Published in final edited form as:

IEEE Trans Nanobioscience. 2010 December ; 9(4): 232–241. doi:10.1109/TNB.2010.2070805.

Probability Theory-based SNP Association Study Method for Identifying Susceptibility Loci and Genetic Disease Models in Human Case-Control Data

Xiguo Yuan,

School of Computer Science & Engineering, Xidian University, Xi'an, 710071, P. R. China
(phone: (001)571-277-3208)

Junying Zhang, and

School of Computer Science & Engineering, Xidian University, Xi'an, 710071, P. R. China

Yue Wang

The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, VA, 22203, USA

Xiguo Yuan: xiguoyuan@mail.xidian.edu.cn; Junying Zhang: jyzhang@mail.xidian.edu.cn; Yue Wang: yuewang@vt.edu

Abstract

One of the most challenging points in studying human common complex diseases is to search for both strong and weak susceptibility single-nucleotide polymorphisms (SNPs) and identify forms of genetic disease models. Currently, a number of methods have been proposed for this purpose. Many of them have not been validated through applications into various genome datasets, so their abilities are not clear in real practice. In this paper, we present a novel SNP association study method based on probability theory, called ProbSNP. The method firstly detects SNPs by evaluating their joint probabilities in combining with disease status and selects those with the lowest joint probabilities as susceptibility ones, and then identifies some forms of genetic disease models through testing multiple-locus interactions among the selected SNPs. The joint probabilities of combined SNPs are estimated by establishing Gaussian distribution probability density functions, in which the related parameters (i.e., mean value and standard deviation) are evaluated based on allele and haplotype frequencies. Finally, we test and validate the method using various genome datasets. We find that ProbSNP has shown remarkable success in the applications to both simulated genome data and real genome-wide data.

Index Terms

Association study; SNPs; probability theory; Gaussian distribution; case-control

I. Introduction

Recently, much research effort on the identification of susceptibility SNP markers has been devoted to genome-wide association studies (GWAS). GWAS has been considered as a perfect way to fully understand the genetic variations underlying common complex diseases from the viewpoint of both theoretical and practical concerns [1], [2]. By far, a lot of academic achievements have been obtained via GWAS, such as Kawasaki Disease [3], type 2 diabetes [4], [5], and late-onset Alzheimer's disease (LOAD) [6]. Most of their findings show high significance levels associated with diseases (e.g., SNP 'rs201825' showed association with LOAD with $p\text{-value}=6.1\times 10^{-7}$). However, there are still several doubts in some of available methods that have been used as GWAS. For one thing, few validations

have been implemented on the methods. *e.g.*, they are not tested by using simulated genome data, which can provide criteria to evaluate performances. For another, the interactions between identified SNPs are not clear, *i.e.*, how the SNPs confer susceptibility to diseases. Apart from these two issues, there are many never explored complex diseases that need to be carved out. Thus, the challenge to both theory exploration and practice in identifying causal SNPs is very huge, and developing more powerful association analysis methods is indispensable and urgent.

Currently, various approaches have been proposed for testing multiple causal SNPs jointly. One kind of intriguing and popular strategies is to use tag SNP [7], and haplotype block structures [8], [9], to carry out association studies. This method may be capable of mapping haplotypes to diseases, but the accuracy is difficult to be assessed if the haplotype partitioning approaches are not designed well and usually the possible number of haplotype partitions is astronomical. In addition, the definitions of haplotype block and the selection of tag SNPs are very difficult, even the classical dynamic programming algorithm for haplotype block partitioning [10] is likely to result in a loss of around 15% of haplotype information. A recently developed Multipoint linkage disequilibrium mapping [11], is a virtual-variant approach that uses the haplotype-block information to increase the power of detection and localization of association with un-typed SNPs. This approach only considers the interactions between neighbouring SNPs. Multifactor Dimensionality Reduction (MDR) approach [12], has been designed to search for various interactions including single locus and multi-locus interactions. This approach fully takes into account most possible interactions, and its result is the multi-loci interaction that has the highest accuracy to classify the target individuals into cases and controls. Theoretically, MDR is a classifier that is built based on the selected markers, and the resultant interactions may not be the real disease models because different combinations of loci may be used to build different classifiers with high classification accuracy. Aiming at both joint analysis of multi-loci and reduction of genotyping effort, Jochen et al. proposed an entropy-based SNP selection method [13]. It succeeded in saving up to 30% of genotyping load in practice. However, this method does not directly bridge the relationships between SNPs and disease status.

