Kim et al. BMC Bioinformatics 2013, 14:58
http://www.biomedcentral.com/1471-2105/14/1/58

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

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

Jinseog Kim[1†], Insuk Sohn[2†], Dae-Soon Son[3], Dong Hwan Kim[4], Taejin Ahn[3] and Sin-Ho Jung[5*]

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.

*Correspondence: sinho.jung@duke.edu
†Contributed equally
[5]Department of Biostatistics and Bioinformatics, Duke University, NC 27710, USA
Full list of author information is available at the end of the article

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, glcoxphSNP R packages (http://datamining.dongguk.ac.kr/Rlib/glcoxphSNP) 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: $\lambda_2(t)/\lambda_1(t) = \lambda_1(t)/\lambda_0(t)$, 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 $(c_0, c_1, c_2) = (0, 0, 1)$ for a recessive model, $(0, 1, 1)$ 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 mode for shrinkage priors based on a stochastic search on a penalized likelihood function.

We propose a MaxTest for survival endpoints. Here we assign numeric scores to three genotypes based on their genetic model: $(c_0, c_1, c_2) = (0, 0, 1)$ for a recessive model, $(c_0, c_1, c_2) = (0, 1, 1)$ for a dominant model, and $(c_0, c_1, c_2) = (0, 1, 2)$ for a multiplicative model. For a given score c_g assigned to the genotype g $(= 0, 1, 2)$ of an SNP, we consider the Cox propotional hazard model,

$$\lambda_g(t) = \lambda_0(t)\exp(\beta c_g).$$

For patient i $(= 1, ..., n)$, let T_i and C_i denote the survival and censoring times, respectively. We observe that $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ indicates the indicator function. In addition, for $t \geq 0$, let $Y_i(t) = I(X_i \geq t)$ and $N_i(t) = \delta_i I(X_i \leq t)$ denote the at-risk and event processes, respectively. For a given score, we set $z_i = c_g$ if patient i has genotype g. Let $s_k(t) = \sum_{i=1}^n z_i^k Y_i(t)$, $k = 0, 1, 2$. Then, the partial score test statistic by Cox [19],

$$W = n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \{z_i - s_1(t)/s_0(t)\}dN_i(t)$$

is asymptotically normal with mean 0 and variance that can be consistently estimated by

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\int_0^\infty \left\{\frac{s_2(t)}{s_0(t)} - \frac{s_1^2(t)}{s_0^2(t)}\right\} dN_i(t)$$

under $H_0 : \lambda_0(t) = \lambda_1(t) = \lambda_2(t)$ [see, e.g., [20]].

By combining the statistics with respect to the three candidate genetic models, we can derive a MaxTest statistic. Let W_l and $\hat{\sigma}_l^2$ denote the test statistic and the variance estimator with respect to genetic model l $(= 1, 2, 3)$.

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, $\sigma_{ll'} = \text{cov}(W_l, W_{l'})$ is consistently estimated by

$$\hat{\sigma}_{ll'} = n^{-1} \sum_{i=1}^{n} \int_0^{\infty} \left\{ z_{li} - \frac{s_{l1}(t)}{s_{l0}(t)} \right\} \left\{ z_{l'i} - \frac{s_{l'1}(t)}{s_{l'0}(t)} \right\} \frac{Y_i(t)}{Y(t)} dN(t)$$

where z_{li} and $s_{lk}(t)$ denote z_i and $s_k(t)$, respectively, for genetic model l; $Y(t) = \sum_{i=1}^{n} Y_i(t)$, $N(t) = \sum_{i=1}^{n} N_i(t)$. Let $\hat{\Sigma} = (\hat{\rho}_{ll'})_{1 \le l,l' \le 3}$, where $\hat{\rho}_{ll'} = \hat{\sigma}_{ll'}/\hat{\sigma}_l\hat{\sigma}_{l'}$. 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 $W^2 = S^T \hat{\Sigma}^{-1} S$, 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 long-rank test based on the one-way ANOVA in [22] or the test based on the Wilcoxon Rank-Sum test in [23], which require no specific genetic model assumptions. In particular, the ANOVA-type test is a reasonable option if the monotone trend in genotypes g = 0, 1, and 2 is doubtful.

