Parallel repulsive logic regression with biological adjacency

DAISUKE YONEOKA*

*Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital,
262 Danny Thomas Place, Mail Stop 735, Memphis, TN 38105, USA*
blue.sky.sea.dy@gmail.com

CINDY IM

*School of Public Health, University of Alberta, 3-300 Edmonton Clinic Health Academy, 11405 - 87 Ave,
Edmonton, AB T6G 1C9, Canada*

YUTAKA YASUI

*Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital,
262 Danny Thomas Place, Mail Stop 735, Memphis, TN 38105, USA*

SUMMARY

Logic regression, an extension of generalized linear models with Boolean combinations of binary variables as predictors, is a useful tool in exploring interactions among single-nucleotide polymorphisms (SNPs) in genome-wide association studies. However, since the search space defined by all possible combinations of SNPs, their complements, and logical operators in Boolean expressions can be exceedingly large in such studies, objective function optimization is slow and likely to be trapped in many local solutions, resulting in model over-fitting. We introduce a new search algorithm, parallel repulsive logic regression (PRLR), to efficiently estimate parameters of a logic regression to find a best model within the large space of SNP interactions by incorporating: (i) relevant biological adjacency matrix between SNPs to define similarity of estimation paths or trees, which are derived from physical SNP positions on chromosomes and/or memberships in biological gene pathways; and (ii) two repulsive forces to counter the similarity between and within estimation paths considered in parallel, which are introduced as penalty terms in the objective function. We compare our method's performance for identifying biologically-meaningful SNP interactions through simulations and with real genetic-epidemiological data. PRLR's detection-accuracy measures outperform existing approaches, especially in terms of positive predictive value and sensitivity for detecting SNP–SNP interactions.

Keywords: Biological adjacency; Childhood cancer; Genome-wide association studies; Logic regression; Repulsive force; Simulated annealing.

1. INTRODUCTION

Single-nucleotide polymorphisms (SNPs) are relatively common variants at single nucleotides which account for approximately 0.1% of the human genome. One of the study goals of genetic association

*To whom correspondence should be addressed.

studies is to detect interactions between SNPs that are associated with an outcome, using a limited number of study samples. In this article, we focus on the logic regression methodology (Ruczinski *and others*, 2003; Kooperberg and Ruczinski, 2005), which has been applied to detect interactions between SNPs associated with outcomes in different settings (Ruczinski *and others*, 2001; Fritsch and Ickstadt, 2007; Schwender and Ickstadt, 2007).

Logic regression considers predictors that are binary and is used to find Boolean combinations of predictors that are associated with an outcome variable of interest (Ruczinski *and others*, 2001, 2003). Logic regression generally includes K logic trees, each of which corresponds to a Boolean expression of binary predictors. Each tree contains several leaves; each leaf indicates a binary predictor or its complement, and each knot of the tree is a logical operator, *AND* or *OR*. For the estimation of logic regression model predictors and their associated parameters, Ruczinski *and others* (2003) proposed a simulated annealing (SA) method with the properties of a Markov chain, for alternating, splitting, deleting a leaf or an operator, or growing or pruning a tree branch as a step in the optimization procedure. From the Bayesian perspective, Kooperberg and Ruczinski (2005) proposed Monte Carlo logic regression (MCLR) to sample from the posterior distribution of logic regression models by using the Markov Chain Monte Carlo (MCMC) technique. MCLR uses a geometric prior distribution for the number of trees to control the model size. Fritsch and Ickstadt (2007) extended MCLR to full Bayesian logic regression (FBLR) by incorporating a prior distribution for regression coefficients, which may be chosen to be Gaussian to regularize how much larger models are penalized. These methods have succeeded in detecting credible SNP interactions associated with outcomes (Ruczinski *and others*, 2001, 2003; Kooperberg and Ruczinski, 2005; Fritsch and Ickstadt, 2007; Schwender and Ickstadt, 2007).

