BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 4 2009, pages 497–503
doi:10.1093/bioinformatics/btn641

Genetics and population analysis

ATOM: a powerful gene-based association test by combining optimally weighted markers

Mingyao Li[1,*], Kai Wang[2], Struan F. A. Grant[2], Hakon Hakonarson[2] and Chun Li[3,*]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, [2]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104 and [3]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

Received on August 20, 2008; revised and accepted on December 11, 2008

Advance Access publication December 15, 2008

Associate Editor: Martin Bishop

ABSTRACT

Background: Large-scale candidate-gene and genome-wide association studies genotype multiple SNPs within or surrounding a gene, including both tag and functional SNPs. The immense amount of data generated in these studies poses new challenges to analysis. One particularly challenging yet important question is how to best use all genetic information to test whether a gene or a region is associated with the trait of interest.

Methods: Here we propose a powerful gene-based Association Test by combining Optimally Weighted Markers (ATOM) within a genomic region. Due to variation in linkage disequilibrium, different markers often associate with the trait of interest at different levels. To appropriately apportion their contributions, we assign a weight to each marker that is proportional to the amount of information it captures about the trait locus. We analytically derive the optimal weights for both quantitative and binary traits, and describe a procedure for estimating the weights from a reference database such as the HapMap. Compared with existing approaches, our method has several distinct advantages, including (i) the ability to borrow information from an external database to increase power, (ii) the theoretical derivation of optimal marker weights and (iii) the scalability to simultaneous analysis of all SNPs in candidate genes and pathways.

Results: Through extensive simulations and analysis of the *FTO* gene in our ongoing genome-wide association study on childhood obesity, we demonstrate that ATOM increases the power to detect genetic association as compared with several commonly used multi-marker association tests.

Contact: mingyao@mail.med.upenn.edu; chun.li@vanderbilt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Human genetics research has accelerated in the last decade owing to our increased understanding of the human genome. With the completion of the International HapMap Project (2005, 2007), the development of high-throughput genotyping technology and rapid decline in genotyping costs, multiple large-scale candidate-gene and genome-wide association studies are being conducted to identify novel genetic risk factors for complex human diseases. In such studies, several SNPs within a gene, including both functional and tag SNPs, are genotyped. The immense amount of data generated in these studies poses new challenges to data analysis. One particularly challenging yet important question is how to best use all genetic information to test whether a gene or a region is associated with the trait of interest.

The most commonly used approach for detecting genetic association is the single marker analysis in which each marker is analyzed individually for association with the trait. This simple approach has led to the discovery of novel disease susceptibility genes for many diseases, including age-related macular degeneration (Klein *et al*., 2005), inflammatory bowel disease (Duerr *et al*., 2006) and type 2 diabetes (Saxena *et al*., 2007; Scott, 2007; Zeggini *et al*., 2007) among many others. Despite these early successes, single marker tests may be inefficient when each single marker carries a small to moderate amount of association information about the trait or when there is allelic heterogeneity. In this situation, the association might not be detected when markers are analyzed individually, while combining genotypes from neighboring markers may provide a more powerful test.

Several tests have been developed to combine information from multiple markers. For example, for case–control data, one could perform a logistic regression with marker genotypes included as covariates and simultaneously test for significance of the main effects and possibly interactions among markers. Although this method can be more powerful in some situations than testing each marker individually, it might suffer from low power due to high degrees of freedom. Alternatively, one could use haplotype-based methods, which are attractive since genomic variations in humans are structured into haplotypes (Clark, 2004; Schaid, 2004). Haplotype-based methods utilize a preliminary step that infers the haplotypes for each individual by substituting expected haplotype probability (Epstein and Satten, 2003; Schaid *et al*., 2002). These methods then assess the relationship between the trait and the markers through an overall test of association across haplotypes. For example, in a genetics study of age-related macular degeneration, due to strong linkage disequilibrium (LD) in the *CFH* gene, haplotype analysis detected a stronger association signal than single marker analysis and joint-genotype analysis did (Li *et al*., 2006a). Other than these traditional haplotype analysis methods, clustering-based haplotype analysis methods have also

*To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/bioinformatics/article/25/4/497/248790 by guest on 30 April 2025

been developed (Li and Jiang, 2005; Su *et al*., 2008). However, when the signal is mainly driven by a single marker or when LD in the region is weak or moderate, the power of haplotype analysis might be low. With the goal of combining information across multiple markers while reducing the degrees of freedom, Wang and Elston (2007) proposed a weighted score test in which information from multiple correlated SNPs is compressed using a Fourier transformation and then a test statistic with one degree of freedom is constructed. Furthermore, two other recent studies proposed using principal components of marker genotypes as covariates in multiple regression analysis (Gauderman *et al*., 2007; Wang and Abbott, 2007).

All of these methods focus on the analysis of markers genotyped in the study sample without utilizing additional information provided by other resources. However, several external databases have been generated that can provide additional information on the LD structure in the marker region. For example, the HapMap data contain a complete set of genotypes for 269 individuals on >3.9 million SNPs across the genome. This dataset provides the most comprehensive information about LD structure in the human genome to date. To improve the power over existing methods, here we propose a novel gene-based association test that combines optimally weighted marker genotypes. The optimal weights are analytically derived and can be calculated according to a reference LD database such as the HapMap, other publicly available dense SNP genotype data or resequencing data on a subset of the study subjects. Unlike imputation-based methods (Li *et al*., 2006b; Marchini *et al*., 2007; Nicolae, 2006), which test one marker (either genotyped or imputed) at a time, a unique feature of our method is its ability to combine all information in a region without the requirement of explicitly imputing the untyped markers. We demonstrate that our approach can increase the power to detect genetic association for a wide range of LD structures. We also apply the proposed method to the *FTO* gene using data from our ongoing genome-wide association study on childhood obesity (Grant *et al*., 2008), and demonstrate that our method performs favorably against other multi-marker association tests.

