aggregate the information from m_j markers in gene j by defining a weighted genotype score

$$S_i^{(j)} = \frac{1}{m_j} \sum_{l_j=1}^{m_j} \frac{\Delta_{l_j}^{(j)}}{p_{l_j}^{(j)} q_{l_j}^{(j)}} g_{i,l_j}^{(j)} = \frac{1}{m_j} \sum_{l_j=1}^{m_j} w_{l_j}^{(j)} g_{i,l_j}^{(j)}, \tag{8}$$

and then assess the interaction between the two QTLs by examining the cross product of their scores, $S_i^{(1)} \times S_i^{(2)}$. In situations in which the trait locus is in complete or strong LD with a genotyped marker or is itself genotyped, the score in equation (8) may not work well as the other markers will simply add noise and dilute the association signal. Given this consideration, an alternative weighted genotype score is

$$S_i^{(j)} = w_{l_{j,max}}^{(j)} g_{i,l_{j,max}}^{(j)}, \tag{9}$$

where $l_{j,\max}$ is the genotyped marker that has the strongest LD with the trait locus as measured by r^2.

Estimation of weights

In the previous sections, we assumed the trait loci are known. In real data analysis, the locations of the trait loci are unknown. It is reasonable to assume that each of the known polymorphisms in the gene, either genotyped or untyped in the study sample, is equally likely to be the trait locus. For each such locus, we can estimate the weights for all the genotyped markers and calculate a score for the locus. Following Li *et al*,[13] we propose to estimate the weight using LD information obtained from a reference database such as that generated by the HapMap, other publicly available dense SNP data sets or resequencing data from a subset of the study sample. Suppose M_j markers are available for gene j in the reference data set, and they are a superset of the markers genotyped in the study sample. If marker k_j $(1 \le k_j \le M_j)$ in the reference data set is the trait locus, then the weight for marker l_j in the study sample is

$$w_{k_j,l_j}^{(j)} = \frac{\Delta_{k_j,l_j}^{(j)}}{p_{l_j}^{(j)} q_{l_j}^{(j)}}, \tag{10}$$

where $\Delta_{k_j,l_j}^{(j)}$ is the LD coefficient between markers k_j and l_j, and $p_{l_j}^{(j)}$ and $q_{l_j}^{(j)}$ are allele frequencies at marker l_j in gene j. These quantities can be estimated from the reference data set. For individual i and each marker k_j in the reference data set, we can calculate a weighted genotype score

$$S_{i,k_j}^{(j)} = \frac{1}{m_j} \sum_{l_j=1}^{m_j} w_{k_j,l_j}^{(j)} g_{i,l_j}^{(j)}, \tag{11}$$

or an alternative weighted genotype score

$$S^{(j)}_{i,k_j} = w^{(j)}_{k_j,l_{j,\max}} g^{(j)}_{i,l_{j,\max}}, \tag{12}$$

where $l_{j,\max}$ is the genotyped marker that has the strongest LD with marker k_j.

We note that the weighted genotype scores in equations (11) and (12) share similarity with imputation dosage scores, and they can be considered as a simple version of the multilocus LD-based imputation dosage scores obtained from software packages such as MACH and IMPUTE. Although using pairwise-LD information only, the weighted genotype scores in equations (11) and (12) provide an intuitive justification of why incorporating external LD information may provide power gain for association testing. In the following sections, we will consider both the pairwise LD-based weighted genotype scores and multilocus LD-based imputation dosage scores in the testing procedure.

Gene-based interaction analysis

Once we have calculated the scores, either weighted genotype scores in equation (11) or (12) or imputation dosage scores in MACH or IMPUTE, for each marker in the reference data set, we can then test for gene–gene interaction based on the scores $(S^{(j)}_1, ..., S^{(j)}_{M_j})$ for gene j, where $S^{(j)}_{k_j} = (S^{(j)}_{1,k_j}, ..., S^{(j)}_{n,k_j})^T$ and n is the total number of individuals in the study. As the trait loci are unknown, a simple test of interaction could be to include all pairwise interactions of the imputation dosage scores in a regression framework and then test for their overall significance. However, this approach may suffer from low power due to the large number of degrees of freedom involved. To efficiently aggregate all information while reducing the degrees of freedom, we propose to test for gene–gene interaction using PCs obtained from the scores. Without loss of generality, for gene j, we order its PCs such that $PC^{(j)}_1$ has the largest variance and $PC^{(j)}_2$ has the second largest variance and so on.

Once the PCs are computed, we can then test for gene–gene interaction by conducting a regression analysis with a set of selected PCs and their interactions as covariates. As the PCs are ordered by the magnitude of explained variance, for each gene, we select the first several PCs that explain a prespecified fraction of the total variance. Suppose L_j PCs are selected for gene j. For a binary trait, we can fit the data with the following logistic regression model (see Supplementary Material)

$$\begin{aligned}\text{logit}\,[P(Y=1|\text{genotypes})] \\ = \alpha + \sum_{l_1=1}^{L_1} \beta^{(1)}_{l_1} PC^{(1)}_{l_1} + \sum_{l_2=1}^{L_2} \beta^{(2)}_{l_2} PC^{(2)}_{l_2} + \sum_{l_1=1}^{L_1}\sum_{l_2=1}^{L_2} \beta_{l_1,l_2} PC^{(1)}_{l_1} PC^{(2)}_{l_2}.\end{aligned} \tag{13}$$

