OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Education

Chapter 10: Mining Genome-Wide Genetic Markers

Xiang Zhang[1], Shunping Huang[2], Zhaojun Zhang[2], Wei Wang[3]*

[1] Department of Electrical Engineering and Computer Science, Case Western Reserve University, Ohio, United States of America, [2] Department of Computer Science, University of North Carolina at Chapel Hill, North Carolina, United States of America, [3] Department of Computer Science, University of California at Los Angeles, California, United States of America

Abstract: Genome-wide association study (GWAS) aims to discover genetic factors underlying phenotypic traits. The large number of genetic factors poses both computational and statistical challenges. Various computational approaches have been developed for large scale GWAS. In this chapter, we will discuss several widely used computational approaches in GWAS. The following topics will be covered: (1) An introduction to the background of GWAS. (2) The existing computational approaches that are widely used in GWAS. This will cover single-locus, epistasis detection, and machine learning methods that have been recently developed in biology, statistic, and computer science communities. This part will be the main focus of this chapter. (3) The limitations of current approaches and future directions.

This article is part of the "Translational Bioinformatics" collection for *PLOS Computational Biology*.

1. Introduction

With the advancement of genotyping technology, genome-wide high-density single nucleotide polymorphisms (SNPs) of human and other organisms are now available [1,2]. The goal of genome-wide association studies (GWAS) is to seek strong associations between phenotype and genetic variations in a population that represent (genomically proximal) causal genetic effects. As the most abundant source of genetic variation, millions of SNPs have been genotyped across the entire genome. Analyzing such large amount of markers poses great challenges to traditional computational and statistical methods. In this chapter, we introduce the basic concept of genome-wide association study, and discuss recently developed methods for GWAS.

Genome-wide association study is an inter-discipline problem of biology, statistics and computer science [3,4,5,6]. In this section, we will first provide a brief introduction to the necessary biological background. We will then formalize the problem and discuss both traditional and recently developed methods for genome-wide analysis of associations.

A human genome contains over 3 billion DNA base pairs. There are four possible nucleotides at each base in the DNA: adenine (A), guanine (G), thymine (T), and cytosine (C). In some locations in the genome, a genetic variation may be found which involves two or more nucleotides across different individuals. These genetic variations are known as *single-nucleotide polymorphism* (SNPs), i.e., a variation of a single nucleotide in the DNA sequence. In most cases, there are two possible nucleotides for a variant. We denote the more frequent one as "0", and the less frequent one as "1". For bases on autosomal chromosomes, there are two parallel nucleotides, which leads to three possible combinations, "00", "01" and "11". These genotype combinations are known as "major homozygous site", "heterozygous site" and "minor heterozygous site" respectively. These genetic variations contribute to the phenotypic differences among the individuals. (A phenotype is the composite of an organism's observable characteristics or traits.) Genome-wide association study (GWAS) aims to find strong associations between SNPs and phenotypes across a set of individuals.

More formally, let $X=\{X_1,X_2,\cdots,X_N\}$ be the set of N SNPs for M individuals in the study, and Y be the phenotype of interest. The goal of GWAS is to find SNPs (markers) in X, that are highly associated with Y. There are several challenging issues that need to be addressed when developing an analytic method for GWAS [7,8].

Scalability Most GWAS datasets consist of a large number of SNPs. Therefore the algorithms for GWAS need to be highly scalable. For example, for a typical human GWAS, the dataset may contain up to millions SNPs and involve thousands of individuals. Inefficient methods may consume a large amount of computational resources and time to find highly associated SNPs.

Missing markers Even with the current dense genotyping technique, many genetic variants are still not genotyped. Current methods usually assume genetic linkage to enhance the power. Imputation, which tries to impute the unknown markers by using existing SNPs databases, is another popular approach to handle missing markers. The well known related projects include the International HapMap project [9] and the 1000 Genomes Project [10].

Complex traits One approach in GWAS is to test the association between the trait and each marker in a genome, which is successful in detecting a single gene related disease. However, this approach may have problems in finding markers associated with complex traits. This is because that complex traits are affected by multiple genes, and each gene may only have a weak association with the phenotype. Such markers with low marginal effects are hard to detect by the single-locus methods.

Citation: Zhang X, Huang S, Zhang Z, Wang W (2012) Chapter 10: Mining Genome-Wide Genetic Markers. PLoS Comput Biol 8(12): e1002828. doi:10.1371/journal.pcbi.1002828

Editors: Fran Lewitter, Whitehead Institute, United States of America and Maricel Kann, University of Maryland, Baltimore County, United States of America

Published December 27, 2012

Funding: This work was supported by the following grants: NSF IIS-1162369, NSF IIS-0812464, NIH GM076468 and NIH MH090338. The funders had no role in the preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: weiwang@cs.ucla.edu

What to Learn in This Chapter

- The background of Genome-wide association study (GWAS).
- The existing computational approaches that are widely used in GWAS. This will cover single-locus, epistasis detection, and machine learning methods.
- The limitations of current approaches and future directions.

In the remainder of the chapter, we will first discuss the single-locus methods. We will then study epistasis detection (multi-locus) approaches which are designed for association studies of complex traits. For epistasis detection, we will mainly focus on exact two-locus association mapping methods.

2. Single-Locus Association Mapping

As the rapid development of high-throughput genotyping technology, millions of SNPs are now available for genome-wide association studies. Single-locus association test is a traditional way for association studies. Specifically, for each SNP, a statistical test is performed to evaluate the association between the SNP and the phenotype. A variety of tests can be applied depending on the data types. The phenotype involved in a study can be case-control (binary), quantitative (continuous), or categorical. We categorize the statistical tests based on what kind of phenotypes they can be applied on.

2.1 Problem Formalization

Let $\{X_1, \cdots, X_N\}$ be a set of N SNPs for M individuals and $X_n = \{X_{n1}, \cdots, X_{nM}\}$ $(1 \le n \le N)$. We use 0, 1, 2 to represent the homozygous major allele, heterozygous allele, and homozygous minor allele respectively. Thus we have that $X_{nm} \in \{0,1,2\}$ $(1 \le n \le N, 1 \le m \le M)$. Let $Y = \{y_1, \cdots, y_M\}$ be the phenotype. Note that the values that Y can take depend on its type.

2.2 Case-Control Phenotype

In a case-control study, the phenotype can be represented as a binary variable with 0 representing controls and 1 representing cases.

