BMC Genetics

METHODOLOGY ARTICLE Open Access

Genome-wide Two-marker linkage disequilibrium mapping of quantitative trait loci

Jie Yang[1], Wei Zhu[2], Jiansong Chen[2], Qiao Zhang[2] and Song Wu[2*]

Abstract

Background: In a natural population, the alleles of multiple tightly linked loci on the same chromosome co-segregate and are passed non-randomly from generation to generation. Capitalizing on this phenomenon, a group of mapping methods, commonly referred to as the linkage disequilibrium-based mapping (LD mapping), have been developed recently for detecting genetic associations. However, most current LD mapping methods mainly employed single-marker analysis, overlooking the rich information contained within adjacent linked loci.

Results: We extend the single-marker LD mapping to include two linked loci and explicitly incorporate their LD information into genetic mapping models (tmLD). We establish the theoretical foundations for the tmLD mapping method and also provide a thorough examination of its statistical properties. Our simulation studies demonstrate that the tmLD mapping method significantly improves the detection power of association compared to the single-marker based and also haplotype based mapping methods. The practical usage and properties of the tmLD mapping method were further elucidated through the analysis of a large-scale dental caries GWAS data set. It shows that the tmLD mapping method can identify significant SNPs that are missed by the traditional single-marker association analysis and haplotype based mapping method. An R package for our proposed method has been developed and is freely available.

Conclusions: The proposed tmLD mapping method is more powerful than single marker mapping generally used in GWAS data analysis. We recommend the usage of this improved method over the traditional single marker association analysis.

Keywords: Genetic mapping, Linkage disequilibrium mapping, Linked loci, Genome wide association study

Background

Most economically, biologically and clinically important traits, such as those linked to poplar growth, cancer development and dental caries risk, are inherently complex in terms of their polygenic control and sensitivity to the environment [1]. The number of genes involved in these traits is typically large, each exerting a small effect and acting singly or interactively with others in a complicated network. For this reason, the genetic analysis of complex traits has been very difficult. However, a profound understanding of the genetic control mechanisms of complex traits is crucial to economy and life. Therefore, the development of more powerful and complex genetic mapping methods has become increasingly urgent.

In recent years, with the advancement of new DNA-based biotechnologies, such as single-nucleotide polymorphism (SNP) arrays, genome-wide association studies (GWAS) have become feasible to dissect the phenotypic variation of a complex trait into individual genetic components. Particularly, SNP arrays have gained popularity due to their cost-effectiveness: in year 2011 alone, 1068 GWAS were performed, each with at least 100,000 SNPs genotyped (www.genome.gov/gwastudies). Based on the most recent summary data of dbSNP database (www.ncbi.nlm.nih.gov/projects/SNP), there are ~$38 million (about 1 percent of the total genome) of validated SNPs in human genome. However, even the densest SNP array on the market can only accommodate ~1 million SNPs, and hence a great percentage of SNPs is not able to be sampled in a real genetic study. Fortunately, SNPs in the genome are not independent from each other, *i.e.* they are locally connected and form the so-called linkage disequilibrium (LD) blocks. Because of this unique correlation

* Correspondence: songwu@ams.sunysb.edu
[2]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11790, USA
Full list of author information is available at the end of the article

structure, the sampled genetic markers carry partial information about the unsampled SNPs and may be used for genomewide association analyses.

LD is a phenomenon arising from the co-inheritance of alleles at nearby loci on the same chromosome, and is defined as the deviation of the observed frequency of a haplotype from random association [2]. Historically, LD analysis was developed to quantify the genetic structure and the diversity of natural populations [3-5]. Many efforts have been put into developing dense maps of molecular markers for a wide variety of species. For example, LD structures have been estimated in human [6] as well as Holstein cattle [7], sheep [8] and dog [9]. With some regularity conditions [2], it can be shown that a LD value between any two loci decays with generations at the recombination rate between them:

$$\boldsymbol{D}^{(\boldsymbol{t}+1)} = (1-\boldsymbol{r})\boldsymbol{D}^{(\boldsymbol{t})} \quad (1)$$

where $D^{(t+1)}$ is the LD value at generation $t+1$ and r is the recombination rate between the two loci. Therefore, the LD value approaches to zero gradually at a geometric rate of 1-r. The larger the r, the faster the rate of convergence. According to Equation (1), if a significant $D^{(t+1)}$ value can be detected in the current generation, it implies r must be very small, almost close to 0, under the assumption that the initial LD was generated long time ago (*i.e.* t is large). This assumption is plausible because it does take a long time for mutations/LD to be spread in a population. Therefore, the principle of linkage disequilibrium decaying with generation builds up an alternative mapping strategy [10,11], which provides an important tool for the fine mapping of genes affecting a quantitative trait.

The LD mapping based on a single marker has been greatly studied [12-14]. However, little effort has been put on the LD mapping with multiple markers. Motivated by the seminal work of interval mapping proposed by Lander and Botstein in 1989 [15], in which genetic mapping was performed based on two neighboring genetic markers in controlled experiments, we propose to develop a new LD mapping framework that utilizes two SNP markers in a natural population. The new model explicitly incorporates the LD information between two markers into the mapping analysis, and thus we expect the analysis based on two markers is more powerful than that based on a single marker in a natural population just as Lander and Botstein have discovered in the controlled experiment. In the following sections, we first laid out the modeling framework for the two-marker LD mapping (tmLD), with details on parameter estimation and hypothesis testing. We then further elucidated our method through extensive simulation studies. Finally, we applied our method to a GWAS dental caries data set, followed by some discussions.