For a quantitative trait, we can fit the data with a linear regression model. Under the null hypothesis of no interaction between the two genes, $\beta_{l_1,l_2} = 0$ for l_1=1,…, L_1 and l_2=1,…, L_2. We can test this null hypothesis by a likelihood ratio test, and the corresponding test statistic is approximately distributed as a χ^2 distribution with $L_1 \times L_2$ degrees of freedom. We call this test as a *global* test. Alternatively, we can conduct pairwise interaction analysis between all selected PCs and choose the statistic for the most significant pair as the test statistic and evaluate its significance by Bonferroni correction. We call this test as a *pairwise* test. In our analyses, we used 90% threshold for the fraction of variance as it generally provides better power than other variance thresholds in scenarios we considered. We note that for binary traits, the null hypothesis tested by logistic regression is only an approximation to the null hypothesis that $(f_{2,2}–f_{2,0})–(f_{0,2}–f_{0,0})=0$, and thus the weighted genotype scores we derived earlier may not be 'optimal'. However, as shown in the Supplementary Material, this approximation is probably valid as long as the interaction effects are not too strong and the disease is not common.

RESULTS

In this section, we evaluate the performance of the gene-based interaction tests for binary traits and compare with SNP-based interaction test. We considered four gene-based interaction tests: (1) ATOM-AVG, which uses weighted genotype scores from equation (11); (2) ATOM-MAX, which uses weighted genotype scores from equation (12); (3) MACH, which uses imputation dosage scores from MACH; and (4) PCA, which uses genotyped markers only. For each test *T*, we considered two versions: (1) the global version, which tests for the joint interaction effect of all selected PCs; and (2) the pairwise version, which tests for the pairwise interaction among all selected PCs. Significance for the pairwise version is adjusted by Bonferroni correction. For the SNP-based interaction analysis, we only considered the pairwise version as the power of the global version is extremely low due to the large number of degrees of freedom.

Comparison of type I error and power based on simulated data

We simulated data based on the LD structures of two genes *CHI3L2* (Figure 1) and *PTPN22* (Figure 2), both are located on chromosome 1 but are in linkage equilibrium with each other. For each gene, we considered common SNPs with minor allele frequency $\geq$0.05 and selected tagSNPs using the program Tagger[18] with pairwise tagging at $r^2 \geq 0.8$. We identified 25 common SNPs for *CHI3L2* and selected seven tagSNPs; for *PTPN22*, 29 common SNPs and 9 tagSNPs. We assumed that only the tagSNPs were genotyped and available for analysis, a common scenario in both candidate gene and GWAS studies. To simulate case–control data with LD, we first estimated the haplotype frequencies of the tagSNPs for each gene, and then simulated the genotype data according to the estimated haplotype frequencies. We considered two situations: (1) each gene has only one disease locus; and (2) each gene has two disease loci.

For the first situation, we designated one locus in each gene as the disease locus, and the case–control status for individual i was simulated according to the following model

$$\begin{aligned}\text{logit}\,[P(Y_i=1|g^{(1)}_{i,D_1}, g^{(2)}_{i,D_2})] =& \alpha + 0.2g^{(1)}_{i,D_1} + 0.2g^{(2)}_{i,D_2} \\ &+ 0.3g^{(1)}_{i,D_1} g^{(2)}_{i,D_2},\end{aligned}$$

where α is determined in a way such that the overall disease prevalence is 5%. Power was estimated based on 1000 replicate data sets each consisting of 2000 cases and 2000 controls and significance was assessed at the 1% level. The type I error rate was evaluated based on 10 000 data sets by setting the interaction effect in the above logit model to 0. Since gene-based interaction tests based on ATOM and MACH require external LD information, we simulated 60 individuals (mimicking the HapMap CEU samples) as a reference data set and then calculated the weighted genotype scores or the imputation dosage scores using the LD information estimated from these 60 individuals.

As the performance of different tests may vary depending on whether the disease loci are genotyped or not, we considered three scenarios: (1) both disease loci are genotyped; (2) only one of the disease loci is genotyped; and (3) both disease loci are untyped. A thorough evaluation of all tests would require consideration of 25×29=725 combinations. To avoid extensive simulations for all marker combinations, we classified the markers in each gene into three categories according to LD levels. Specifically, a marker is classified into the 'strong LD' category if five or more markers in the gene have $r^2 > 0.8$ with it; a marker is in the 'moderate LD' category if three to five markers in the gene have $r^2 > 0.8$ with it; the rest are in the 'weak LD' category. On the basis of this classification, markers in *CHI3L* fall into either strong or weak LD categories. By classifying markers in this manner, we were able to investigate the performance of various tests under a wide range of settings, yet avoided simulations of all marker combinations.