A contingency table records the frequencies of different events. Table 1 is an example contingency table. For a SNP X_n and a phenotype Y, and we use O_{ij} to denote the number of individuals whose X_n equals i and Y equals j. Also, we have $O_{i.} = \sum_j O_{ij}$ and $O_{.j} = \sum_i O_{ij}$. The total number of individuals $S = \sum_{i,j} O_{ij}$.

Many tests can be used to assess the significance of the association between a single SNP and a binary phenotype. The test statistics are usually based on the contingency table. The null hypothesis is that there is no association between the rows and columns of the contingency table.

2.2.1 Pearson's χ^2 test. Pearson's χ^2 test can be used to test a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution [11].

The value of the test statistic is

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where $E_{ij} = \frac{O_{i.} O_{.j}}{S}$. The degree of freedom is 2.

2.2.2 G-test. G-test is an approximation of the log-likelihood ratio. The test statistic is

$$G = 2 \sum_i \sum_j O_{ij} \cdot ln(\frac{O_{ij}}{E_{ij}}),$$

where $E_{ij} = \frac{O_{i.} O_{.j}}{S}$.

The null hypothesis is that the observed frequencies result from random sampling from a distribution with the given expected frequencies. The distribution of G is approximately that of χ^2, with the same degree of freedom as in the corresponding χ^2 test. When applied to a reasonable size of samples, the G-test and the χ^2 test will lead to the same conclusions.

2.2.3 Fisher exact test. When the sample size is small, the Fisher exact test is useful to determine the significance of the association. The p-value of the test is the probability of the contingency table given the fixed margins. The probability of obtaining such values in Table 1 is given by the hypergeometric distribution:

$$p = \frac{\binom{O_{.0}}{O_{00}} \binom{O_{.1}}{O_{01}} \binom{O_{.2}}{O_{02}}}{\binom{S}{O_{0.}}} = \frac{(O_{.0}! O_{.1}! O_{.2}!)(O_{0.}! O_{1.}!)}{S!(O_{00}! O_{01}! O_{02}! O_{10}! O_{11}! O_{12}!)}$$

Most modern statistical packages can calculate the significance of Fisher tests. The actual computation performed by the existing software packages may be different from the exact formulation given above because of the numerical difficulties. A simple, somewhat better computational approach relies on a gamma function or log-gamma function. How to accurately compute hypergeometric and binomial probabilities remains an active research area.

Table 1. Contingency table for a single SNP X_n and a phenotype Y.

	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

doi:10.1371/journal.pcbi.1002828.t001

2.2.4 Cochran-Armitage test. For complex traits, contributions to disease risk from SNPs are widely considered to be roughly additive. In other words, the heterozygous alleles will have an intermediate risk between two homozygous alleles. Cochran-Armitage test can be used in this case [12,5]. Let the test statistic of U be the following:

$$U = O_{1.} O_{0.} \sum_{i=0}^{2} i \cdot (\frac{O_{1i}}{O_{1.}} - \frac{O_{0i}}{O_{0.}})$$

After substitution, we get

$$U = S \cdot (O_{11} + 2O_{12} - O_{1.} \cdot (O_{.1} + 2O_{.2})$$

The variance of U under the null hypothesis can be computed as

$$Var(U) = \frac{(S - O_{1.}) O_{1.}}{S} [S(O_{.1} + 4O_{.2}) - (O_{.1} + 2O_{.2})^2]$$

Notice that for a large sample size S, we have $\frac{U}{\sqrt{Var(U)}} \sim N(0,1)$, hence $\frac{U^2}{Var(U)} \sim \chi_1^2$.

2.2.5 Summary. There is no overall winner of the introduced tests. Cochran-Armitage test may not be the best if the risks are deviated from the additive model. Meanwhile, χ^2 test, G-test, and Fisher exact test can handle the full range of risks, but they will unavoidably lose some power in the detection of additive ones. Different tests may be applied on the same data to detect different effects.

2.3 Quantitative Phenotype

In addition to case-control phenotypes, many complex traits are quantitative. This type of study is also often referred to as the quantitative trait locus (QTL) analysis. The standard tools for testing the association between a single marker and a continuous outcome are analysis of variance (ANOVA) and linear regression.

2.3.1 One-way ANOVA. The F-test in one-way analysis of variance is used to assess whether the expected values of a quantitative variable within several predefined groups differ from each other.

For each SNP X_n, we can divide all the individuals into three groups according to their genotypes. Let $Y_i'(i\in\{0,1,2\})$ be a subset of phenotypes of which the individuals have the genotypes equal to i. We represent the number of phenotypes in Y'_i as M_i, and we have $Y_i'=\{y_{i1},\cdots,y_{iM_i}\}$. Notice that $\bigcup_{i=0}^{2} Y_i'=Y$ and $\sum_{i=0}^{2} M_i=M$

The total sum of squares (SST) can be divided into two parts, the between-group sum of squares (SSB) and the within-group sum of squares (SSW):

$$SST=\sum_{m=1}^{M}(y_m-\overline{Y})^2=\sum_{i=0}^{2}\sum_{m=1}^{M_i}(y'_{im}-\overline{Y})^2,$$

$$SSB=\sum_{i=0}^{2}(\overline{Y_i'}-\overline{Y})^2, \quad and$$

$$SSW=SST-SSB=\sum_{i=0}^{2}\sum_{m=1}^{M_i}(y'_{im}-\bar{Y}_i')^2,$$

where

$$\overline{Y}=\frac{1}{M}\sum_{m=1}^{M}y_m \quad and \quad \bar{Y}_i'=\frac{1}{M_i}\sum_{m=1}^{M_i}y'_{im}\,.$$

The formula of F-test statistic is $F=\frac{SSB}{SSW}$, and F follows the F-distribution with 2 and S-3 degrees of freedom under the null hypothesis, i.e., $F\sim F_{(2,S-3)}$.

2.3.2 Linear regression. In the linear regression model, a least-squares regression line is fit between the phenotype values and the genotype values [11]. For simplicity, we denote the genotypes of a single SNP to be $x_1,x_2,\cdots,x_M$. Based on the data $(x_1,y_1),\cdots,(x_M,y_M)$, we need to fit a line in the form of $Y=a+bx$.