However, applying these methods to high-dimensional data, e.g., data evaluated in genome-wide association studies (GWAS) which include millions of SNPs, is challenging. Logic regression in the GWAS context requires searching a large space defined by all possible combinations of SNP predictors, their complements, and logical operators in Boolean expressions. One practical approach entails restricting the large search space by considering “candidate genes” only. Kooperberg and Ruczinski (2005) restricted the search space to 89 SNPs in 62 candidate genes that were previously associated with heart disease. Schwender and Ickstadt (2007) used logic regression to analyze 40 SNPs belonging to the DNA repair pathway or the xenobiotic and drug metabolism pathway. Clark *and others* (2007) limited the search space to a candidate haplotype block where there is no genetic recombination and SNPs are in high linkage disequilibrium with each other.

In contrast to “candidate gene” studies, we consider GWAS without restricting the search space for SNPs. In this case, the number of SNPs under consideration is large and it is difficult to search the entire space within a limited time. It is also likely that the search may be trapped in several local solutions due to the explosion of combinations as the number of SNPs in the model increases. To address this issue, this study introduces two ideas: (i) biological adjacency, to define the similarity of estimation paths and trees; and (ii) parallel implementation, to consider repulsive forces generated by the biological adjacency-based similarities between and within several estimation paths. The motivation for (i) is to introduce external biological information to aid the stochastic search of the large space of SNPs. In the original implementation of logic regression, all possible moves in the space of binary predictors can be considered equally: there is no metric for the space of binary predictors. However, SNPs are ordered on the genome with respect to their physical positions, and SNPs that are proximal tend to have correlated genotypes. This correlation between SNPs may not only reflect patterns of inheritance and selection over time, but may also signal that such SNPs are involved in interactions that regulate the expression of genes (Phillips, 2008). Furthermore, individual SNPs can be mapped to genes based on their physical positions, and genes may belong to biological pathways that are presented in archives such as Gene Ontology

(Ashburner and others, 2000) and KEGG (Kanehisa and others, 2014). SNPs of genes that belong to the same biological pathway are likely to have related biological functions. Thus, the biological information that links two SNPs can be informative for defining their potential similarity, which can be used as the metric of the large search space. It would be advantageous to incorporate such external biological knowledge in logic regression modeling (Afreixo and others, 2009). Our second innovation (ii), which we will refer to as parallel repulsive logic regression (PRLR), is designed to explore SNP–SNP interactions effectively, especially when the number of SNPs is large. It is well known that once a Markov chain is trapped in a local region, it is difficult to escape from that region within a finite time, especially when the dimension of the search space is large. To avoid this problem, the PRLR algorithm parallelly processes several estimation paths for a logic regression model with repulsive forces that interact with each other in each step of the Markov chain during the optimization process. This repulsive system imposes a stronger penalty on closer (or similar) estimation paths or trees based on the biological adjacency defined in (i). Due to the repulsive system present in the parallel runs, different estimation paths can explore dissimilar regions in the large space of SNPs compared with a single logic regression model run.

The rest of the article is organized as follows. In Section 2, we explain the idea of biological adjacency induced by the physical locations of SNPs and biological gene pathways. The biological weighted adjacency matrix controls the similarity of estimation paths and trees during the optimization process. Then, we propose the PRLR algorithm. In Section 3, the performance of PRLR is evaluated with simulation studies and is compared with existing approaches. PRLR is illustrated through the use of genetic-epidemiological data for bone mineral density in Section 4. Lastly, Section 5 provides discussion, including limitations.

2. METHODS

In this section, we first introduce the logic regression model, and then discuss defining biological adjacency based on the physical locations of SNPs and biological gene pathways. Next, we extend logic regression to PRLR for use in GWAS. For the efficient search of the large space of SNPs, we introduce two repulsive forces that use biological adjacency: (F1), the repulsive force between parallel estimation paths for a logic regression model; and (F2), the repulsive force between trees within a current estimation path. Under this repulsive system, PRLR can explore SNP–SNP interactions more efficiently compared with conventional logic regression approaches.

