Prediction of a time-to-event trait using genome wide SNP data

Jinseog Kim1†, Insuk Sohn2†, Dae-Soon Son3, Dong Hwan Kim4, Taejin Ahn3 and Sin-Ho Jung5*

Abstract

Background: A popular objective of many high-throughput genome projects is to discover various genomic markers associated with traits and develop statistical models to predict traits of future patients based on marker values.

Results: In this paper, we present a prediction method for time-to-event traits using genome-wide single-nucleotide polymorphisms (SNPs). We also propose a MaxTest associating between a time-to-event trait and a SNP accounting for its possible genetic models. The proposed MaxTest can help screen out nonprognostic SNPs and identify genetic models of prognostic SNPs. The performance of the proposed method is evaluated through simulations.

Conclusions: In conjunction with the MaxTest, the proposed method provides more parsimonious prediction models but includes more prognostic SNPs than some naive prediction methods. The proposed method is demonstrated with real GWAS data.

Background

A genome-wide association study (GWAS) involves an examination of the entire genome, typically single-nucleotide polymorphisms (SNPs), of different individuals to determine whether any variant is associated with a particular clinical outcome. Many researchers have considered the design and analysis of GWASs with respect to binary clinical outcomes such as case/control or response/non-response ones [1-5].

In clinical cancer research, the primary endpoint of interest is usually a time-to-event trait subject to censoring. In CALGB 80803, for example, germline DNAs are collected, together with time to progression and overall survival data, from 352 advanced pancreatic cancer patients. One objective of an SNP correlative study is to discover SNP markers that are correlated with such time-to-event endpoints.

One of the first objectives of a statistical analysis in a GWAS is the discovery of SNP markers that are associated with a particular trait. The major statistical issue in marker discovery is multiple testing to avoid enlarged type I error probability due to the large number of univariate tests [6,7]. Each prognostic SNP has two or three possible outcomes depending on its genetic model, and the efficiency of a statistical method in associating it with a trait is maximized when the true genetic model is known. For most SNPs, however, the true genetic model is unknown. To identify the true genetic model of each SNP and optimize the association analysis, many researchers have considered some candidate genetic models for a given trait and derived a null distribution of the maximum of test statistics specific to individual genetic models [8,9]. This test is referred to as the MaxTest. These methods have been developed for binary traits such as case/control or response/non-response ones. We develop a MaxTest to identify the genetic model of each SNP when the trait is a survival endpoint, e.g., the time to tumor progression or death.

Another major objective of a GWAS is to predict a trait of interest by using SNPs. Prediction methods using microarray data have been widely investigated [10-12], but cannot be directly applied to SNP-based predictions. The number of SNP markers in genome-wide SNP data far exceeds that of gene markers (or probes) in microarray data, e.g., 1M vs. 20K. In addition, although gene expression data in microarray studies are continuous, genome-wide SNP data are discrete, taking only three different values at most and showing different values depending on the genetic model.
This paper presents a method for predicting a survival outcome that uses genome-wide SNP data but can be easily modified for any type of trait, including binary or continuous outcomes. The proposed method uses the gradient lasso method [13], which has been developed for microarray data. Some investigators fit a prediction model while ignoring the genetic model of each SNP [14]. We also propose a MaxTest associating between a time-to-event trait and a SNP accounting for its possible genetic model and identifies the genetic model of each candidate prognostic SNP by using the proposed MaxTest before fitting a prediction model. The simulation results show that this procedure improves prediction efficiency and prognostic power. For computational efficiency, nonsignificant SNPs are excluded using the MaxTest before starting the gradient lasso. For the facilitation of the proposed MaxTest and prediction method, glcophSNP R packages (http://datamining.dongguk.ac.kr/Rlib/glcoxpathSNP) are provided.

Methods
Genetic Models of SNPs
Suppose that the genotype for an SNP is encoded as AA, AB, or BB. Let \( g \) denote the number of copies of the B allele. That is, \( g = 0, 1 \) or \( 2 \) if the genotype is AA, AB, or BB, respectively. Let \( \lambda_g(t) \) denote the hazard function of genotype \( g \). Without loss of generality, assume that B is the risk allele in the sense that having B increases the risk of an event. More specifically, assume that \( \lambda_0(t) \leq \lambda_1(t) \leq \lambda_2(t) \) for all \( t \geq 0 \). (Note that for some specific diseases, this may not be an appropriate genetic model.) We now consider the following three popular genetic models:

- **Recessive model**: \( \lambda_0(t) = \lambda_1(t) < \lambda_2(t) \).
- **Dominant model**: \( \lambda_0(t) < \lambda_1(t) = \lambda_2(t) \).
- **Multiplicative model**: \( \lambda_2(t)/\lambda_1(t) = 1 + \gamma \lambda_0(t)/\lambda_1(t) \), or equivalently \( \lambda_1(t) = \gamma \lambda_0(t) \) and \( \lambda_2(t) = \gamma^2 \lambda_0(t) \) for \( \gamma > 0 \).