We have the sums of squares as follows:

$$SS_{xx}=\sum_{i=1}^{M}(x_i-\overline{x})^2, SS_{yy}=\sum_{i=1}^{M}(Y_i-\overline{Y})^2,$$

$$and \quad SS_{xy}=\sum_{i=1}^{M}(x_i-\overline{x})(Y_i-\overline{Y})$$

where $\overline{x}=\frac{1}{M}\sum_{i=1}^{M}x_i$ *and* $\overline{Y}=\frac{1}{M}\sum_{i=1}^{M}y_i$

To achieve least squares, the estimator of b is $\frac{SS_{xy}}{SS_{xx}}$. To evaluate the significance of the obtained model, a hypothesis testing for $b=0$ is then applied.

2.4 Multiple Testing Problem

In a typical GWAS, the test needs to be performed many times. We should pay attention to a statistical issue known as the multiple testing problem. In the remainder of this section, we will discuss the multiple testing problem and how to effectively control error rate in GWAS.

Type 1 error rate, is the possibility that a null hypothesis is rejected when it is actually true. In other words, it is the chance of observing a positive (significant) result even if it is not. If a test is performed multiple times, the overall Type 1 Error rate will increase. This is called the multiple testing problem.

Let α be the type 1 error rate for a statistical test. If the test is performed n times, the experimental-wise error rate α' is given by

$$\alpha'=1-(1-\alpha)^n.$$

For example, if $\alpha=0.05$ and $n=20$, then $\alpha'=1-(1-0.05)^{20}=0.64$. In this case, the chance of getting at least one false positive is 64%.

Because of the multiple testing problem, the test result may not be that significant even if its p-value is less than a significant level α. To solve this problem, the nominal p-value need to be corrected/adjusted.

2.5 Family-Wise Error Rate Control

For the single-locus test, we denote the p-value for a association test of a SNP X_i and a phenotype Y to be $p(X_i,Y)$, and the corrected p-value to be $p'(X_i,Y)$. Family-wise error rate (FWER), or the experiment-wise error rate, is the probability of at least one false association. We use α' to denote family-wise error rate, and it is given by

$$\alpha'=P(\text{reject } H_0|H_0)=P(\text{reject at least one of } H_i(1\leq i\leq n)|H_0),$$

where n is the total number of tests and H_0 is the hypothesis that all the $H_i(1\leq i\leq n)$ are true.

Many methods can be used to control FWER. Bonferroni correction is a commonly used method, in which p-values need to be enlarged to account for the number of comparisons being performed. Permutation test [13] is also widely used to correct for multiple testing in GWAS.

2.5.1 Bonferroni correction. In Bonferroni correction, the p-value of a test is multiplied by the number of tests in the multiple comparison.

$$p'(X_i,Y)=p(X_i,Y)*N$$

Here the number of tests is the number of SNPs N in a study. Bonferroni correction is a single-step procedure, in which each of the p-values is independently corrected.

2.5.2 Permutation tests. In the permutation test, data are reshuffled. For each permutation, p-values for all the tests are re-calculated, and the minimal p-value is retained. After K permutations, we get totally K minimal p-values. The corrected p-value is given by the proportion of minimal p-values which is less than the original p-value.

Let $\{Y_1,\cdots,Y_k\}$ be the set of K permutations. For each permutation $Y_k(1\leq k\leq K)$, the minimal p-value p_{Y_k} is given by

$$p_{Y_k}=min\{p(X_i,Y_k)|1\leq i\leq n\}.$$

Then we have the corrected p-value

$$p'(X_i,Y)=\frac{\#\{p_{Y_k}<p(X_i,Y)|1\leq k\leq K\}}{K}.$$

The permutation method takes advantage of the correlation structure between SNPs. It is less stringent than Bonferroni correction.

2.6 False Discovery Rate Control

False discovery rate (FDR) controls the expected proportion of type 1 error among all significant hypotheses. It is less conservative than the family-wise error rate. For example, if 100 observed results are claimed to be significant, and the FDR is 0.1, then 10 of results are expected to be false discoveries.

One way to control the FDR is as follows [14]. The p-values of SNPs and the phenotype are ranked from smallest to largest. We denote the ordered p-values to be $p_1,\cdots,p_N$. Starting from the largest p-value to the smallest, the original p-value is multiplied by the total number of SNPs and divided by its rank. For the i^{th} p-value p_i, its corrected p-value p_i' is given by

$$p_i'=p_i*(\frac{N}{i}).$$

In this section, we have discussed commonly used methods in single-locus study, the multiple testing problem and how to control error rate in GWAS. In the next section, we will introduce methods used for two-locus association studies. We will focus on one class work that finds exact solution when searching for SNP-SNP interactions in GWAS.

3. Exact Methods for Two-Locus Association Study

The vast number of SNPs has posed great computational challenge to genome-wide association study. In order to understand the underlying biological mechanisms of complex phenotype, one needs to consider the joint effect of multiple SNPs simultaneously. Although the idea of studying the association between phenotype and multiple SNPs is straightforward, the implementation is nontrivial. For a study with total N SNPs, in order to find the association between n SNPs and the phenotype, a brute-force approach is to exhaustively enumerate all $\binom{N}{n}$ possible SNP combinations and evaluate their associations with the phenotype. The computational burden imposed by this enormous search space often makes the complete genome-wide association study intractable. Moreover, although permutation test has been considered the gold standard method for multiple testing correction, it will dramatically increase the computational burden because the process needs to be performed for all permuted data.

In this section, we will focus on the recently developed exact method for two-locus epistasis detection. Different from the single-locus approach, the goal of two-locus epistasis detection is to identify interacting SNP-pairs that have strong association with the phenotype. FastANOVA [15] is an algorithm for two-locus ANOVA (analysis of variance) test on quantitative traits and FastChi [16] for two-locus chi-square test on case-control phenotypes. COE [17] is a general method that can be applied in a wide range of tests. TEAM [18] is designed for studies involving a large number of individuals such as human studies. In this subsection, we will discuss these algorithms, and their strengths and limitations.

3.1 The FastANOVA Algorithm

FastANOVA utilizes an upper bound of the two-locus ANOVA test to prune the search space. The upper bound is expressed as the sum of two terms. The first term is based on the single-SNP ANOVA test. The second term is based on the genotype of the SNP-pair and is independent of permutations. This property allows to index SNP-pairs in a 2D array based on the genotype relationship between SNPs. Since the number of entries in the 2D array is bound by the number of individuals in the study, many SNP-pairs share a common entry. Moreover, it can be shown that all SNP-pairs indexed by the same entry have exactly the same upper bound. Therefore, we can compute the upper bound for a group of SNP-pairs together. Another important property is that the indexing structure only needs to be built once and can be reused for all permutated data. Utilizing the upper bound and the indexing structure, FastANOVA only needs to perform the ANOVA test on a small number of candidate SNP-pairs without the risk of missing any significant pair. We discuss the algorithm in further detail in the following.