2.1. Logic regression model

Logic regression estimates regression coefficients and Boolean combinations of binary predictors simultaneously (Ruczinski and others, 2003; Fritsch and Ickstadt, 2007). For example, a Boolean expression can be expressed as $X_{(1)} \wedge X_{(2)}^c \vee X_{(6)} \wedge X_{(4)}^c$, where $X_{(j)}$ is a binary variable ($X_{(j)} \in \{0, 1\}$) indicating a SNP at location j ($j = 1, \dots, P$), $\wedge$ and $\vee$ denote *AND* and *OR* logical operators, respectively, and $X_{(j)}^c$ denotes the complement of $X_{(j)}$. Figure 1 shows the tree-type representation of the Boolean expression. A logic regression model takes the generalized linear model form $g(E[Y]) = \beta_0 + \sum_{k=1}^K \beta_k L_k$, where L_k is a tree (i.e., a Boolean expression), $g()$ is an appropriate link function, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_K)^T$ is a vector of unknown regression coefficients, and K is the number of trees (i.e., the number of Boolean expressions) in the model. We consider the case of the binary outcome variable ($Y \in \{0, 1\}$), and therefore, g takes the logistic form $\text{logit}(x) = \log(x/(1-x))$. There are several methods to estimate the regression parameters $\boldsymbol{\beta}$ and the Boolean expressions $\mathbf{L} = (L_1, \dots, L_K)^T$, including the SA method (Ruczinski and others, 2003) and the MCMC method (Kooperberg and Ruczinski, 2005; Fritsch and Ickstadt, 2007). In implementing PRLR, we adopt the SA method for regression coefficient and logic tree estimation.

Fig. 1. Tree-type representation of $X_1 \wedge X_2^c \vee X_6 \wedge X_4^c$.

2.2. Biological adjacency of SNPs

To define the biological adjacency of SNPs, assume that there are P SNPs ordered on the DNA sequence with respect to their physical locations and G genes correlated each other, which are illustrated by the pathway-network of genes as shown in Figure 2 (upper left), where $P = g_5$ and $G = 5$. Each individual SNP is assumed to be mapped to one gene. We use the explicit expression $X_{(j)}^{(g)}$ ($j = 1, \dots, P_g, g = 1, \dots, G$) when it is useful to emphasize the membership of j th SNP in the g th gene. Based on these notations, the biological adjacency can be divided into two parts: (i) the adjacency based on the physical locations of SNPs; and (ii) the adjacency based on underlying biological gene pathways. Overall, the biological adjacency can be considered as the metric of the large search space to define the similarity of estimation paths and trees.

The first biological adjacency, defined by the physical locations of SNPs, can be formulated by using a weighting function based on two SNP locations s and s' . The weighting function $C_{g,g'}(s, s')$ characterizes the spatial distance and takes isotropic form as $C_{g,g'}(s, s') = C_{g',g}(s', s)$. This weighting function represents biological similarity/redundancy of the two SNPs due to physical proximity (e.g., correlation of their genotypes and functions). In this study, we adopt the Gaussian class of the weighting function with an additional parameter σ as

$$C_{g,g'}(s, s', \sigma) = \begin{cases} \exp\left(-\frac{|s - s'|^2}{\sigma^2}\right) & (\text{if } g = g') \\ 0 & (\text{otherwise}). \end{cases}$$

Alternatively, values of $C_{g,g'}(s, s')$ can be specified directly from available biological data, e.g., from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), instead of treating them as a function of physical distance. Given the weighting function $C_{g,g'}(s, s', \sigma)$, the restricted weighted adjacency matrix for the g th gene, Σ_{gg} , is given by

$$\Sigma_{gg} = \begin{pmatrix} 0 & C_{g,g}(s_1, s_2, \sigma) & \dots & C_{g,g}(s_1, s_{P_g}, \sigma) \\ C_{g,g}(s_2, s_1, \sigma) & 0 & \dots & C_{g,g}(s_2, s_{P_g}, \sigma) \\ \vdots & \vdots & \ddots & \vdots \\ C_{g,g}(s_{P_g}, s_1, \sigma) & C_{g,g}(s_{P_g}, s_2, \sigma) & \dots & 0 \end{pmatrix}.$$

By stacking Σ_{gg} to form a block diagonal matrix, the restricted weighted adjacency matrix Σ_{re} can be constructed as shown in Figure 2 (lower left).

Fig. 2. Upper: Pathway network of genes (left) and its matrix representation N (right). Lower: Biological weighted adjacency matrix Σ constructed by restricted weighted adjacency matrix Σ_{re} and pathway network matrix N .