2 METHODS

We consider the problem of genetic association analysis with multiple markers in a gene. Our goal is to develop an association test that best uses genetic information contributed by all markers while minimizing its degrees of freedom. We present a multi-marker Association Test by combining Optimally weighted Markers (ATOM) through analytical derivations, and then evaluate its performance using simulations and analysis of real data. We will first present our analytical solutions for quantitative and binary traits assuming the trait locus is known, and then extend the method to the more practical situation in which the trait locus is unknown.

2.1 Quantitative trait

Suppose the quantitative trait of interest, *Y*, is influenced by a diallelic quantitative trait locus (QTL), with *alleles* T and t. Let p_T and p_t denote the corresponding allele frequencies. If the genetic effect of the QTL is additive, then the mean of the trait value given genotype g_T can be described as

$$E(Y|g_T) = \alpha_T + \beta_T g_T, \tag{1}$$

where $g_T \in \{0, 1, 2\}$ counts the number of allele T. In genetic association analysis, we wish to test the null hypothesis H_0: $\beta_T = 0$. However, since g_T may not be directly observed, the test of association is often accomplished through examination of association with genetic markers. Assume a diallelic marker is genotyped in the region of the QTL with *alleles* A and *a*. Let p_A and p_a denote the corresponding allele frequencies, and $g \in \{0, 1, 2\}$ denote the genotype that counts the number of allele A. We will show that if (1) holds then the means model at the marker will be

$$E(Y|g) = \alpha + \beta g. \tag{2}$$

Equation (2) allows indirect assessment of association with the QTL by testing $H_0 : \beta = 0$.

The slopes, β_T and β, reflect the strength of association of the trait with the QTL and the marker, respectively. Their relationship depends on the LD between the QTL and the marker. To explicitly derive their relationship, we note that

$$\begin{aligned} E(Y|g) &= \sum_{g_T=0}^{2} E(Y|g_T)P(g_T|g) = \sum_{g_T=0}^{2} (\alpha_T + \beta_T g_T)P(g_T|g) \\ &= \alpha_T + \beta_T H(g) \end{aligned} \tag{3}$$

where $H(g) = \sum_{g_T=0}^{2} g_T P(g_T|g)$. We can show that when both the marker and the QTL follow Hardy–Weinberg equilibrium in the population, $H(g) = 2(p_T - \Delta/p_a) + [\Delta/(p_A p_a)]g$, where $\Delta = p_{AT} - p_A p_T$ is the LD coefficient between the QTL and the marker. Therefore, by Equation (3), the means model at the marker becomes a linear function of *g*:

$$E(Y|g) = \alpha_T + 2\beta_T\left(p_T - \frac{\Delta}{p_a}\right) + \beta_T \frac{\Delta}{p_A p_a} g. \tag{4}$$

Comparing (4) with Equation (2), we get

$$\alpha = \alpha_T + 2\beta_T\left(p_T - \frac{\Delta}{p_a}\right), \tag{5}$$

$$\beta = \beta_T \frac{\Delta}{p_A p_a}. \tag{6}$$

Thus, the slope at the marker, β, and the slope at the QTL, β_T, differ by a factor $\Delta/(p_A p_a)$, which is a function of the LD coefficient between the two loci and the marker allele frequencies. It can be easily shown that $|\Delta/(p_A p_a)| \le 1$, and thus the magnitude of the effect at the marker is always smaller than that at the trait locus. Clearly, when the QTL and the marker are in perfect LD, $\beta = \beta_T$; when they are in linkage equilibrium, $\beta = 0$; when they are in incomplete LD, $\beta = 0$ if and only if $\beta_T = 0$; and the test of H_0: $\beta = 0$ is an indirect test of H_0: $\beta_T = 0$. If we define a weighted genotype score at the marker, $g^* = (\Delta/p_A p_a)g$, then the corresponding means model will be $E(Y|g^*) = \alpha_T + \beta_T g^*$, with the slope being the same as that in Equation (1). This fact is employed below to combine information from multiple markers. The relationship (6) between the slopes also implies that the power of the test on the marker depends on the strength of LD between the QTL and the marker.

In general, suppose *m* diallelic markers are genotyped in the region of interest, with *alleles* 1_j and 0_j at marker *j* ($1 \le j \le m$) and allele frequencies p_j and q_j, respectively. The above derivations suggest that for individual *i* and marker *j*, we may consider the weighted genotype score, $g^*_{i,j} = \Delta_j / p_j q_j g_{i,j}$, where $\Delta_j = p_T 1_j - p_T p_j$ is the LD coefficient between the QTL and marker *j*, and $g_{i,j}$ denotes the number of allele 1_j carried by individual *i*. Then the means model at marker *j* will be $E(Y|g^*_{i,j}) = \alpha_T, j + \beta_T g^*_{i,j}$, and all *m* markers will share a common slope slope β_T. This motivated us to aggregate the information from the *m* markers by defining a score

$$S_i = \frac{1}{m}\sum_{j=1}^{m} w_j g_{i,j} \tag{7}$$

for individual *i*, where $w_j = \Delta_j / p_j q_j$. In general, the stronger LD between the trait and marker loci, the higher magnitude the weight is. Thus, this aggregate score as defined in (7) effectively allocates weights to markers according to their levels of LD with the trait locus. It allows us to capture association information contributed by all *m* markers and to reduce the dimension from *m* to 1. In a later section, we will describe how to estimate the weights. We note that the score in (7) is different from imputation based methods (Li *et al*., 2006b; Marchini *et al*., 2007), which calculate the probability of each possible genotype for each of the missing or unknown genotypes in the study sample.