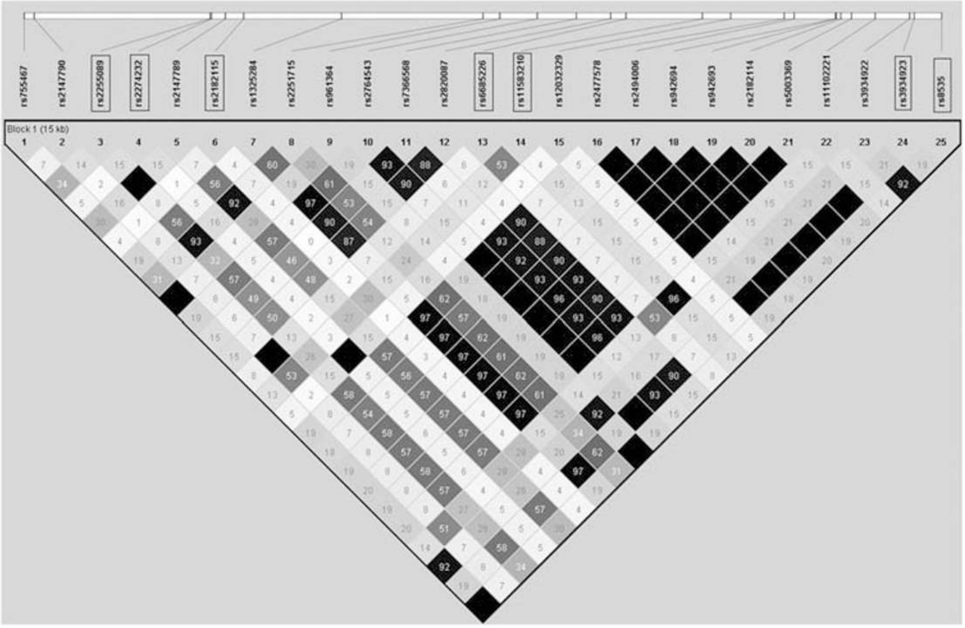

Figure 1 LD structure of *CHI3L2* on chromosome 1 in the HapMap CEU samples. Displayed is estimated r^2 for 25 SNPs with MAF ≥ 0.05. SNPs within the black boxes are tagSNPs selected using the Tagger program at r^2 threshold of 0.8.

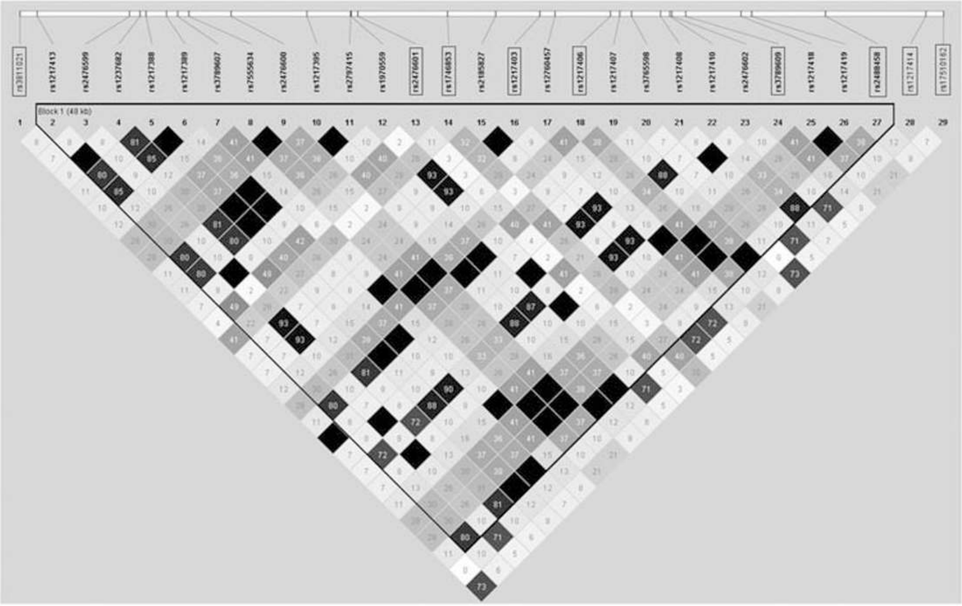

Figure 2 LD structure of *PTPN22* on chromosome 1 in the HapMap CEU samples. Displayed is estimated r^2 for 29 SNPs with MAF ≥ 0.05. SNPs within the black boxes are tagSNPs selected using the Tagger program at r^2 threshold of 0.8.

Table 1 displays the estimated type I error rates under two-locus interaction model. The type I error rates of all tests are under control. Not surprisingly, for each test, the pairwise version of the test is more conservative than the global version due to the correction of a large number of pairwise comparisons. Table 2 shows the estimated power. As expected, when there is a single disease locus in each gene, $T_{\text{SNP-pairwise}}$ consistently outperforms the other tests. Among the other tests we considered, $T_{\text{ATOM-AVG-global}}$, $T_{\text{ATOM-MAX-global}}$ and $T_{\text{MACH-global}}$, which incorporate external LD information, offer better power, followed by $T_{\text{PCA-global}}$. We note that the powers of ATOM- and MACH-based tests are similar, despite that MACH is much more computationally intensive. For example, it requires ~ 210 s to finish one simulation for MACH-based tests with 2000 cases and 2000 controls; however, the required computing time for ATOM-based tests is only ~ 5 s, 40 times faster.

Table 1 Type I error rates (%) under a two-locus interaction model in which one locus in *CHI3L2* interacts with one locus in *PTPN22*

Disease locus in CHI3L2		*Disease locus in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD category	*SNP*	*LD category*	*SNP*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G		G										
W	3	W	13	0.82	1.04	0.79	1.15	0.70	1.10	1.00	1.08	0.80
S	24	S	27	0.72	1.10	0.94	0.93	0.68	1.03	0.91	1.13	1.03
G		U										
W	3	S	5	0.65	1.09	0.81	1.10	0.66	1.00	0.98	1.17	0.77
S	24	M	10	0.73	1.16	0.95	0.99	0.51	1.20	0.82	0.98	0.88
U		G										
W	5	S	27	0.71	1.18	0.82	1.15	0.57	1.04	0.90	1.09	0.96
S	7	W	24	0.74	1.06	0.97	1.12	0.84	1.08	1.13	0.91	0.75
U		U										
W	5	M	12	0.60	0.85	0.86	1.00	0.48	0.86	0.91	0.93	0.77
S	7	W	7	0.72	0.98	1.05	1.14	0.61	1.00	1.06	0.92	0.91

Abbreviations: G/U, genotyped/imputed; S/M/W, strong/moderate/weak LD category.