Bearing the idea of identifying substantial susceptibility SNPs and disease models, in this paper, we develop an alternative approach, probability theory based SNP association study, called ProbSNP, to estimate SNPs across genome and select those with the lowest joint probabilities in combining with disease status as the susceptibility SNPs. Single-locus disease models are determined during this selection process. Furthermore, based on the selected SNPs, the method identifies two-locus disease models through testing any two SNPs combined with disease status. In order to increase the accuracy of the measurement of joint probabilities, we introduce Gaussian distributions to evaluate them, since we know that the Gaussian is often a good model for the actual probability distribution. Here, Gaussian distribution probability density functions are built based on allele and haplotype frequencies.

Currently, no authors can absolutely guarantee that their findings are the correct loci contributing to diseases, primarily because there are no standard criteria for their references so that both false-positive error (type I error) and false-negative error (type II error) rates are difficult to evaluate. This is a common problem for many researchers. To fully display and illustrate the performance of our method, we firstly apply it to simulated genome data. The main contribution from simulated data is that it can provide known answers for comparisons with identified results so as to guide us to evaluate results, regulating and improving the method. In the simulation studies, our method is able to exactly detect the ground truth SNPs and models with very low both type error rates. We also apply our method to a real genome-wide dataset for studying lung adenocarcinoma, finding out 92 SNPs and 2 single-locus and 8 two-locus models that have higher confidence to confer susceptibility to the disease. Most

of the identified SNPs are located in significant regions, which encompass or closely adjacent to tumor related genes. In these studies, we implement the method for several times using different input data sampled from the target data and get a series of results. We find that these results are highly consistent. We know that the essential causal SNPs underlying one type disease are not varying from one input data to another. Thus, the results indicate that ProbSNP is stable and robust to identify potential susceptibility SNPs.

II. METHODS

A. Principle and overview of ProbSNP

In this section, we consider a probability theory-based method to measure the associations between genetic variations and disease status. We believe that the lower probability one event is expected to occur, the more important the event is if it really occurs. From another perspective, lower probability of an event will contribute to larger amount of information for observations. This can be explained by referring to entropy information theory [14], [15], which is a good measure of information for variables.

For the purpose of SNP association mapping, we regard the disease status (class label) as a special SNP, termed 'sSNP'. This assumption is made based on the viewpoint of data representations, *e.g.*, common SNP alleles can be represented as '1' (major allele) or '2' (minor allele), disease status can also be represented as '1' (control) or '2' (case). Thus, disease status can be combined with SNPs conveniently on the same platform, and the state of the combinations can be observed and evaluated from the viewpoint of joint probability theory.

Fig. 1 gives the flow chart of ProbSNP for identifying susceptibility SNPs and disease models. The method uses joint probability to evaluate one SNP at a time by combining with 'sSNP', and selects the SNPs with the lowest joint probabilities as the susceptibility ones. Although many authors previously thought that marker-by-marker testing for association studies can lead to a locus-specific probability of type I error, and such error can easily be inflated when large number of SNP markers are tested simultaneously and independently [16], [17], our method directly treat SNPs in combining with disease status based on joint probability. This strategy can detect both strong and weak susceptibility SNPs, and both direct and indirect effective factors, since ProbSNP is sensitive to susceptibility SNPs including real causal SNPs and those in high linkage disequilibrium (LD) with them. Moreover, we implement the method for several times using various input data. The false-positives can be observed by comparing the results.

After the selection of SNPs, ProbSNP identifies some forms of interactions within the SNPs. ProbSNP only focus on identifying single-locus and two-locus disease models. This is primarily because some bias might be introduced when evaluating the probabilities of combinations involved with a large number of SNPs. For single-locus models, we determine them during the SNP selection process (this process is detailed in the next text). For two-locus models, we make combinations of any two selected SNPs and combine them with 'sSNP', the combinations with the lowest joint probabilities are regarded as disease models (detailed in the next text). The process of identifying two-locus disease models, at a certain degree, is similar to that of SNPs identification. In our method, the interactions of any two combined SNPs within the selected common SNPs will be fully considered. This is different from the set-association approach proposed by Hoh et al.[16], who primarily paid attention to the interactions of SNP markers with the highest values that were calculated by combining allelic association and Hardy-Weinberg disequilibrium.

B. Estimation of probabilities

For two given bi-allelic SNPs, SNP i and SNP j , the joint probability of them is estimated through building Gaussian distribution probability density function. We assume that, the major allele frequencies of the two SNPs are p_i and p_j , respectively, and the haplotype frequency of '1 1' (here, we use '1' to represent the major allele and '2' to represent the minor allele) in the combination of the two SNPs is p_{ij} . Next we use the haplotype frequency of '1 1' to reflect the state of the combination of SNP i and SNP j , given p_i and p_j . We know that there are four types of haplotypes in the combination of SNP i and SNP j , '1 1', '1 2', '2 1', and '2 2' (Fig. 2). If the frequency of '1 1' is determined, then the frequencies of other three haplotypes can also be determined, given p_i and p_j , and accordingly the state of the combination of SNP i and SNP j can be determined (here, the state of combination is described by the types of haplotypes and haplotype frequencies). Therefore, the frequency of '1 1' can be used to reflect the state of the combination. Of course, knowing any one of the other three haplotype frequencies ('1 2', '2 1', and '2 2') can also determine the state of the combination, i.e., the other three haplotype frequencies can also be used to reflect the state of the combination. Here, we use the frequency of '1 1' in our method.

