A Tournament Approach to the Detection of Multiple Associations in Genome-wide Studies with Pedigree Data

Zehua Chen[1], Jiahua Chen[2] and Jianjun Liu[3]

[1]*National University of Singapore*

[2]*University of Waterloo*

[3]*Genome Institute of Singapore*

Correspondence Author:

Zehua Chen

Department of Statistics & Applied Probability

National University of Singapore

3 Science Drive 2

Singapore 117543

Email: stachenz@nus.edu.sg

Abstract

Genome-wide association studies become feasible and promising with the availability of densely spaced markers over the whole genome. The data from genome-wide association studies typically consists of information on a huge number of markers with a relatively small sample size. The currently existing methods, which only account for marginal gene effects, are either ineffective or unmanageable in computation when extended to assess joint gene effects. There is an urgent need for efficient statistical methods that can assess joint gene effects and, at the same time, are computationally feasible. In this paper, we develop a tournament approach of this nature. The approach combines four ingredients together: a) variance component model, b) non-quadratic penalized likelihood, c) permutation aggregating, and d) model selection criteria. With these ingredients properly combined, the approach assesses the joint gene effects in stages which mimic rounds of competitions in a tournament, hence the name of the approach. The tournament approach is applied to a real data set containing quantitative trait values and genotypes of 2155 SNPs of 16 pedigrees with 233 individuals. It is demonstrated by simulation studies that the tournament approach is powerful in detecting multiple associations and at the same time incurs low false discovery rate. The tournament approach will provide a powerful tool for genome-wide association studies, especially, when there are a huge number of markers.

Keywords: Genome-wide association study, Penalized likelihood, Permutation aggregating, QTL mapping, Variance component model.

1 Introduction

With the advance of biotechnology, rapid collection of huge amount of molecular biological data is becoming the norm. Genetic markers, especially single nucleotide polymorphisms (SNPs), are becoming available in tens or hundreds of thousands. As the markers are sufficiently closely spaced to allow the detection of the linkage disequilibrium (LD) with any etiological variant, a genome-wide association study is becoming feasible and promising.

The analysis of the data resulted from genome-wide studies poses great challenges to statisticians and statistical geneticists. Since etiological variants for common diseases and complex traits are generally large in number and small in individual effects, they are overwhelmed in the sea of the huge amount of markers. Traditional statistical methods are inefficient in this situation. New statistical methods must be developed. The development of statistical methods for genome-wide association studies is still in its infancy.

The main approaches currently available for genome-wide association studies can roughly be classified into three categories. The first category consists of strategies of multiple tests based on single locus statistics or two loci statistics. They include methods using Bonferroni correction to control the family-wise type I error rate, see Marchini, Donnelly and Cardon (2005), and methods to control a more appropriate measure, the false discovery rate (FDR), see Benjamini and Hochberg (1995), and Storey and Tibshirani (2003). The second category consists of strategies that pool together the strength of single locus statistics to increase the power of detecting genes with significant contributions. These include the sum-statistic method developed by Hoh, Wille and Ott (2001), see also Hoh and Ott (2003), and the method using truncated product of p-values, see Zaykin et al. (2002) and Dudbridge and Koeleman

(2003). The third category consists of strategies, which have mainly been used for microarray analysis, that treat gene effects as random variables and use Bayesian approach or mixture model techniques, see Ishwaran and Rao (2003), Kauermann and Eilers (2004). There are also a few other approaches, say, the combinatorial-partitioning method (CPM) proposed by Nelson et al. (2001), the multifactor dimensionality reduction method modified from CPM by Ritchie et al. (2001) and Martin et al. (2006). All these methods just mentioned have met some success in certain particular situations and none of them dominates the others. The statistics used in the above mentioned approaches only summarize marginal gene effects such as single locus and two loci marginal effects. Consequently, they may fail to capture some precious joint effects of various genes in the data. A surge of new statistical methodologies for genome-wide association studies is yet to come.

We confine ourselves to association studies of quantitative traits with SNPs in this paper. The complexity of genome-wide association studies is tremendously increased. The fundamental issue, however, remains the same: among all the SNPs under investigation, which of them are responsible or are in LD with the QTL that are responsible for the observed variation in the quantitative traits of interest? For a given set of SNPs, the effects of the SNPs on the quantitative trait can be assessed jointly by an appropriate statistical model. The problem of genome-wide association study then amounts to the problem of model selection. However, at the scale of tens or hundreds of thousands SNPs, traditional model selection procedures can not be expected to work. The all-subset selection procedure is well known to be instable, see Breiman (1996a). The backward selection procedure is infeasible since the number of SNPs is much larger than the number of observations. The full model can achieve a perfect goodness-of-fit with any sufficiently large subset of SNPs no matter whether

or not they have any effect on the quantitative trait of concern. The consequence of this is that even the most important SNP may appear very ordinary at the presence of all the others and be removed from the model at very early stages. The forward or stepwise selection procedures are well known for their greedy nature; that is, the virtue of a variable (in the current context, a SNP) is assessed only against the variables already included in the model, not considered in its synergetic role among all the variables. As a consequence, relatively unimportant variables might be selected but more important variables might be missed. The weakness of the traditional model selection procedures become even more prominent at the scale of genome-wide studies. Besides, the computation task involved in those procedures is also prohibitive at such a scale.

In this paper, we propose a model selection approach for the genome-wide association study with pedigree data. Because of its similarity to the competitions in a tournament, we refer to this approach as the tournament approach. The basic idea of the tournament approach is as follows. The SNPs are assessed and selected jointly by penalized likelihood models in stages. The stages are similar to rounds of competitions. At each stage, the SNPs entered into this stage are divided into groups. The SNPs in each group are jointly assessed by a penalized likelihood model and a specified number of them are selected. These are similar to parallel matches in a round of competitions. The SNPs selected in the current stage then enter the next stage. At the last stage (the final), all the SNPs entered the final are jointly assessed and ranked by a penalized likelihood model. The SNPs are then grouped into nested subsets. For each of the subsets, an un-penalized likelihood model is fitted and a model selection criterion is applied to assess these models. There are four ingredients in the tournament approach: (a) a variance-component model to account for the co-

variance structures of the pedigrees, (b) a non-quadratic penalized likelihood used for selecting SNPs, (c) a permutation aggregating process adopted to ensure fair-play, and (d) a model selection criterion for the assessment of final models. Unlike the approaches based on marginal SNP effects, the tournament approach makes the selection of SNPs based on their joint effects. The information missing in those single locus and two loci statistics are recaptured in the tournament approach. In addition, the tournament approach is applicable for any number of SNPs, small or huge, with much less computational difficulties.