For a chosen score \( c_g \), we consider a proportional hazard model (PHM), \( \lambda_g(t) = \lambda_0(t) \exp(\beta c_g) \). Then Cox’s partial maximum likelihood test has optimal power with \( (c_0, c_1, c_2) = (0, 0.1, 0) \) for a recessive model, \( (0.1, 1, 1) \) for a dominant model, and \( (0, 1, 2) \) for a multiplicative model [15]. Note that the PHM is invariant to the linear transformation of the covariate \( (c_0, c_1, c_2) \).

MaxTest
Suppose that we want to test whether an SNP is associated with a given clinical outcome. The test statistic is dependent on the true genetic model of the SNP. At the time of testing, however, we usually have no knowledge of the true genetic model. In this case, a naive approach is to conduct all statistical tests by assuming different genetic models and choose the lowest p-value as the measurement of the association. This approach can lead to an enlarged Type I error because of multiple tests. To adjust for multiple tests, investigators have proposed a method considering the maximum of test statistics with respect to all candidate genetic models under consideration, namely the MaxTest.

Many studies have considered the MaxTest for binary clinical outcomes. Zheng et al. [8] propose a robust ranking method when the underlying genetic model is unknown, namely the MAX-rank test. Conneely and Boehnke [16] propose a method for computing p-values that adjusts for correlated tests and show that the method can improve the accuracy of permutation tests with greater computational efficiency. Li et al. [17] propose a method for approximating the p-value for the MaxTest with or without covariates adjusted for, namely the P-rank test. Li et al. [9] compare the results of the MAX-rank and P-rank tests. Hoggart et al. [18] formulate the problem as variable selection in a logistic regression analysis including a covariate for each SNP and find the posterior ranking method when the underlying genetic model is unknown, namely the MAX-rank test.
Then we can define the proposed MaxTest as $Q = \max(|T_1|, |T_2|, |T_3|)$, where $T_L = W_l/\hat{\sigma}_l$. Under $H_0$, $\sigma_H = \text{cov}(W_i, W_F)$ is consistently estimated by

$$\hat{\sigma}_H = n^{-1} \sum_{i=1}^{n} \int_{0}^{\infty} \left\{ z_l - \frac{s_l(t)}{s_H(t)} \right\} \left\{ z_l - \frac{s_l(t)}{s_H(t)} \right\} \frac{Y_l(t)}{Y(t)} \cdot dN(t)$$

where $z_l$ and $s_l(t)$ denote $z_l$ and $s_l(t)$, respectively, for genetic model $l$; $Y(t) = \sum_{i=1}^{n} Y_i(t)$, $N(t) = \sum_{i=1}^{n} N_i(t)$. Let $\hat{\Sigma} = (\hat{\sigma}_{ij})_{1 \leq l, f \leq 3}$, where $\hat{\sigma}_{lf} = \hat{\sigma}_{lf}/\hat{\sigma}_H$. Then we can obtain the critical value of $Q$ by a numerical method or a simulation method from the $N(0, \hat{\Sigma})$ distribution. This is a survival trait counterpart for the MaxTest with a binary trait, as discussed in several studies [9, 21].