Theoretically, the joint probability of SNP i and SNP j is the probability of the occurrence of the state of the combination of SNP i and SNP j . Now the problem of the joint probability estimation for SNP i and SNP j can be reduced to the assessment of the probability of the occurrence of the haplotype frequency of '1 1', p_{ij} , given p_i and p_j . To assess it, we firstly explore the distribution of the haplotype frequency in the combination of SNP i and SNP j . As we know, if the major alleles are randomly distributed in the panels of SNP i and SNP j , then the haplotype frequency of '1 1' is expected to be $p_i p_j$. Based on the major allele frequencies, p_i and p_j , we have randomly generated a large number of samples for the combination, and we have found that the haplotype frequency of '1 1' in the sample population follows a Gaussian distribution with mean value $\mu = p_i p_j$. For example, when $p_i = 0.6$ and $p_j = 0.7$, we randomly generate 10000 samples for the combination of SNP i and SNP j . The frequencies of '1 1' of these samples and their distributions are shown in Fig. 3, from which we can see that most of them are around 0.42 (i.e., 0.6×0.7) and the frequency follows a Gaussian distribution. After calculation, the mean value and standard deviation of such frequencies in the sample population are 0.42 and 0.0157, respectively. Thus, it is reasonable to set $\mu = p_i p_j$ to build Gaussian distribution for the haplotype frequency of '1 1'.

In the study of genome data, we should establish a Gaussian distribution for each common SNP in combining with 'sSNP' to estimate the joint probability. The standard deviation can be estimated by simulating a population of combinations based on the major allele frequencies. For saving computational time, we estimate the standard deviation as Equation 1, since the number of '1 1' approximately follows a binomial distribution when the population is large enough.

$$\sigma = \sqrt{\frac{\mu(1-\mu)}{s}} \quad (1)$$

where s is the size of the studied population. Therefore, given the major allele frequencies p_i and p_j , the Gaussian distribution probability density function can be obtained as below.

$$f(x|\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

where x is the frequency of '1 1' in the combination of SNP i and SNP j . The purpose of building the distribution is to estimate the probability of the occurrence of the haplotype frequency of '1 1'. For this, we design a probability function (Equation 3) by dividing the probability density function using $f(\mu)$.

$$P(x) = \frac{f(x|\mu, \sigma)}{\max f(x|\mu, \sigma)} = e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (3)$$

By far, the joint probability of the combination of SNP i and SNP j can be achieved, given p_i , p_j and p_{ij} , i.e., $P(\text{SNP } i, \text{SNP } j) = P(p_{ij})$, where p_{ij} is the haplotype frequency of '1 1'.

For a clear understanding, we simulate a population of 1000 samples to make an example for estimating the probability of the combination of a common SNP (SNP k) and 'sSNP'. In the simulated population, the major allele frequencies of 'sSNP' and SNP k are 0.360 (p_s) and 0.613 (p_k), respectively, and the haplotype frequency of '1 1' in the combination of 'sSNP' and SNP k is 0.232. According to Equation 2 and 3, we get a Gaussian distribution probability density function and the evaluated probability function shown in Fig. 4(a) and Fig. 4(b), respectively. Consequently, the probability of the occurrence of the haplotype ('1 1') frequency can be calculated according to Equation 3 by setting $x = 0.232$, $\mu = p_s p_k = 0.2207$, and $\sigma = 0.0131$ (calculated by Equation 1). After calculation, $P(\text{sSNP}, \text{SNP } k)$ is 0.6888.

In fact, it is not difficult to realize that the central feature of the estimation of probability is to observe the deviations of actual allelic associations from expectations or statistical allelic associations, and the greater the deviation is, the lower probability of the joint SNPs appears. Of course, for estimating the probability of each SNP combination, we should establish a Gaussian distribution probability density function and its corresponding probability function. Although we have only described the probability estimation of two combined SNPs, actually, for the combinations of more SNPs, the estimation of joint probabilities is similar.