The second biological adjacency, defined by gene pathways/networks, can be formulated by using a (symmetric) network matrix N . It represents gene relationships in the same biological pathways: SNPs of genes in the same biological pathway likely have related biological functions. Pathway information can be expressed as an undirected graph, where nodes and edges indicate the genes and the dependencies of the genes, respectively. Based on the graph, we can set up the adjacent matrix N (Figure 2 (upper)). The simplest form for the elements of N is a scalar constant for all SNPs between the two genes if the graph indicates the two genes are in the same pathway; note that other types of gene network representation, such as incidence, Laplacian, and Tutte matrices, can be used. Lastly, by aggregating these two matrices (Σ_{re} and N) and adding a smoothing parameter $\gamma \geq 0$ to regulate the relationship between Σ_{re} and N , the biological weighted adjacency matrix Σ can be obtained (Figure 2 (lower)) as $\Sigma = \Sigma_{re} + \gamma N =$

$\text{diag}(\mathbf{\Sigma}_{11}, \dots, \mathbf{\Sigma}_{GG}) + \gamma N$, where $\text{diag}(\mathbf{\Sigma}_{11}, \dots, \mathbf{\Sigma}_{GG})$ indicates a block diagonal matrix whose diagonal entries are $\mathbf{\Sigma}_{11}, \dots, \mathbf{\Sigma}_{GG}$.

2.3. Parallel repulsive logic regression (PRLR) algorithm

Based on the biological weighted adjacency matrix $\mathbf{\Sigma}$, we estimate the parameters in the logic regression model of Section 2.1 by the proposed PRLR algorithm. Consider M attempts (estimation paths) of parameter estimation of the logic regress model ($m = 1, \dots, M$), N samples ($i = 1, \dots, N$), P SNPs ($j = 1, \dots, P$), and a set of observed data $\mathcal{D} = \{X_{(1)i}, \dots, X_{(P)i}, Y_i\}_{i=1}^N$. Unless otherwise stated, for simplicity, the subscripts indicating SNP locations are omitted in the following sections (i.e., $X_{(j)}$ is written as X). In the m th estimation path, there are K^m trees, which are denoted as $\mathbf{L}^m = (L_1^m, \dots, L_{K^m}^m)^T$. Each tree L_k^m ($k = 1, \dots, K^m$) has l_k^m leaves and consists of a set of binary variables $\mathbf{X}_k^m = (X_{k,1}^m, \dots, X_{k,l_k^m}^m)^T$, a set of complements $\mathbf{C}_k^m = (C_{k,1}^m, \dots, C_{k,l_k^m}^m)^T$ where $C_{k,l}^m = \{\cdot, \cdot^c\}$ ($l = 1, \dots, l_k^m$), and a set of operators $\mathbf{O}_k^m = (O_{k,1}^m, \dots, O_{k,l_k^m-1}^m)^T$ where $O_{k,l}^m = \{\wedge, \vee\}$ ($l = 1, \dots, l_k^m - 1$). As a whole, the configuration of the k th tree of the m th estimation path is

$$\phi(\mathbf{L}_k^m) = (X_{k,1}^m, C_{k,1}^m, O_{k,1}^m, X_{k,2}^m, \dots, O_{k,l_k^m-1}^m, X_{k,l_k^m}^m, C_{k,l_k^m}^m),$$

where $\phi(\cdot)$ is an operator to indicate the configuration of the k th tree.

Since individual samples are assumed independent, the objective function as the m th likelihood function over a dataset of N samples is given by

$$\begin{aligned} & \mathcal{L}_m(\boldsymbol{\beta}^m, \mathbf{L}^m | \mathcal{D}) \\ &= \prod_{i=1}^N \left[\left\{ \frac{\exp(\beta_0^m + \sum_{k=1}^{K^m} \beta_k^m L_{k,i}^m)}{1 + \exp(\beta_0^m + \sum_{k=1}^{K^m} \beta_k^m L_{k,i}^m)} \right\}^{Y_i} \left\{ 1 - \frac{\exp(\beta_0^m + \sum_{k=1}^{K^m} \beta_k^m L_{k,i}^m)}{1 + \exp(\beta_0^m + \sum_{k=1}^{K^m} \beta_k^m L_{k,i}^m)} \right\}^{1-Y_i} \right], \end{aligned} \quad (2.1)$$

where $\boldsymbol{\beta}^m = (\beta_0^m, \beta_1^m, \dots, \beta_{K^m}^m)^T$. Since the logic regression model uses the generalized linear modeling framework, it is possible to extend our method for other types of response variables by changing this likelihood.