Methods

Two-marker LD (tmLD) mapping

In the tmLD mapping framework, we assume a dichotomous quantitative trait locus (QTL, $\mathcal{Q}$) of alleles Q and q that is causal but unobserved, and the allele frequencies of Q and q are expressed as p_2 and 1-p_2. Suppose that this QTL is genetically associated with two genotyped SNP markers, $\mathcal{M}_1$ and $\mathcal{M}_2$, of two alleles M_1 and m_1, and M_2 and m_2, with corresponding frequencies of p_1 and 1-p_1, and p_3 and 1-p_3, respectively. Further suppose the three linked SNPs in a tandem order, $\mathcal{M}_1$, $\mathcal{Q}$ and $\mathcal{M}_2$ at loci 1, 2 and 3, and the recombination rates between $\mathcal{M}_1$ and $\mathcal{Q}$, between $\mathcal{Q}$ and $\mathcal{M}_2$, and between $\mathcal{M}_1$ and $\mathcal{M}_2$ are r_{12}, r_{23} and r_{13}, respectively. The three SNPs form 8 possible haplotypes: M_1QM_2 (111), M_1Qm_2 (110), M_1qM_2 (101), M_1qm_2 (100), m_1QM_2 (011), m_1Qm_2 (010), m_1qM_2 (001), m_1qm_2 (000). To describe the linkage disequilibrium among them, their frequencies can be represented as follows using four trigenic disequilibria parameters D_{12}, D_{23}, D_{13} and D_{123} (Additional file 1):

$$\boldsymbol{p}_{\boldsymbol{ijk}} = \boldsymbol{p}_1^{\boldsymbol{i}}(1-\boldsymbol{p}_1)^{1-\boldsymbol{i}}\,\boldsymbol{p}_2^{\boldsymbol{j}}(1-\boldsymbol{p}_2)^{1-\boldsymbol{j}}\,\boldsymbol{p}_3^{\boldsymbol{k}}(1-\boldsymbol{p}_3)^{1-\boldsymbol{k}} + \boldsymbol{D}_{\boldsymbol{ijk}} \quad (2)$$

and $D_{ijk} = \frac{1}{2}[(-1)^{|i-j|}D_{12} + (-1)^{|j-k|}D_{23} + (-1)^{|i-k|}D_{13} - (-1)^{|i+j+k-1|}D_{123}]$ where $i, j, k = 0, 1$, D_{12}, D_{23}, D_{13} have exactly the same meaning as those in digenic disequilibria models for loci at positions 1/2, 2/3 and 1/3; and D_{123} is an additional trigenic disequilibria parameter for three loci together. Model (1) implies that D_{12}, D_{23}, D_{13} all geometrically decay with generations. It can be shown that with some reasonable assumptions, the D_{123} decreases with generations at a rate of (1-r_{13}) and therefore also changes very slowly with time (Additional file 2). Hence, significant D_{12}, D_{23}, and D_{123} at current generation imply $r_{12}$and r_{23} are very small, which form the basis for LD mapping using two genetic markers.

Likelihood function

Suppose there is a random sample of size n drawn from a natural human population at Hardy–Weinberg equilibrium. In this sample, multiple polymorphic sites, e.g. single nucleotide polymorphism (SNP), are genotyped, aiming at the identification of QTL affecting a continuous trait. The relationship between the observed phenotypic values and their expected means, determined by QTL genotypes, can then be described by the following model,

$$\boldsymbol{y}_{\boldsymbol{i}} = \sum\nolimits_{\boldsymbol{j}=0}^{2} \xi_{\boldsymbol{ij}}\boldsymbol{\mu}_{\boldsymbol{j}} + \boldsymbol{e}_{\boldsymbol{i}}, \quad \boldsymbol{i} = 1, \ldots, \boldsymbol{n} \quad (3)$$

Where y_i is the phenotypic values for subject i, ξ_{ij} is an indicator variable defined as 1 if subject i, which contains markers ($\mathcal{M}_{i1}$, $\mathcal{M}_{i2}$), has a QTL genotype j (j = 2 for

QQ, 1 for *Qq* and 0 for *qq*) and 0 otherwise, μ_j is the expected phenotypic value for QTL genotype *j*, and e_i is the error term reflecting the polygenic effects of other unlinked genes and the environmental effect, which can be assumed to follow $N(0, \sigma^2)$ if *y* is continuous. The conditional probability of subject *i* with its given markers carrying a certain QTL genotype *j*, $\pi_{j|i=P(\mathcal{Q}=j|\mathcal{M}_{i1},\mathcal{M}_{i2})}$ or $P(\xi_{ij}=1)$, can be calculated from Table 1. Therefore, the likelihood of the quantitative trait (y) and molecular markers ($\mathcal{M}_1, \mathcal{M}_2$) for one putative QTL ($\mathcal{Q}$) and can be constructed by a mixture model:

$$L(\Omega_p, \Omega_q | y, \mathcal{M}_1, \mathcal{M}_2) = \prod_{i=1}^{n} \sum_{j=0}^{2} \pi_{j|i} f_j(y_i | \Omega_q),$$

where Ω_p is a vector of the population genetic parameters $(p_1, p_2, p_3, D_{12}, D_{23}, D_{13}, D_{123})$ that is used to describe frequencies of haplotypes formed by markers and QTL and subsequently $\pi_{j|i}$ s, Ω_q is a vector of the quantitative genetic parameters that define genotype-specific traits, which contains ($\mu_j, j = 1, 2, 3$, and σ) for a continuous trait that is assumed to be normally distributed, and $f_j(\cdot)$ is the probability density function for QTL genotype *j*.

The likelihood function provides a model for obtaining the maximum likelihood estimates of the unknown parameters (Ω_p, Ω_q), which can be achieved by differentiating the log-likelihood with respect to each unknown parameter, setting the derivatives equal to zero and then solving the equations. The log-likelihood function of the phenotypic values is given by

$$\ell = \log[L(\Omega_p, \Omega_q | y, \mathcal{M}_1, \mathcal{M}_2)] = \sum_{i=1}^{n} \log\left[\sum_{j=0}^{2} \pi_{j|i} f_j(y_i | \Omega_q)\right]$$

Computational algorithms

Within the maximum likelihood estimation framework, an efficient EM algorithm can be implemented to obtain the MLEs of (Ω_p, Ω_q), and is summarized into the following steps:

Step 1. Give initial values for the unknown parameters (Ω_p, Ω_q);

Step 2. E step – Calculate the posterior probabilities for each subject *i* to carry a particular QTL genotype *j* using the equation $\Pi_{j|i} = \frac{\pi_{j|i} f_j(y_i|\Omega_q)}{\sum_{j=0}^{2} \pi_{j|i} f_j(y_i|\Omega_q).}$