Downloaded from https://academic.oup.com/bioinformatics/article/25/4/497/248790 by guest on 30 April 2025

2.2 Binary trait

For a binary trait such as disease status, the strength of association with the disease can be measured by genotype relative risks or risk (i.e. penetrance) differences between different genotypes. The latter is analogous to the slope in (1) for quantitative traits. Assume the binary trait, Y (1 affected, 0 unaffected), is influenced by a diallelic disease locus with *alleles D* and *d*. Let $g_D \in \{0, 1, 2\}$ denote the genotype with respective copies of allele D at the disease locus. Again, we assume a diallelic marker is genotyped in the disease locus region with *alleles* A and a, allele frequencies p_A and p_a, and genotypes $g \in \{0, 1, 2\}$ coded as the number of copies of allele A. Let $f_{g_D} = P(Y = 1|g_D)$ be the penetrance of genotype g_D at the disease locus, and $\varphi_g = P(Y = 1|g)$ denote the 'penetrance' of marker genotype g, an indirect effect due to LD with the disease locus. Since $E(Y|g_D) = P(Y = 1|g_D)$ and $E(Y|g) = P(Y = 1|g)$, thus under the additive model assumption, i.e. $f_1 - f_0 = f_2 - f_1$, the relationship between the disease locus and the test marker can be derived in a similar fashion as for quantitative traits:

$$\begin{aligned} \varphi_0 &= K - \frac{1}{p_a}[2\Delta(f_1 - f_0)], \\ \varphi_1 &= K - \frac{1}{p_A p_a}[\Delta(p_A - p_a)(f_1 - f_0)], \\ \varphi_2 &= K + \frac{1}{p_A}[2\Delta(f_1 - f_0)], \end{aligned} \tag{8}$$

where $K = f_0 p_d^2 + 2f_1 p_D p_d + f_2 p_D^2$ is the disease prevalence. These relationships imply that

$$\varphi_1 - \varphi_0 = \varphi_2 - \varphi_1 = (f_1 - f_0)\frac{\Delta}{p_A p_a}. \tag{9}$$

Equation (9) indicates that if the genetic model at the disease locus is additive, then the corresponding model at the marker locus is also additive, and the risk difference at the marker locus and the risk difference at the disease locus differ by the same factor, $\Delta / p_A p_a$, as that in (6) for quantitative traits. This suggests that the weighting scheme as defined in (7) can also be used for binary traits to appropriately aggregate information from all the markers.

2.3 Estimation of weights

In the above derivations, the weight for each marker depends on its LD with the unobserved QTL or disease locus. To use ATOM in practice, we need to estimate the weight using available information. Since the location of the trait locus is unknown, one may assume that each known polymorphism in the region, either genotyped or untyped in the study sample, is likely to be the trait locus.

The estimation of weight requires knowledge of LD among genetic variations in the gene of interest. Such information can be obtained from a reference dataset such as that generated by the International HapMap Project, other publicly available dense SNP datasets, or resequencing data from a subset of the study sample. Suppose M markers are available in the reference dataset, which is a superset of the m markers genotyped in the study sample. If marker k ($1 \le k \le M$) in the reference dataset is the trait locus, then the weight for marker j ($1 \le j \le m$) in the study sample is

$$w_j^k = \Delta_j^k \Big/ p_j q_j \tag{10}$$

where Δ_j^k is the LD coefficient between markers k and j, and p_j and q_j are allele frequencies at marker j. These quantities can be estimated from the reference dataset. Then for each marker k in the reference dataset, we can calculate a score $S_{i,k} = \frac{1}{m}\sum_{j=1}^{m} \frac{\Delta_j^k}{p_j q_j} g_{i,j}$ for individual i. These scores will be used in subsequent association analysis.

When the reference dataset is not available (i.e. $M = m$), the weight in Equation (10) can be estimated based on genotypes in the study sample for quantitative traits or based on genotypes in the controls for case–control studies. Alternatively, we can use the observed genotypes in the study sample directly, and this is equivalent to the principal components analysis (PCA). Our results indicate that these two approaches have similar performance (data not shown).

2.4 Test of genetic association

Once we have calculated the scores for each marker in the reference dataset, one analysis strategy might be testing the markers individually for genetic association using the aggregate scores $S_k = (S_{1,k}, \ldots, S_{n,k})$, where n is the total number of individuals in the study sample. This is similar in spirit to the imputation-based methods (Li *et al*., 2006; Marchini *et al*., 2007; Nicholae, 2006) in which one can test for genetic association for untyped markers by borrowing strength of LD among genotyped markers. Although each of the M tests aggregates information across all m genotyped markers in the study sample, they might provide redundant information due to LD; moreover, since M is typically much greater than m, the lack of adequate adjustment of multiple testing might lead to loss of power. To efficiently aggregate all available information, we propose to conduct an association test using the principal components obtained from the scores S_k ($1 \le k \le M$).

The central idea of the PCA is to reduce the dimensionality of a dataset consisting of a large number of correlated variables, while retaining as much as possible the variation present in the dataset. The principal components will be linear combinations of the scores, with the l-th component being $PC_l = \sum_{k=1}^{M} e_{l,k} S_k$, where $e_l = (e_{l,1}, \ldots, e_{l,M})$ satisfies $\sum_{k=1}^{M} e_{l,k}^2 = 1$ and is chosen to maximize the variance of the component that is orthogonal to all previous components. The PCA transforms the original set of M correlated scores $\{S_k\}$ into a set of m uncorrelated principal components $\{PC_k\}$. Without loss of generality, we can order the principal components by the magnitude of variance so that PC_1 has the largest variance, and PC_2 has the second largest variance, etc. Let $\lambda_1 \ge \lambda_2 \ge \lambda_M$ denote their corresponding eigenvalues. Then the variance of PC_l is λ_l, and the fraction of total variance in the data that can be explained by PC_l is $\lambda_l / (\lambda_1 + \ldots + \lambda_M)$.