Table 2 Comparison of power (%) under a two-locus interaction model in which one locus in *CHI3L2* interacts with one locus in *PTPN22*

Disease locus in CHI3L2		*Disease locus in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD category	*SNP*	*LD Category*	*SNP*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G		G										
W	3	W	13	40.6	23.9	21.0	29.4	24.2	18.5	6.1	20.1	17.7
W	3	M	16	60.3	51.9	52.1	56.1	48.7	54.2	29.2	56.6	40.3
W	3	S	27	71.6	65.6	66.0	67.9	65.6	68.3	40.2	66.3	58.1
S	24	W	13	36.4	21.6	14.6	24.9	18.7	17.6	5.6	16.3	17.8
S	24	M	16	55.9	44.5	38.9	47.4	39.3	51.3	26.3	50.0	46.5
S	24	S	27	68.0	61.5	52.1	58.7	55.0	64.7	38.7	66.7	72.1
			Mean	55.5	44.8	40.8	47.4	41.9	45.8	24.4	46.0	42.1
U		G										
W	5	W	13	40.1	20.3	18.1	27.6	24.0	18.6	14.5	19.7	14.8
W	5	M	16	60.1	51.4	54.7	58.6	50.5	54.3	37.1	56.5	41.2
W	5	S	27	72.7	66.7	62.2	69.1	63.2	67.5	54.9	66.3	58.5
S	7	W	14	80.1	73.7	65.5	72.6	60.1	75.0	76.9	70.0	56.1
S	7	M	16	55.3	46.3	41.2	50.8	39.2	52.4	47.7	50.1	46.0
S	7	S	27	68.6	60.2	50.8	59.5	52.8	62.7	68.4	66.8	72.8
			Mean	62.8	53.1	48.8	56.4	48.3	55.1	49.9	54.9	48.2
G		U										
W	3	W	17	73.6	69.3	58.7	71.7	59.3	70.4	52.1	67.3	56.8
W	3	M	3	61.0	52.8	52.0	57.4	50.1	56.3	38.6	57.7	42.5
W	3	S	5	74.3	67.5	62.8	69.7	63.5	69.0	55.2	65.9	57.8
S	24	W	17	67.4	59.4	42.2	59.8	48.2	63.4	63.4	65.8	67.5
S	24	M	15	58.5	50.3	42.9	52.4	40.4	54.2	49.1	52.8	46.5
S	24	S	5	66.5	59.8	50.9	58.5	51.8	63.2	67.9	63.6	68.4
			Mean	66.9	59.8	51.6	61.6	52.2	62.8	54.4	62.2	56.6
U		U										
W	5	W	17	74.1	66.5	60.7	72.5	60.1	69.3	53.7	68.9	59.1
W	5	M	12	61.9	55.5	53.8	60.7	50.5	56.3	36.9	54.8	39.9
W	5	S	5	72.9	65.0	62.6	69.8	63.6	65.8	53.6	67.5	61.3
S	7	W	7	66.4	59.0	45.7	61.0	51.2	62.4	66.6	63.0	66.1
S	7	M	3	55.6	47.1	40.7	50.5	40.5	53.3	47.2	54.3	50.7
S	7	S	5	66.1	61.5	46.6	57.6	51.7	63.6	68.7	62.0	69.0
			Mean	66.2	59.1	51.7	62.0	52.9	61.8	54.5	61.8	57.7

Abbreviations: G/U, genotyped/imputed; S/M/W, strong/moderate/weak LD category.

Table 3 Type I error rates (%) under a four-locus interaction model in which two loci in *CHI3L2* interact with two loci in *PTPN22*

Disease loci in CHI3L2		*Disease loci in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD Category	*SNPs*	*LD Category*	*SNPs*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G, G		G, G										
W, W	4, 14	M, S	16, 27	0.83	1.10	0.91	1.12	1.01	1.18	1.08	1.05	1.00
W, S	4, 24	W, M	1, 16	0.76	0.99	0.92	1.06	0.98	0.98	0.94	0.97	1.03
G, U		G, U										
W, W	2, 4	W, M	1, 3	0.73	1.05	1.00	1.03	0.92	1.04	0.77	1.17	1.03
W, S	4, 7	M, M	3, 18	0.62	1.28	0.92	1.21	1.05	1.13	0.94	1.08	1.09
U, U		U, U										
W, S	15, 1	W, W	7, 20	0.82	1.15	0.98	1.18	1.05	1.13	0.97	1.19	1.06
W, S	15, 11	M, S	23, 19	0.73	0.87	0.97	0.96	0.99	0.88	0.95	1.07	1.01

Abbreviations: G/U, genotyped/imputed; S/M/W, strong/moderate/weak LD category.