Let $\{X_1,X_2,\cdots,X_N\}$ be the set of SNPs of M individuals ($X_i\in\{0,1\},1\leq i\leq N$) and $Y=\{y_1,y_2,\cdots,y_M\}$ be the quantitative phenotype of interest, where y_m ($1\leq m\leq M$) is the phenotype value of individual m.

For any SNP X_i ($1\leq i\leq N$), we represent the F-statistic from the ANOVA test of X_i and Y as $F(X_i,Y)$. For any SNP-pair (X_iX_j), we represent the F-statistic from the ANOVA test of (X_iX_j) and Y as $F(X_iX_j,Y)$.

The basic idea of ANOVA test is to partition the total sum of squared deviations SS_T into between-group sum of squared deviations SS_B and within-group sum of squared deviations SS_W:

$$SS_T=SS_B+SS_W.$$

In our application of the two-locus association study, Table 2 and Table 3 show the possible groupings of phenotype values by the genotypes of X_i and (X_iX_j) respectively.

Let A, B, a_1, a_2, b_1, b_2 represent the groups as indicated in Table 2 and Table 3. We use $SS_B(X_i,Y)$ and $SS_B(X_iX_j,Y)$ to distinct the one locus (i.e., single-SNP) and two locus (i.e., SNP-pair) analyses. Specifically, we have

$$SS_T(X_i,Y)=SS_B(X_i,Y)+SS_W(X_i,Y),$$

$$SS_T(X_iX_j,Y)=SS_B(X_iX_j,Y)+SS_W(X_iX_j,Y).$$

Table 2. Grouping of Y by X_i.

$X_i=1$	$X_i=0$
group A	group B

doi:10.1371/journal.pcbi.1002828.t002

The F-statistics for ANOVA tests on X_i and (X_iX_j) are:

$$F(X_i,Y)=\frac{M-2}{2-1}\times\frac{SS_B(X_i,Y)}{SS_T(X_i,Y)-SS_B(X_i,Y)},\qquad(1.1)$$

$$F(X_iX_j,Y)=\frac{M-g}{g-1}\times\frac{SS_B(X_iX_j,Y)}{SS_T(X_iX_j,Y)-SS_B(X_iX_j,Y)},\qquad(1.2)$$

where g in Equation (1.2) is the number of groups that the genotype of (X_iX_j) partitions the individuals into. Possible values of g are 3 or 4, assuming all SNPs are distinct: If none of groups A, B, a_1, a_2, b_1, b_2 is empty, then $g=4$. If one of them is empty, then $g=3$.

Let $T=\sum_{y_m\in Y}y_m$ be the sum of all phenotype values. The total sum of squared deviations does not depend on the groupings of individuals:

$$SS_T(X_i,Y)=SS_T(X_iX_j,Y)=\sum_{y_m\in Y}y_m^2-\frac{T^2}{M}.$$

Let $T_{group}=\sum_{y_m\in group}y_m$ be the sum of phenotype values in a specific group, and n_{group} be the number of individuals in that group. $SS_B(X_i,Y)$ and $SS_B(X_iX_j,Y)$ can be calculated as follows:

$$SS_B(X_i,Y)=\frac{T_A^2}{n_A}+\frac{T_B^2}{n_B}-\frac{T^2}{M},$$

Table 3. Grouping of Y by X_iX_j.

	$X_i=1$	$X_i=0$
$X_j=1$	group a_1	group b_1
$X_j=0$	group a_2	group b_2

doi:10.1371/journal.pcbi.1002828.t003

$$SS_B(X_iX_j,Y)=\frac{T_{a_1}^2}{n_{a_1}}+\frac{T_{a_2}^2}{n_{a_2}}+\frac{T_{b_1}^2}{n_{b_1}}+\frac{T_{b_2}^2}{n_{b_2}}-\frac{T^2}{M}.$$

Note that for any group of A, B, a_1, a_2, b_1, b_2, if $n_{group}=0$, then $\frac{T_{group}^2}{n_{group}}$ is defined to be 0.

Let $\{y_m|y_m \in A\}=\{y_{A_1},y_{A_2},\cdots,y_{A_{n_A}}\}$ be the phenotype values in group A. Without loss of generality, assume that these phenotype values are arranged in ascending order, i.e.,

$$y_{A_1}\leq y_{A_2}\leq\cdots\leq y_{A_{n_A}}.$$

Let $\{y_m|y_m \in B\}=\{y_{B_1},y_{B_2},\cdots,y_{B_{n_B}}\}$ be the phenotype values in group B. Without loss of generality, assume that these phenotype values are arranged in ascending order, i.e.,

$$y_{B_1}\leq y_{B_2}\leq\cdots\leq y_{B_{n_B}}.$$

We have the overall upper bound on $SS_B(X_iX_j,Y)$:

Theorem 1 *(Upper bound of $SS_B(X_iX_j,Y)$)*

$$SS_B(X_iX_j,Y)\leq SS_B(X_i,Y)+R_1(X_iX_jY)+R_2(X_iX_jY).$$

The notations in the bound can be found in Table 4. The upper bound in Theorem 1 is tight. The tightness of the bound is obvious from the derivation of the upper bound, since there exists some genotype of SNP-pair (X_iX_j) that makes the equality hold.

We now discuss how to apply the upper bound in Theorem 1 in detail. The set of all SNP-pairs is partitioned into non-overlapping groups such that the upper bound can be readily applied to each group. For every X_i $(1\leq i\leq N)$, let $AP(X_i)$ be the set of SNP-pairs

$$AP(X_i)=\{(X_iX_j)|i+1\leq j\leq N\}.$$

For all SNP-pairs in $AP(X_i)$, n_A, T_A, n_B, T_B and $SS_B(X_i,Y)$ are constants. Moreover, l_{a_1}, u_{a_1} are determined by n_{a_1}, and l_{b_1}, u_{b_1} are determined by n_{b_1}. Therefore, in the upper bound, n_{a_1} and n_{b_1} are the only variables that depend on X_j and may vary for different SNP-pairs (X_iX_j) in $AP(X_i)$.