Once the principal components are computed, we can then test for genetic association by conducting a regression analysis, linear regression for quantitative traits and logistic regression for binary traits, with a set of selected principal components as covariates. Since the principal components are ordered by the magnitude of explained variance, we could select either a predetermined number of principal components or the first several principal components so that their total fraction of explained variance exceeds a prespecified threshold. In both scenarios, we could carry out a likelihood ratio test to test the null hypothesis that all regression coefficients of the selected principal components are equal to zero. The degrees of freedom would be the same as the number of the selected components. Ideally, we would like to achieve as large a fraction of the total explained variance as possible while using as few degrees of freedom as possible. However, our simulations suggested that the optimal number of principal components varies with LD structures, as does the optimal fraction of explained variance (data not shown). As such, we opt to use several thresholds for fractions of variance, each resulting in a test statistic, and choose the maximum statistic, denoted as T_{ATOM}. Significance of T_{ATOM} will be evaluated by permutations. Specifically, we permute the phenotypes of the subjects H times, and for each permutation h, we record the maximum test statistic $T_{ATOM}^{(h)}$. Then the permutation-based P-value can be estimated as $P = \sum_{h=1}^{H} I\{T_{ATOM}^{(h)} \ge T_{ATOM}\} \Big/ H$. It is worth noting that logistic regression is mainly used as a modeling tool to test for genetic association. Since the weights are derived from a 'linear' relationship on the penetrance scale, the weights may not be optimal in the logistic regression framework. However, we note that for small effect sizes, this is unlikely to make much difference.

Our method is different from traditional PCA-based approaches (Gauderman *et al*., 2007; Wang and Abbott, 2007) that operate directly on the marker genotypes $\{g_j\}_{j=1}^{m}$ observed on the study sample. Instead, our method operates on the scores $\{S_k\}_{k=1}^{M}$ for all markers in the region. A principal component in the traditional PCA approaches is a linear combination of marker genotypes, with the coefficients (i.e. weights) determined solely by the correlation structure among the genotyped markers. In contrast, the weights in our aggregate scores are optimally derived and are based on the additional LD information for the region. This may help achieve a better allocation of marker information than in the traditional PCA-based approaches.

Downloaded from https://academic.oup.com/bioinformatics/article/25/4/497/248790 by guest on 30 April 2025

Our method shares similarity with imputation-based methods such as TUNA (Nicolae, 2006) and SNPTEST (Marchini *et al*., 2007) in that all these methods rely on external LD information to derive a 'score' for the untyped marker. However, unlike TUNA and IMPUTE which impute untyped marker using multi-locus LD, our method is based on pairwise LD between the putative trait locus and every genotyped marker. Such an approach is computationally simple, and as shown in the Section 3, it is also statistically powerful.

3 RESULTS

In this section, we evaluate the power of T_{ATOM} for binary traits, and compare it with other multi-marker tests based on both simulated data and analysis of the *FTO* gene using data from our ongoing genome-wide association study on childhood obesity (Grant *et al*., 2008). Methods that we considered for comparison include T_{JOINT}—a joint genotype test with all markers included as covariates in logistic regression; T_{HAPLO}—the haplo.score test (Schaid *et al*., 2002) with rare haplotypes with frequency <0.05 pooled together; T_{PCA}—the traditional PCA-based test that operates directly on the marker genotypes (Gauderman *et al*., 2007; Wang and Abbott, 2007); T_{WE}—a Fourier transformation-based test (Wang and Elston, 2007); T_{TUNA}—a method that estimates allele frequencies for uptyped markers in a reference dataset based on observed genotypes (Nicolae, 2006); and T_{SNPTEST}—a method that relies on imputed genotypes from IMPUTE (Marchini *et al*., 2007). Since TUNA and SNPTEST examine each marker in the region (either genotyped or untyped) individually, to obtain a test that assesses the overall association, we picked the maximum statistic among all markers as the test statistic. For T_{ATOM} and T_{PCA}, the maximum statistic was chosen among tests for five different thresholds (80%, 90%, 95%, 99% and 99.9%) for the percentage of explained variation. For these four tests—T_{TUNA}, T_{SNPTEST}, T_{ATOM} and T_{PCA}, significance was evaluated based on 1 00 000 permutations. Since the significance level of these tests is determined by permutations of the phenotypes, the type I error rates are under control and we therefore omit the results of the type I error rate evaluations. Significance of T_{JOINT}, T_{HAPLO} and T_{WE} was evaluated based on their asymptotic null distributions. Our simulation results show that the type I error rates of these three tests are under control (data not shown). For T_{WE}, we used the original allele coding in the analysis due to two reasons: (i) Wang and Elston (2007) did not provide an algorithm to recode *alleles*, and (ii) for the situations we considered, there are a large number of SNPs in the gene, and hence exhaustive search of the best allele coding is computationally infeasible.

3.1 Comparison of power based on simulated data

We simulated data based on the LD structures of two genes: *CHI3L2* on chromosome 1 and *CDH17* on chromosome 8, for which the LD structures were estimated using the genotypes of the CEPH (Utah residents with ancestry from northern and western Europe) (CEU) sample of the International HapMap Project. The *CHI3L2* gene contains a single LD block (Fig. 1) and has the most significant evidence of association with *cis* regulatory elements in a previously published genotype-expression association study (Cheung *et al*., 2005). The *CDH17* gene contains two LD blocks (Fig. 2); this allows us to evaluate the performance of our method when markers fall in different LD blocks. For each gene, we considered common SNPs with MAF ≥ 0.05 and then selected tagSNPs using the Tagger program (de Bakker *et al*., 2005) using pairwise tagging at $r^2 \geq 0.8$. We identified 25 and 86 common SNPs for *CHI3L2* and *CDH17*, respectively, among which 7 and 29 tagSNPs were selected. We assumed only the tagSNPs were genotyped and available for analysis, a common scenario in both candidate-gene and genome-wide association studies.