We can construct an alternative test based on the quadratic form $W^T S^{-1} W$, where $S = (T_1, T_2, T_3)^T$. In addition to recessive, dominant, and multiplicative genetic models, we can consider other models to develop a test statistic to measure the relationship between an SNP and a survival trait. For example, we may consider the most recent paper that proposes an algorithm using quadratic programming [26] provides the entire penalization path for the Cox model in a forward stagewise manner. Because it requires matrix inversions only for active sets, it is faster and more stable than other methods. Sohn et al. [13] provide a comparative analysis using real and simulated data and show that the gradient lasso algorithm outperforms glm-path in analyzing high-dimensional survival data in terms of the sparsity, predictability, and computational efficiency of the final prediction model. Therefore, the following gradient lasso algorithm can be a useful alternative for fitting the Cox model to predict the survival time of patients based on high-dimensional SNP data:

1. **Initialize**: $\beta = 0$ and $k = 0$.
2. **Do** until convergence
   
   (a) **Addition**:
   
   (i) Compute the gradient $\nabla l(\beta)$.
   (ii) Find the $j$ maximizing $|\partial l(\beta)/\partial \beta_j|$ for $j = 1, \ldots, p$ and $\gamma = -s \times \text{sign}(\partial l(\beta)/\partial \beta_j)$.
   (iii) Let $v$ be a $p$-dimensional vector such that its $j$-th element $\gamma$ and other elements are zeros.
   (iv) Find $\hat{\alpha} = \arg\min_{\alpha \in [0,1]} l((1-\alpha)\beta + \alpha v)$.
   (v) Update $\beta = (1 - \hat{\alpha})\beta + \hat{\alpha} v$.

   (b) **Deletion**:
   
   (i) Calculate $h_\sigma = -\nabla l(\beta_\sigma) + \theta_\sigma \nabla l(\beta_\sigma)^T \theta_\sigma / |\sigma|$, where $\sigma = \{ j : \beta_j \neq 0 \}$.
   (ii) Find $\hat{\delta} = \arg\min_{\delta : \delta \in [U]} l(\beta + \delta h)$, where $U = \min_{\delta \in \sigma} \{|-\beta_k / h_k : \beta_k h_k < 0| \}$.
   (iii) Update $\beta = \beta + \hat{\delta} h$.

   (c) Set $m = m + 1$.

3. **Return** $\beta$.

**Proposed algorithm for predicting a survival trait**

We propose a new algorithm for fitting a Cox regression model using SNP data. The proposed algorithm consists of the following four steps: We (i) select significant SNPs
by the MaxTest, as in Section “Example using real data,” (ii) convert these SNPs into corresponding scores by genetic models identified by the MaxTest, (iii) standardize these scores, and (iv) apply the gradient lasso algorithm [13] to selected SNPs. We summarize the algorithm in greater detail as follows:

1. Read in the clinical data \( \{(X_i, d_i), i = 1, ..., n\} \) and SNP data \( \{s_{ij}, ..., s_{jm}\}, i = 1, ..., n\) where \( s_{ij} \) denotes the number of B alleles for SNP \( j (= 1, ..., m) \).
2. For SNP \( j (= 1, ..., m) \), calculate the variance and covariance matrix \( \hat{\Sigma}_j \), and generate the null distribution of the MaxTest as follows.
   (a) For \( b = 1, ..., B = (100, 000, \text{say}) \), generate \( (t_{1j}^{(b)}, t_{2j}^{(b)}, t_{3j}^{(b)}) \) from \( N(0, \hat{\Sigma}_j) \).
   (b) Let \( q_j^{(b)} = \max(|t_{1j}^{(b)}|, |t_{2j}^{(b)}|, |t_{3j}^{(b)}|) \) for \( b = 1, ..., B \).
3. For SNP \( j (= 1, ..., m) \),
   (a) Using original data, calculate the test statistics \( (T_{1j}, T_{2j}, T_{3j}) \), the MaxTest statistic \( q_j = \max(|T_{1j}|, |T_{2j}|, |T_{3j}|) \), and two-sided p-values \( p_{1j}, p_{2j}, p_{3j} \) from the marginal test for respective genetic models.
   (b) Approximate the p-value of the MaxTest by \( p_j = B^{-1} \sum_{b=1}^{B} I(q_j^{(b)} \geq q_j) \).
4. SNP screening: Select \( J (< < m) \) SNPs with \( p_j < \alpha \) for specified \( \alpha = 0.01 \), say).
5. For selected SNPs \( j (= 1, ..., J) \), identify the genetic model \((1, 2, 3)\) by the lowest marginal p-value from \( p_{1j}, p_{2j}, p_{3j} \) or the largest test statistic from \( T_{1j}, T_{2j}, T_{3j} \).
6. For patient \( i (= 1, ..., n) \), define covariates \( (z_{i1}, ..., z_{ij}) \) by the identified genetic model and the corresponding score.
7. Standardize the covariates:
   \[
   z'_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j},
   \]
   where \( \bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij} \) and \( s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2 \).
8. Apply the gradient lasso to the Cox regression model with response data \( \{(X_i, d_i), i = 1, ..., n\} \) and standardized covariates \( \{z'_{i1}, ..., z'_{ij}\}, i = 1, ..., n\).