Step 3. M step – Solve the log-likelihood equations for each parameter based on observed data and $\Pi_{j|i}$ to obtain its estimate. To estimate the quantitative genetic parameters (Ω_q), their expressions in closed forms can be derived based on the estimation equations. For the estimates of the population genetic

Table 1 Joint zygote probabilities of the QTL genotypes at QTL Q and two-marker genotypes at markers M1 and M2, as expressed in terms of zygote configurations in a natural population

Marker			Joint marker-QTL genotype frequency		
Genotype		**Frequency**	***qq* (0)**	***Qq* (1)**	***QQ* (2)**
$m_1m_1m_2m_2$	(00)	p_{00}^2	p_{000}^2	$2p_{010}p_{000}$	p_{010}^2
			(n_{000})	(n_{010})	(n_{020})
$m_1m_1M_2m_2$	(01)	$2p_{01}p_{00}$	$2p_{001}p_{000}$	$2p_{011}p_{000} + 2p_{010}p_{001}$	$2p_{010}p_{011}$
			(n_{001})	(n_{011})	(n_{021})
$m_1m_1M_2M_2$	(02)	p_{01}^2	p_{001}^2	$2p_{011}p_{001}$	p_{011}^2
			(n_{002})	(n_{012})	(n_{022})
$M_1m_1m_2m_2$	(10)	$2p_{00}p_{10}$	$2p_{100}p_{000}$	$2p_{110}p_{000} + 2p_{100}p_{010}$	$2p_{110}p_{010}$
			(n_{100})	(n_{110})	(n_{120})
$M_1m_1M_2m_2$	(11)	$2p_{11}p_{00}$ + $2p_{10}p_{01}$	$2p_{101}p_{000} + 2p_{100}p_{001}$	$2p_{111}p_{000} + 2p_{110}p_{001}$ + $2p_{101}p_{010} + 2p_{100}p_{011}$	$2p_{111}p_{010} + 2p_{110}p_{011}$
			(n_{101})	(n_{111})	(n_{121})
$M_1m_1M_2M_2$	(12)	$2p_{11}p_{01}$	$2p_{101}p_{001}$	$2p_{111}p_{001} + 2p_{101}p_{011}$	$2p_{111}p_{011}$
			(n_{102})	(n_{112})	(n_{122})
$M_1M_1m_2m_2$	(20)	p_{10}^2	p_{100}^2	$2p_{110}p_{100}$	p_{110}^2
			(n_{200})	(n_{210})	(n_{220})
$M_1M_1M_2m_2$	(21)	$2p_{11}p_{10}$	$2p_{101}p_{100}$	$2p_{111}p_{100} + 2p_{110}p_{101}$	$2p_{110}p_{111}$
			(n_{201})	(n_{211})	(n_{221})
$M_1M_1M_2M_2$	(22)	p_{11}^2	p_{101}^2	$2p_{111}p_{101}$	p_{111}^2
			(n_{202})	(n_{212})	(n_{222})

parameters (Ω_p), another inner layer of EM algorithm can be employed.

Step 4. Repeat the E and M steps until the estimates converge to stable values. The estimates at convergence are the MLEs of parameters.

The detailed derivation for the EM algorithm is given in Additional file 3.

Hypothesis testing

In general, the hypothesis testing of QTL mapping includes two steps: (1) the existence of QTL and (2) their locations. The focus of this study is on the second step, assuming that sufficient evidences for the existence of QTL have been collected to enable a large-scale genotyping study. Then the hypotheses for the tmLD method can be formulated as follows:

$$H_0: \text{The QTL is not associated with two SNP markers, } i.e.\ D_{12}=D_{23}=D_{123}=0; H_1: \text{Not } H_0$$

The estimates of the parameters under the null hypotheses can be obtained with the same EM algorithm derived for the alternative hypotheses, but with a constraint that all subjects have the same posterior probability. A likelihood ratio test (LRT) statistics can be constructed and computed to draw the inference about whether a QTL may be associated with given markers. Under the H_0, the LRT statistics asymptotically follows a χ^2-distribution with three degrees of freedom.

Results

Simulation settings

Extensive Monte Carlo simulation experiments were performed to examine the statistical properties of the proposed tmLD mapping method. Since in a genome-wide scan, a QTL must be located between some pair of markers, in the experimental design of simulations, we considered two scenarios as illustrated in Figure 1: (1) the QTL is assumed to be unobserved, but it is in LD with two adjacent SNPs; and (2) the QTL is assumed to be one of the genetic markers and therefore genotyped.

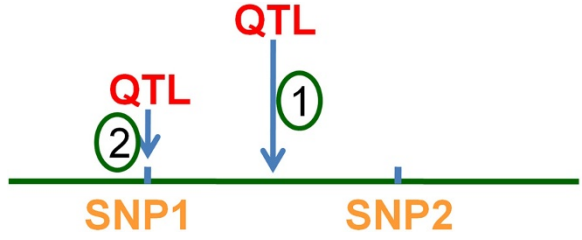

Figure 1 Two simulation settings. (1) QTL is unobserved but in linkage equilibrium with two adjacent SNPs. (2) QTL is observed as one of the SNP markers.

Let us randomly choose a sample of n subjects from a human population at Hardy-Weinberg equilibrium. In this population, one QTL is segregating and is inferred by a pair of markers. The allele frequencies of the markers ($\mathcal{M}_1$ and $\mathcal{M}_2$) and QTL ($\mathcal{Q}$) and their linkage disequilibria values are given as follows: $p_1 = 0.5$ for allele M_1 of $\mathcal{M}_1$; $p_2 = 0.5$ for allele Q of $\mathcal{Q}$; $p_3 = 0.5$ for allele M_2 of $\mathcal{M}_2$. The LD parameters among the markers and QTL loci are given as: $D_{12} = 0.05$, $D_{13} = 0.15$, $D_{23} = 0.05$ and $D_{123} = 0.04$. For subjects who carry QTL genotype j, their phenotypic values were simulated based on Model (3), with $\mu_2 = 10$, $\mu_1 = 5$, $\mu_0 = 0$. The variances in phenotypic values were calculated based on different heritability values (H^2). H^2 quantifies the genetic contribution from the QTL to the overall trait and $H^2 = 0$ implies that the means for three QTL genotype groups are the same, which are set to be 0. With the above given parameters and design, we simulated the phenotypic and marker information by assuming different sample sizes ($N = 100$, 250, 500, 1000, 1500, 2000, 2500, 3000), and different heritability values ($H^2 = 0$, 0.05, 0.1, 0.2, 0.3, 0.4). Each simulation setting is carried out 1000 times for the evaluation of power and type I error.