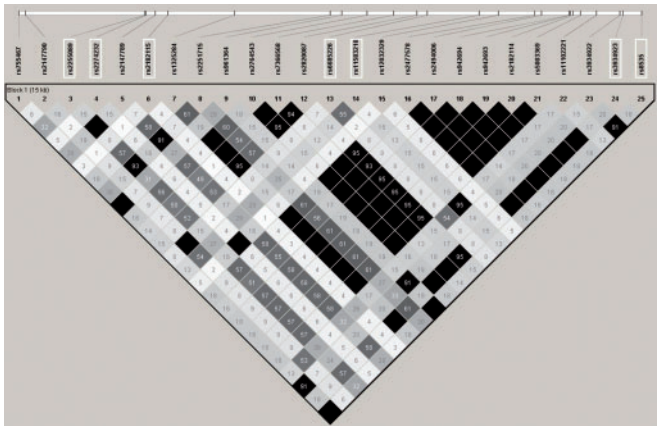

Fig. 1. LD structure of the *CHI3L2* gene on chromosome 1 in the HapMap CEU sample. Displayed is the estimated r^2 for 25 SNPs with minor allele frequency (MAF) ≥ 0.05. SNPs within the white boxes are tagSNPs selected using the Tagger program at r^2 threshold of 0.8.

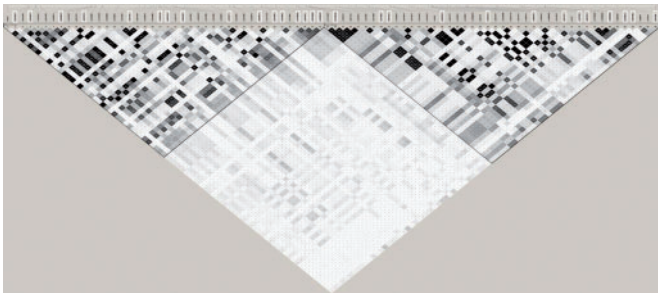

Fig. 2. LD structure of the *CDH17* gene on chromosome 8 in the HapMap CEU sample. Displayed is the estimated r^2 for 86 SNPs with MAF ≥ 0.05. SNPs within the white boxes are tagSNPs selected using the Tagger program at r^2 threshold of 0.8.

The performance of genetic association analyses may vary depending on the position of the disease locus relative to the genotyped markers, whether the disease variant is genotyped, and whether it sits in the center or the boundary of a block. To evaluate the effect of the disease locus position on the performance of the methods, we repeated the simulations with each common SNP designated as the true disease locus. We assumed an additive model at the disease locus with disease prevalence 5% and sibling recurrence risk ratio 1.02, corresponding to a genotype relative risk of 1.69 when the disease allele frequency is 0.05, and 1.39 when the disease allele frequency is 0.5. The case–control status of each individual was determined by comparing a random number with the penetrances at the disease locus. Power was estimated based on 1000 replicate datasets each consisting of 750 cases and 750 controls, and significance was assessed at the 1% level. Our method uses LD

Downloaded from https://academic.oup.com/bioinformatics/article/25/4/497/248790 by guest on 30 April 2025

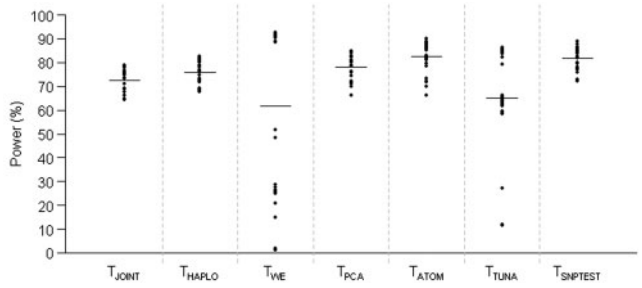

Fig. 3. Comparison of power for binary trait when LD structure is simulated based on the *CHI3L2* gene. Results are based on 1000 replicates of 750 cases and 750 controls. Significance is assessed at the 1% level.

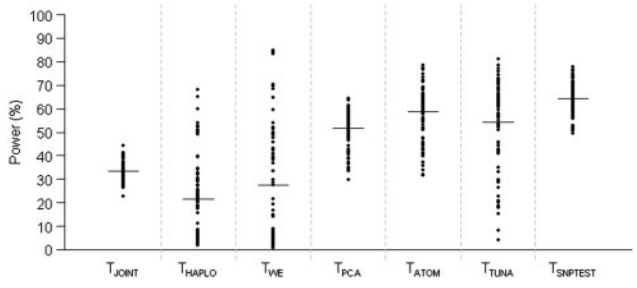

Fig. 4. Comparison of power for binary trait when LD structure is simulated based on the *CDH17* gene. Results are based on 1000 replicates of 750 cases and 750 controls. Significance is assessed at the 1% level.

information from a reference dataset, which may not be accurate due to random variation. To take this variation into account, we simulated 60 individuals as a reference dataset and calculated the aggregate scores using the LD information reestimated from these 60 individuals. The reference dataset was simulated for each of the 1000 replicates.