Results and discussion
Simulation study
We provide a simulation study. The data generation scheme is as follows: We generate SNP data \( z_1, ..., z_m \) from \( N(0, 1) \) random variables with an AR(1) correlation structure with the autocorrelation coefficient \( \rho (\geq 0) \), \( x_1, ..., x_m \). Due to linkage disequilibrium, SNPs which lay in close vicinity within chromosomes tend to have a stronger association. In this sense, an AR(1) correlation structure is a reasonable correlation structure for the continuous random variables generating SNP data. Let \( x_1 = \epsilon_1 \) and \( x_j = \rho x_{j-1} + \sqrt{1 - \rho^2} \epsilon_j \) for \( j = 2, ..., m \), where \( \epsilon_1, ..., \epsilon_m \sim \text{IID} N(0,1) \) random numbers. The cutoff values for \( x_j \) for generating \( z_j \) determine allele frequency. For each SNP, let \( f_1, f_2, f_3, f_4 \) \((f_1 + f_2 + f_3 = 1)\) denote the frequency of AA, AB, and BB genotypes, respectively, where B denotes the risk allele. Note that marginally \( x_j \sim N(0,1) \). The true model for the survival times is given as

\[
\Lambda(t) = \Lambda_0(t) \exp \left( \sum_{j=1}^{D} \beta_j z_j \right),
\]
where \( D \) denotes the number of prognostic SNPs.

For the experiment, we set \( m = 1000, n = 200, D = 6, \rho = 0 \) or \( 0.3, \beta_j = 0.8 \) \((j = 1, ..., D)\), and a uniform censoring distribution for 15% or 30% of censoring. All six prognostic SNPs have \((f_1, f_2, f_3) = (0.25, 0.25, 0.5)\). SNP 1 and SNP 4 have a dominant model; SNP 2 and SNP 5, a recessive model; and SNP 3 and SNP 6, a multiplicative model. Each of the remaining 994 SNPs has \((AA, AB, BB)\) with \((f_1, f_2, f_3) = (1/3, 1/3, 1/3)\).

To evaluate the performance of the proposed method, we generate 200 random samples and divide them into a training set (100 samples) and a test set (100 samples). We calculate the MaxTest p-value of each SNP by using \( B=100,000 \) permutations from the training set and identify the genetic model for each SNP. We select SNPs with p-values less than \( \alpha = 0.01 \) and convert selected SNPs into corresponding scores by their genetic models. We apply the gradient lasso to the selected SNPs to fit the prediction model. Let \( \hat{S}_j \) \((j = 1, ..., K)\) be included in the fitted prediction model with corresponding regression estimates \( \hat{\beta}_1, ..., \hat{\beta}_K \). Then we can define the risk score for sample \( i \) as \( r_i = \hat{\beta}_1 z_{i1} + \cdots + \hat{\beta}_K z_{iK} \). Using the median risk score from the test set as a cutoff value, we divide the patients in the test set into high- and low-risk groups. We apply a two-sample log-rank test to compare the survival distribution between these two risk groups. We repeat this procedure 100 times and count the number of SNPs and that of prognostic SNPs included in each fitted prediction model by the gradient lasso. We summarize the distribution of log-rank p-values from the test set, and for comparison purposes, we consider the prediction methods by assuming that all \( m \) SNPs have the same genetic model.

Table 1 reports the mean number of SNPs and that of prognostic SNPs included in the fitted prediction model, recovery rate, and the means (and standard deviations) of the log-rank p-value from the test set for the proposed method and the methods assuming a recessive, dominant,
or multiplicative model for all SNPs under various simulation settings. We define recovery rate as the ratio of mean number of selected prognostic SNP to the mean number of selected SNP as in Sohn et al. [13]. The proposed method tends to result in prediction models with a smaller number of SNPs but a larger number of prognostic SNPs than the approaches assuming a specific genetic model for all SNPs (i.e., recessive, dominant, and multiplicative methods in the table). The recovery rates of the proposed method are higher than those of the methods based on a pre-specified model (recessive, dominant, and multiplicative). Among the three methods assuming a specific genetic model for all SNPs, the one assuming a multiplicative model shows the best performance in terms of the number of prognostic SNPs included in the final prediction model. In addition, the proposed method outperforms the recessive, dominant, and multiplicative methods in terms of the log-rank p-value and results in fitted prediction models with a smaller number of SNPs but a larger number of prognostic SNPs with 15% compared to 30% censoring. According to sample size ($n$) and the effect size ($\beta$), the mean number of SNPs and that of prognostic SNPs selected by the proposed method at $\rho=0$ and 30% censoring is shown in Table 2. The mean number of prognostic SNPs a little bit increases as $\beta$ increase and the mean number of prognostic SNPs increases as $n$ increases. The recovery rate increases as $\beta$ or $n$ increases.