Type I error evaluation and power comparison

Simulated data were used to compare our proposed tmLD method with single-marker based association analyses, including the single-marker LD mapping method (smLD) and single-marker based association test (smAT), and two-marker based haplotype analysis (haplo). The smLD was performed as described in Additional file 4. The smAT is a simple linear regression model with phenotypic trait as response variable and marker genotypes as categorical independent variable. The haplotype analysis was conducted as described in [16]; briefly, the haplotype that yields the best model fitting among those formed by two markers is used in comparison with tmLD.

Under the simulation scenario 1, where the QTL is in LD phase with both markers, the results suggest that the association analysis based on two markers is significantly higher than the single- marker based and also haplotype based methods. Figure 2 shows that as the heritability increases, the power of each method increases correspondingly as expected. When $H^2 = 0$, which suggests no QTL effects, all methods maintained the nominal type I error (0.05); when $H^2 \neq 0$, the two-marker association performed consistently better than others, and as expected, the power increased with the sample size.

Under the simulation scenario 2, where the QTL is set to be the marker 1, the most powerful test is the single marker association method using marker 1, and the power of the single marker association based on marker 2 is significantly lower (Figure 3). However, the tmLD analysis is almost as powerful as the optimal test, particularly when the sample size is reasonably large (N > 1000). This demonstrates that

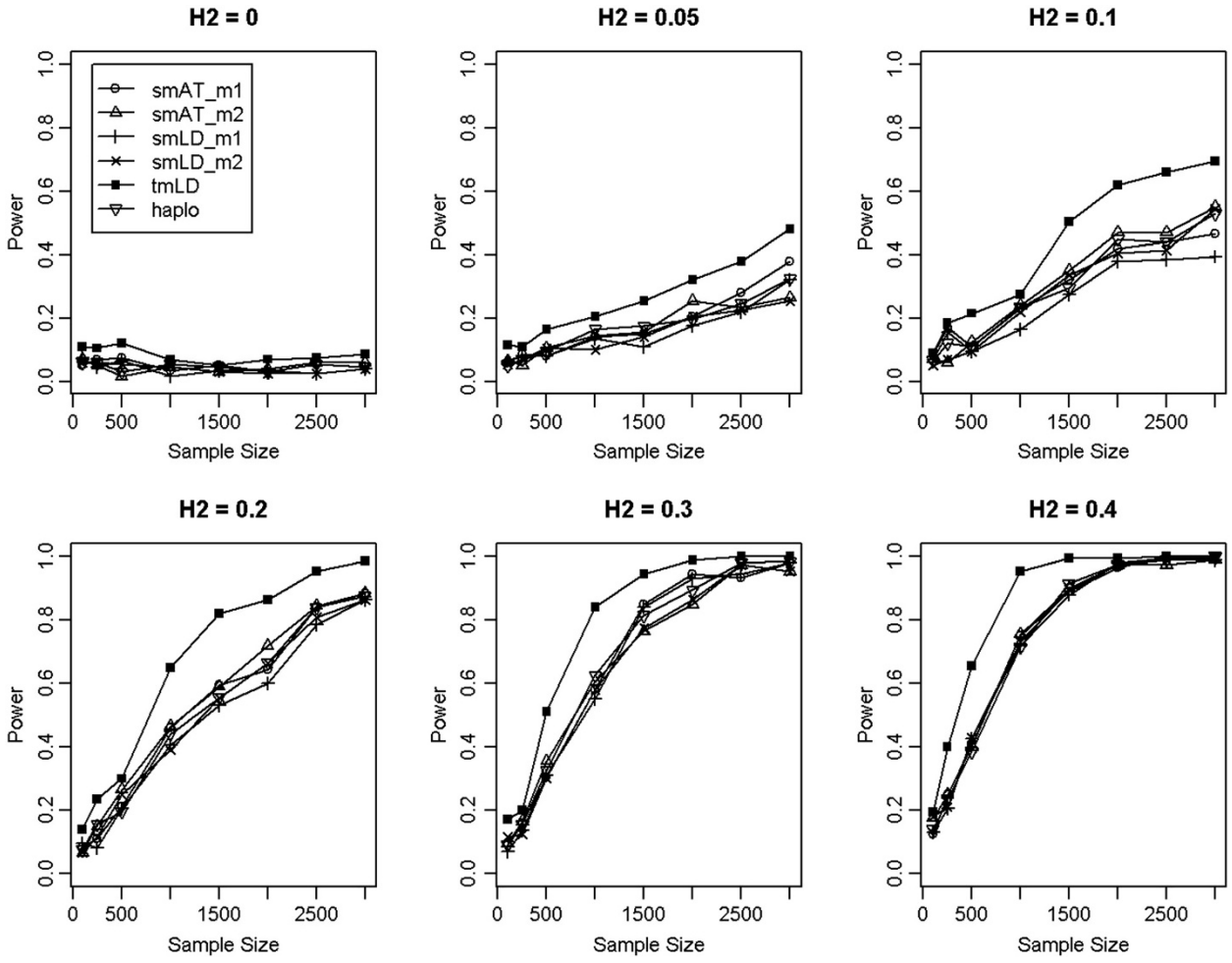

Figure 2 Power comparison when QTL is in linkage disequalibria with both marker 1 and marker 2. The power curves were constructed under different heritability (H^2). smAT_m1 and smAT_m2 denote the single-marker association analyses for marker 1 and marker 2, respectively; smLD_m1 and smLD_m2 denote single-marker LD mapping using marker 1 and marker 2, respectively; and haplo is for the two-marker based haplotype analysis.

even when the QTL is indeed sampled in a genomic study, our proposed model is as good as the optimal test. These simulation results demonstrate the power advantage and robustness of our proposed method comparing with existing methods based on single marker. Its practical usage was further elucidated in a real GWAS data set.