Cox model with a lasso penalty

In an analysis using SNP data, we may face a problem in which the number of SNPs exceeds the size of data, that is, $m \gg p$, which frequently occurs even when a smaller number of SNPs are selected through a prescreening step. This may lead to a serious collinearity problem when directly applying the partial likelihood estimation to the Cox model. To address this problem, Tibshirani[24] estimates the parameters of the Cox model under the L_1 constraint as follows:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}, s} l(\boldsymbol{\beta}), \text{ subject to } \sum_{j=1}^{m} |\beta_j| \le s,$$

where $l(\cdot)$ is the partial likelihood function [19].

The above optimization problem is suitable for reducing the dimension of covariates but is computationally difficult because the L_1 objective function is not differentiable. To address this computational problem, previous studies have proposed many algorithms [13,24-26]. Tibshirani[24] proposes an algorithm using quadratic programming within an iterative procedure. Gui and Li[25] propose an LAS-Cox procedure applying the Choleski decomposition and the LARS procedure. However, these algorithms can be computationally burdensome and sometimes fail to converge to an optimum because they involve quadratic programming and/or matrix inversions. Sohn et al. [13] propose glcoxph for the Cox model by using the gradient lasso algorithm in [27]. This glcoxph employs a coordinate-wise gradient decent with a deletion step and requires only univariate optimization in each iteration. Its convergence speed is almost independent of the number of input variables, and it does not require a matrix inversion, which makes it scalable to high-dimensional data and allows it to converge to a global optimum faster. glmpath [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] provided a comparative analysis using real and simulated data and show that the gradient lasso algorithm outperforms glmpath 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**: $\boldsymbol{\beta} = \mathbf{0}$ and $k = 0$.
2. **Do** until convergence
   (a) **Addition:**
      (i) Compute the gradient $\nabla l(\boldsymbol{\beta})$.
      (ii) Find the $\hat{j}$ maximizing $|\partial l(\boldsymbol{\beta})/\partial\beta_j|$ for $j = 1, \ldots, p$ and $\hat{\gamma} = -s \times \text{sign}(\partial l(\boldsymbol{\beta})/\partial\beta_{\hat{j}})$.
      (iii) Let $\boldsymbol{\nu}$ be a p-dimensional vector such that its $\hat{j}$-th element $\hat{\gamma}$ and other elements are zeros.
      (iv) Find $\hat{\alpha} = \arg\min_{\alpha \in [0,1]} l\left((1-\alpha)\boldsymbol{\beta} + \alpha\boldsymbol{\nu}\right)$.
      (v) Update $\boldsymbol{\beta} = (1-\hat{\alpha})\boldsymbol{\beta} + \hat{\alpha}\boldsymbol{\nu}$.
   (b) **Deletion:**
      (i) Calculate $\boldsymbol{h}_\sigma = -\nabla l(\boldsymbol{\beta}_\sigma) + \theta_\sigma \nabla l(\boldsymbol{\beta}_\sigma)^T \theta_\sigma / |\sigma|$, where $\sigma = \{j : \beta_j \neq 0\}$.
      (ii) Find $\hat{\delta} = \arg\min_{\delta \in [0, U]} l(\boldsymbol{\beta} + \delta\boldsymbol{h})$, where $\boldsymbol{h} = \begin{pmatrix} \boldsymbol{h}_\sigma \\ \mathbf{0} \end{pmatrix}$ and $U = \min_{k \in \sigma}\{-\beta_k/h_k : \beta_k h_k < 0\}$.
      (iii) Update $\boldsymbol{\beta} = \boldsymbol{\beta} + \hat{\delta}\boldsymbol{h}$.
   (c) Set $m = m + 1$.
3. **Return** $\boldsymbol{\beta}$.

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 $\{(X_i, \delta_i), i = 1, ..., n\}$ and SNP data $\{(s_{i1}, ..., s_{im}), i = 1, ..., n\}$, where s_{ij} denotes the number of B alleles for SNP j $(= 1, ..., m)$.
2. For SNP j $(= 1, ..., m)$, calculate the variance and covarince matrix $\hat{\Sigma}_j$, and generate the null distribution of the MaxTest as follows.
   (a) For $b = 1, ..., B$ $(= 100,000$, 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 a α $(= 0.01$, 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 $\bar{z}_j = n^{-1} \sum_{i=1}^{n} z_{ij}$ and $s_j^2 = n^{-1} \sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2$.
8. Apply the gradient lasso to the Cox regression model with response data $\{(X_i, \delta_i), i = 1, ..., n)$ 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 $\epsilon_1, ..., \epsilon_m \sim$ 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 ($f_1 + f_2 + f_3 = 1$) 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

$$\Lambda(t) = \Lambda_0(t) \exp\left(\sum_{j=1}^{D} \beta_j z_{ij}\right),$$

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 $(f_1, f_2, f_3) = (.25, .5, .25)$. 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 SNPs 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,

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	ρ	Genetic model	Mean number of selected SNPs	Mean number of 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 (0.0028)
		Dominant	6.38	3.74	0.59	0.0011 (0.0011)
		Multiplicative	7.55	4.89	0.65	0.0001 (<0.0001)

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 (β), the mean number of SNPs and that of prognostic SNPs selected by the proposed method at ρ=0 and 30% censoring is shown in Table 2. The mean number of prognostic SNPs a little bit increases as β increase and the mean number of prognostic SNPs increases as n increases. The recovery rate increases as β or n increases.