C. SNP selection criterion

In order to decide which SNPs should be selected as susceptibility SNPs, we combine each common SNP with 'sSNP' and estimate the probability of the combination based on Gaussian distribution, as described in previous section, and we suggest a threshold θ (e.g., $\theta = 0.5$) so that the current SNP (SNP i) is selected if and only if

$$P(\text{sSNP}, \text{SNP } i) \leq \theta, (1 \leq i \leq n). \quad (4)$$

If θ is set as a small value, then not many SNPs can be selected. In this case, it is unavoidable to miss some potential susceptibility SNPs, so it is better to extend the selection result by including the SNPs that are in high LD with the selected SNPs. Otherwise, if θ is not so small, most of the SNPs that are strongly or weakly to confer susceptibility to diseases are likely to be included in the result. To sum up, if θ is well assigned, even the SNP selection is implemented marker-by-marker in our method, it usually can find out almost all the SNPs related with diseases and controls type I error rate to an acceptable level. In addition, the forms of potential multi-loci interactions among the selected SNPs will be identified in the next step.

D. Identification of genetic disease models

After the selection of susceptibility SNPs, we further use allelic association information and joint probability theory to identify some forms of interactions between them (i.e., to explore some structures of genetic disease models). For this purpose, we will deal with multiple SNPs at a time by combining them with ‘sSNP’. While in the realistic problems, it is well understood that the structures of disease models are very complicated, and both the number of underlying disease models and the number of SNPs involved with each model are not clear. By far, no experts can give the exact answers to these problems.

With these considerations, we try to identify various kinds of disease models. To avoid the bias of estimation of probabilities for large length of haplotypes, we only focus on searching for single-locus and two-locus disease models. Single-locus models can be identified in the SNP selection process. They are the SNPs that have the lowest probabilities in combined with ‘sSNP’ (e.g., the top one or two SNPs can be regarded as single-locus models). In this section, we emphasize the identification of two-locus disease models. Fig. 5 illustrates the process of disease models identification. In order to find out all possible two-locus disease models, we test any two SNPs among the selected SNPs by combining them with ‘sSNP’.

For example, if there are m selected SNPs, we should make $m(m-1)/2$ possible combinations with ‘sSNP’ (e.g., SNP 2 and SNP 3 are combined with ‘sSNP’, shown in Fig. 5). For all these possible combinations, we evaluate their probabilities through building Gaussian distribution probability density functions based on allele frequencies and haplotype frequencies. This evaluation is similar to that of single SNP combined with ‘sSNP’, which has been described in previous text. After evaluating the probabilities of all the combinations, we regard the combination with the lowest probability as the first two-locus disease model. For the rest susceptibility SNPs, we implement the same step to identify the second disease model, the third one, and the others.

As for the termination criterion for detecting disease models, we set a parameter k as the number of two-locus disease models that want to identify, or suggest a threshold δ so that another two-locus interaction with the lowest probability can be included if

$$p(\text{another model}) - p(\text{the first model}) \leq \delta \quad (5)$$

III. RESULTS

A. Application to simulated data

To validate the performance of our method, we implement it using a simulated genome dataset, which was generated by a recently developed simulator at <http://genomeld.sourceforge.net/>. The dataset consists of 2000 individuals, including 1276 cases and 724 controls. Each individual was simulated for 1000 SNPs. Three disease models (one single-locus model and two two-locus models) involved 5 susceptibility SNPs were embedded in the data. Clearly, the purpose of our method in this study is to find out the 5 susceptibility SNPs and identify the disease models. Of course, this study is not a genome-wide association study, but it serves for showing the ability of the method. For simplicity, we will report results using SNP ID numbers. The SNP ID numbers and minor allele frequencies (MAF) of the 5 SNPs and the disease models are given in Table I and Table II, respectively.

For studying this simulated data, we set the related parameters as $\theta = 0.48$ and $k=3$, and implement ProbSNP for three times using the three input datasets with different

cardinalities. Specifically, the sizes of the three input datasets are 1500, 1800, and 2000, respectively. The first two datasets are generated by randomly sampling from the target data. The SNP selection results are given in Fig. 6, which illustrates the three sequences of the selected SNPs and their corresponding probabilities by increasing value. We can see that the three results contain 22, 21, and 20 SNPs, respectively, and all the 5 embedded SNPs (circled ones) are included in each result. Moreover, all the three sequences are composed of almost the same SNPs. This indicates that our method shows stable and robust in identifying SNPs in this study. Such results imply a fact that the real casual SNPs are not varying from one input data to another one. Therefore, it is not exaggerated to say that a powerful SNP selection method should have the ability to find out the same SNPs with different input data, or else, users would be difficult to decide which result should be reliable.