Table 4. Notations for the bounds.

Symbols	Formulas
l_{a_1}	$\sum_{i=1}^{n_{a_1}} y_{A_i}$
u_{a_1}	$\sum_{i=n_A-n_{a_1}+1}^{n_A} y_{A_i}$
$R_1(X_iX_jY)$	$\frac{\max\{(n_Al_{a_1}-n_{a_1}T_A)^2,(n_Au_{a_1}-n_{a_1}T_A)^2\}}{n_{a_1}(n_A-n_{a_1})n_A}$
l_{b_1}	$\sum_{i=1}^{n_{b_1}} y_{B_i}$
u_{b_1}	$\sum_{i=n_B-n_{b_1}+1}^{n_B} y_{B_i}$
$R_2(X_iX_jY)$	$\frac{\max\{(n_Bl_{b_1}-n_{b_1}T_B)^2,(n_Bu_{b_1}-n_{b_1}T_B)^2\}}{n_{b_1}(n_B-n_{b_1})n_B}$

doi:10.1371/journal.pcbi.1002828.t004

Note that n_{a_1} is the number of 1's in X_j when X_i takes value 1, and n_{b_1} is the number of 1's in X_j when X_i takes value 0. It is easy to prove that switching n_{a_1} and n_{a_2} does not change the F-statistic value and the correctness of the upper bound. This is also true if we switch n_{b_1} and n_{b_2}. Therefore, without loss of generality, we can always assume that n_{a_1} is the smaller one between the number of 1's and number of 0's in X_j when X_i takes value 1, and n_{b_1} is the smaller one between the number of 1's and number of 0's in X_j when X_i takes value 0.

If there are m 1's and $(M-m)$ 0's in X_i, then for any $(X_iX_j)\in AP(X_i)$, the possible values that n_{a_1} can take are $\{0,1,2,\cdots,\lfloor m/2\rfloor\}$. The possible values that n_{b_1} can take are $\{0,1,2,\cdots,\lfloor(M-m)/2\rfloor\}$.

To efficiently retrieve the candidates, the SNP-pairs (X_iX_j) in $AP(X_i)$ are grouped by their (n_{a_1},n_{b_1}) values and indexed in a 2D array, referred to as $Array(X_i)$.

Suppose that there are 32 individuals, and the genotype of X_i consists of half 0's and half 1's. Thus for the SNP-pairs in $AP(X_i)$, the possible values of n_{a_1} and n_{b_1} are $\{0,1,2,\cdots,8\}$. Figure 1 shows the 9×9 array, $Array(X_i)$, whose entries represent the possible values of (n_{a_1},n_{b_1}) for the SNP-pairs $(X_iX_j)\in AP(X_i)$. The entries in the same column have the same n_{a_1} value. The entries in the same row have the same n_{b_1} value. The n_{a_1} value of each column is noted beneath each column. The n_{b_1} value of each row is noted left to each row. Each entry of the array is a pointer to the SNP-pairs $(X_iX_j)\in AP(X_i)$ having the corresponding (n_{a_1},n_{b_1}) values.

For any SNP X_i, the maximum number of the entries in $Array(X_i)$ is $(\lceil\frac{M}{4}\rceil+1)^2$. The proof of this property is straightforward and omitted here. In order to find candidate SNP-pairs, we scan all entries in $Array(X_i)$ to calculate their upper bounds. Since the SNP-pairs indexed by the same entry share the same (n_{a_1},n_{b_1}) value, they have the same upper bound. In this way, we can calculate the upper bound for a group of SNP-pairs together. Note that for typical genome-wide association studies, the number of individuals M is much smaller than the number of SNPs N. Therefore, the additional cost for accessing $Array(X_i)$ is minimal compared to performing ANOVA tests for all pairs $(X_iX_j)\in AP(X_i)$.

For multiple tests, permutation procedure is often used in genetic analysis for controlling family-wise error rate. For genome-wide association study, permutation is less commonly used because it often entails prohibitively long computation times. Our FastANOVA algorithm makes permutation procedure feasible in genome-wide association study.

Let $Y'=\{Y_1,Y_2,\cdots,Y_K\}$ be the K permutations of the phenotype Y. Following the idea discussed above, the upper bound in Theorem 1 can be easily incorporated in the algorithm to handle the permutations. For every SNP X_i, the indexing structure $Array(X_i)$ is independent of the permuted phenotypes in Y'. The correctness of this property relies on the fact that, for any $(X_iX_j)\in AP(X_i)$, n_{a_1} and n_{b_1} only depend on the genotype of the SNP-pair and thus remain constant for different phenotype permutations. Therefore, for each X_i, once we build $Array(X_i)$, it can be reused in all permutations.

3.2 The FastChi Algorithm

As our initial attempt to develop scalable algorithms for genome-wide association study, FastANOVA is specifically designed for the ANOVA test on quantitative phenotypes. Another category of phenotypes is generated in case-control study, where the phenotypes are binary variables representing disease/non-disease individuals. Chi-square test is one of the most commonly used statistics in binary phenotype association

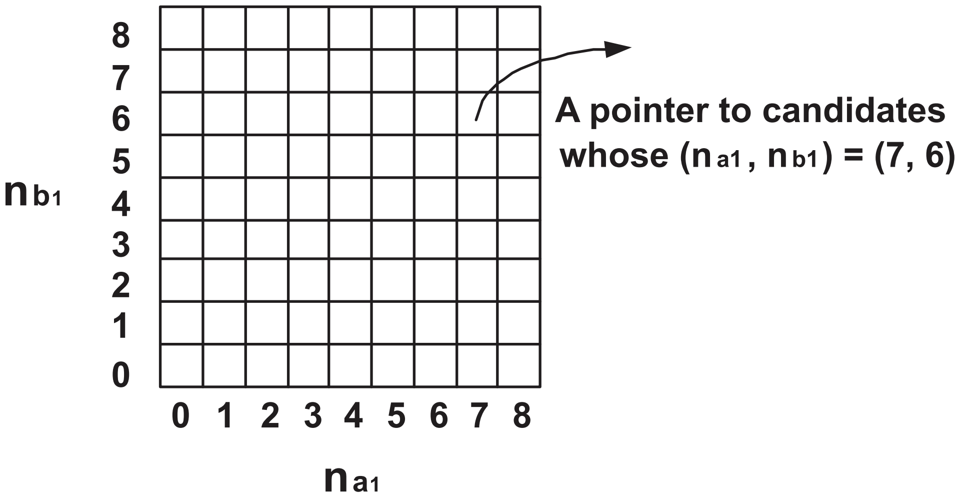

Figure 1. The index array $Array(X_i)$ **for efficient retrieval of the candidate SNP-pairs.**
doi:10.1371/journal.pcbi.1002828.g001

study. We can extend the principles in FastANOVA for efficient two-locus chi-square test. The general idea of FastChi is similar to that of FastANOVA, i.e., re-formulating the chi-square test statistic to establish an upper bound of two-locus chi-square test, and indexing the SNP-pairs according to their genotypes in order to effectively prune the search space and reuse redundant computations. Here we briefly introduce the FastChi algorithm.