Real data example

Dental caries or cavities, more commonly known as tooth decay, is one of the most common chronic disorders in humans, affecting approximately 40% children and adolescents and 90% adults in the US. The etiology and pathogenesis of dental caries have been determined to be multifactorial, such as environmental factors related to social behaviors [17]. However, it is also apparent that some individuals are very susceptible to caries while some others are more resistant, almost irrelevant to the environmental risk factors they are exposed to, suggesting that genetic factors may play prominent roles in the caries development. Supported by evidence in both human and animal studies [18-21], the caries heritability has been estimated to be between 30-60%. The most compelling evidence come from the twin studies that the significant resemblance of dental caries lies within monozygotic but not dizygotic twin pairs [22,23]. So it is without question that in addition to environmental factors, genetic components also profoundly influence the dental caries trait. To understand the genetic mechanisms of the dental caries, a GWAS study has been conducted and the dataset has been deposited in dbGaP (Study Accession: phs000095.v2.p1). Here we will apply our proposed model to analyze this caries GWAS dataset, in which 1843 adults were genotyped with a large panel of SNPs (610,000). We carried out the analysis using the caries outcomes that have been well defined in other GWAS studies, i.e. the D1MFT index

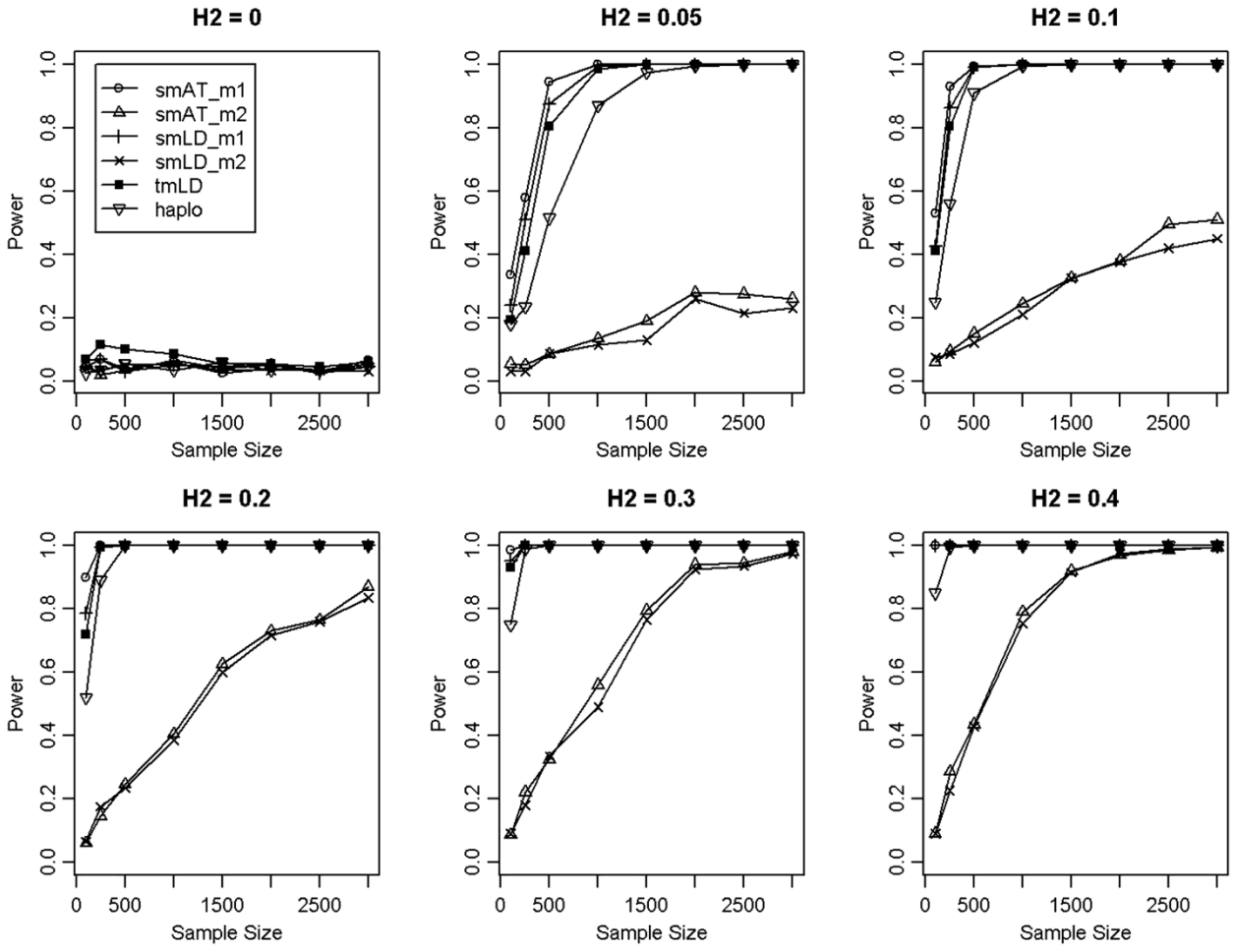

Figure 3 Power comparison when QTL is at the exact position of the marker 1. The power curves were constructed under different heritability (H^2). The tmLD model performs almost identically with the true model even when the QTL is the marker 1. smAT_m1 and smAT_m2 denote the single-marker association analyses for marker 1 and marker 2, respectively; smLD_m1 and smLD_m2 denote single-marker LD mapping using marker 1 and marker 2, respectively; and haplo is for the two-marker based haplotype analysis.

which quantifies the total permanent tooth caries with white spots.

smAT, smLD, haplo and tmLD association methods were applied to the data. After removing SNPs that do not satisfy HWE (p-values < 10^{-7}) and also SNPs with minor allele frequency less than 0.1, the number of SNPs that were included in the analysis is 443,175. To compare the performance of all methods, we plotted out the association signals at each SNP locus. Figures 4 and 5 show the Manhattan plots of the -log10(p-values) from smAT and tmLD methods, respectively, and the dashed red line corresponds to the genome-wide Bonferroni threshold (1.1E-7). SNPs that passed this threshold are considered to be significant and were tabulated in Table 2. For the haplo and smLD methods, since no significant SNP was identified by these two methods, their Manhattan plots were not shown. Particularly, the tmLD model identified two significant genes, CNTN5 and COL4A2, which have been shown from other studies to be associated with dental related phenotypes in other studies [24], validating the findings of our model biologically. None of the other three methods (smAT, smLD or haplo) found these two genes. The smAT identified another significant locus. However, gene annotation shows that it is not related to any known genes, so its biological implication remains unclear.