For a clear overview of the probability distribution, we depict the probabilities of the 1000 SNPs in the case of input data with 2000 individuals in Fig. 7. It can be seen that most of the SNPs fall into the interval $[0.8, 1]$, while the causal SNPs and their neighbours are below around 0.5. Apart from the 5 ground truth SNPs, some other SNPs (e.g., 46, 47, etc.) have also been included in the results. After a further exploration, we find that the false positives are very near to and in high LD with the ground truth SNPs, see Fig. 8, which was plotted by Haploview software by showing D' values [18]. In some sense, the high LD related SNPs may also have some kind of relationships with phenotypes, so ProbSNP usually can detect LD related SNPs except for ground truth.

From the SNP selection results in the simulated dataset, we may conclude that our method has exactly found out the embedded susceptibility SNPs with no false-negatives. Moreover, the stable results make us have higher confidence to believe that the proposed method is powerful for identifying susceptibility SNPs in the study of complex diseases. As for the identification of the embedded disease models, we firstly select some representative SNPs in each block (shown in Fig. 8) for testing the interactions between the selected SNPs. Here, we do not deal with the combinations of the SNPs that are in fully LD, because the fully LD related SNPs can be regarded as a unity, and the state of the interactions between them can be predicated by a representative of them. According to the definition of LD [19], we know that a particular SNP at one site can predict other nearby SNPs within the same fully LD block.

With these considerations, in this study, we select SNPs 45, 50, 53, 371, 373, 413, 417, 438, and 439 as the representatives for disease models identification. For the identification of single-locus disease models, we get that SNP 438 has the lowest probability in combining with 'sSNP' during the SNP selection process. Thus, we regard 438 as a single-locus disease model. For identifying two-locus disease models, we set the related parameter k as 3, i.e., three such models will be selected. After the evaluation of the probabilities of any two SNPs within the selected representatives in combining with 'sSNP', we select the three combinations with the lowest probabilities as the disease models. The resultant models, including single-locus and two-locus models, and their corresponding probabilities are listed in Table. 3.

When compared with Table II, Table III shows two models (438 and (373, 413)) that have been found out by our method. However, because we have not considered the combinations of SNPs in fully LD, the model involved with 49 and 50 could not be identified. In fact, when we test the interaction of the two SNPs 49 and 50, we find that its joint probability in combining with 'sSNP' is $5.6851e-004$, which is very near to that of model (50, 53) shown in Table III. This means that any two combined SNPs within block 9 (Fig. 8(a)) might be regarded as disease models, and the identified model (50, 53) can be regarded as a typical one of them.

B. Comparison with other approaches

Through 100 replications of simulation of 2000 samples and 1000 SNPs under the 5 susceptibility SNPs (49, 50, 373, 413, and 438), we compared the statistical power of our method against two commonly used methods, chi-square test and MDR. The false positive rate (FPR) and true positive rate (TPR) were counted at different cutoffs of the number of selected SNPs and then averaged over the 100 replications of simulation. The receiver operating characteristic (ROC) curves are presented in Fig. 9. The ROC curves indicate that our method is more powerful than chi-square and MDR, especially when FPR is less than around 0.02. In analysis of real high-resolution SNP data, since a FPR lower than 0.02 is usually required due to the need to control false discovery rate (FDR). Therefore, our method will be more powerful in practice.

C. Application to real genome-wide data

We apply the method into a real genome-wide dataset to study the factors influencing lung adenocarcinoma. This dataset was generated with the STY chip of the Affymetrix Human Mapping array and is publicly available at <http://www.broad.mit.edu/cancer/pub/tsp/index.html>. More details can be found elsewhere [20]. To fit the input to the method, we translate the dataset into a matrix with rows corresponding to individuals and columns corresponding to SNPs. We use '1', '2', and '0' to represent major allele, minor allele, and missing data, respectively. This dataset consists of 290 individuals including 191 cases and 99 controls, each individual was genotyped for 238, 304 SNP markers.

In this study, we set the parameters as $\theta = 0.1$ and $k=8$, and we report results with SNP indices. The corresponding marker names are given in Supplementary Table I. We implement ProbSNP using three input datasets randomly sampled from the target data. Data_250, data_270 and data_290 include 250, 270 and 290 individuals, respectively. The results are given in Fig. 10, where result(250), result(270) and result(290) include 85, 87 and 92 SNPs, respectively. From the three sequences of results, we can see that they are almost the same except for several SNPs in differences. This means that ProbSNP is able to get stable results in studying real genome-wide data, making high confidence to believe that the selected SNP are those contributing to lung adenocarcinoma. In addition, we give an overview of the probability distribution of the 238, 304 SNPs in Fig. 11, where we can see that only the selected SNPs have probabilities below 0.1, while most of others are falling into the interval [0.4, 1].