For SNP X_i, we represent the chi-square test value of X_i and the binary phenotype Y as $\chi^2(X_i,Y)$. For any SNP-pair X_i and X_j, we use $\chi^2(X_iX_j,Y)$ to represent the chi-square test value for the combined effect of (X_iX_j) with Y. Let A,B,C,D represent the following events respectively: $Y=0 \wedge X_i=0$; $Y=0 \wedge X_i=1$; $Y=1 \wedge X_i=0$; $Y=1 \wedge X_i=1$. Let O_{event} denote the observed value of an event. T_1, T_2, S_1, S_2, $\mathcal{R}_1$, and $\mathcal{R}_2$ represent the formulas shown in Table 5. We have the upper bound of $\chi^2(X_iX_j,Y)$ stated in Theorem 2.

Theorem 2 *(Upper bound of* $\chi^2(X_iX_j, Y)$*)*

$$\chi^2(X_iX_j,Y) \leq \chi^2(X_i,Y) + T_1S_1\mathcal{R}_1 + T_2S_2\mathcal{R}_2.$$

For given phenotype Y and SNPX_i, $\chi^2(X_i,Y)$, T_1, S_1, T_2, and S_2 are constants. $\mathcal{R}_1$ and $\mathcal{R}_2$ are the only variables that depend on X_j and may vary for different SNP-pairs $(X_iX_j) \in AP(X_i)$. (Recall that $AP(X_i)=\{(X_iX_j)|i+1 \leq j \leq N\}$.) Thus for a given X_i, we can treat equation $\chi^2(X_i,Y)+T_1S_1\mathcal{R}_1+T_2S_2\mathcal{R}_2=\theta$ as a *straight line* in the 2-D space of $\mathcal{R}_1$ and $\mathcal{R}_2$. The ones whose $(\mathcal{R}_1(X_iX_j),\mathcal{R}_2(X_iX_j))$ values fall below the line can be pruned without any further test.

Suppose that there are 32 individuals, X_i contains half 0's, and half 1's. For the SNP-pairs in $AP(X_i)$, the possible values of $\mathcal{R}_1$ (and $\mathcal{R}_2$) are $\{\frac{0}{16},\frac{1}{15},\frac{2}{14},\frac{3}{13},\frac{4}{12},\frac{5}{11},\frac{6}{10},\frac{7}{9},\frac{8}{8}\}$. Figure 2 shows the 2-D space of $\mathcal{R}_1$ and $\mathcal{R}_2$. The blue stars represent the values that $(\mathcal{R}_1,\mathcal{R}_2)$ can take. The line $\chi^2(X_i,Y)+T_1S_1\mathcal{R}_1+T_2S_2\mathcal{R}_2=\theta$ is plotted in the figure. Only the SNP-pairs whose $(\mathcal{R}_1,\mathcal{R}_2)$ values are in the shaded region are subject to two-locus Chi-square test.

Similar to FastANOVA, in FastChi, we can index the SNP-pairs in $AP(X_i)$ according to their genotype relationships, i.e., by the values of $(\mathcal{R}_1,\mathcal{R}_2)$. Experimental results demonstrate that FastChi is an order of magnitude faster than the brute force alternative.

3.3 The COE Algorithm

Both FastANOVA and FastChi rework the formula of ANOVA test and Chi-square test to estimate an upper bound of the test value for SNP pairs. These upper bounds are used to identify candidate SNP pairs that may have strong epistatic effect. Repetitive computation in a permutation test is also identified and performed once those results are stored for use by all permutations. These two strategies lead to substantial speedup, especially for large permutation test, without compromising the accuracy of the test. These approaches guarantee to find the optimal solutions. However, a common drawback of these methods is that they are designed for specific tests, i.e., chi-square test and ANOVA test. The upper bounds used in these methods do not work for other statistical tests, which are

Table 5. Notations used in the derivation of the upper bound for two-locus Chi-square test.

Symbols	Formulas
T_1	$\frac{M^2}{(O_A+O_B)(O_A+O_C)(O_C+O_D)}$
S_1	$\max\{O_A^2,O_C^2\}$
$\mathcal{R}_1$	$\min\{\left[\frac{O_{X_j=1}}{O_{X_j=0}}\|X_i=0\right],\left[\frac{O_{X_j=0}}{O_{X_j=1}}\|X_i=0\right]\}$
T_2	$\frac{M^2}{(O_A+O_B)(O_B+O_D)(O_C+O_D)}$
S_2	$\max\{O_B^2,O_D^2\}$
$\mathcal{R}_2$	$\min\{\left[\frac{O_{X_j=1}}{O_{X_j=0}}\|X_i=1\right],\left[\frac{O_{X_j=0}}{O_{X_j=1}}\|X_i=1\right]\}$