For complex diseases, it might be an oversimplification to consider only one disease locus per gene. To evaluate the performance of different tests under a more complicated setting, we considered a model in which two loci in *CHI3L2* interact with two loci in *PTPN22*. Specifically, we simulated case–control status according to the model:

$$\begin{aligned}\mathrm{logit}[P(Y_i=1|g^{(1)}_{i,D_1},g^{(1)}_{i,D_2},g^{(2)}_{i,D_1},g^{(2)}_{i,D_2})] = {} & \alpha+0.2(g^{(1)}_{i,D_1}+g^{(1)}_{i,D_2}\\ & +g^{(2)}_{i,D_1}+g^{(2)}_{i,D_2})+0.3(g^{(1)}_{i,D_1}g^{(2)}_{i,D_1}\\ & +g^{(1)}_{i,D_1}g^{(2)}_{i,D_2}+g^{(1)}_{i,D_2}g^{(2)}_{i,D_1}+g^{(1)}_{i,D_2}g^{(2)}_{i,D_2}).\end{aligned}$$

Again, the overall disease prevalence was set at 5% by adjusting the value of α. For type I error estimation, we set the coefficient for the interaction effect at 0.

As shown in Table 3, the type I error rates of all interaction tests are under control. Table 4 shows the power comparison results. These results indicate that all gene-based interaction tests outperform the SNP-based test. For example, the power advantage of $T_{\text{ATOM-MAX-global}}$ over $T_{\text{SNP-pairwise}}$ as measured by mean power difference ranged from 12.7 to 27.3%. This is much higher than the mean power difference (4.1–9.7%) between the two tests under the simpler disease models in Table 2. This indicates that SNP-based interaction analysis is not sufficient when multiple loci in a gene interact with multiple loci in another gene. Among all gene-based tests we considered, $T_{\text{MACH-global}}$ is generally the most powerful test, followed by $T_{\text{ATOM-MAX-global}}$, $T_{\text{ATOM-AVG-global}}$ and $T_{\text{PCA-global}}$. It is worth noting that the power of $T_{\text{ATOM-MAX-global}}$ is only slightly lower than $T_{\text{MACH-global}}$ despite that MACH is much more computationally intensive. The pairwise versions of these three tests are typically less powerful than the global versions of the tests. Moreover, our results clearly indicate the advantage of incorporating external LD information in the analysis. The power gain of $T_{\text{ATOM-MAX-global}}$ over $T_{\text{PCA-global}}$ as measured by the mean power difference ranged from 3.3 to 7.9%, and the power gain of $T_{\text{MACH-global}}$ over $T_{\text{PCA-global}}$ ranged from 4.6 to 10.7%.

Application to the HDL data set

We applied the three gene-based interaction tests to an ongoing candidate gene study on subjects with extreme levels of HDL-C. In this study, 625 subjects of European ancestry with HDL >90th percentile were considered as cases and 606 subjects with HDL <30th percentile were considered as controls. All study subjects were genotyped using the IBC 50K SNP array.[19] Our previous SNP pairwise interaction analysis on this data set reveals that a number of SNPs in *CETP* significantly interact with several SNPs in *BCAT1*. It is well known that *CETP* promotes the transfer of cholesteryl esters from HDL to low-density lipoprotein, and individuals that are genetically deficient for *CETP* often have extremely high HDL levels.[20,21] In a recent GWAS on biochemical traits, *BCAT1* is shown to be significantly associated with serum albumin concentration.[18] As albumin is correlated with HDL,[22] it is possible that *CETP* and *BCAT1* interact in modulating the level of HDL-C.

Figures 3 and 4 display the LD structures of *CETP* and *BCAT1* estimated using the HDL data set. We downloaded genotype data at these two genes for the CEU samples from the HapMap website. For *CETP*, there are 31 common SNPs in the HapMap, whereas the HDL data set has 57, with 27 common SNPs in both data sets. As the HapMap data set does not provide much additional LD information, for ATOM-based tests, we calculated the weighted genotype scores using the LD information provided by the 57 SNPs in the HDL controls. For *BCAT1*, 164 common SNPs are in the HapMap and 79 are in the HDL data set, with 56 in both. For the 164 SNPs in the HapMap, we calculated their weighted genotype scores using the LD information provided by the HapMap; for the 23 SNPs in the HDL data set but not in the HapMap, we used their observed genotypes in the HDL data set.

The *BCAT1* SNPs are in several LD blocks with weak LD between some of the blocks, requiring 23 PCs to explain 90% of the variance. Testing interaction using all SNPs in *BCAT1* may have low power due to the large number of degrees of freedom. To reduce the dimensionality, we divided the SNPs in *BCAT1* into four blocks (Figure 4) and tested interaction between *CETP* and each of the four blocks. We found significant interaction between *CETP* and the third block of *BCAT1*. The *P*-value of $T_{\text{ATOM-AVG-global}}$ is 0.0034. In comparison, the *P*-values of $T_{\text{ATOM-MAX-global}}$, $T_{\text{MACH-global}}$ and $T_{\text{PCA-global}}$ are 0.22, 0.25 and 0.072, respectively. The *P*-values of the pairwise versions of the four tests are 0.035, 0.062, 0.38 and 0.029, respectively. The *P*-value of $T_{\text{SNP-pairwise}}$ is 0.078. Compared with other gene-based interaction tests, $T_{\text{ATOM-AVG-global}}$ clearly revealed stronger evidence of association.

DISCUSSION

We have proposed a PC framework for gene-based interaction analysis. Our tests are based on the aggregation of information from weighted genotype scores using pairwise LD information or imputation dosage scores using multilocus LD information in a gene.