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 ρ = 0 and censoring = 30%

n	β	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.

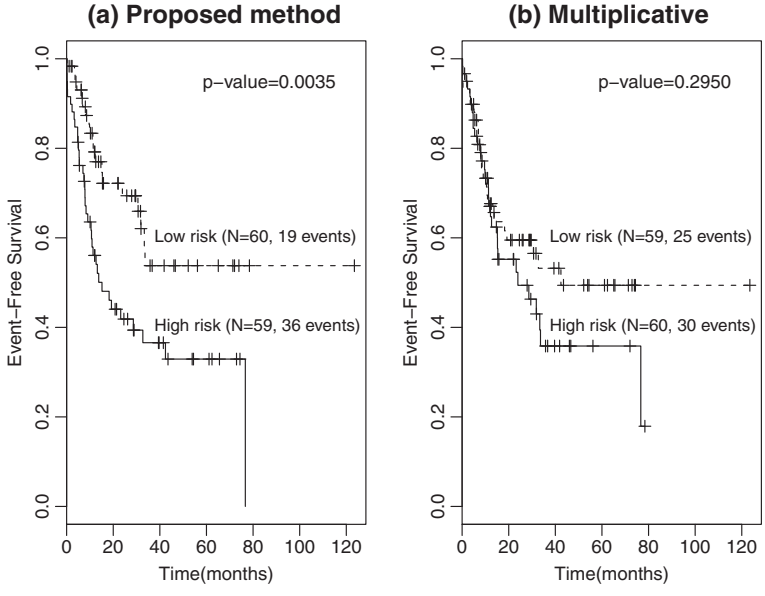

Figure 1 Kaplan-Meier curves for high- and low-risk groups classified by the LOOCV procedure.