Table 1 Mean numbers of SNPs and prognostic SNPs included in fitted prediction models, recovery rate and means/standard deviations of the log-rank p-value from test sets for the proposed method and methods assuming recessive, dominant, or multiplicative models for all SNPs

| Censoring | \(\rho\) | Genetic model | Mean number of selected SNPs | Mean number of selected prognostic SNPs | Recovery rate | Mean (SD) p-value of the log-rank test |
|-----------|---------|---------------|-------------------------------|----------------------------------------|--------------|--------------------------------------|
| 30%       | 0       | Proposed      | 6.72                          | 5.05                                   | 0.75         | <0.0001 (<0.0001)                    |
|           |         | Recessive     | 8.03                          | 4.01                                   | 0.50         | 0.0052 (0.0018)                      |
|           |         | Dominant      | 6.66                          | 3.85                                   | 0.58         | <0.0001 (<0.0001)                    |
|           |         | Multiplicative| 7.72                          | 4.95                                   | 0.64         | 0.0004 (0.0003)                      |
|           | 0.3     | Proposed      | 6.51                          | 4.83                                   | 0.74         | 0.0001 (0.0001)                      |
|           |         | Recessive     | 7.73                          | 3.83                                   | 0.50         | 0.0045 (0.0016)                      |
|           |         | Dominant      | 6.58                          | 3.66                                   | 0.56         | 0.0011 (0.0007)                      |
|           |         | Multiplicative| 7.52                          | 4.72                                   | 0.63         | 0.0006 (0.0004)                      |
| 15%       | 0       | Proposed      | 6.65                          | 5.18                                   | 0.78         | <0.0001 (<0.0001)                    |
|           |         | Recessive     | 8.59                          | 4.19                                   | 0.49         | 0.0028 (0.0011)                      |
|           |         | Dominant      | 6.69                          | 3.88                                   | 0.58         | <0.0001 (<0.0001)                    |
|           |         | Multiplicative| 7.96                          | 4.98                                   | 0.63         | 0.0005 (0.0005)                      |
|           | 0.3     | Proposed      | 6.37                          | 4.98                                   | 0.78         | <0.0001 (<0.0001)                    |
|           |         | Recessive     | 7.88                          | 3.94                                   | 0.50         | 0.0048 (0.0008)                      |
|           |         | Dominant      | 6.38                          | 3.74                                   | 0.59         | 0.0011 (0.0011)                      |
|           |         | Multiplicative| 7.55                          | 4.89                                   | 0.65         | 0.0001 (<0.0001)                     |

Table 2 Mean number of SNPs and prognostic SNPs included in the fitted prediction models, recovery rate and means/standard deviations of the log-rank p-values from the test set for the proposed method at \(\rho = 0\) and censoring = 30%

| Sample size \(n\) | \(\beta\) | Mean number of selected SNP | Mean number of selected prognostic SNP | Recovery rate | Mean (SD) p-value of the log-rank test |
|-------------------|---------|----------------------------|---------------------------------------|--------------|--------------------------------------|
| 200               | 0.8     | 6.72                       | 5.05                                  | 0.75         | <0.0001 (<0.0001)                    |
|                   | 1       | 6.13                       | 5.18                                  | 0.85         | <0.0001 (<0.0001)                    |
|                   | 2       | 5.60                       | 5.17                                  | 0.92         | <0.0001 (<0.0001)                    |
| 300               | 0.8     | 6.18                       | 5.53                                  | 0.89         | <0.0001 (<0.0001)                    |
| 400               | 0.8     | 5.89                       | 5.72                                  | 0.97         | <0.0001 (<0.0001)                    |
Example using real data
We apply the proposed method to the GWAS data in Choi et al. [28], who provide a GWAS of 119 patients with normal karyotype acute myeloid leukemia (AML-NK) by using Affymetric Genome-Wide Human SNP Arrays 6.0 (San Diego, CA, USA). We exclude those SNPs with missing genotype data for any patient. We also exclude those SNPs with only one genotype for the 119 patients. The final data set for the analysis includes \( m = 251,748 \) autosomal SNPs from \( n = 119 \) patients. The primary endpoint in this analysis is event-free survival (EFS), which is defined as the interval between the registration and the end of induction chemotherapy for patients showing no complete response (CR), a relapse after achieving a CR to induction chemotherapy, or death.