Subsequently, we use ProbSNP to identify some forms of disease models underlying lung adenocarcinoma based on the selected SNPs. Within the 92 SNPs resulted from data_290, we identified 10 models including 2 single-locus and 8 two-locus models (Table IV). As for why to select the first two SNPs in the sequence (Fig. 10) as single-locus models, this is because they have the lowest probabilities in combining with 'sSNP', and moreover, Fisher's P value is decreasing to the lowest point at the second SNP (Fig. 12) among a certain number of SNPs (e.g., 12 SNPs) ranking in front of the sequence. As for the 8 two-locus models, we identified them because these two-locus interactions have the lowest joint probabilities with 'sSNP'. In Table IV, we can note that the probability increases rapidly from $2.9515e-101$ to $1.6742e-006$. In fact, the probability then steps into almost a constant level after the 8th model (this is not illustrated here). Among the 8 models, the model (156826, 227192) shows the highest significance level of $1.3e-03$ associated with lung adenocarcinoma. The Fisher's P-values were calculated by using a web software tool, SHEsis [21].

When compared with previously reported results [20], [22], many of the identified SNPs are located in significant regions, including 1p36, 5p14.3, 5p15, 7q11.22, 8q24, 9p21, 10q23, 12q13.1, 14q12, 20q13.2 and 20q13.31. These regions encompass or closely adjacent to tumor related genes (i.e., oncogenes or tumor suppressor genes), such as TERT, CDH12, MYC, PTEN, CALN1, CDKN2A and CDKN2B. These results indicate that our method is applicable to real data. For a further understanding of the result from our method, we make a comparison with chi-square test method in studying lung adenocarcinoma. The result of chi-square method is obtained by applying Bonferroni corrected p-value <0.05 , and is presented in Supplementary Table II. We find that there are only a few overlaps between the two methods, and most of the SNPs identified by chi-square method are not located in tumor related regions reported previously. Therefore, our method outperforms chi-square method in this study.

IV. DISCUSSION

In this paper, we present a novel SNP association study method based on probability theory to identify susceptibility SNPs and disease models. The method has been validated through applications into simulated genome data, exactly finding out the ground truth SNPs, and also has shown success in studying a real genome-wide dataset under a GWAS setting. It furnishes 92 SNPs and 2 single-locus and 8 two-locus models that are likely to confer susceptibility to lung adenocarcinoma. One of the central features to this method is its high efficient and stable performance, i.e., it is able to generate the consistent results using different input datasets randomly sampled from the target data. Such results, in turn, make us have high confidence to believe the findings to be the real susceptibility SNPs.

In our method, the most important contribution to the results from the simulated and real genome data sets is the use of probability theory. In the test of SNPs, we directly combine them with the disease status, 'sSNP', and then evaluate the probabilities of such combinations by building Gaussian distribution probability density functions. If the probabilities are very low, then the SNPs are likely to be included in results. As for the theoretical basis to this selection strategy, it can be discussed from the following two points. For one thing, it is well-known that lower probability of one event usually produces more information for observations. This means that the event with lower probability usually plays significance role in its environment. Another point is that, for example, a common SNP, SNP_i , combined with 'sSNP', if SNP_i has not any relationship with 'sSNP', it is easy to expect that the major alleles ('1') of the two SNPs will distribute in the combined panel in a random way, i.e., the haplotype frequency of '1 1' is expected to be around the multiplying of the both major allele frequencies (e.g., p , q are the major allele frequencies of SNP_i and 'sSNP', the multiplying is pq). In this case, the combination probability of SNP_i and 'sSNP' will be very large. Inversely, if SNP_i has some relations with 'sSNP', the distributions of both major alleles are not under a random way, so that the probability of the combination will be relatively small.

Of course, there is no absolute guarantee that this method correctly identified all the SNPs contributing to lung adenocarcinoma. However, the results are really fit to our expectations. We also believe the results after observing the behaviours of the method's applications to simulated datasets. As for the identification of disease models, in the simulated dataset, most of the embedded disease models have been detected, while in studying lung adenocarcinoma, 10 models have been detected, in which the model (156826, 227192) shows the highest significance level of $1.3e-03$ associated with lung adenocarcinoma.

However, several problems need to be addressed here. First of all, in the study of the real genome-wide data, genotype errors and missing data affect our method seriously.

Obviously, this is a common problem for many researchers. However, improving quality control technology in laboratory will be efficient for reducing genotyping error rate. As for missing data, there are some existing methods, such as GERBIL algorithm [23] and EM algorithm [24], are good tools for imputing them. Secondly, our method has only focused on identifying d -locus disease models with small d due to the limitation of data. In the case of higher-dimensional models, e.g., a five-locus model, there would be 2^6 possible haplotypes in the combination of the five loci and sSNP (disease status), and the averaged frequency of each possible haplotype in a population with size of 2000 is around 0.015. Such frequency is too low to be used to reasonably evaluate the probability of the combination of the six markers because there are usually some random events that can introduce such frequency to some particular haplotypes. Therefore, it would result in much bias if the method identified higher-dimensional models based on the probability. Theoretically, for discovering higher-dimensional models with the proposed method, more data should be collected.