We have applied the tournament approach to a real data set consisting of 16 pedigrees with 233 individuals. The data set contains, for each individual, the value of the trait of interest and the genotypes of 2155 SNPs spread over 23 chromosomes. The heritability of the trait is estimated from the data as 0.48. The tournament approach detects four SNPs which are countable for the heritability. A numerical assessment shows that the detection is highly significant. We also used the pedigree structures and the SNP genotypes of the real data set as the setting for a simulation study. The simulation study demonstrates that the tournament approach is powerful to detect true QTL and at the same time incurs a low false discovery rate.

The details of the tournament approach are described and discussed in §2. The background of the real data set and its analysis are given in §3. The simulation study is discussed in §4. Some further discussion is presented in §5. Computational issues are addressed in the Appendix.

2 The tournament approach

The tournament approach is described in detail in this section. We first discuss the ingredients and then the procedure of the tournament approach.

The variance component model. To facilitate our discussion, we begin with some basics of quantitative genetics. Let Y be a quantitative trait whose variation can be attributed to genetic factors. Then we have $Y = \mu_0 + g + \epsilon$, where μ_0 is the overall mean, g is the genetic effect with variance σ_g^2 and ϵ is the non-genetic effect with variance σ_e^2. It is assumed that g and ϵ are independent. If there are more than one QTL contributing to Y, g is decomposed as $g = \sum_k g_k$, where g_k is the effect of QTL k with variance σ_k^2. Also the g_k's are commonly assumed independent. If there are dependent QTL, they can be pooled together and be considered as a compound QTL. Thus, $\sigma_g^2 = \sum_k \sigma_k^2$. Now consider two individuals j and l with quantitative trait values Y_j and Y_l given, respectively, by

$$Y_j = \mu_0 + \sum_k g_{jk} + \epsilon_j \quad \text{and} \quad Y_l = \mu_0 + \sum_k g_{lk} + \epsilon_l.$$

Then the covariance between Y_j and Y_l is given by

$$\mathrm{Cov}(Y_j, Y_l) = \sum_k \mathrm{Cov}(g_{jk}, g_{lk}).$$

It follows from the theory of population genetics, see Lange (2002), that

$$\mathrm{Cov}(g_{jk}, g_{lk}) = 2\sigma_k^2 \Phi_{jl},$$

where Φ_{jl} is the kinship coefficient between individual j and l which is a function of the biological relationship between j and l. The kinship coefficient is indeed the probability that an allele selected at random from j and an allel selected at random from the same autosomal locus of l are identical by descend. If j and l are biologically un-related, $\Phi_{jl} = 0$. We have,

$$\mathrm{Var}(Y_j) = \mathrm{Var}(Y_l) = \sigma_g^2 + \sigma_e^2,$$
$$\mathrm{Cov}(Y_j, Y_l) = 2\sigma_g^2 \Phi_{jl}.$$

The ratio $\sigma_g^2/(\sigma_g^2+\sigma_e^2)$ is referred to as the heritability in genetics. If the genotypes at QTL k are known for j and l, then conditioning on these genotypes, we have

$$\begin{aligned}
\mathrm{Var}(Y_j|g_{jk}) &= \mathrm{Var}(Y_l|g_{lk}) = (\sigma_g^2-\sigma_k^2)+\sigma_e^2,\\
\mathrm{Cov}(Y_j,Y_l|g_{jk},g_{lk}) &= 2(\sigma_g^2-\sigma_k^2)\Phi_{jl}.
\end{aligned}$$

If the genotypes at all QTL are known then the above variances and convariance reduce to

$$\begin{aligned}
\mathrm{Var}(Y_j|g_j) &= \mathrm{Var}(Y_l|g_l) = \sigma_e^2,\\
\mathrm{Cov}(Y_j,Y_l|g_j,g_l) &= 0.
\end{aligned}$$

Let $\boldsymbol{y}_i = (Y_{i1},\ldots,Y_{in_i})^t$ be the vector of trait values of n_i individuals in pedigree i. Then the variance-covariance matrix of $\boldsymbol{y}_i$ is given by

$$\Sigma_i = \sigma_g^2 A_i + \sigma_e^2 I,$$

where $A_i = 2(\Phi_{jl})$. Furthermore, if $\boldsymbol{y}_i$ is assumed to follow a multivariate normal distribution, the model is referred to as the variance component model which was first considered by Amos (1994). The variance component model is used in the tournament approach to model the correlations among the individuals from the same pedigree. In the variance component model considered by Amos, there are also additional variance components due to common household or other attributable common non-genetic factors. In this paper, we confine ourselves to the simpler form described above. However, if necessary, other components can be easily added into the model, which will not affect the procedure of the tournament approach.

The genotypes at an SNP on an autosome or on the sex chromosome for female take the form: A-A, A-B and B-B, and are coded as 0, 1, and 2 respectively. The genotypes at an SNP on the sex chromosome for male take the form A and B and are

coded as 0 and 1 respectively. Let x_k denote the genotype code of SNP k, the effect of this SNP, if any, can be represented by $\beta_k x_k$. Here, for the sake of convenience, we implicitly assumed that the dominant effect of an SNP is negligible.