Figure 3 shows power comparison results when the LD structure was simulated based on the *CHI3L2* gene. The results indicate that T_{ATOM} consistently outperformed the non-imputation based methods except T_{WE}. Although T_{WE} was more powerful than T_{ATOM} under some situations, its performance was highly unstable, with much lower power than the other methods for 11 (44%) SNPs. This is not surprising since the performance of T_{WE} depends critically on whether the SNPs are positively correlated. If negative correlation exists for some SNP pairs, then the association signals can be canceled out, leading to loss of power. Although Wang and Elston (2007) recommended recoding the SNP genotypes to avoid this, no detailed procedure was provided. Furthermore, given the large number of SNPs we considered, it is not computationally feasible to search for the best allele coding to get the maximal number of positively correlated loci. Among the three imputation-based methods, T_{ATOM} has similar power as T_{SNPTEST}, despite that the T_{SNPTEST} is much computationally intensive; both T_{ATOM} and T_{SNPTEST} are more powerful than T_{TUNA} for all markers that we considered. We also calculated the average power of the seven tests across all 25 disease loci in the gene; the average power of T_{ATOM} was 83%, compared with 81%, 65%, 78%, 62%, 76% and 73% for T_{SNPTEST}, T_{TUNA}, T_{PCA}, T_{WE}, T_{HAPLO} and T_{JOINT}, respectively. These results show that T_{ATOM} performs consistently well under all circumstances we considered.

Figure 4 shows power comparison results when the LD structure was simulated based on the *CDH17* gene, which has two LD blocks (Fig. 2). Given the complex LD structure and the large number of tagSNPs in the gene, standard approaches such as T_{HAPLO} and T_{JOINT} may not perform well. Not surprisingly, our results show that both T_{HAPLO} and T_{JOINT} had relatively low power for the disease loci in this gene, with average power 21% and 33%, respectively. Again, T_{WE} exhibited highly variable performance and its average power was only 28%. In contrast, the average power of T_{PCA} and T_{ATOM} were 52% and 59%, respectively. Of the 86 disease loci, T_{ATOM} outperformed T_{PCA} for 78 loci, again demonstrating the advantage of incorporating external LD information for the whole region in genetic association analysis. Among the three imputation-based methods, T_{SNPTEST} appears to be the most powerful. Of the 86 disease loci, T_{SNPTEST} outperformed T_{ATOM} for 69 loci, and outperformed T_{TUNA} for 58 loci. The average power of T_{SNPTEST}, T_{ATOM} and T_{TUNA} was 64%, 59% and 54%, respectively. The relatively low average power of T_{TUNA} is due to its greater variability than T_{ATOM}.

We also investigated the effect of allelic heterogeneity on the performance of T_{ATOM} and other tests. We simulated datasets assuming two disease loci existed in the *CDH17* gene, each in a different LD block (Table 1). We varied the disease loci in the two LD blocks. For each pair of disease loci, an additive model was used, with the disease prevalence fixed at 5%. The results suggest that T_{ATOM} performed consistently well in this large gene when two disease loci contributed independently to the disease risk (Table 1). Across all disease models we considered, the average power of T_{ATOM} was 68%, compared with 61%, 63%, 54%, 49%, 42% and 41% for T_{PCA}, T_{SNPTEST}, T_{TUNA}, T_{WE}, T_{JOINT} and T_{HAPLO}, respectively.

3.2 Application to childhood obesity data

To illustrate the effectiveness of our method in real data, we applied all five multi-marker association tests to an ongoing genome-wide association study on childhood obesity, which is a major health problem in modern societies. Specifically, we investigated the strength of association between obesity and the genotypes at the *FTO* gene, which is known to be associated with childhood obesity (Grant *et al*., 2008). All study subjects were recruited from the greater Philadelphia area from 2006 to 2007 at The Children's Hospital of Philadelphia. After excluding subjects with potential measurement error or Mendelian causes of extreme obesity, we analyzed 394 Caucasian obese children as cases and 2133 Caucasian controls. All subjects were biologically unrelated and were aged between 2 and 18 years old. The study was approved by the Institutional Review Board of The Children's Hospital of Philadelphia.

The SNP that was previously reported to be associated with obesity, rs9939609, was not included on the Illumina 550 K BeadChip, which was used for our genotyping. However, two SNPs on the chip, rs8050136 and rs3751812, were in complete LD with rs9939609 in the HapMap CEU sample. We also identified eight additional common SNPs on the SNP chip that are in the same LD region, and extracted the genotype data of these 10 SNPs. We used the HapMap CEU data as the external reference dataset, which contains 58 common SNPs in the *FTO* gene region. The LD structure of the 58 SNPs is displayed in Figure 5.

Downloaded from https://academic.oup.com/bioinformatics/article/25/4/497/248790 by guest on 30 April 2025

Table 1. Comparison of power (%) for two-locus models when LD structure is simulated based on the CDH17 gene