The proposed PRLR algorithm is developed to search the space more comprehensively by performing an all-at-once iterating parallel run. PRLR performs parallel M Markov chains for the estimation of the regression coefficients $\boldsymbol{\beta}^m$ and the Boolean expressions $\mathbf{L}^m$ (the estimation procedure is described in Section 2.3.2). Our method simultaneously performs M parameter estimation ($\{\boldsymbol{\beta}^m, \mathbf{L}^m\}_{m=1}^M$) to find the best one. There are repulsive forces (F1) between M estimation paths and (F2) between K^m trees within the m th path so that each trajectory of Markov chains can search different regions efficiently by avoiding overlap (Figure 3). To estimate the m th sets of parameters, the PRLR algorithm is derived by solving the optimization problem of the objective function for logic regression (i.e., the likelihood function in equation (2.1)) with restrictions:

$$\begin{aligned} & \min_{\boldsymbol{\beta}^m, \mathbf{L}^m} \{-\log \mathcal{L}_m(\boldsymbol{\beta}^m, \mathbf{L}^m | \mathcal{D})\} \quad \text{subject to} \quad \underbrace{\psi_1(\mathbf{L}^m | \mathbf{L}^{(-m)}) \geq t_1}_{\text{(F1) Repulsive force between estimation paths}} \quad \text{and} \\ & \underbrace{\psi_2(L_1^m, \dots, L_{K^m}^m) \geq t_2}_{\text{(F2) Repulsive force within a estimation path}}, \end{aligned} \quad (2.2)$$

Fig. 3. A schematic picture of PRLR model.

where $\mathbf{L}^{(-m)}$ is the set of trees except those in the m th estimation path, $\psi_1(\mathbf{L}^m | \mathbf{L}^{(-m)})$ is a function for (F1) representing the similarity of the tree-structures between the m th estimation path and other paths, $\psi_2(L_1^m, \dots, L_{K^m}^m)$ is a function for (F2) representing the similarity of the tree-structures within the m th path, and $t_1, t_2 > 0$ can regularize the degrees of the two repulsive forces (F1) and (F2). In this study, the repulsive forces (F1) and (F2) are modeled by

$$(F1) : \psi_1(\mathbf{L}^m | \mathbf{L}^{(-m)}) = \max_{m' \neq m} D_1(\mathbf{L}^m, \mathbf{L}^{m'}) = \max_{m' \neq m} 2 \sum_{i \in \mathcal{X}^m} \sum_{j \in \mathcal{X}^{m'}} \Sigma[i, j],$$

$$(F2) : \psi_2(L_1^m, \dots, L_{K^m}^m) = \max_k \max_{k' \neq k} D_2(L_k^m, L_{k'}^m) = \max_k \max_{k' \neq k} 2 \sum_{i \in \mathcal{X}_k^m} \sum_{j \in \mathcal{X}_{k'}^m} \Sigma[i, j],$$

where $\mathcal{X}^m$ and $\mathcal{X}_k^m$ are a set of SNPs included in the m th estimation path and in k th tree of the m th estimation path, respectively, and $\Sigma[i, j]$ is the $[i, j]$ element of the biological weighted adjacency matrix Σ .

From duality and Karush–Kuhn–Tucker conditions, we can rewrite the problem in the penalized forms of the objective function as

$$\min_{\beta^m, \mathbf{L}^m} \left[-\log \mathcal{L}_m(\beta^m, \mathbf{L}^m | \mathcal{D}) + \lambda_1 \max_{m' \neq m} D_1(\mathbf{L}^m, \mathbf{L}^{m'}) + \lambda_2 \max_k \max_{k' \neq k} D_2(L_k^m, L_{k'}^m) \right] \quad (2.3)$$