The variance-component-model ingredient of the tournament approach is now described as follows. Suppose that the data set consists of n pedigrees and the trait values of pedigree i are represented by $\boldsymbol{y}_i = (Y_{i1}, \ldots, Y_{in_i})^t$. Let $\mathcal{S}$ be a set of SNPs. Let $X_{i\mathcal{S}}$ denote the matrix of genotypes of pedigree i with its rows corresponding to individuals and columns corresponding to the SNPs in $\mathcal{S}$, its first column being a vector of 1's. Then, conditioning on $\mathcal{S}$,

$$\boldsymbol{y}_i|\mathcal{S} \sim N(X_{i\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}, \Sigma_{i\mathcal{S}}),$$

where $\Sigma_{i\mathcal{S}} = \nu_{\mathcal{S}} A_i + \sigma_e^2 I$, $\nu_{\mathcal{S}}$ being the genetic variance conditioning on $\mathcal{S}$. It should be noted that, for the SNPs in $\mathcal{S}$ which are not QTL, the corresponding components in $\boldsymbol{\beta}_{\mathcal{S}}$ are zeros. If there is no QTL in $\mathcal{S}$ then $\nu_{\mathcal{S}} = \sigma_g^2$. The more QTL the $\mathcal{S}$ contains, the smaller the $\nu_{\mathcal{S}}$. Let

$$\Sigma_{\mathcal{S}} = \begin{pmatrix} \Sigma_{1\mathcal{S}} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \Sigma_{n\mathcal{S}} \end{pmatrix}, X_{\mathcal{S}} = \begin{pmatrix} X_{1\mathcal{S}} \\ \cdots \\ X_{n\mathcal{S}} \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \cdots \\ \boldsymbol{y}_n \end{pmatrix}.$$

Denote by $\boldsymbol{v}$ the vector $(\nu_{\mathcal{S}}, \sigma_e^2)^t$. Then the likelihood function of the data is given by

$$\begin{aligned} L(\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{v}|\mathcal{S}) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{n_i/2}|\Sigma_{i\mathcal{S}}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{y}_i - X_{i\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}})^t \Sigma_{i\mathcal{S}}^{-1}(\boldsymbol{y}_i - X_{i\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}})\} \\ &= \frac{1}{(2\pi)^{N/2}|\Sigma_{\mathcal{S}}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{y} - X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}})^t \Sigma_{\mathcal{S}}^{-1}(\boldsymbol{y} - X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}})\}, \qquad (1) \end{aligned}$$

where $N = \sum_{i=1}^{n} n_i$. This model will be referred to as the un-penalized likelihood model in the tournament approach.

Non-quadratic penalized likelihood. In recent years, a new class of model selection procedures using non-quadratic penalized likelihood has emerged. In the penalized

likelihood approach, a penalty function imposed on the linear parameters is added to the log likelihood function. The penalty function is deliberately constructed so that, by tuning a parameter in the penalty function, the zero components of the linear parameters should be estimated as zeros when the penalized log likelihood is maximized. Let $P_\lambda(|\boldsymbol{\beta}_\mathcal{S}|)$ be the penalty function where λ is the tuning parameter. The penalized likelihood is given by

$$l_p(\boldsymbol{\beta}_\mathcal{S}, \boldsymbol{v}|\mathcal{S}, \lambda) = \ln L(\boldsymbol{\beta}_\mathcal{S}, \boldsymbol{v}|\mathcal{S}) - NP_\lambda(|\boldsymbol{\beta}_\mathcal{S}|). \tag{2}$$

Usually, $P_\lambda(|\boldsymbol{\beta}_\mathcal{S}|)$ is of the form $\sum_{k\in\mathcal{S}} p_\lambda(|\beta_k|)$. If $p_\lambda(|\beta|) = \lambda|\beta|$, $P_\lambda(|\boldsymbol{\beta}_\mathcal{S}|)$ is the penalty function used in LASSO proposed by Tibshiranni (1996). Another popular choice is more conveniently specified by its derivative function

$$p'_\lambda(|\beta|) = \lambda\left\{I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\right\}$$

for some choice of $a > 2$. This penalty is proposed by Fan and Li (2001) and will be referred to as the SCAD penalty. It is seen that the SCAD penalty becomes a constant when β is larger than $a\lambda$. Thus, it does not discriminate between "large" fitted values of β. This property is useful to allow the fitted values of important SNPs not being influenced much when the penalty increases, and therefore more likely to be retained in the model. Because of this property, the SCAD penalty is used in the tournament approach. The penalized maximum likelihood estimates of the coefficients have positive probability to contain zero fitted components. When the penalty increases, the number of zero fitted coefficients increases. By tuning the parameter λ in the penalty function, a model containing practically any given number of covariates can be selected. The selection is based on the joint effect of all the SNPs instead of on the marginal effects of the SNPs. This feature is essential for the non-quadratic penalized likelihood to be an ingredient of the tournament approach.

The non-quadratic penalized likelihood approach has only been used so far for model selection in the case that the number of linear parameters is not much larger than and usually smaller than the number of observations. It is used in the following way. All the covariates are entered into the model, a cross-validation procedure is used to tune the parameter λ such that a subset of the covariates that best explain the data can be selected. The estimation of the parameters and the selection of covariates are done simultaneously in this procedure. In the genome-wide association studies, we are confronted with a completely different situation — the number of covariates is much larger than the number of observations. We use the penalized likelihood in a different way in the tournament approach. It is used to select a prespecified number of SNPs with non-zero fitted coefficients by tuning the parameter λ at each stage of the tournament procedure. The tuned value of λ determines in certain sense the qualification of SNPs for their further participation in the tournament.

Permutation aggregating. In principle, the penalized likelihood approach can be applied for any number of covariates even when the number of covariates far exceeds the number of observations. However, when the number of covariates becomes too large, there will be some practical obstacles. For example, the numerical computation in this case involves a very large Hessian matrix, and it is severely ill conditioned. The computation becomes possible only for relatively large values of the tuning parameter λ. But then the fitted values of the linear parameters will be overly influenced by the penalty, not much by their own virtue. Some covariates might hence be removed from the model prematurely. In the end, the truly important covariates can be easily missed. The huge size of Hessian matrix also inflates the amount of computation, and destabilize the numerical precision.

To overcome the difficulties discussed above, we adopt the following strategy. At

the begining of the tournament, the SNPs are divided into nearly equally numbered subsets. Then the penalized likelihood maximization is applied to each of the subsets to select a pre-specified number of SNPs with non-zero fitted parameters. If the total number of SNPs pooled together from all the subsets is still large, then they are subject to further division and selection, until a group of finalists close to a targeted number is reached.