Table 4 Comparison of power (%) under a four-locus interaction model in which two loci in *CHI3L2* interact with two loci in *PTPN22*

Disease loci in CHI3L2		*Disease loci in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD category	*SNPs*	*LD category*	*SNPs*	pairwise	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G, G		G, G										
W, W	4, 14	W, W	1, 13	25.5	44.8	30.6	51.9	22.7	47.8	22.2	53.2	25.7
W, W	4, 14	W, M	1, 16	44.8	68.5	39.5	74.1	33.7	72.5	52.3	73.3	41.2
W, W	4, 14	M, M	16, 18	59.7	71.7	50.0	75.0	56.6	74.8	61.3	75.7	65.7
W, W	4, 14	M, S	16, 27	18.8	58.0	39.7	67.3	52.3	69.5	57.7	73.3	56.7
W, W	4, 14	W, S	24, 27	39.9	60.7	37.0	67.3	36.0	66.6	32.8	68.7	39.8
W, S	4, 24	W, W	13, 28	42.8	55.5	62.1	60.5	71.8	47.7	46.7	45.4	37.5
W, S	4, 24	W, M	1, 16	77.3	66.2	62.1	72.9	60.8	72.5	56.9	75.2	46.9
W, S	4, 24	M, M	16, 18	73.8	69.7	64.3	74.4	66.7	75.4	62.8	75.1	67.8
W, S	4, 24	M, S	16, 27	42.3	60.6	68.7	68.3	80.5	73.8	66.5	73.3	64.9
W, S	4, 24	W, S	24, 27	71.2	60.6	62.1	67.7	58.1	67.7	39.4	71.7	49.9
			Mean	49.6	61.6	51.6	67.9	53.9	66.8	49.9	68.5	49.6
G, U		G, U										
W, W	3, 15	W, W	1, 7	84.6	94.9	93.8	96.9	95.4	98.1	88.8	98.4	86.7
W, W	2, 4	W, M	1, 3	31.6	58.0	34.3	50.1	23.0	51.3	28.6	53.2	24.3
W, W	2, 4	M, M	3, 18	48.6	67.8	48.4	63.6	44.7	63.5	48.2	64.5	50.6
W, W	3, 15	M, S	16, 19	82.2	96.1	92.7	97.6	95.3	98.2	94.6	98.4	94.2
W, W	2, 4	W, S	1, 22	30.1	60.2	37.8	49.5	22.8	50.1	25.6	55.3	25.3
W, S	4, 7	W, W	1, 7	40.9	58.2	67.9	67.8	77.3	71.9	55.1	73.4	57.0
W, S	4, 7	W, M	1, 3	74.7	66.0	61.4	72.5	62.7	73.0	58.3	74.7	48.5
W, S	4, 7	M, M	3, 18	75.9	69.8	63.9	74.4	60.8	76.4	62.1	77.1	70.7
W, S	4, 7	M, S	16, 19	40.3	59.8	67.8	68.1	77.3	71.8	62.4	72.5	66.2
W, S	4, 7	W, S	1, 19	78.5	66.5	64.6	73.8	61.6	75.4	56.8	75.2	56.1
			Mean	58.7	69.7	63.3	71.4	62.1	73.0	58.1	74.3	58.0
U, U		U, U										
W, W	1, 15	W, W	7, 20	20.8	56.8	54.1	67.0	57.9	70.0	41.5	70.8	39.3
W, W	1, 15	W, M	20, 3	46.4	66.3	49.7	72.9	45.0	71.6	43.3	71.9	31.6
W, W	1, 15	M, M	3, 8	29.0	46.1	25.5	52.2	24.9	51.8	21.2	74.3	59.2
W, W	1, 15	M, S	3, 22	18.0	54.5	51.5	61.9	54.7	63.3	45.6	63.2	41.7
W, W	1, 15	W, S	7, 19	41.8	62.4	53.9	69.1	39.3	67.6	28.4	70.8	33.2
W, S	15, 11	W, W	7, 20	41.9	59.1	68.2	65.7	75.2	70.4	53.7	74.2	56.0
W, S	15, 11	W, M	7, 15	74.7	62.7	56.0	71.1	63.8	71.6	60.3	72.0	59.6
W, S	15, 11	M, M	3, 8	76.0	68.9	63.5	75.1	64.9	75.6	62.8	76.6	71.3
W, S	12, 2	M, S	12, 19	17.7	63.0	55.4	56.3	45.9	64.9	37.6	62.2	37.6
W, S	15, 11	M, S	23, 19	45.4	65.5	71.9	74.0	80.6	76.7	66.2	76.3	67.3
			Mean	41.2	60.5	55.0	66.5	55.2	68.4	65.1	71.2	49.7

Abbreviations: G/U, genotyped/imputed; S/M/W, strong/moderate/weak LD category.

To reduce dimensionality, the scores within a gene are further summarized into PCs and then used in a regression framework for interaction analysis. By extensive simulations under various settings and the analysis of a real data set, we demonstrated that gene-based interaction tests are a powerful alternative to the conventional SNP-based interaction test and to approaches that do not incorporate external LD information.

The gene-based interaction tests consider each gene as a testing unit and tests for interaction at the gene level. Compared with methods that operate at the marker level, a key advantage of gene-based interaction tests lies in their ability to capture all potential risk conferring variants in a gene. This makes gene-based interaction tests particularly attractive when multiple disease loci in a gene interact with multiple disease loci in another gene. We note that when a single locus in a gene interacts with a single locus in another gene, or when some of the interaction effects are weak when more than two loci interact, the SNP-based interaction test may perform well, as such a simple test can capture the interaction effect more effectively than the gene-based interaction tests.