doi:10.1371/journal.pcbi.1002828.t005

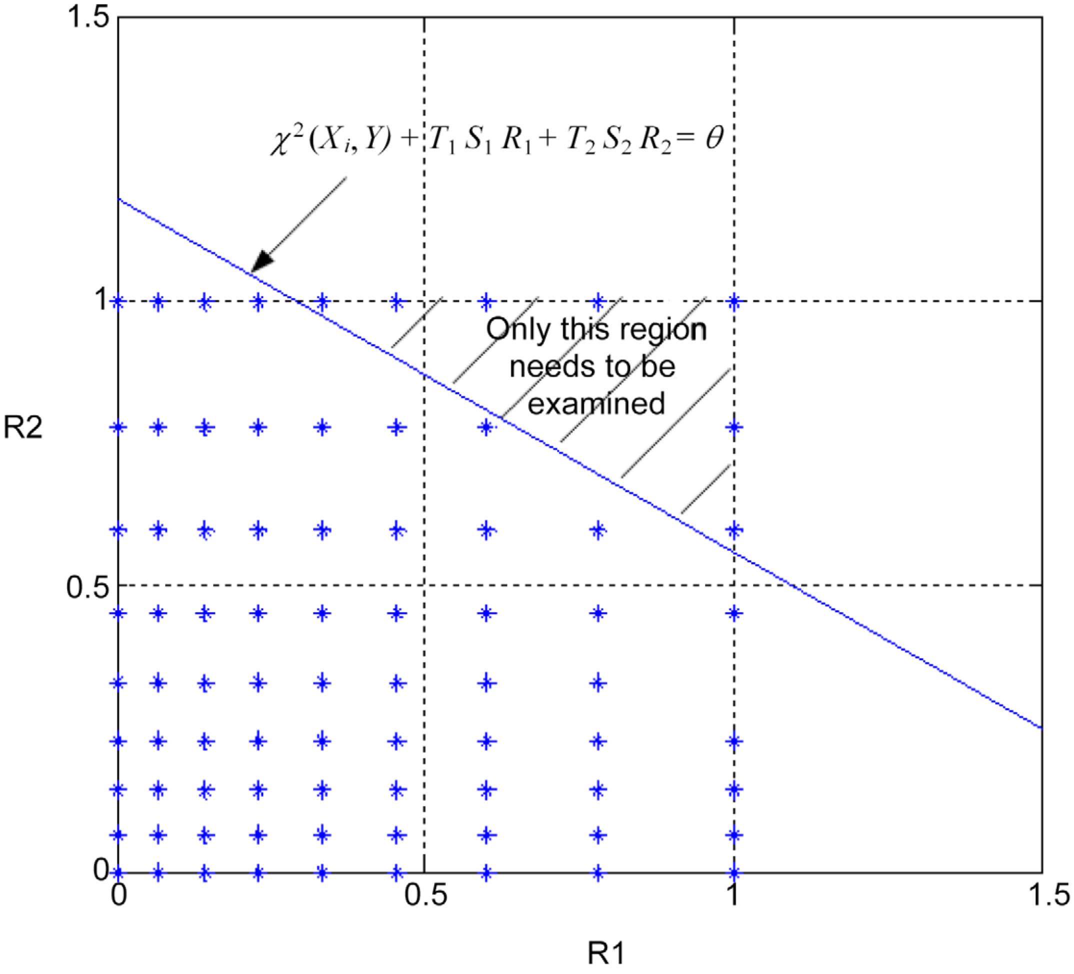

Figure 2. Pruning SNP-pairs in $AP(X_i)$ using the upper bound.
doi:10.1371/journal.pcbi.1002828.g002

also routinely used by researchers. In addition, new statistics for epistasis detection are continually emerging in the literature. Therefore, it is desirable to develop a general model that supports a variety of statistical tests.

The COE algorithm takes the advantage of convex optimization. It can be shown that a wide range of statistical tests, such as chi-square test, likelihood ratio test (also known as G-test), and entropy-based tests are all convex functions of observed frequencies in contingency tables. Since the maximum value of a convex function is attained at the vertices of its convex domain, by constraining on the observed frequencies in the contingency tables, we can determine the domain of the convex function and get its maximum value. This maximum value is used as the upper bound on the test statistics to filter out insignificant SNP-pairs. COE is applicable to all tests that are convex.

3.4 The TEAM Algorithm

The methods we have discussed so far provide promising alternatives for GWAS. However, there are two major drawbacks that limit their applicability. First, they are designed for relatively small sample size and only consider homozygous markers (i.e., each SNP can be represented as a $\{0,1\}$ binary variable). In human study, the sample size is usually large and most SNPs contain heterozygous genotypes and are coded using $\{0,1,2\}$. These make previous methods intractable. Second, although the family-wise error rate (FWER) and the false discovery rate (FDR) are both widely used for error controlling, previous methods are designed only to control the FWER. From a computational point of view, the difference in the FWER and the FDR controlling is that, to estimate FWER, for each permutation, only the maximum two-locus test value is needed. To estimate the FDR, on the other hand, for each permutation, all two-locus test values must be computed.

To address these limitations, TEAM is proposed for efficient epistasis detection in human GWAS. TEAM has several advantages over previous methods. It supports to both homozygous and heterozygous data. By exhaustively computing all two-locus test values in permutation test, it enables both FWER and FDR controlling. It is applicable to all statistics based on the contingency table. Previous methods are either designed for specific tests or require the test statistics satisfy certain property. Experimental results demonstrate that TEAM is more efficient than existing methods for large sample studies.

TEAM incorporates the permutation test for proper error controlling. The key idea is to incrementally update the contingency tables of two-locus tests. We show that only four of the eighteen observed frequencies in the contingency table need to be updated to compute the test value. In the algorithm, we build a minimum spanning tree [19] on the SNPs. The nodes of the tree are SNPs. Each edge represents the genotype difference between the two connected SNPs. This tree structure can be utilized to speed up the updating process for the contingency tables. A majority of the individuals are pruned and only a small portion are scanned to update the contingency tables. This is advantageous in human study, which usually involves thousands of individuals. Extensive experimental results demonstrate the efficiency of the TEAM algorithm.

As a summary of the exact two-locus algorithms, FastANOVA and FastChi are designed for specific tests and binary genotype data. The COE algorithm is a more general method that can be applied to all convex tests. The TEAM algorithm is more suitable for large sample human GWAS.

4. Multifactor Dimensionality Reduction

Multifactor dimensionality reduction (MDR) [20] is a data mining method to identify interactions among discrete variables for binary outcomes. It can be used to detect high-order gene-gene and gene-environment interactions in case-control studies. By pooling multi-locus SNPs into two groups, one classified as high-risk and the other classified as low risk, MDR effectively reduces the predictors from n dimensions to one dimension. Then, the one-dimensional variable is evaluated through cross-validation. The steps are repeated for all other n factor combinations, and the factor model which has the lowest prediction error is chosen as the 'best' n factor model. Its detailed steps are as follows:

- Divide the set of factors into 10 equal subsets.
- Select a set of n factors from the pool of all factors in the training set
- Create a contingency table for these n factors by counting the number of cases and controls in each combination.
- Compute the case-control ratio in each combination. Label them as “high-risk if it is greater than a certain threshold, and otherwise, it is marked as “low-risk”.
- Use the labels to classify individuals. Compute the misclassification rate.
- Repeat previous steps for all combinations of n factors across 10 training and testing subsets.
- Choose the model whose average misclassification rate is minimized and cross-validation consistency is maximized as the “best” model.