However, this one-time partition might be erratic; that is, the finalist group is dependent of the particular initial partition. To reduce this erratic nature, a permutation aggregating procedure is applied for the selection of the finalists. The permutation aggregating procedure is as follows. Instead of just one partition, the SNPs are randomly permuted and partitioned for a given number of times. For each partition, the procedure described in the last paragraph is repeated. In the end, the SNPs that appear in the finalist groups most often are finally selected to enter the final.

The idea of permutation aggregating is the same as the bootstrap aggregating (Bagging) proposed by Breiman (1996b). The only difference is that bootstrapping in bagging is here replaced by random permutation.

Model selection criterion. The last ingredient of the tournament approach is the model selection criterion. At the final stage, a sequence of values of λ are tuned such that, for a pre-specified number m, with the first λ value, m SNPs with non-zero fitted coefficients are selected, with the second λ value, $m-1$ SNPs with non-zero fitted coefficients are selected, and so on, until, with the last λ value, only one SNP with non-zero fitted coefficient is selected. The SNPs selected by a particular λ value form a subset. An un-penalized likelihood model of form (1) is fitted with each subset. A model-selection-criterion ingredient is used to assess these models. In principle,

any model selection criteria such as AIC, BIC, CV or GCV can be used, see Akaike (1973), Schwarz (1978), and Craven and Wahba (1979). However, in the case that the number of covariates is huge, which is typical in genome-wide association study, both AIC and BIC are too liberal; that is, these criteria tend to select a model with a larger than necessary number of covariates. This phenomenon has been observed by Broman and Speed (2002). CV and GCV are theoretically close to AIC. Besides, they need more computaions. Here we introduce a **modified** information criterion (MIC) which is defined below:

$$MIC(k) = -2\sup\{L(\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{v}|\mathcal{S}) : \mathcal{S} \text{ contains } k \text{ SNPs}\} + k(\log N + \log P) \qquad (3)$$

where P is the total number of SNPs under consideration over the whole genome. A heuristic justification of MIC is as follows. Since $\sup\{L(\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{v}|\mathcal{S})\}$ is the maximum of $\binom{P}{k}$ χ^2-random variables, it follows that $\sup\{L(\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{v}|\mathcal{S})\} - \log\binom{P}{k}$ has essentially an asympotic extreme value distribution. The quantity $\log\binom{P}{k}$ is the amount in $\sup\{L(\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{v}|\mathcal{S})\}$ inflated by taking the maximum of the χ^2-random variables. Therefore, the superimum must be adjusted by subtracting $\log\binom{P}{k}$. It follows from Sterling's formula that $\log\binom{P}{k} \approx k\log(P - k) - \log(k!) \approx k\log P$ for large P. **When some SNPs are in linkage disequilibrium, as long as the equivalent number of independent SNP is still in the order of P, this size is still well motivated**. In the original BIC criterion, replacing the log likelihood by the adjusted superimum log likelihood then gives rise to the criterion MIC.

We now put all the ingredients together and describe the complete procedure of the tournament approach. Let M be a pre-determined subset size such that the maximization of the penalized likelihood is stable. From our experience, it seems appropriate to take M to be half of the number of observations. Let m be a pre-determined number for the SNPs to be selected from each subset. The m is determined

large enough to retain all important SNPs and small enough to reduce the burden of further computation. We take m to be two tenth of M. The procedure of the tournament approach goes as follows.

Pre-matches

Round 1: Let $\mathcal{S}^1$ be the set of all SNPs in the data set. Partition $\mathcal{S}^1$ at random into subsets of nearly equal size M to yield

$$\mathcal{S}^1 = \mathcal{S}_l^1 \cup \cdots \cup \mathcal{S}_J^1,$$

where J is the integer such that $[JM]$ is the total number of SNPs. For each subset $\mathcal{S}_j^1$, maximize $l_p(\boldsymbol{\beta}_{\mathcal{S}_j^1}, \boldsymbol{v}|\mathcal{S}_j^1, \lambda)$ by tuning λ so that m SNPs in $\mathcal{S}_j^1$ have non-zero fitted coefficients. Let $\mathcal{S}_{j*}^1$ denote the set of these m SNPs. Form the set

$$\mathcal{S}^2 = \mathcal{S}_{1*}^1 \cup \cdots \cup \mathcal{S}_{J*}^1.$$

Round r: Repeat the same procedure as in round 1 with the set $\mathcal{S}^r$ generated from the previous round when necessary.

Pre semi-final: The pre semi-final begins when there are nearly only M SNPs left from the previous rounds of competitions. From these M SNPs, a group of C finalists (C is pre-specified and is not necessarily the same as m) are selected through the maximization of the penalized likelihood by tuning the parameter λ.

Permutation aggregating

Repeat the pre-matches B times, say $B = 100$. From all the SNPs that entered the finalist groups at least once, select the C SNPs that entered the finalist groups most often. These C SNPs form the eventual group of finalists.

The Final

Ranking and grouping: The C finalists are ranked in the following way. First, in the maximization of the penalized likelihood including all the C finalists, tune the parameter λ such that λ_1 is the smallest value that produces less than C SNPs that have non-zero fitted coefficients. There could be less than $C-1$ SNPs with non-zero fitted coefficients. The SNPs with non-zero fitted coefficients are grouped together and the group is referred to as the λ_1-level group. The SNPs with zero fitted coefficients receive rank C and are excluded from further competitions. Next, consider the maximization of the penalized likelihood with the λ_1-level group, increase the value of the tuning parameter until the first time the number of SNPs with non-zero fitted coefficients differs from the λ_1-level group. Suppose the value of the tuning parameter achieving this state is λ_2. The SNPs with non-zero fitted coefficients at this step are grouped together and the group is referred to as the λ_2-level group. The SNPs with zero fitted coefficients at this step receive rank $C-1$. The process continues this way until eventually the value of the tuning parameter becomes so large that none of the SNPs remained will have a non-zero fitted coefficient.

Model fitting and Selection: For each of the groups at different λ-levels, fit the un-penalized likelihood model (1) and compute the model selection criterion MIC. The group which achives the smallest MIC value is selected as the significant group of SNPs.