Another advantage of gene-based interaction analysis over the conventional SNP-based interaction analysis is that it requires much less number of tests. For example, for the IBC data with 50 000 SNPs genotyped in 2000 candidate genes, the conventional SNP pairwise interaction analysis will involve ~1.25 billion tests. In contrast, using gene-based interaction analysis, the number of tests is reduced to 2 million. For large-scale candidate gene and GWAS data sets, gene-based interaction tests can be used as a screening tool. After a pair of significant interacting genes is identified, one can then conduct further investigation to evaluate which SNPs within the genes significantly interact.

Our method concerns with gene-based tests of interaction effect. We note that there exist gene-based methods that jointly test for the main effect and the interaction effect.[23,24] Although the goals of these tests are slightly different from ours, they all aim to incorporate information contributed by multiple markers in a gene. How to

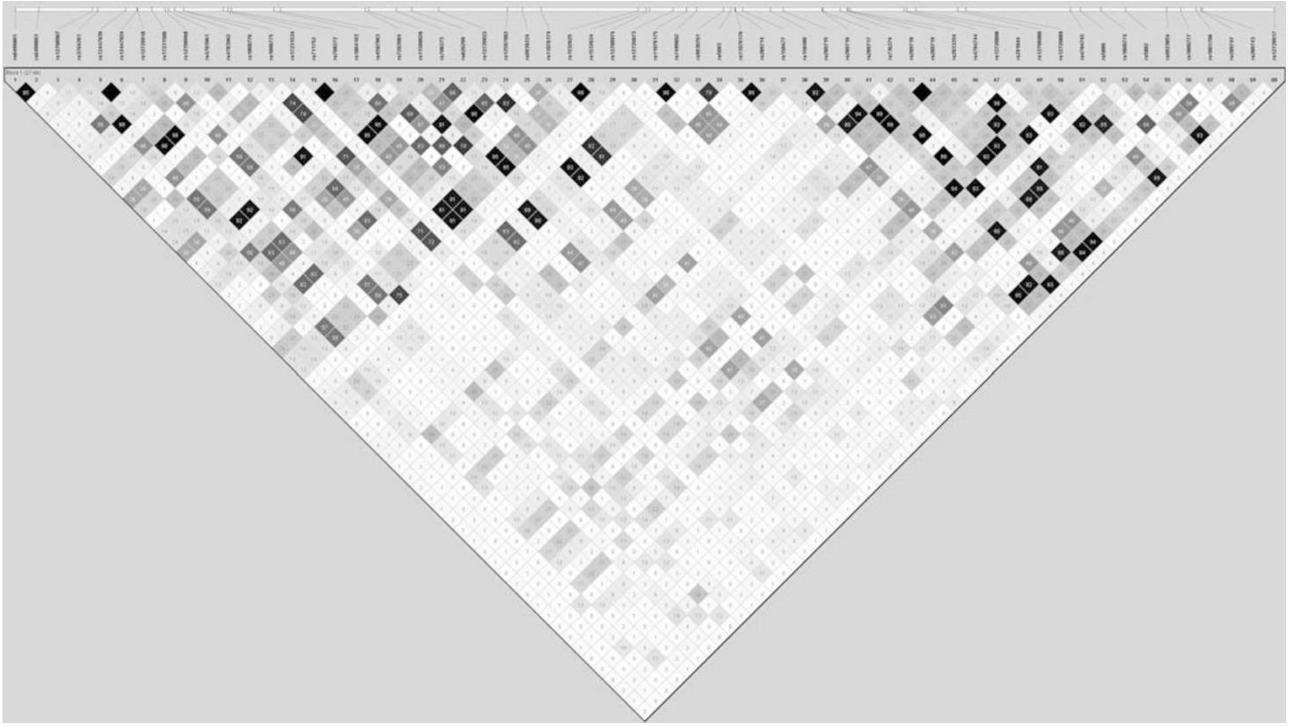

Figure 3 LD structure of *CETP* on chromosome 16 in the HDL data set. Displayed is estimated r^2 for 57 SNPs with MAF ≥0.05 based on the controls.