Discussion

It is well recognized that naturally occurring variations in most complex disease traits have a genetic basis and consequently many GWAS studies have been conducted in the past few years. In analyzing these data, a phenomenon, called "missing heritability", has been observed that the detected genetic variants can explain only a small portion of the heritability of phenotypic traits while a majority part remains mysterious [25]. Part of the reason may be attributed to the lack of power in current methods. Thus, developing

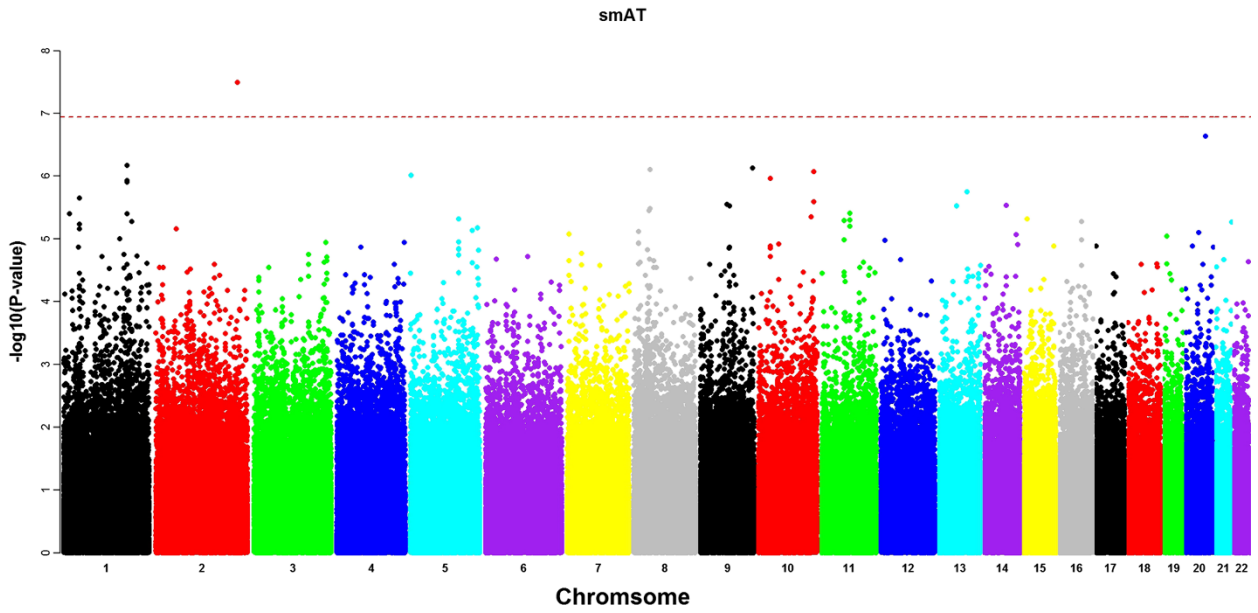

Figure 4 The Manhattan plot for GWAS scanning using the single marker association analysis. The x-axis displays the genomic coordinate of SNPs and the y-axis shows the negative base-10 logarithm of the association p-value for each SNP.

novel and powerful methods to better detect significant genes has been of great interest. Currently the routine GWAS analyses seek single-marker association between SNPs and phenotype, and when a significant association is detected, it implies that there might be some SNP(s) in linkage that are causal. Note that it cannot imply the test SNP itself is causal because there is no guarantee that the truly causal SNPs would have been genotyped. Since the interpretation of a significant association relies on the linkage concept, it is sensible to directly incorporate the LD information into association models. Additionally, due to the structure of LD blocks, a causal SNP is usually in linkage with multiple neighboring SNPs, all of which carry partial information about it. So in this sense, a new model that can incorporates more genetic information of linked SNPs should draw better inferences about the causal SNP.