Apart from the significant group, the output of the above procedure also contains the ranks of the finalists and the MIC value of each λ-level group. From these results,

one can judge the relative importance of the SNPs selected and draw guidelines for further investigation or confirmation study.

The computational issue of the tournament approach is dealt with in the Appendix.

3 Analysis of the Real Data

We have applied the tournament approach to a real data set. The data consists of the trait values together with the genotypes at 2155 SNPs spread over 23 chromosomes of 16 pedigrees with a total of 233 individuals. The 16 pedigrees are a part of the reference pedigrees that were originally collected from Utah, USA by the Centre dEtude du Polymorphisme Humain (CEPH). B lymphocytes from the blood samples of these pedigrees were transformed into immortalized lymphoblastoid cell lines (LCLs) by Epstein-Barr Virus (EBV). The trait of interest is a measure of the mRNA expression level of the EBNA-3A gene in LCLs. EBNA-3A is one of the EBV genes that are expressed in the LCLs and are important for the transformation of B lymphocytes. The genotype data of the 2155 SNPs in the 233 individuals of the 16 pedigrees are extracted from a larger data set at the SNP Consortium that has been used for constructing a linkage map of the human genome (Matise et al. 2003).

In this section, we give a detailed analysis on the output of the tournament procedure with this data.

The data set contains a large number of missing values. At some SNPs, the genotypes of all the individuals in a family are missing. These SNPs are removed from our consideration. There are also SNPs at which the genotypes of a pedigree are only partially missing. For these SNPs, We imputed the missing values by a random sampling from the un-missing genotypes of the pedigree. This might have

caused some genetically incompatible genotypes in these pedigrees. However, these incompatibility does not severely affect the result of the statistical analysis. There are also SNPs at which all the individuals have the same genotype. Since these SNPs are non-informative, they are removed from the analysis. There are also a few missing trait values. For the individuals with missing trait values, their trait values are imputed by a random sampling from their pedigree members as well.

After this preliminary treatment, 741 SNPs are removed and 1414 SNPs are left for the analysis. Some visual inspection reveals that the removed SNPs are scattered sporadically. Thus, the information loss in terms of association study is not too serious. Table 1 gives both the original number of SNPs as well as the number of SNPs left on each chromosome. Also given in Table 1 are the ranges of indices of the SNPs in the original data set.

The tournament approach is applied to these 1414 SNPs. At the first round, the SNPs are randomly divided into 14 subsets of equal size 101, and 20 SNPs are selected from each subset. At the second round, the SNPs selected from the first round are randomly divided into two subsets of equal size and 30 SNPs are selected from each subset. Then 30 finalists are determined from these 60 SNPs. This process is repeated 100 times. The SNPs which entered the finalist lists most frequently are given in Table 2. The order of the SNPs reflects their frequencies in descending order. The cutoff point is determined as the frequency has a big drop after the 38th SNP.

In the data set, the locations of the SNPs on the chromosomes are known. In order to see whether or not the locations can supply more information for the detection of associations, we adopted another strategy in the tournament approach. Instead of partitioning the SNPs at random in the first round, each chromosome is taken as a natural subset. A set of around 15 SNPs are selected from each chromosome at the

Table 1: Distribution of the SNPs over the chromosomes

Chromosome	Index Range	Original Number	Number Left
1	1-188	188	119
2	945-1117	173	118
3	1314-1466	153	97
4	1467-1586	120	81
5	1587-1712	126	80
6	1713-1836	124	89
7	1837-1954	118	76
8	1955-2064	108	76
9	2063-2155	93	53
10	189-302	114	84
11	303-385	83	52
12	386-479	94	62
13	480-547	68	49
14	548-639	82	51
15	630-687	58	35
16	688-760	73	47
17	761-823	63	41
18	824-903	80	52
19	904-944	41	28
20	1118-1189	72	44
21	1190-1232	43	29
22	1233-1280	48	32
23	1281-1313	33	20

first round. At the second round, the SNPs selected at the first round are divided at random into two subsets of equal size. Then for each subset, 30 SNPs are selected. At the third round, the 60 SNPs from the second round are screened by the penalized likelihood and 28 finalists are selected. The indices of these 28 SNPs together with their chromosomes are given in Table 3.

Since the number of SNPs is not uniformly distributed over the 23 chromosomes. We were concerned that the above strategy may give SNPs in chromosomes with

Table 2: The SNPs selected from the aggregation of 100 random permutations (the SNPs with * are the frequent finalists in Table 3, and the SNPs with ** appear in Table 3 but are not the frequent finalists)

Order	1	2	3	4	5	6	7	8	9	10
SNP	1985*	1836*	847*	393	83*	1999*	462	2079*	1832	1762*
Order	11	12	13	14	15	16	17	18	19	20
SNP	819*	846	1373	426**	681*	1750*	1847	389	290**	1
Order	21	22	23	24	25	26	27	28	29	30
SNP	1040*	928	602*	278*	1231*	1787*	2055**	333**	868	2151
Order	31	32	33	34	35	36	37	38		
SNP	124	687	125**	957	1232	78	925	77*		

fewer competitors unwarranted advantage of being finally selected. For example, 15 SNPs are choosing out of 20 SNPs on chromosome 23, but out of 119 SNPs on chromosome 1. Is it possible for less important SNPs to be selected from chromosome 23 and for more important SNPs to be missed on chromosome 1? However, from the results given in Table 3, it does not seem to be the case. To further clear our doubt, we randomly chose only 50 SNPs from those chromosomes with more than 50 SNPs and went through the tournament procedure with the above strategy. This was repeated 100 times and resulted 100 finalist lists. The SNPs in Table 3 with the * sign are the ones which most frequently entered the 100 lists of finalists. Note that the count of these frequent finalists is 18 out of 28. This demonstrates that the imbalanced set size at the first stage does not seem to have much influence on the final results.