As for our future work, we intend to introduce conditional probability theory to improve the method, and explore novel ways to establish probability density function based on allelic association and Hardy-Weinberg disequilibrium (HWD), and try to find more sophisticated strategies to test multiple-locus models. Additionally, studying various common complex diseases through applying for public genome data sets, such as WTCCC (<http://www.wtccc.org.uk/>) and Malaria GEN (<http://www.malariagen.net/access>), is our principal goal in developing powerful SNP association analysis methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the National Science Fund of China under Grant nos. 60933009, 60371044, and by the US National Institutes of Health under Grants GM085665 and HL090567.

References

1. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* Feb;2005 6:109–18. [PubMed: 15716907]
2. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* Feb;2005 6:95–108. [PubMed: 15716906]
3. Burgner D, Davila D, Breunis WB, Ng SB, Li Y, Bonnard C, et al. A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease. *PLoS Genet* Jan;2009 5:e1000319. [PubMed: 19132087]
4. Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PI, Chen H, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* Jun;2007 316:1331–6. [PubMed: 17463246]
5. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren W, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* Jun;2007 316:1341–5. [PubMed: 17463248]
6. Abraham R, Moskvin V, Sims R, Hollingworth P, Morgan A, Georgieva L, et al. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med Genomics* Sep; 2008 1:1–44. [PubMed: 18237448]
7. Stram DO. Tag SNP selection for association studies. *Genet Epidemiol* Dec;2004 27:365–74. [PubMed: 15372618]
8. Zhang K, Calabrese P, Nordborg M, Sun F. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* Dec;2002 71:1386–94. [PubMed: 12439824]

9. Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res* May;2004 14:908–16. [PubMed: 15078859]
10. Zhang K, Deng M, Chen T, Waterman MS, Sun F. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci U S A* May;2002 99:7335–9. [PubMed: 12032283]
11. Zheng M, McPeck MS. Multipoint linkage-disequilibrium mapping with haplotype-block structure. *Am J Hum Genet* Jan;2007 80:112–25. [PubMed: 17160899]
12. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* Feb;2003 19:376–82. [PubMed: 12584123]
13. Hampe J, Schreiber S, Krawczak M. Entropy-based SNP selection for genetic association studies. *Hum Genet* Dec;2003 114:36–43. [PubMed: 14505034]
14. Shannon CE. A mathematical theory of communication. *Bell systems Technical journal* Oct;1948 27:379–423. 623–656.
15. Gray, RM. *Entropy and Information Theory*. Springer-Verlag; 1990.
16. Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* Dec;2001 11:2115–9. [PubMed: 11731502]
17. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* Sep; 1996 273:1516–7. [PubMed: 8801636]
18. Barrett JC, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* Jan;2005 21:263–5. [PubMed: 15297300]
19. The International HapMap Consortium. A haplotype map of the human genome. *Nature* Oct;2005 437:1299–320. [PubMed: 16255080]
20. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhi R, et al. Characterizing the cancer genome in lung adenocarcinoma. *Nature* Dec;2007 450:893–8. [PubMed: 17982442]
21. Shi YY, He L. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. *Cell Res* Feb;2005 15:97–8. [PubMed: 15740637]
22. Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD, Chen K. CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics* Feb;2010 26:464–9. [PubMed: 20031968]
23. Kimmel G, Shamir R. GERBIL: Genotype resolution and block identification using likelihood. *Proc Natl Acad Sci U S A* Jan;2005 102:158–62. [PubMed: 15615859]
24. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc Ser B (Methodological)* 1977;39:1–38.

Biographies

Xiguo Yuan was born in China, on Dec. 17, 1983, and received B.S. and M.S. degree in computer applications from Wuhan University of Science & Technology, Wuhan, Hubei, China, in 2005 and 2008 respectively.

He is currently with the School of Computer Science & Engineering, Xidian University, and The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University.

Junying Zhang received her PhD degree in Signal and Information Processing from Xidian University, Xi'an, China, in 1998. From 2001 to 2002 and in the first half of 2007 respectively, she was a visiting scholar at the department of electrical engineering and computer science, the Catholic University of America, Washington, DC, USA, and at the Bradley department of electrical and computer engineering, Virginia Polytechnic Institute and State University, USA. She is currently a Professor in the School of Computer Science and Engineering in Xidian University, Xi'an, China. Her research interests cover from intelligent information processing, machine learning and causation learning, and their applications to genome-wide association study, network motifs in biological networks, cancer related bioinformatics, medical image processing and pattern recognition.