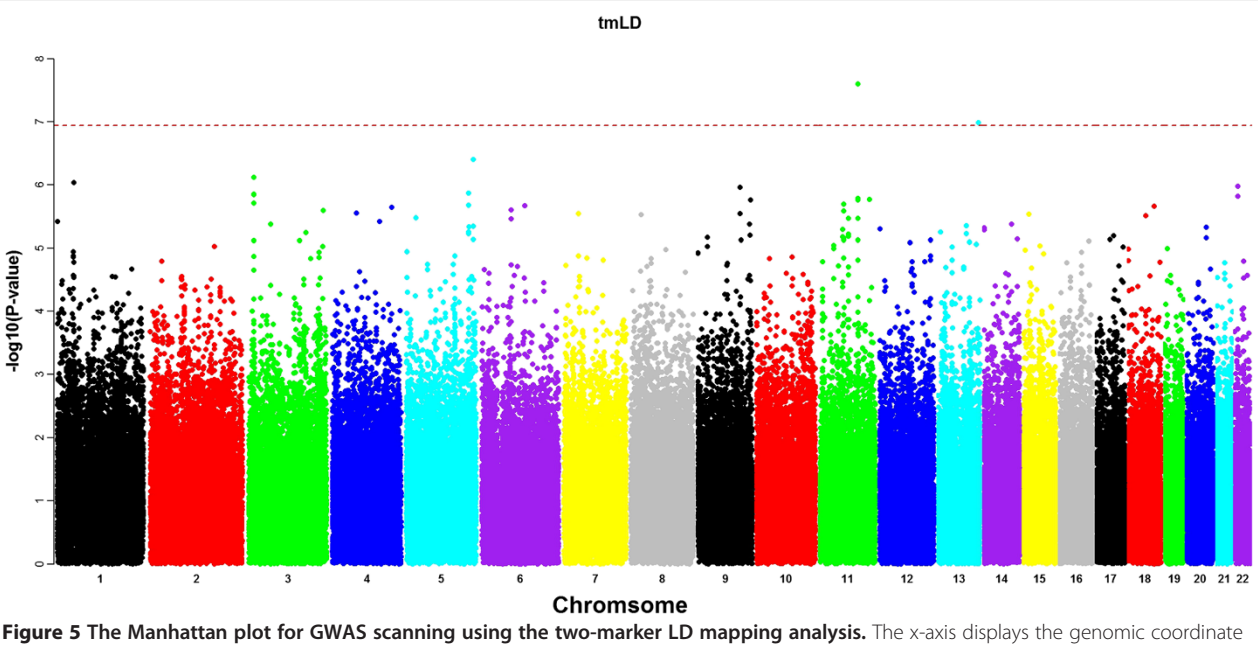

Figure 5 The Manhattan plot for GWAS scanning using the two-marker LD mapping analysis. The x-axis displays the genomic coordinate of SNPs and the y-axis shows the negative base-10 logarithm of the association p-value for each SNP.

Table 2 List of significant SNPs with p-value < 1.1e-7 in the Caries dataset

SNP ID	Gene	Chr	Coordinate	Allele	MAF	P_{smAT}	P_{smLD}	P_{haplo}	P_{tmLD}
rs7607421	–	2	220500564	C/T	0.390	**3.2E-08**[*]	2.1E-04	2.0E-06	6.9E-05
rs10790497	CNTN5	11	98539071	A/G	0.346	8.8E-01	8.2E-01	1.7E-03	**2.6E-08**[‡]
rs7319311	COL4A2	13	109828579	A/G	0.326	5.8E-02	2.7E-02	2.8E-02	**1.0E-07**[‡]

P_{smAT}, P_{smLD}, P_{haplo}, P_{tmLD}: p values for corresponding methods. [*]Significant SNPs identified by smAT. [‡]Significant SNPs identified by tmLD.

In this article, we proposed a novel statistical method by considering two SNPs simultaneously. Our model is built upon the general LD mapping framework, and extends the previous methods based on single-marker LD. The simulation studies demonstrated that our new methods dramatically improved the detection power of the underlying QTLs. This is intuitively reasonable since our model can capture the linkage information between SNP markers, and hence has more power to detect the particular QTL that are in LD with both markers. Furthermore, the simulation studies indicated that even when the underlying QTL is indeed genotyped and is one of the markers, the performance of the tmLD analysis is nearly identical to that of the optimal test resulting from the causal SNP, suggesting the robustness of our model.

We applied our model to a GWAS date set that aimed to understand the genetic mechanisms of the dental caries. The data set contains a large cohort of 1,843 subjects as well as a very large number of SNPs (443,175). This shows that both our proposed method and the corresponding software package in R can be well applied to a typical GWAS data set. In addition, we also observed that the association analyses based on the single-marker and the two-marker models yielded different profiles of significant SNPs. This is somewhat expected since their assumptions are different. For the tmLD method, we assume that both markers must obey HWE and have to be in LD with the casual SNP. It might be possible that some SNPs would violate these assumptions and become unsuitable to the tmLD. In this sense, the single and two-marker analyses may be complementary to each other, and therefore it might be beneficial to use both methods in analyzing a real data set.

Sometimes population structure may be a concern in a GWAS analysis if subpopulations indeed exist in the sample, as it may lead to spurious associations. Several well-known methods developed to account for population structure [26] can be incorporated into our LD mapping framework to address this issue. For instance, the principal component analysis (PCA) can be applied to correct for stratifications [27]. That is, we may first apply PCA on the genotype data and then choose the first few large principal components to be included in the Model (3) as additional covariates. With slight modifications, the computation algorithms and hypothesis testing described in the Method section can be readily applied.

In this work, we generalized the single marker LD analysis to a more general LD mapping framework using two adjacent markers. There are several ongoing works worthy of further investigation. First, the model can be easily extended to other types of phenotypic data, such as case–control binary and count data. Second, currently the two adjacent markers were used for the analysis; however, it is possible that another two markers in the same LD block might have better power, so it would be very interesting to determine how to choose the best SNP pair. Third, typically, one LD block may contain several SNPs, and if there exists one causal SNP within the LD block, it would be very interesting to see if we can summarize all SNPs in one LD block to make even better inference about the unobserved QTL.

Conclusions

The proposed tmLD model is a novel mapping method that can simultaneously consider two linked SNPs in a natural population. Through the extensive simulation studies, the tmLD method demonstrates better power than single-marker mapping strategies traditionally used in GWAS association analysis. The practical usage of the tmLD method was also shown in the analysis of a large-scale dental GWAS dataset. Hence, we recommend the usage of this improved method over the traditional single-marker association analysis.

Software availability

http://www.ams.sunysb.edu/~songwu/software.html.

Additional files

Additional file 1: Representation of three-loci haplotypes with four LD parameters.

Additional file 2: Derivation of how D_{123} may change with time.

Additional file 3: Derivation of the EM algorithm used to find MLEs for a mixture model.

Additional file 4: Single-marker based LD mapping.

Abbreviations

LD: Linkage disequilibrium; SNP: Single-nucleotide polymorphism; QTL: Quantitative trait loci; GWAS: Genome-wise association study; smAT: Single-marker association test; smLD: Single-marker linkage disequilibrium method; tmLD: Two-marker linkage disequilibrium method; haplo: Two-marker based haplotype analysis; MAF: Minor allele frequency; HWE: Hardy-Weinberg equilibrium.

Competing interests
No competing interests exist for any author.

Authors' contributions
JY conceived of the study, performed the statistical analysis and drafted the manuscript. WZ conceived of the study and drafted the manuscript. JC and QZ performed the statistical analysis and drafted the manuscript. SW conceived of the study, performed the statistical analysis, drafted the manuscript and developed the R package. All authors have read and approved the final manuscript.