It is interesting to note that among the 18 frequent finalists in Table 3, 17 of them are among the first 30 frequent finalists in Table 2. A total of 21 finalists in Table 3 are among the first 30 frequent finalists in Table 2. This indicates that the random partitioning strategy in the tournament approach does not loss any information which

Table 3: The selected finalist SNPs with chromosomes treated as natural subsets

Chromosome	SNP indices
1	77*, 83*, 125, 185
2	1025, 1040*
4	1565
6	1750*, 1762*, 1769, 1787*, 1836*
7	1839
8	1985*, 1999*, 2055
9	2079*
10	278*, 290
11	333
12	389*, 426, 462*
14	602*
15	681*
17	819*
18	847*
21	1231*

could have been carried by the position structure of the SNPs.

The SNPs in Table 2 are brought into the final of the tournament procedure. For the purpose of comparison, after ranking and grouping, the MIC as well as AIC, BIC are computed for the ranked groups. These values for the first 8 highly ranked SNPs and their corresponding groups are given in Table 4. The values in each column correspond to the model containing the SNPs in the first row from the leftmost to the one in that column. For example, the first model contains SNP 1836 only, and the second model contains SNPs 1836 and 393, and so on.

As expected, both AIC and BIC have not reached their minimum yet when all the 8 SNPs are included in the model. On the other hand, the MIC has its minimum at the model containing the first 4 SNPs. Therefore, the first 4 SNPs are selected as singificant ones. These 4 SNPs were recommended for further boilogical confirmation

Table 4: The final result of the tournament procedure with the real data set

SNPs	None	1836	393	1231	1985	1999	1	847	681
AIC	233.8	216.9	207.7	194.9	181.3	176.2	174.3	169.8	157.6
BIC	233.8	220.3	214.6	205.3	195.1	193.5	195.0	183.9	185.2
MIC	233.8	227.6	229.1	227.0	224.1	229.8	238.5	244.7	243.2

study. In addition to these 4 SNPs, the other SNPs can also be taken for confirmation study if resources are available. A simulation study for the justification of the use of MIC will be provided in the sequel.

It is of interest to know how many SNPs will be selected by the tournament procedure if none of the SNPs has significant contribution to the variation of the quantitative trait. Here we prerent some results of a random permutation study regarding this question. The issue will be tackled more systematically in the next section. The quantitative trait values are randomly permuted 200 times, thus cutting off any possible association of the trait values with the SNP genotypes. For each permutation, a slightly simplified version of the tournament procedure is carried out. In the simplified version, the permutation aggregating step is skipped for the sake of less amount of computation. The frequencies of different number of SNPs being selected by the AIC, BIC and MIC criteria are given in Table 5.

It is seen from Table 5 that neither AIC nor BIC can be used as the model selection criterion in the tournament procedure. Although none of the SNPs affects the quantitative trait, the AIC and BIC still have a high probability to select a large number of SNPs. This is especially the case for AIC which selects more than 7 SNP in all the 200 replicates. However, on the contrast, the MIC is quite efficient in eliminating the non-significant SNPs. There are 82% of times the MIC selects none

Table 5: The frequencies of different number of SNPs being selected by tournament procedure with randomly permuted quantitative traits.

No.of SNPs	0	1	2	3	4	5	6	7	8	≥ 9
AIC	0	0	0	0	0	0	0	0	3	197
BIC	0	11	19	16	27	24	19	14	20	51
MIC	164	33	3	0	0	0	0	0	0	0

of the SNPs, 16.5% of times it selects only one SNP, 1.5% of times it selects two SNPs. The MIC does not select more than two SNPs. In other words, if there is no genetic effect at all, the probability that MIC will falsely select more than two SNPs is almost zero. Thus we can interpret that the selection of four SNPs by MIC in the real data example is extremly significant.

4 Simulation studies

In this section, we discuss some simulation studies which were designed to assess the performance of the tournament approach.

In the usual simulation study, we need to generate pedigree data. However, the generation of SNP genotypes over the whole genome for a given pedigree structure is difficult. To avoid this difficulty, we adopted the following strategy. We retain the same pedigree structure together with all the SNP genotypes of the real data set throughout the simulation studies. Then the QTL and trait values are generated as follows. In each simulation repetition, we randomly select a fixed number of SNPs out of the 1414 SNPs in the real data set and take them as if they are the true QTL to form

$$\boldsymbol{y}_{\mathcal{S}} = X_{\mathcal{S}}\boldsymbol{\beta} + \epsilon_{\mathcal{S}},$$

where $X_{\mathcal{S}}$ consists of the columns in the design matrix of the real data set correspond-

ing to the selected SNPs, the $\boldsymbol{\beta}$ is a vector of given coefficients, and $\epsilon_{\mathcal{S}}$ is generated as a vector of independent identically distributed random errors with a normal distribution of zero mean and a given variance. The randomly selected SNPs and hence the matrix $X_{\mathcal{S}}$ and the error vector $\epsilon_{\mathcal{S}}$ change from repetition to repetition. The QTL effects $\boldsymbol{\beta}$ remain the same throughout each simulation study. It is assumed here that the genetic variation is completely accounted for by the selected SNPs. Therefore, given $X_{\mathcal{S}}$, the residuals are independent.

We considered two settings of the simulation. In both settings, we fix the number of randomly selected SNPs at 10. In the first setting, the vector $\boldsymbol{\beta}$ is taken as

$$\boldsymbol{\beta}^t = (2, -.31, -.23, .42, -.32, -.33, -.26, .41, .29, -.35, -.69),$$

where the first component is the intercept. The variance of the error term is taken as 0.6. These values are two times a set of fitted values obtained in our exploratory analysis of the real data set. This setting represents the situation where none of the SNPs dominates the others in their genetic contributions to the quantitative trait. With this setting, an average heritability is around 40%. We did not insist on matching the heritability in the real data set, because to do so not only the coefficient values but also the structure of the genotypes of selected SNPs must match those in the real data, which is far from trivial.

In the second setting, the vector $\boldsymbol{\beta}$ is taken as

$$\boldsymbol{\beta}^t = (2, -1.56, -1.09, 1.22, -.06, -.08, -.012, .067, .047, .07, .05).$$

In this setting, the effects of the first three SNPs are markedly more prominent than the others. This represents the case where a few major QTL dominate the contribution to the variation of the quantitative trait. Together with the other 7 SNPs, they contribute around 66% of the variation in the simulated trait values.