We employ the leave-one-out cross-validation (LOOCV) procedure to evaluate the predictive performance of the proposed method for the data set. From a training set of size \( n - 1 = 118 \), we calculate the MaxTest p-value of each SNP based on \( B = 100,000 \) permutations, select \( J \) candidate SNPs with p-values less than \( \alpha = 0.01 \) by MaxTest, and apply the gradient lasso to candidate SNPs to obtain a prediction model. Using the median risk score for the patients in the training set, we allocate those patients who are left out for the validation to the high- or low-risk group. We repeat this procedure \( n \) times and calculate the log-rank p-value to compare the EFS between the two risk groups. Figure 1(a) shows the Kaplan-Meier curves for the high- and low-risk groups classified by the LOOCV procedure. The five-year EFS rate for the low-risk group (n=60, 53.8%) is much higher than that for the high-risk group (n=59, 32.9%) with the estimated hazard ratio of 0.446 (95% CI, 0.256-0.778), and the log-rank p-value is 0.0035.

A standard approach may be to fit a prediction model assuming a multiplicative genetic models for all SNPs, e.g. Tan et al. [29]. We analyzed this data set using the same method as above except that all SNPs were assumed to have a multiplicative model. Figure 1(b) displays the LOOCV results. Note that the fitted prediction models do not significantly partition the test set into high- and low-risk groups by ignoring the possible genetic models.

We also apply our prediction procedure to the whole data set with \( n = 119 \). Using \( \alpha = 0.01 \), we select \( J = 1122 \) candidate SNPs, among which 444 (39.6%) are shown to have a recessive model, 463 (41.3%) a dominant model, and 215 (19.2%) a multiplicative model. By applying the gradient lasso to the selected 1122 SNPs, we obtain the final prediction model including \( k = 24 \) SNPs. Table 3 lists the RS IDs, the chromosome numbers, the base-pair position and the gene name of the 24 SNPs included in the final prediction model. For each of the 24 SNPs, we report the genetic model (=1 for recessive model, =2 for dominant model, and =3 for multiplicative model) identified by the MaxTest, the marginal MaxTest p-value and number of times (frequency) that each SNP is included in the prediction models during the LOOCV. Note that the first four SNPs in Table 2 are included in all 119 prediction models during LOOCV.

![Figure 1](http://www.biomedcentral.com/1471-2105/14/1/58) Kaplan-Meier curves for high- and low-risk groups classified by the LOOCV procedure.
Table 3 List of 24 SNPs selected by the proposed method from the whole data set of 119 samples, their MaxTest p-values, genetic models, the number of times selected by prediction models fitted during the LOOCV procedure