with regularization parameter $\lambda_1, \lambda_2 \geq 0$. By minimizing the objective function (i.e., restricted negative log-likelihood function), the (weighted) maximum likelihood estimates of β^m and $\mathbf{L}^m$ can be obtained. Given the biological weighted adjacency matrix Σ , the first regularization term, which corresponds to (F1), imposes a larger penalty to the objective function for a estimation path that is closer to other estimation paths. The second term, which corresponds to (F2), imposes a larger penalty for a set of trees that are

closer to the current trees within the m th estimation path. The parameters λ_1 and λ_2 control the strengths of the repulsive forces: the larger λ s produce stronger repulsion forces, and if $\lambda_1 = \lambda_2 = 0$, the PRLR reduces to the standard logic regression model. Note that there is no repulsive force between leaves within a tree: l_k^m leaves in the k th tree of the m th model can freely search all possible Boolean expressions within the restrictions induced by the repulsive forces. Due to the repulsive system, our method encourages a search of dissimilar areas defined by the biological weighted adjacency matrix.

2.3.1. Additional penalty for tree complexity. Multiple studies have reported model over-fitting when fitting a single logic regression model (Ruczinski and others, 2003; Kooperberg and Ruczinski, 2005; Fritsch and Ickstadt, 2007). To avoid over-fitting with PRLR (i.e., to avoid picking up many irrelevant leaves and trees), a penalty for tree-complexity can be considered in each model, similar to the pruning process used in regression tree theory (Mansour, 1997). Specifically, we include another penalty term for the tree-complexity, measured by the average number of leaves per tree with $\alpha \geq 0$ as its parameter. As α increases, the optimized logic regression model tends to have a smaller average number of leaves per tree. With this additional penalty for complexities of trees, the equation (2.3) can be written in the extended form as

$$\min_{\beta^m, L^m} \left[-\log \mathcal{L}_m(\beta^m, L^m | \mathcal{D}) + \lambda_1 \max_{m' \neq m} D_1(L^m, L^{m'}) + \lambda_2 \max_k \max_{k' \neq k} D_2(L_k^m, L_{k'}^m) + \alpha \frac{\sum_{k=1}^{K^m} l_k^m}{K^m} \right]. \quad (2.4)$$

Note that it is possible to incorporate a penalty term for the number of trees to avoid over-fitting. We applied here a restriction on the range of K (e.g., $K \leq 5$), which is equivalent to assuming a discrete uniform penalty on K with a restricted support (e.g., $\{1, 2, 3, 4, 5\}$) in our optimization method.

2.3.2. Estimation procedure using the simulated annealing (SA) method. In the optimization procedure to obtain m th set of estimates of β^m and L^m to minimize the objective function (2.4), we apply the SA method by using the properties of a Markov chain as described by Ruczinski and others (2003). In the Markov chain, neighbors of a certain tree should be defined as the trees that can be reached from the tree by a single “move”. As with Ruczinski and others (2003), we consider six possible moves based on the current sets of trees $L^m = (L_1^m, \dots, L_{K^m}^m)^T$:

- (1) *Tree birth*: Adding a new tree $L_{K^m+1}^m$ consisting of randomly chosen SNPs. The number of leaves in the new tree ($l_{K^m+1}^m$) is randomly sampled from $U(1, l_{\max})$, where $U(a, b)$ denotes the discrete uniform distribution with the support defined by the two parameters a (minimal value) and b (maximum value), and $l_{\max}$ is the maximum number of leaves, which should be defined *a priori* by users. In addition, a new set of SNPs ($X_{K^m+1}^m$), complements ($C_{K^m+1}^m$) and operators ($O_{K^m+1}^m$) for the new tree $L_{K^m+1}^m$ are randomly sampled.
- (2) *Tree death*: Removing a randomly chosen tree L_k^m ($k = 1, \dots, K^m$).
- (3) *Leaf birth*: Choosing one tree L_k^m ($k = 1, \dots, K^m$) randomly, and adding one new SNP $X_{(j)} \notin X_k^m$ to a randomly chosen leaf l ($l = 1, \dots, l_k^m$) in the tree. The new SNP $X_{(j)}$, complement and operator are randomly sampled.
- (4) *Leaf death*: Choosing one tree L_k^m ($k = 1, \dots, K^m$), and removing one SNP $X \in X_k^m$ from randomly chosen leaf l ($l = 1, \dots, l_k^m$) in the tree.
- (5) *Change complement*: Choosing one tree L_k^m ($k = 1, \dots, K^m$), and changing an element of complement $C_{l,k}^m$ at randomly chosen position ($1, \dots, l_k^m$) in the tree.