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
rs13333329	16	1695776	CRAMP1L	3	0.00015	117
rs2132183	3	84966867	LOC440970,CADM2	3	0.00149	117
rs1950400	14	27105035	MIR4307,NOVA1	2	0.00040	115
rs2155777	11	133290007	OPCML	3	0.00142	113
rs1677914	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,ACTC1	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	72888187	TRHDE	2	0.00070	52
rs16852300	2	167414424	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 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: **Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies.** *Nat Genet* 2006, **38**(2):209–213.
5. Sluis SVD, Dolan CV, Neale MC, Posthuma D: **Power calculations using exact data simulation: a useful tool for genetic study designs.** *Behav Genet* 2008, **38**(2):202–211.
6. Westfall PH, Young SS: *Resampling-based Multiple, Testing: Examples and Methods for Pvalue Adjustment*. New York: Wiley; 1993.
7. Storey JD: **A direct approach to false discovery rates.** *J R Stat Soc, Ser B* 2002, **64:**479–498.
8. Zheng G, Freidlin B, Gastwirth JL: **Comparison of robust tests for genetic association using case-control studies.** *IMS Lecture Notes-Monograph Series 2nd Lehmann Symposium - Optimality* 2006, **49:**253–265.
9. Li Q, Zheng G, Li Z, Yu K: **Efficient approximation of p-values of the maximum of correlated tests, with applications to genome-wide association studies.** *Ann Human Genet* 2008, **72:**397–406.
10. Bair E, Tibshirani R: **Semi-supervised methods to predict patient survival from gene expression data.** *PLoS Biol* 2004, **2:**511–522.
11. Gui J, Li H: **Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.** *Bioinformatics* 2005, **21:**3001–3008.
12. Kaderali L, Zander T, Faigle U, Wolf J, Schultze JL, Schrader R: **CASPAR: a hierarchical Bayesian approach to predict survival times in cancer from gene expression data.** *Bioinformatics* 2006, **22:**1495–1502.
13. Sohn I, Kim J, Jung SH, Park C: **Gradient lasso for Cox proportional hazards model.** *Bioinformatics* 2009, **25:**1775–1781.
14. Kooperberg C, LeBlanc M, Obenchain V: **Risk Prediction Using Genome-Wide Association Studies.** *Genet Epidemiol* 2010, **34:**643–652.
15. Owzar K, Li Z, Cox N, Jung SH: **Power and sample size calculations for SNP association Studies with censored time-to-event outcomes.** *Genet Epidemiol* 2012, **36:**538–548.
16. Conneely KN, Boehnke M: **So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests.** *American J Hum Genet* 2007, **81:**1158–1168.
17. 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.
18. 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:**e1000130.
19. Cox DR: **Regression Models and Life Tables (with Discussion).** *J R Stat Soc, Ser B* 1972, **34:**187–220.
20. Fleming TR, Harrington DP: *Counting Processes and, Survival Analysis*. New York: Wiley; 1991.
21. Freidlin B, Zheng G, Li Z, Gastwirth JL: **Trend tests for case-control studies of genetic markers: power, sample size and robustness.** *Hum Hered* 2002, **53**(3):146–152.
22. Jung SH, Hui S: **Sample size calculations to compare K different survival distributions.** *Lifetime Data Anal* 2002, **8:**361–373.
23. Jung SH, Owzar K, George SL: **A multiple testing procedure to associate gene expression levels with survival.** *Stat Med* 2005, **24:**3077–3088.
24. Tibshirani R: **The lasso method for variable selection in the Cox model.** *Stat Med* 1997, **16:**385–395.
25. Gui J, Li H: **Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.** *Bioinformatics* 2005, **21:**3001–3008.
26. Park MY, Hastie T: **L1 regularization path algorithm for generalized linear models.** *J R Stat Soc B* 2007, **69:**659–677.
27. Kim J, Kim Y, Kim Y: **A gradient-based optimization algorithm for lasso.** *J Comput Graph Stat* 2008, **17:**994–1009.
28. Choi H, Jung C, Kim S, Kim H-J, K, Y-K, Kim T, Zhang Z, Shin E-S, Lee J-E, Sohn SK, Moon JH, Kim SH, Kim KH, Mun Y-C, Kim H, Park J, Kim J, Kim D: **Genome-wide genotype-based risk model for survival in acute myeloid leukemia patients with normal karyotype.** 2012. In submition.
29. Tan XL, Moyer AM, Fridley BL, Schaid DJ, Niu N, Batzler AJ, Jenkins GD, Abo RP, Li L, Cunningham JM, Sun Z, Yang P, Wang L: **Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy.** *Clin Cancer Res* 2011, **17:**5801–5811.
30. Berman DM, Gilman AG: **Mammalian RGS proteins: barbarians at the gate.** *J Biol Chem* 1998, **273**(3):1269–1272.
31. Maity 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 GTPase-activating protein activity.** *J Biol Chem* 2011, **286**(2):1409–1419.
32. Berman DM, Wang Y, Liu Z, Dong Q, Burke LA, Liotta LA, Fisher R, Wu X: **A functional polymorphism in RGS6 modulates the risk of bladder cancer.** *Cancer Res* 2004, **64**(18):6820–6826.
33. Baines CP, Molkentin JD: **Adenine nucleotide translocase-1 induces cardiomyocyte death through upregulation of the pro-apoptotic protein Bax.** *J Mol Cell Cardiol* 2009, **46**(6):969–977.
34. Malorni W, Farrace MG, Matarrese P, Tinari A, Ciarlo L, Mousavi-Shafaei P, D'Eletto M, Di Giacomo G, Melino G, Palmieri L, Rodolfo C, Piacentini M: **The adenine nucleotide translocator 1 acts as a type 2 transglutaminase substrate: implications for mitochondrial-dependent apoptosis.** *Cell Death Differ* 2009, **16**(11):1480–1492.
35. Chen L, Su L, Li J, Zheng Y, Yu B, Yu Y, Yan M, Gu Q, Zhu Z, Liu B: **Hypermethylated FAM5C and MYLK in serum as diagnosis and pre-warning markers for gastric cancer.** *Dis Markers* 2012, **32**(3):195–202.
36. Kuroiwa T, Yamamoto N, Onda T, Shibahara T: **Expression of the FAM5C in tongue squamous cell carcinoma.** *Oncol Rep* 2009, **22**(5):1005–1011.

doi:10.1186/1471-2105-14-58
Cite this article as: Kim *et al.*: Prediction of a time-to-event trait using genome wide SNP data. *BMC Bioinformatics* 2013 **14**:58.