Block1	Block2	T_{ATOM}	T_{PCA}	$T_{SNPTEST}$	T_{TUNA}	T_{WE}	T_{JOINT}	T_{HAPLO}
SNP2	SNP52	61.5	56.3	51.8	52.4	6.1	35.3	13.0
SNP2	SNP60	71.9	63.1	60.8	47.9	32.9	47.3	35.7
SNP2	SNP75	66.6	65.1	69.6	68.3	3.7	42.6	17.9
SNP2	SNP82	19.7	24.3	28.3	24.9	1.0	16.7	5.8
SNP2	SNP86	63.4	57.7	64.8	39.7	23.7	35.9	38.9
SNP5	SNP52	87.4	80.9	82.1	78.3	46.2	58.4	51.8
SNP5	SNP60	85.3	69.6	69.0	91.1	83.2	45.6	58.3
SNP5	SNP75	67.9	59.2	56.9	56.7	35.3	37.8	14.4
SNP5	SNP82	32.8	28.8	33.8	31.6	20.7	20.5	12.3
SNP5	SNP86	76.2	64.8	63.9	56.2	76.4	42.0	60.6
SNP10	SNP52	59.7	53.4	42.3	43.3	13.5	31.8	16.4
SNP10	SNP60	55.8	49.0	41.6	25.2	45.9	33.1	28.3
SNP10	SNP75	41.1	36.3	25.6	23.1	8.1	22.9	13.2
SNP10	SNP82	28.3	28.2	62.2	20.0	1.9	17.1	6.2
SNP10	SNP86	58.2	54.6	48.8	27.2	37.3	37.5	41.5
SNP15	SNP52	87.7	75.8	83.7	68.7	82.6	53.0	58.2
SNP15	SNP60	96.6	89.3	89.6	68.6	99.2	67.5	86.8
SNP15	SNP75	86.9	78.0	85.3	72.7	78.0	53.3	63.4
SNP15	SNP82	51.9	44.3	52.2	35.7	59.3	29.3	47.0
SNP15	SNP86	92.6	81.7	91.1	63.7	95.1	63.7	90.1
SNP23	SNP52	88.9	79.8	77.3	78.8	74.9	56.1	52.2
SNP23	SNP60	92.8	82.1	82.7	76.5	95.2	65.6	71.8
SNP23	SNP75	79.8	67.4	62.3	67.8	69.7	49.1	39.5
SNP23	SNP82	55.8	51.6	56.9	54.2	52.0	34.4	27.5
SNP23	SNP86	90.8	82.4	90.5	70.4	93.9	64.1	81.3
Mean power		68.0	61.0	62.9	53.7	49.4	42.4	41.3

Results are based on 1000 replicates of 750 cases and 750 controls. Significance is assessed at the 1% level.

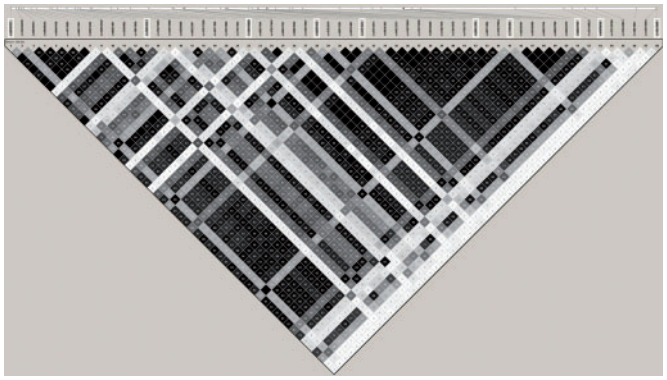

Fig. 5. LD structure of the *FTO* gene in the HapMap CEU sample. Displayed is the estimated r^2 for SNPs with MAF ≥ 0.05. SNPs within the white boxes are present on Illumina's HumanHap550 BeadChip, which were genotyped in our childhood obesity samples.

We analyzed the data using all seven tests. Significance of T_{ATOM}, T_{PCA}, $T_{SNPTEST}$ and T_{TUNA} was evaluated based on 100 000 permutations. Similarly to what we observed in simulations, the result from T_{ATOM} was the most significant among all the tests we considered. The P-value of the T_{ATOM} was 0.0082, which is much smaller than that of T_{PCA} ($P = 0.0195$), providing further evidence that incorporation of external LD information can increase the power to detect genetic association. In contrast, except for T_{TUNA} , which has P-value 0.0094, all the other tests were non-significant at the 0.05 level ($T_{SNPTEST}$: $P = 0.0727$; T_{JOINT}: $P = 0.1356$; T_{HAPLO}: $P = 0.0543$; T_{WE}: $P = 0.0547$).

4 DISCUSSION

We have developed a powerful gene-based association test that combines information from nearby markers to test genetic association for both quantitative and binary traits. Our method is based on optimal aggregation of information from genotyped markers in a gene. Since different markers often associate with the trait locus at different levels, to appropriately apportion their contributions, we assigned a weight to each marker that is proportional to the amount of information it captures about the underlying trait locus. We described a weighting scheme that allows estimation of weight from a reference dataset such as the HapMap data, other publicly available dense SNP data, or resequencing data for a subset of the study subjects. The virtue of our test lies in the ability to borrow strength from nearby markers while reducing the degrees of freedom. Through simulations and real data analysis based on a wide range of LD structures, we demonstrated that it can increase the power to detect genetic association compared with several commonly used multi-point association tests.

Our weighting scheme shares some similarity with recently proposed imputation-based tests in that these approaches use external LD information (Li *et al*., 2006b; Marchini *et al*., 2007; Nicholae, 2006). Unlike the imputation-based methods, which test each marker individually, our method tests the whole gene as a unit and thus can borrow strength from adjacent markers while reducing the degrees of freedom.

We propose to assign weights to markers based on pairwise LD coefficients estimated from a reference dataset. Alternatively, one could consider using high-dimensional LD information among all markers. For example, following our analytical derivations in the Section 2, given genotypes at m markers, for a quantitative trait Y, it can be shown that

$$E(Y|g,\cdots,g_m) = \alpha_T + \beta_T \sum_{g_r=0}^{2} g_T P(g_T|g,\cdots,g_m).$$

Therefore, one could define a 'predictive score' $S_T(g_1,\ldots,g_m) = \sum_{g_T=0}^{2} g_T P(g_T|g_1,\ldots,g_m)$ for each subject, where the high-dimensional LD information $P(g_T|g_1,\ldots,g_m)$ can be estimated from a reference dataset. This is similar in spirit to the methods proposed by Nicolae (2006) and Zaitlen *et al*. (2007). However, despite the increased complexity in computation, our preliminary results indicate that such an analysis does not offer a power gain over our approach, probably due to higher data sparseness and higher variability in the estimation of multi-marker LD than the pairwise LD measure, especially when the estimations are based on a relatively small sample such as the HapMap. In addition, as shown by our simulation results, TUNA (Nicolae, 2006) is less powerful than ATOM for situations that we considered.