- (6) *Change operator*: Choosing one tree L_k^m ($k = 1, \dots, K^m$), and changing an element of operator $O_{l,k}^m$ at randomly chosen position $(1, \dots, l_k^m - 1)$ in the tree.

First, we generate M parallel estimation paths and K^m trees in the initial m th estimation path: we first decide the number of trees and leaves in a tree, and then randomly sample the sets of SNPs, complements and operators in one tree from a discrete uniform distribution $U(1, P)$ and a binomial distribution $B(0.5)$, respectively. The values of M should be decided based on the data size. At each iteration in the m th estimation path, the objective function is updated sequentially by comparing the score of the objective function between the old and new states. The acceptance probability depends on T , which reflects at which point in time the chain is and is usually referred to as a temperature. The new state is accepted with probability $p = \min\{1, \exp(v_{old} - v_{new})/T\}$, where v is the value of the objective function. Then, different Markov chains and different trees within each Markov chain move towards dissimilar regions of the large space. This property allows us to search a much wider solution space efficiently with an all-at-once interacting parallel simulation than with a single method. Additionally, the SA algorithm requires the starting (highest) temperature, the finishing (lowest) temperature, and the cooling schedule, which determines the length of the chains we run. Example settings for the temperatures and cooling schedule are provided in the simulation section. After Markov chain iterations, we can obtain M sets of estimated parameters $\{\beta^m, L^m\}_{m=1}^M$. Based on these sets, a final set of parameter estimates can be obtained by selecting the set that provides the minimal value of the negative log-likelihood function in equation (2.4).

Lastly, since the goal of logic regression is to select the best model that gives the lowest out-of-sample prediction error for an independent dataset, a cross-validation method can be used to decide the values of the regularization parameters λ_1, λ_2 , and α . An average prediction error $RCV(\lambda_1, \lambda_2, \alpha)$ for all minimizing models can be calculated as a function of λ_1, λ_2 , and α . The optimal parameters $\lambda_{1,CV}, \lambda_{2,CV}$, and α_{CV} produce the lowest $RCV(\lambda_1, \lambda_2, \alpha)$. The regularization parameter σ in the weighting function $C_{g,g'}(s, s', \sigma)$ and the smoothing parameter γ in Σ should be decided *a priori* or by other model selection methods (Tibshirani and Hastie, 1987).

3. SIMULATION STUDY

3.1. Simulation settings

We report on the performance of several algorithms for the discovery of “true” SNP–SNP interaction associations among the large number of possible SNP combinations in a simulation study. We assessed the performance of conventional approaches including standard logic regression (Logic), logistic regression with the Lasso penalty (Lasso), FBLR, and our proposed algorithm (PRLR). The dataset for the simulations contained 2000 SNPs ($j = 1, \dots, 2000$), where each binary SNP covariate was sampled from a binomial distribution $X_j \sim B(\pi_j)$, and π_j was independently sampled from a uniform distribution $U(0, 1)$. The sample size varied from 500 to 5000 ($N = 500, 1000$, or 5000). To make a biological weighted adjacency matrix Σ , the locations of SNPs were consecutively determined by drawing the distance to the next SNP from the uniform distribution $U(1, 100)$ sequentially from the 3' end. Further, it was assumed that there were 10 genes ($g = 1, \dots, 10$), and SNPs were mapped to the 10 genes by $\{\text{SNPs mapped to the } g\text{th gene}; 10(g-1)\% \leq \text{SNP location} - A < 10g\% \text{ of } D\}$, where A is the location of the SNP closest to the 3' end and D is the distance between the SNPs closest and furthest from the 3' end plus $\epsilon < 1$. The variables indicating biological pathways (i.e., $r_{gg'}$ in the network matrix N as in Figure 2 (upper)) took 1 if the g th and g' th genes had a biological relationship, or 0 otherwise. The biological relationships were defined as follows: there is no gene–gene relationship in Scenario 1, 2, and 3, and thus all elements of N were set to 0; there is one gene–gene relationship (gene 3 - 4) in Scenario 4 and thus we put 1 on $N[i, j], i \neq j, (i, j) \in \{S_3, S_4\}$, where $N[i, j]$ indicates the $[i, j]$ element of N and S_i indicates the set of SNPs that are mapped to the i th gene; and we put 0 on all the other elements, and there are three gene–gene