Acknowledgements
The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that have helped improve the quality of the paper significantly. This work is partly supported by the FUSION award from the Stony Brook University to SW.
The dataset used in the real data example was obtained from dbGaP through dbGaP accession number [phs000095]. Funding support for collecting this dataset was provided by the National Institute of Dental and Craniofacial Research (NIDCR, grant number U01-DE018903). Data and samples were provided by: (1) the Center for Oral Health Research in Appalachia (NIDCR R01-DE 014899); (2) the University of Pittsburgh School of Dental Medicine (SDM) DNA Bank and Research Registry (NIH/NCRR/CTSA Grant UL1-RR024153); (3) the Iowa Fluoride Study and the Iowa Bone Development Study (NIDCR R01-DE09551and R01-DE12101); and (4) the Iowa Comprehensive Program to Investigate Craniofacial and Dental Anomalies (NIDCR, P60-DE-013076).

Author details
[1]Department of Preventive Medicine, Stony Brook University, Stony Brook, NY 11790, USA. [2]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11790, USA.

Received: 28 August 2013 Accepted: 31 January 2014
Published: 8 February 2014

References
1. Lynch M, Waslsh B: *Genetics and analysis of quantitative traits.* Sunderland, MA: Sinauer Associates, Inc.; 1998.
2. Wu RL, Ma CX, Casella G: *Statistical Genetics of Quantitative Traits: Linkage, Map and QTL.* New York: Springer-Verlag; 2007.
3. Lewontin RC: **The interaction of selection and linkage. I. General considerations; heterotic models.** *Genetics* 1964, **49**(1):49–67.
4. Hedrick PW: **Gametic disequilibrium measures: proceed with caution.** *Genetics* 1987, **117**(2):331–341.
5. Weir BS: **Genetic data analysis II.** Sunderland, MA: Sinauer Associates; 1996.
6. Kruglyak L: **Genetic isolates: separate but equal?** *Proc Natl Acad Sci U S A* 1999, **96**(4):1170–1172.
7. Farnir F, Grisart B, Coppieters W, Riquet J, Berzi P, Cambisano N, Karim L, Mni M, Moisio S, Simon P, Wagenaar D, Vilkki J, Georges M: **Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14.** *Genetics* 2002, **161**(1):275–287.
8. McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J: **Linkage disequilibrium in domestic sheep.** *Genetics* 2002, **160**(3):1113–1122.
9. Liu T, Todhunter RJ, Lu Q, Schoettinger L, Li HY, Littell RC, Burton-Wurster N, Acland GM, Lust G, Wu RL: **Modeling extent and distribution of zygotic disequilibrium: implications for a multigenerational canine pedigree.** *Genetics* 2006, **174**(1):439–453.
10. Lou XY, Casella G, Todhunter RJ, Yang MCK, Wu RL: **A general statistical framework for unifying interval and linkage disequilibrium mapping: toward high-resolution mapping of quantitative traits.** *J Am Stat Assoc* 2005, **100**(469):158–171.
11. Weiss KM, Clark AG: **Linkage disequilibrium and the mapping of complex human traits.** *Trends Genet* 2002, **18**(1):19–24.
12. Wu R, Ma CX, Casella G: **Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations.** *Genetics* 2002, **160**(2):779–792.
13. Wu R, Zeng ZB: **Joint linkage and linkage disequilibrium mapping in natural populations.** *Genetics* 2001, **157**(2):899–909.
14. Wang Z, Wu R: **A statistical model for high-resolution mapping of quantitative trait loci determining HIV dynamics.** *Stat Med* 2004, **23**(19):3033–3051.
15. Lander ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using rflp linkage maps.** *Genetics* 1989, **121**(1):185–199.
16. Wu S, Yang J, Wang C, Wu R: **A general quantitative genetic model for haplotyping a complex trait in humans.** *Curr Genomics* 2007, **8**(5):343–350.
17. Ditmyer MM, Dounis G, Howard KM, Mobley C, Cappelli D: **Validation of a multifactorial risk factor model used for predicting future caries risk with Nevada adolescents.** *BMC Oral Health* 2011, **11**:18.
18. Boraas JC, Messer LB, Till MJ: **A genetic contribution to dental-caries, occlusion, and morphology as demonstrated by twins reared apart.** *J Dent Res* 1988, **67**(9):1150–1155.
19. Bretz WA, Corby PM, Schork NJ, Robinson MT, Coelho M, Costa S, Melo MR, Weyant RJ, Hart TC: **Longitudinal analysis of heritability for dental caries traits.** *J Dent Res* 2005, **84**(11):1047–1051.
20. Bretz WA, Corby PMA, Melo MR, Coelho MQ, Costa SM, Robinson M, Schork NJ, Drewnowski A, Hart TC: **Heritability estimates for dental caries and sucrose sweetness preference.** *Arch Oral Biol* 2006, **51**(12):1156–1160.
21. Goodman HO, Luke JE, Rosen S, Hackel E: **Heritability in dental caries, certain oral microflora and salivary components.** *Am J Hum Genet* 1959, **11**(3):263–273.
22. Bretz WA, Corby PMA, Hart TC, Costa S, Coelho MQ, Weyant RJ, Robinson M, Schork NJ: **Dental caries and microbial acid production in twins.** *Caries Res* 2005, **39**(3):168–172.
23. Liu H, Deng H, Cao CF, Ono H: **Genetic analysis of dental traits in 82 pairs of female-female twins.** *Chin J Dent Res* 1998, **1**(3):12–16.
24. Bueno DF, Sunaga DY, Kobayashi GS, Aguena M, Raposo-Amaral CE, Masotti C, Cruz LA, Pearson PL, Passos-Bueno MR: **Human stem cell cultures from cleft lip/palate patients show enrichment of transcripts involved in extracellular matrix modeling by comparison to controls.** *Stem Cell Rev* 2011, **7**(2):446–457.
25. Zuk O, Hechter E, Sunyaev SR, Lander ES: **The mystery of missing heritability: genetic interactions create phantom heritability.** *Proc Natl Acad Sci U S A* 2012, **109**(4):1193–1198.
26. Wu C, DeWan A, Hoh J, Wang Z: **A comparison of association methods correcting for population stratification in case–control studies.** *Ann Hum Genet* 2011, **75**(3):418–427.
27. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**(8):904–909.

doi:10.1186/1471-2156-15-20
Cite this article as: Yang *et al.*: Genome-wide Two-marker linkage disequilibrium mapping of quantitative trait loci. *BMC Genetics* 2014 **15**:20.