| RS ID    | Chr | Position | Gene name       | Genetic model g | P-value   | Frequency |
|----------|-----|----------|-----------------|-----------------|-----------|-----------|
| rs1030254| 16  | 60696651 | LOC644649, CDH8, LOC729159 | 3               | 0.00009   | 119       |
| rs1030252| 16  | 60696869 | LOC644649, CDH8, LOC729159 | 2               | 0.00010   | 119       |
| rs10798122| 1   | 187584699| PLA2G4A, FAM5C     | 1               | 0.00048   | 119       |
| rs10026207| 4   | 186039201| HELT, SLC25A4      | 3               | 0.00233   | 119       |
| rs13333229| 16  | 1695776  | CRAMP1L           | 3               | 0.00015   | 117       |
| rs2132183| 3   | 84966867 | LOC440970, CADM2   | 3               | 0.00149   | 117       |
| rs1905400| 14  | 27105035 | MIR4307, NOVA1     | 2               | 0.00040   | 115       |
| rs2155777| 11  | 133290007| OPCML            | 3               | 0.00142   | 113       |
| rs16797194| 12  | 78274425 | NAV3             | 2               | 0.00283   | 106       |
| rs1476847| 18  | 9834599  | RAB31            | 1               | 0.00029   | 102       |
| rs7614596| 3   | 84986027 | LOC440970, CADM2   | 2               | 0.00020   | 100       |
| rs2648117| 4   | 186787096| SORBS2           | 3               | 0.00856   | 90        |
| rs1851317| 15  | 35077786 | GJD2, AACTC1      | 1               | 0.00999   | 88        |
| rs3790217| 20  | 19441650 | SLC24A3          | 2               | 0.00728   | 85        |
| rs4902990| 14  | 72618432 | RGS6             | 2               | 0.00004   | 81        |
| rs9482583| 6   | 125318379| RNF217           | 3               | 0.00847   | 79        |
| rs3020444| 14  | 64791013 | ESR2             | 3               | 0.00288   | 77        |
| rs10851869| 15  | 74331083 | PML              | 2               | 0.00036   | 65        |
| rs11986200| 8   | 22698209 | PEBP4            | 1               | 0.00222   | 63        |
| rs11260756| 1   | 16759616 | SPATA21          | 1               | 0.00827   | 63        |
| rs4968415| 17  | 60264240 | MED13, TBC1D3P2  | 1               | 0.00075   | 62        |
| rs12416722| 11  | 133300460| OPCML           | 1               | 0.00067   | 59        |
| rs626266| 12  | 7288187 | TRHDE            | 2               | 0.00070   | 52        |
| rs16852300| 2   | 167414242| SCN7A, XIRP2     | 3               | 0.00513   | 33        |

The RGS6 gene (rs4902990) is associated with treatment outcomes in AML-NK patients. RGS6, a regulator of G-protein signaling 6, modulates the G-protein function in the signaling pathway by activating the intrinsic GTPase activity of alpha subunits [30,31]. An SNP on RGS6 has been found to modulate the risk of bladder cancer [32]. In addition, it is known that RGS6 induces apoptosis through a mitochondrial-dependent pathway, which implies that RGS6 may be involved in cancer progression [29]. Further, membrane drug transporters, including SLC25A4 (rs10026207) and SLC24A3 (rs3790217), are known to be associated with event-free survival. SLC25A4, solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator; ANT1), member 4, is known to interact with the Bcl-2-associated X protein, which is involved in the apoptosis pathway [33,34]. The rs10798122 SNP on family with sequence similarity 5, member C, FAM5C, is selected by the proposed model. A loss of hypermethylated FAM5c is known to be associated with the development of tongue squamous cell carcinoma or gastric cancer [35,36].

Conclusions
We have proposed a prediction method for a survival endpoint using SNPs. The paper also proposes a MaxTest to screen out nonprognostic SNPs and identify genetic model of prognostic SNPs. The simulation results indicate substantial prognostic power for the proposed prediction method. Noteworthy is that, in conjunction with the MaxTest, the proposed method provides more parsimonious prediction models with more prognostic SNPs than those prediction methods ignoring the true genetic model of prognostic SNPs. We apply real GWAS data to patients with acute myeloid leukemia and find that the proposed method provides a prediction model that can efficiently classify the patients into high- and low-risk groups by using a small number of SNPs that are known to be biologically informative. Although the proposed method is limited to the prediction of time-to-event traits, it can be easily extended to a wide range of traits, including dichotomous or continuous ones.
Competing interests
The authors declare no conflict of interest

Author contributions
JK and IS performed the statistical analysis and wrote the manuscript. DS and TA supported the research. DHK provided a biological interpretation of the statistical analysis. SJ proposed the research project. All authors read and approved the final manuscript.

Acknowledgements
This research was supported by a grant from the National Cancer Institute, CA142538.

Author details
1 Department of Statistics and Information Science, Dongguk University, Gyeongju 780-714, Korea. 2 Samsung Cancer Research Institute, Samsung Medical Center, Seoul 137-710, Korea. 3 In Vitro Diagnostics Lab, Bio Research and Development Center, Samsung Advanced Institute of Technology, Suwon 449-712, Korea. 4 Department of Medical Oncology and Hematology, Princess Margaret Hospital, University of Toronto, Toronto ON, Canada. 5 Department of Biostatistics and Bioinformatics, Duke University, NC 27710, USA.