Table 6: The average number of SNPs selected by the model selection criteria AIC, BIC and MIC over 200 repititions (numbers in parenthese are standard deviations).

	Setting 1			Setting 2		
	AIC	BIC	MIC	AIC	BIC	MIC
Correct	6.37 (1.50)	6.35 (1.50)	5.73(1.73)	3.00(.07)	3.00(.07)	3.00(.07)
Incorrect	4.61 (1.50)	4.05 (1.69)	0.82(0.96)	7.67(.65)	3.36(2.25)	0.13(0.39)

The tournament approach is applied in both settings in the same way except that the finalists consist of 30 SNPs in the first setting but the finalists consist of only 15 SNPs in the second setting. To save the amount of computation, the permutation aggregating is not implemented in the simulation.

All the three criteria — AIC, BIC and MIC — are applied in the final stage of the tournament approach. The average numbers of SNPs correctly chosen and incorrectly chosen under both settings are given in Table 6.

The average number of correctly selected SNPs with all the three model selection criteria are very close. In the first setting, around 60% of the SNPs which have minor effects on the quantitative trait are identified. In the second setting, almost all the three major SNPs are identified. But the AIC and BIC suffer high false discovery rates: 42% and 39% respectively in the first setting, 72% and 53% respectively in the second setting. The MIC, on the other hand, controls the false discovery rate quite well. The false discovery rates with MIC are only 13% and 4% respectively in the first and second setting. In the case of false discovery, the AIC always tend to dicover a large number of false SNPs while the numbers of false SNPs discovered by BIC are spread out, which is reflected by the standard deviations of the false discovery numbers given in the Table.

Table 7: Number of SNPs selected when no SNPs are significant.

Number of SNPs	0	1	2	3	4	5	6	7	8	≥ 9
AIC	0	0	0	0	0	0	0	0	0	200
BIC	0	2	8	15	23	20	16	23	17	76
MIC	158	36	6	0	0	0	0	0	0	0

Table 7 contains the simulation results when no SNPs have significant contribution to the variation of the quantitative trait. We set the value of $\boldsymbol{\beta}$ at the order of 0.01 to make the situation more realistic. The results are similar to those presented in Table 5. The AIC and BIC fail to eliminate non-significant SNPs while the MIC guards against the non-significant SNPs very well.

In summary, the simulation results demonstrate that the tournament approach with the MIC model selection criterion has high power and low false discovery rate in detecting major SNPs. Even in the case of minor SNPs, the approach still has desiorable power and false discovery rate, noting that the sample size is not very large. The tournament approach provides a promising powerful tool for gonome-wide association studies.

5 Conclusion and discussion

We have developed an effective tournament approach for the data analysis of genome-wide association studies. Although this approach is developed with the specific genetic application in mind, it can be adapted for any model selection problems with a huge number of covariates. Through the real data analysis and the simulation studies, it is demonstrated that the tournament approach is statistically efficient in identifying genes with significant contributions towards the variation of the quantitative trait at

relatively very low computational cost. These properties are particularly important in the analysis of genetic data containing tens or even hundreds of thousands markers.

The tournament approach is apprently more general than what we have presented in this paper. We can supplement this approach with many other attachments. For example, we can further improve the stability of the method by introducing permutation aggregating in every pre-match rounds. In the presentation of this paper, the dominant effects of the SNPs are assumed negligible. If this is not the case, the approach can be rectified easily by introducing two variables for each SNP in the model with a modest increment in computational amount. Further, the tournament approach can also accommodate the analysis of epistasis effects of genes. In this aspect, more research is needed to moderate the substantial increment in computational amount. There is still much room left for further studies of this approach, which will take our continous effort in our further research.

Appendix

In this appendix, we deal with the computational issue of the tournament approach.

The first point to note is that the design matrix X must be standardized for the tournament approach; that is, each column of X must be normalized to have mean zero and standard error 1. This step is necessary for the penalized likelihood to be meaningful.

Since the penalty function $p_\lambda(\beta)$ is not smooth at 0, the commonly used Newton-Raphson method is not applicable for the maximization of the penalized log likelihood. Following Fan and Li (2001), for non-zero β_{k0}, we approximate $p_\lambda(\beta)$ at the vicinity of β_{k0} by

$$p_\lambda(|\beta_k|) \approx p_\lambda(|\beta_{k0}|) + \frac{1}{2}\left\{\frac{p'_\lambda(|\beta_{k0}|)}{|\beta_{k0}|}\right\}(\beta_k^2 - \beta_{k0}^2).$$

Let

$$G_\lambda(\beta_0) = \text{Diag}\left\{\frac{p'_\lambda(|\beta_{k0}|)}{|\beta_{k0}|}, k = 1, 2, \ldots\right\}.$$

By deleting the columns of X_i with zero initial coefficients, up to a constant, the penalized log likelihood becomes

$$l_p(\beta) = -\sum_{i=1}^{n}(y_i - X_i\boldsymbol{\beta})^t\Sigma_i^{-1}(y_i - X_i\boldsymbol{\beta}) - \boldsymbol{\beta}^t G_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}. \tag{4}$$

In principle, (4) can now be maximized by the Newton-Raphson method. Since the Newton-Raphson method can not guarantee the positiveness of the variance components, we adopt the following coordinate ascent algorithm for the maximization. The approximate penalized log likelihood (3) is estimated alternatively with respect to $\boldsymbol{\beta}$ and (σ_g^2, σ_e^2). For each given value of σ_g^2 and σ_e^2, $l_p(\beta, \sigma_g^2, \sigma_e^2|X, \lambda)$ is maximized with respect to $\boldsymbol{\beta}$. Let the resultant fitted value of $\boldsymbol{\beta}$ be denoted by $\hat{\boldsymbol{\beta}}$. Then $l_p(\hat{\boldsymbol{\beta}}, \sigma_g^2, \sigma_e^2|X, \lambda)$ is maximized with respect to (σ_g^2, σ_e^2). The iteration is repeated until convergence occurs.