Yue Wang received his B.S. and M.S. degrees in electrical and computer engineering from Shanghai Jiao Tong University in 1984 and 1987 respectively. He received his Ph.D. degree in electrical engineering from University of Maryland Graduate School in 1995. In 1996, he was a postdoctoral fellow at Georgetown University School of Medicine. From 1996 to 2003, he was an assistant and later associate professor of electrical engineering at The Catholic University of America. In 2003, he joined Virginia Polytechnic Institute and State

University and is currently the endowed Grant A. Dove Professor of electrical and computer engineering. Yue Wang became an elected Fellow of The American Institute for Medical and Biological Engineering (AIMBE) and ISI Highly Cited Researcher by Thomson Scientific in 2004. His research interests focus on statistical pattern recognition, machine learning, signal and image processing, with applications to computational bioinformatics and biomedical imaging for human disease research.

Fig. 1.

The flow chart of ProbSNP for identifying susceptibility SNPs and disease models. n is the number of common SNPs across the genome.

Fig. 2.
The combination of SNP i and SNP j , there are four types of haplotypes, '1 1', '1 2', '2 1', and '2 2'.

Fig. 3.

(a) The haplotype frequencies of '1 1' of the randomly generated 10000 combinations of SNP i and SNP j , whose major allele frequencies are 0.6 and 0.7, respectively. (b) The Gaussian distribution probability density function for the haplotype frequency.

Fig. 4.
 (a) Gaussian probability density function. (b) The estimated probability function.

Fig. 5.
Flow chart of the process in identifying two-locus disease models.

Fig. 6.

The SNP selection results for the simulated data. (Upper) Three sequences of the selected SNPs for the three input data sets. The circled SNPs are the susceptibility SNPs. (Below) The probability as function of the SNP number. The three results, result(1500), result(1800), and result(2000) include 22, 21, 20 SNPs, respectively.

Fig. 7.

The probability distribution of the 1000 SNPs. It is clear that four clusters of SNPs (centred around 50, 373, 413, and 438) are below around 0.5.

Fig. 8. LD plots for the SNPs (D' values shown) detected by ProbSNP. The values of the dark black boxes are 100.

Fig. 9. Power comparison between Chi-square test, MDR and ProbSNP (our method). The result is based on 100 replications of simulation under the 5 ground truth. TPR and FPR are averaged over the replications.

Fig. 10.

The SNP selection results for the three input data (data_250, data_270, and data_290). (a) The probability as a function of the SNP number. The SNPs with probabilities below 0.1 have been selected. (b) The three sequences of the selected SNPs, including 85, 87, and 92 SNPs, respectively.

Fig. 11.

The probability distribution of the 238,304 SNPs. The probabilities of most SNPs are falling into the interval $[0.4, 1]$, while few are below 0.1.

Fig. 12.

The Fisher's p-values of the first 12 SNPs in the identified sequences, SNP 216314 shows the lowest p-value at the second point.

TABLE I

THE 5 SNPS ID NUMBERS AND THEIR MAF

ID	MAF	ID	MAF	ID	MAF
49	0.222	373	0.202	438	0.399
50	0.222	413	0.294		

TABLE II
THE 3 DISEASE MODELS AND THEIR INVOLVED SNPS

Model	Involved SNP
Two-locus model	49, 50
Two-locus model	373, 413
Single-locus model	438

TABLE III**IDENTIFIED DISEASE MODELS AND THEIR PROBABILITIES**

Model	Involved SNP	Probability
Single-locus model	438	1.86e-005
Two-locus model	373, 413	2.37e-023
Two-locus model	317, 417	2.00e-019
Two-locus model	50, 53	1.05e-004

TABLE IV

THE IDENTIFIED DISEASE MODELS, AND THEIR PROBABILITIES AND FISHER'S P-VALUES

Model	Involved SNP	Probability	P-value
Single-locus	211909	1.20e-02	2.2e-01
Single-locus	216314	1.41 e-02	5.6e-02
Two-locus	109658, 13238	2.95e-101	8.4e-03
Two-locus	140813, 71676	1.15e-076	2.4e-01
Two-locus	156826, 27192	1.27e-049	1.3e-03
Two-locus	87211, 9228	2.83e-032	5.4e-02
Two-locus	153164, 11209	2.40e-029	4.1e-02
Two-locus	49319, 179783	6.51e-016	8.4e-01
Two-locus	38967, 183885	8.21e-013	6.6e-02
Two-locus	186442, 62440	1.67e-006	3.9e-03