Received: 30 October 2012 Accepted: 12 February 2013
Published: 19 February 2013

References
1. Chen BE, Sakoda LC, Hsing AW, Rosenberg PS: Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. Genet Epidemiol 2006, 30(6):495–507.
2. Gordon D, Finch SJ: Factors affecting statistical power in the detection of genetic association. J Clin Invest 2005, 115(6):1408–1418.
3. Hao K, Xu X, Laird N, Wang X, Xu X: Power estimation of multiple SNP association test of case-control study and application. Genet Epidemiol 2004, 26(1):22–30.
4. Skol AD, Scott LJ, Abecasis GR, Boehnke M: Efficient computation of the max statistic for genome-wide association study results. Hum Mol Genet 2006, 15(16):2026–2032.
5. Westfall PH, Young SS: Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. New York: Wiley; 1993.
6. Storey JD: A direct approach to false discovery rates. J R Stat Soc B 2002, 64:479–498.
7. Zheng G, Freidlin B, Gastwirth JL: Comparison of robust tests for association using case-control study and application. Genet Epidemiol 2004, 26(1):22–30.
8. Skol AD, Scott LJ, Abecasis GR, Boehnke M: Efficient computation of the max statistic for genome-wide association study results. Hum Mol Genet 2006, 15(16):2026–2032.
9. Westfall PH, Young SS: Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. New York: Wiley; 1993.
10. Storey JD: A direct approach to false discovery rates. J R Stat Soc B 2002, 64:479–498.
11. Li Q, Yu K, Li Z, Zheng G: Max-rank: a simple and robust genome-wide scan for case-control association studies. Hum Genet 2008, 123(6):617–623.
12. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ: Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 2008, 4(1):e1000130.
13. Cox DR: Regression Models and Life Tables (with Discussion). J R Stat Soc Ser B 1972, 34:187–220.
14. Fleming TR, Harrington DP: Counting Processes and Survival Analysis. New York: Wiley; 1991.
15. Owzar K, Li Z, Cox N, Jung SH: Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 2009, 25(1):3001–3008.
16. Park MY, Hastie T: L1 regularization path algorithm for generalized linear models. J R Stat Soc B 2007, 69:659–677.
17. Kim J, Kim Y: A gradient-based optimization algorithm for lasso. J Comput Graph Stat 2008, 17:299–1109.
18. Maley R, Regulator of G protein signaling 6 (RGS6) induces apoptosis via a mitochondrial-dependent pathway not involving its GTP-ase activating protein activity. J Biol Chem 2011, 286(2):1409–1419.
19. Berman DM, Gilman AG: Mammalian RGS proteins: baritians at the gate. J Biol Chem 1998, 273(1):1269–1272.
20. Malott B, Yang J, Huang J, Askeland RW, Bera S, Fisher RA: Regulator of G protein signaling 6 (RGS6) induces apoptosis via a mitochondrial-dependent pathway not involving its GTP-ase activating protein activity. J Biol Chem 2011, 286(2):1409–1419.
21. Berman DM, Wang Y, Liu Z, Dong Q, Burke LA, Liootta LA, Fisher R, Wu X: A functional polymorphism in RGS6 modulates the risk of bladder cancer. Cancer Res 2006, 66(18):6820–6826.
22. Baines CP, Molkentin JD: Adeninenucleotidetranslocase-1 induces cardiomyocyte death through upregulation of the pro-apoptotic protein Bak. J Mol Cell Cardiol 2009, 46(6):969–977.
23. Malorni W, Farrace MG, Matarrese P, Tinari A, Ciarlo L, Mousavi-Shafaei P, D’Ercole M, Di Giacomo G, Melino G, Palmeni L, Rodolfo C, Placintini M: The adenine nucleotide translocator 1 acts as a type 2 transglutaminase substrate: implications for mitochondrial-dependent apoptosis. Cell Death Differ 2009, 16(1):1480–1492.
24. Chen L, Su L, Li J, Zheng Y, Yu B, Yu Y, Yan M, Gu Q, Zhu Z, Liu B: Hypermethylated FAM139B and MYCML in serum as diagnosis and pre-warning markers for gastric cancer. Dis Markers 2012, 32(2):195–202.
25. Kuroiwa T, Yamamoto N, Onda T, Shibahara T: Expression of the FAM139B gene in tongue squamous cell carcinoma. Oncol Rep 2009, 22(5):1005–1011.

Cite this article as: Kim et al. Prediction of a time-to-event trait using genome-wide SNP data. BMC Bioinformatics 2013 14:58.