In the step of maximizing $l_p(\beta, \sigma_g^2, \sigma_e^2|X, \lambda)$ with respect to $\boldsymbol{\beta}$, the computation is further simplified by taking the structure of Σ_i into account. By standard matrix theory, we may decompose A_i as $A_i = Q_i^t\Lambda_i Q_i$ for each i with Q_i being an orthogonal matrix, and Λ_i being a diagonal matrix. Consequently, we have $\Sigma_i = Q_i^t[\sigma_g^2\Lambda_i + \sigma_e^2 I_i]Q_i$ where I_i is an identity matrix of order n_i, the size of the ith pedigree. With these, we may write

$$l_p(\beta) = -\sum_{i=1}^{n}(\tilde{y}_i - \tilde{X}_i\boldsymbol{\beta})^t D_i^{-1}(\tilde{y}_i - \tilde{X}_i\boldsymbol{\beta}) - \boldsymbol{\beta}^t G_\lambda(\beta_0)\boldsymbol{\beta}, \tag{5}$$

where $\tilde{y}_i = Q_i^t y_i$, $\tilde{X}_i = Q_i^t X_i$, and $D_i = \sigma_g^2\Lambda_i + \sigma_e^2 I_i$. This simplification avoids repeated computations of the inverse of Σ_i for new values of (σ_g^2, σ_e^2). An explicit solution for $\boldsymbol{\beta}$ is then available.

Next, we consider for a given $\hat{\boldsymbol{\beta}}$ the maximization of $l_p(\hat{\boldsymbol{\beta}}, \sigma_g^2, \sigma_e^2|X, \lambda)$ with respect to (σ_g^2, σ_e^2). Define

$$r = -\log(\sigma_g^2/\sigma_e^2)$$

so that $\sigma_g^2 = \sigma_e^2 \exp(-r)$. Let $\tilde{D}_i = \exp(-r)\Lambda_i + I_i$ and $\tilde{\epsilon}_i = D_i^{-1/2}(\tilde{y}_i - \tilde{X}_i\boldsymbol{\beta})$. It is then seen that maximizing $l_p(\hat{\beta}, \sigma_g^2, \sigma_e^2|X, \lambda)$ is equivalent to maximizing

$$l_p(r, \sigma_e^2) = -\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\tilde{\epsilon}_i^\tau\tilde{\epsilon}_i - N\log\sigma_e^2 - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\log\tilde{d}_{ijj}$$

with $\tilde{d}_{ijj}$ being the diagonal element of $\tilde{D}_i$. It is easily seen that for a given value of r, the function is maximized when

$$\sigma_e^2 = \hat{\sigma}_e^2(r) = N^{-1}\sum_{i=1}^{n}\tilde{\epsilon}_i^\tau\tilde{\epsilon}_i.$$

It turns out then that we only need to choose r to maximize

$$l_p(r) = -N\log\sigma_e^2(r) - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\log\tilde{d}_{ijj}$$

because the other term does not depend on r.

Since $\exp(-r)/[1+\exp(-r)]$ has a range between 0 and 1, a range of r can be set easily for the maximization. For example, we may set the range as $[-10,\ 40]$. The one-dimensional optimization problem can then be solved easily.

References

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, in *Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski. Budapest: Akademiai Kiado, page 267.

Amos, C. I. (1994), Robust varaince-component approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54**, 535-543.

Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate — A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* B, **57**, 289-300.

Breiman, L. (1996a), Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350-2383.

Breiman, L. (1996b), Bagging predictors. *Machine Learning*, **26**(2), 123-140.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Statist. Soc. B*, **64**, 641-656.

Craven, P. and Wahba, G. (1979), Smoothing noisy data with Spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematika*, **31**, 377-403.

Dudbridge, F. and Koeleman, B. P. (2003), Rank truncated product of P-values, with application to genome-wide association scans. *Genet. Epidemiol.* **25**, 360-366.

Fan, J. and Li, R. (2001), Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

Hoh, J., Wille, A. and Ott, J. (2001), Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research* **11**, 2115-2119.

Hoh, J. and Ott, J. (2003), Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* **4**, 701-709.

Ishwaran, H. and Rao, J. S. (2003), Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Am. Stat. Assoc.* **98**, 438-455.

Kauermann, G. and Eilers, P. (2004), Modeling microarray data using a threshold mixture model. *Biometrics*, **60**, 376-387.

Lange, K. (2002), *Mathematical and Statistical Methods for Genetic Analysis*. 2nd ed. Springer.

Marchini, J. Donnelly, P. and Cardon, L. R. (2005), Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*,**37**, 413-417.

Martin, E. R., Ritchie, M. D., Hahn, L., Kang, S. and Moore, J. H. (2006), A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genetic Epidemiology* **30**, 111-123.

Matise, T.C., Sachidanandam, R., Clark, A. G., Kruglyak, L., Wijsman, E., Kakol, J., Buyske, S., Chui, B., Cohen, P., de Toma, C., Ehm, M., Glanowski, S., He, C., Heil, J., Markianos, K., McMullen, I., Pericak-Vance, M. A., Silbergleit, A., Stein, L., Wagner, M., Wilson, A. F., Winick, J. D., Winn-Deen, E. S., Yamashiro, C. T., Cann, H. M., Lai, E., Holden, A. L. (2003), A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am. J. Hum. Genet.* **73**(2), 271-284.

Nelson, M. R., Kardia, S. L. R., Ferrell, R. E. and Sing, C. F. (2001), A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research* **11**, 458-470.

Ritchie, M. D., Hahn, L., Roodi, W. N., Bailey, L. R., Dupont, W. D., Parl, F. F. and Moore, J. H. (2001), Multifactor-dimensionality reductio reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69**, 138-147.

Schwarz, G. (1978), Estimating the dimension of a model. *Annuals of Statistics*, **6**, 461-464.

Storey, J. D. and Tibshirani, R. (2003), Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440-9445.

Tibshirani, R. (1996), Regression shrinkage and selection via the LASSO, *J. Roy. Statist. Soc. Ser. B*, **58**, 267-288.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P.H. and Weir, B. S. (2002), Truncated product method for combining p-values, *Genet. Epidemiol.* **22**, 170-185.