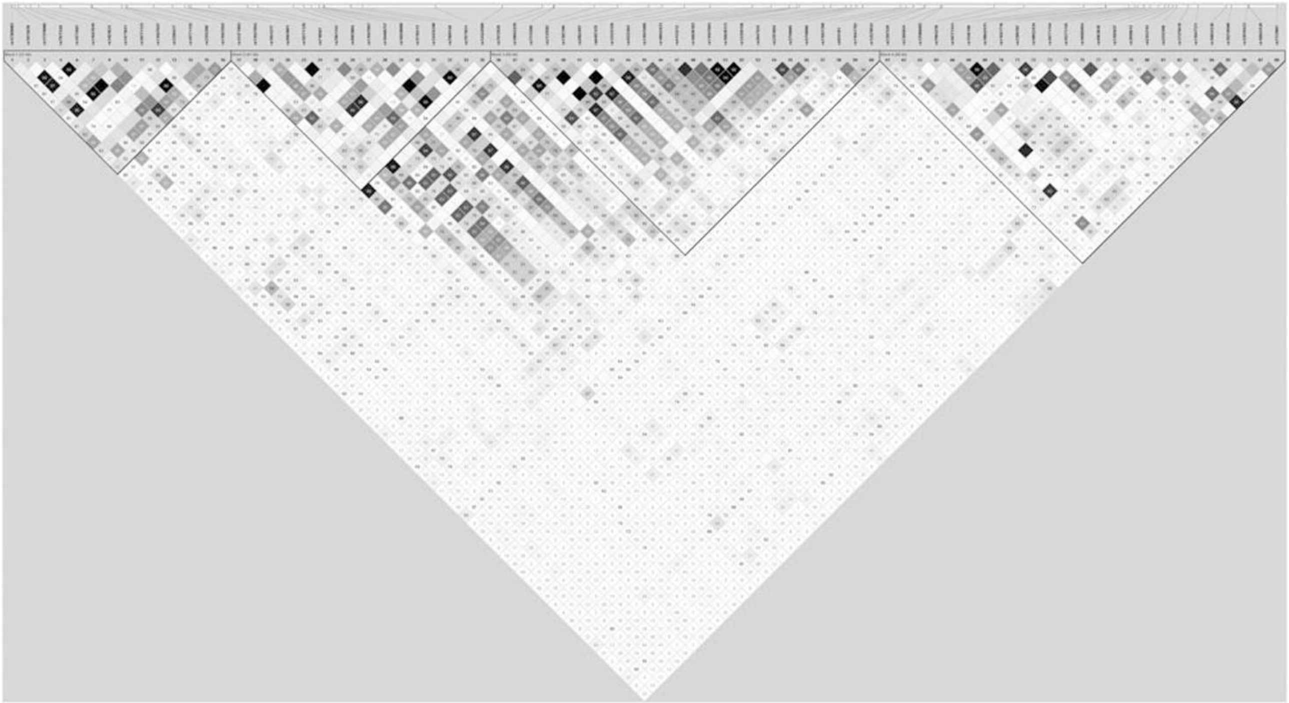

Figure 4 LD structure of *BCAT1* on chromosome 12 in the HDL data set. Displayed is estimated r^2 for 79 SNPs with MAF ≥0.05 based on the controls.

extend the proposed PC framework to jointly test for the main and interaction effects would merit further research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported by Grant R01HG004517 (to ML and CL). We thank Dr Hongzhe Li for helpful discussions. We also thank the reviewers for their constructive comments that greatly improved the paper.

1 Cordell HJ, Todd JA, Bennett ST *et al*: Two-locus maximum LOD score analysis of a multifactorial trait: 7 joint consideration of IDDM2 and IDDM4 with DDM1 in type 1 diabetes. *Am J Hum Genet* 1995; **57**: 920–934.
2 Cox NJ, Frigge M, Nicolae DL *et al*: Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to type 2 diabetes. *Nat Genet* 1990; **21**: 213–215.
3 Howard TD, Koppelman GH, Xu J *et al*: Gene-gene interaction in asthma: *IL4RA* and *IL13* in a Dutch population with asthma. *Am J Hum Genet* 2002; **70**: 230–236.
4 Moore JH, Williams SM: New strategies for identifying gene-gene interactions in hypertension. *Ann Med* 2002; **34**: 88–95.
5 Xu J, Langefeld CD, Zheng SL *et al*: Interaction effect of PTEM and CDKN1B chromosomal regions on prostate cancer linkage. *Hum Genet* 2004; **115**: 255–262.
6 Ochoa MC, Marti M, Azcona C *et al*: Gene–gene interaction between PPARγ2 and ADRβ3 increases obesity risk in children and adolescents. *Int J Obes Relat Metab Disord* 2004; **28**: S37–S41.
7 Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 2003; **4**: 701–709.
8 Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex disease. *Nat Genet* 2005; **37**: 413–417.
9 Neale BM, Sham PC: The future of association studies: gene-based analysis and replication. *Am J Hum Genet* 2004; **75**: 353–362.
10 Gauderman WJ, Murcray C, Gilliland F *et al*: Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol* 2007; **31**: 383–395.
11 Wang K, Abbott D: A principal components regression approach to multilocus genetic association studies. *Genet Epdemiol* 2007; **32**: 108–118.
12 Wei Z, Li M, Rebbeck T *et al*: U-statistics-based tests for multiple genes in genetic association studies. *Ann Hum Genet* 2008; **72**: 821–833.
13 Li M, Wang K, Grant SFA *et al*: A powerful gene-based association test by combining optimally weighted markers. *Bioinformatics* 2009; **25**: 497–503.
14 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
15 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
16 Li Y, Willer CJ, Sanna S, Abecasis GR: Genotype Imputation. *Annu Rev Genomics Hum Genet* 2009; **10**: 387–406.
17 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 2007; **39**: 906–913.
18 Zemunik T, Boban M, Lauc G *et al*: Genome-wide association study of biochemical traits in Korcula Island, Croatia. *Croat Med J* 2009; **50**: 23–33.
19 Keating BJ, Tischfield S, Murray SS *et al*: Concept, Design and Implementation of a Cardiovascular Gene-Centric 50 K SNP Array for Large-Scale Genomic Association Studies. *PLoS ONE* 2008; **10**: e3583.
20 Brown ML, Inazu A, Hesler CB *et al*: Molecular basis of lipid transfer protein deficiency in a family with increased high-density lipoproteins. *Nature* 1989; **342**: 448–451.
21 Inazu A, Brown ML, Hesler CB *et al*: Increased high-density lipoprotein levels caused by a common cholesteryl-ester transfer protein gene mutation. *N Engl J Med* 1990; **323**: 1234–1238.
22 Gillum RF: The association between serum albumin and HDL and total cholesterol. *J Nat Med Assoc* 1993; **85**: 290–292.
23 Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S: Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet* 2006; **79**: 1002–1016.
24 Chapman J, Clayton D: Detecting association using epistatic information. *Genet Epidemiol* 2007; **31**: 894–909.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)