MDR designs a constructive induction method that combines two or more SNPs before testing for association. The power of the MDR approach is that it can be combined with other methodologies including the ones described in this chapter.

5. Logistic Regression

Logistic regression is a statistical method for predicting binary and categorical outcome. It is widely used in GWAS [21,22]. The basic idea is to use linear regression to model the probability of the occurrence of a specific outcome. Logistic regression is applicable to both single-locus and multi-locus association studies and can incorporate covariates and other factors in the model.

Let $Y \in \{0,1\}$ be a binary variable representing disease status (diseased verses non diseased), and $X \in \{0,1,2\}$ be a SNP. The conditional probability of having the disease given a SNP is $\theta(X) = P(Y=1|X)$. We define the logit function to convert the range of the probability from $[0,1]$ to $(-\infty,+\infty)$:

$$logit(X) = ln\frac{\theta(X)}{1-\theta(X)}.$$

The logit can be considered as a latent continuous variable that will be fit to a linear predictor function:

$$logit(X) \sim \beta_0 + \beta * X.$$

To cope with multiple SNP loci and potential covariates, we can modify the above model. For example, in the following model the logit is fit with predictors of SNPs (X_1, X_2) and covariates (Z_1, Z_2):

$$logit(X) \sim \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * \\ X_1 * X_2 + \beta_4 * Z_1 + \beta_5 * Z_2.$$

Although logistic regression can handle complicated models, it may be computationally demanding when the number of predictors is large [23].

6. Summary

The potential of genome-wide association study for the identification of genetic variants that underlying phenotypic variations is well recognized. The availability of large SNP data generated by high-throughput genotyping methods poses great computational and statistical challenges. In this chapter, we have discussed serval computational approaches to detect associations between genetic markers and the phenotypes. For further readings, the readers are encouraged to refer to [11,7,24,25] for discussions about current progress and challenges in large-scale genetic association studies.

7. Exercises

Question 1: The table below contains binary genotype and case-control phenotype data from ten individuals. Give the contingency table and use χ^2 test to compute the association test score.

Genotype	Phenotype
0	1
0	0
1	1
0	0
1	0
0	0
1	1
0	0
1	1
0	0

Question 2: Assuming that we have the following SNP and phenotype data, is the SNP significantly associated with the phenotype? Here, we represent each SNP site as the number of minor alleles on that locus, so 0 and 2 are for major and minor homozygous sites, respectively, and 1 is for the heterozygous sites. We also assume that minor alleles contribute to the phenotype and the effect is additive. In other words, the effect from a minor homozygous site should be twice as large as that from a heterozygous site. You may use any test methods introduced in the chapter. How about permutation tests?

Genotype	Phenotype
1	0.53
2	0.78
2	0.81
1	−0.23
1	−0.73
0	0.81
2	0.27
0	2.59
1	1.84
0	0.03

Question 3: Categorize the following methods in the table. The methods are χ^2 test, G-test, ANOVA, Student's T-test, Pearson's correlation, linear regression, logistic regression.

case – control phenotype	quantitative phenotype

Question 4: Why is it important to study multiple-locus association? What are the challenges?

Answers to the Exercises can be found in Text S1.

Further Reading

- Cantor RM, Lange K, Sinsheimer JS (2008) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Nat Rev Genet 9(11): 855–867.
- Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10(6): 392–404.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461(7265): 747–753.
- Moore JH, Williams SM (2009) Epistasis and its implications for personal genetics. Am J Hum Genet 85(3): 309–320.
- Phillips PC (2010) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Am J Hum Genet 86(1): 6–22.
- Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11: 843–854.

Supporting Information

Text S1 Answers to Exercises
(PDF)

References

1. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, et al. (2004) The collaborative cross, a community resource for the genetic analysis of complex traits. Nat Genet 36: 1133–1137.
2. The International HapMap Consortium (2003) The international hapmap project. Nature 426(6968): 789–796.
3. Saxena R, Voight B, Lyssenko V, Burtt N, de Bakker P, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331–1336.
4. Scuteri A, Sanna S, Chen W, Uda M, Albai G, et al. (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 3(7): e115. doi:10.1371/journal.pgen.0030115
5. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
6. Weedon M, Lettre G, Freathy R, Lindgren C, Voight B, et al. (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Genet 39: 1245–1250.
7. Hirschhorn J, Daly M (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95–108.
8. McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5): 356–369.
9. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The international hapmap project web site. Genome Res 15: 1592.
10. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
11. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7(10): 781–791.
12. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, et al. (2007) Genomewide association analysis of coronary artery disease. N Engl J Med 357: 443–453.
13. Westfall PH, Young SS (1993) Resampling-based multiple testing. Wiley: New York.
14. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57(1): 289–300.
15. Zhang X, Zou F, Wang W (2008) FastANOVA: an efficient algorithm for genome-wide association study. KDD 2008: 821–829.
16. Zhang X, Zou F, Wang W (2009) FastChi: an effcient algorithm for analyzing gene-gene interactions. PSB 2009: 528–539.
17. Zhang X, Pan F, Xie Y, Zou F, Wang W (2010) COE: a general approach for efficient genome-wide two-locus epistatic test in disease association study. J Comput Biol 17(3): 401–415.
18. Zhang X, Huang S, Zou F, Wang W (2010) TEAM: Efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics 26(12): 217–227.
19. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms. MIT Press and McGraw-Hill.
20. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138–147.
21. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11: 2463–2468.
22. Wason J, Dudbridge F (2010) Comparison of multimarker logistic regression models, with application to a genomewide scan of schizophrenia. BMC Genet 11: 80.
23. Yang C, Wan X, Yang Q, Xue H, Tang N, et al. (2011) A hidden two- locus disease association pattern in genome-wide association studies. BMC Bioinformatics 12: 156.
24. Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 4: 701–709.
25. Musani S, Shriner D, Liu N, Feng R, Coffey C, et al. (2007) Detection of gene×gene interactions in genome-wide association studies of human population data. Hum Hered 63(2): 67–84.