Our approach is quite general and can be applied to a wide range of applications including both candidate-gene and genome-wide association studies. For candidate-gene studies, one could use our test to obtain a global P-value for a gene, which can be used to facilitate comparisons across studies that have genotyped different

Downloaded from https://academic.oup.com/bioinformatics/article/25/4/497/248790 by guest on 30 April 2025

sets of markers of the same gene. For genome-wide association studies, with increased marker density in future chips, single-marker analysis might suffer from severe over-correction of multiple comparisons. An alternative approach might be to divide the genome into regions, screen the genome first by multi-point association tests such as our method, and then follow up significant regions through more thorough analysis. In such an approach, the total number of tests during the screening phase would remain the same regardless of how dense genotyping had been carried out.

We recognize that our method relies on permutations for significance assessment, which is time consuming for genome-wide association studies. However, this problem can be solved by dividing the genome into smaller subsets and running our method for each subset on a node in a high-speed computing cluster. Moreover, we can adopt an inverse-sampling method for empirical *P*-value estimation based on the procedures described by North *et al*. (2002) and Hauser *et al*. (2004). This procedure is based on the Poisson approximation to the Binomial distribution for small *P*-values. In this procedure, we can set a high number of maximum permutations but only reach that maximum for small *P*-values, thus increasing computational efficiency.

Our method assumes that the reference dataset, which is used for estimating the weights has similar LD structure as the study sample. Although not completely realistic, various studies have demonstrated genetic similarities across different populations (Conrad *et al*., 2006; Willer *et al*., 2006). In general, we would recommend that investigators compare the LD patterns between the study sample and the reference dataset first, and use our proposed method only when they have similar LD patterns.

Although our method was developed for analysis of markers within a gene or a region, it is readily extendable to pathway-based analysis (Wang *et al*., 2007). For example, if the genes in the same pathway increase the disease risk additively, then we can calculate the scores for markers within each gene as described in Section 2 and then obtain the principal components across all genes in the same pathway. We can then test for association between the pathway-based principal components and the trait of interest. Our simulation results for two-locus models demonstrate that the proposed approach has the potential to perform well when multiple disease variants are present.

In summary, we have developed a novel multi-marker association test by optimally weighting genotyped markers using LD information from a reference dataset. The standard approach for detecting genetic association has been the single-marker approach, which assesses the marginal effect of each marker separately. But this strategy may not be the most powerful if each marker only contributes small to moderate amounts of association information or if there is allelic heterogeneity. By optimally weighting the genotyped markers, our method efficiently captures association signals in the region and thus improves the power of detecting association. With the wide application of large-scale genotyping in current genetics studies, we believe that our method will provide a powerful multi-marker approach to identifying disease loci.

Funding: National Institute of Health (grant R01HG004517, to M.L. and C.L.).

Conflict of Interest: none declared.

REFERENCES

Cheung,V.G. *et al*. (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369.

Clark,A.G. (2004) The role of haplotypes in candidate gene studies. *Genet. Epidemiol.*, **27**, 321–333.

Conrad,D. *et al*. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet*., **38**, 1251–1260.

Duerr,R.H. *et al*. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**, 1461–1463.

de Bakker,R. *et al*. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.

Epstein,M.P. and Satten,G.A. (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *Am. J. Hum. Genet*., **73**, 1316–1329.

Gauderman,W.J. *et al*. (2007) Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.*, **31**, 383–395.

Grant,S.F.A. *et al*. (2008) Association analysis of the *FTO* gene with obesity in children of Caucasian and African ancestry reveals a common tagging SNP. *PLoS ONE*, **v3**, e1746.

Hauser,E.R. *et al*. (2004) Ordered subset analysis in genetic linkage mapping of complex traits. *Genet. Epidemiol*., **27**, 53–63.

Klein,R.J. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Li,J. and Jiang,T. (2005) Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics*, **21**, 4384–4393.

Li,M. *et al.* (2006a) CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat. Genet*., **38**, 1049–1054.

Li,Y. *et al*. (2006b) Markov model for rapid haplotyping and genotype imputation in genome wide studies. *Am. J. Hum. Genet*., **S79**, A2290.

Marchini,J. *et al*. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.

Nicolae,D.L. (2006) Testing untyped alleles (TUNA)—applications to genome-wide association studies. *Genet. Epidemiol*., **30**, 718–727.

North,B.V. *et al*. (2002) A note on the calculation of empirical p values from Monte Carlo procedures. *Am. J. Hum. Genet*., **71**, 439–441.

Saxena,R. *et al*. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, **316**, 1331–1336.

Schaid,D.J. (2004) Genetic epidemiology and haplotypes. *Genet. Epidemiol*., **37**, 317–320.

Schaid,D.J. *et al*. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet*., **70**, 425–434.

Scott,L.J. (2007) A Genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.

Su,S.-Y. *et al*. (2008) Disease association tests by inferring ancestral haplotypes using a hidden Markov model. *Bioinformatics*, **24**, 972–978.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Wang,K. and Abbott,D. (2007) A principal components regression approach to multilocus genetic association studies. *Genet. Epdemiol*., **32**, 108–118.

Wang,K. *et al*. (2007) Pathway-based approaches for analysis of genome-wide association studies. *Am. J. Hum. Genet*., **81**, 1278–1283.

Wang,T. and Elston,R.C. (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet*., **80**, 353–360.

Willer,C.J. *et al*. (2006) Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet. Epidemiol.*, **30**, 180–190.

Zaitlen,N. *et al*. (2007) Leveraging the HapMap correlation structure in association studies. *Am. J. Hum. Genet.*, **80**, 683–691.

Zeggini,E. *et al*. (2007) Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.

Downloaded from https://academic.oup.com/bioinformatics/article/25/4/497/248790 by guest on 30 April 2025