relationship (gene 3 - 4, 3 - 6, and 4 - 6) in Scenario 5 and thus, we put 1 on $N[i, j], i \neq j, (i, j) \in \{S_3, S_4, S_6\}$ and 0 on all the other elements. The binary outcome Y was sampled from the Bernoulli distribution with the probability of $P(Y = 1) = e^\eta / (1 + e^\eta)$. We considered the following five scenarios for η :

Scenario 1: $\eta = 0$

Scenario 2: $\eta = -1 + 5 \times \underbrace{(X_{(505)}^{c(3)} \wedge X_{(511)}^{(3)})}_{\text{Gene}_3}$

Scenario 3: $\eta = -1 + 5 \times \underbrace{(X_{(505)}^{c(3)} \wedge X_{(511)}^{(3)} \vee X_{(494)}^{c(3)} \wedge X_{(519)}^{(3)})}_{\text{Gene}_3}$

Scenario 4: $\eta = -1 + 1 \times \underbrace{(X_{(488)}^{c(3)} \wedge X_{(505)}^{(3)} \vee X_{(519)}^{(3)} \wedge X_{(497)}^{(3)})}_{\text{Gene}_3} + 1 \times \underbrace{(X_{(631)}^{c(4)} \vee X_{(625)}^{(4)} \vee X_{(588)}^{c(4)})}_{\text{Gene}_4}$

Scenario 5: $\eta = -1 + 1 \times \underbrace{(X_{(505)}^{c(3)} \wedge X_{(519)}^{c(3)} \wedge X_{(497)}^{c(3)} \wedge X_{(512)}^{c(3)})}_{\text{Gene}_3} + 1 \times \underbrace{(X_{(613)}^{(4)} \vee X_{(588)}^{c(4)} X_{(618)}^{(4)})}_{\text{Gene}_4}$
 $+ 1 \times \underbrace{(X_{(1070)}^{(6)} \wedge X_{(1080)}^{(6)} \vee X_{(1098)}^{c(6)})}_{\text{Gene}_6},$

where the SNPs in Scenarios 2, 3, 4, and 5 were clustered in regions that corresponded to 0.6%, 1.1%, 3.3%, and 4.2% of the genome length covered by the whole set of SNPs, respectively. By setting the variance parameter σ in the weighting function $C_{g,g'}(s, s')$ and the smoothing parameter γ to 100 and 1, respectively, the biological weighted adjacency matrix Σ was constructed as in Figure 2 (upper).

An advantage of all logic regression-based approaches is that the interaction terms can be created and evaluated automatically as covariates in the model without requiring them to be pre-specified. This is not the case for Lasso. Thus, we did not create or evaluate interactions in Lasso here: rather, we evaluated in Lasso whether the SNPs in the underlying linear predictor were selected or not. While any models including trees of any size can be created theoretically, to limit the time and computations required for the simulation, this simulation's search considered a maximum of 3 trees per model and 5 leaves per tree. For Lasso, we used 10-fold cross-validation to choose values for the smoothing parameter (Friedman and others, 2010). For Logic, we also used 10-fold cross-validation to choose from models with 1 to 3 trees and 1 to 5 leaves per tree. For FBLR, we used a burn-in of 100 000 iterations, proposed by Fritsch and Ickstadt (2007). Other settings were set to the default values in the corresponding R statistical packages (*glmnet* package for Lasso, *LogicReg* package for Logic, and *scrime* package for FBLR). For our approach, PRLR, we ran 5 parallel Markov chains ($M = 5$).

The SA algorithm requires the starting temperature and cooling schedule: we started the first Markov chain at a temperature of 10^4 and ended the run at 10^{-2} as with Ruczinski and others (2003). At each fixed temperature, 10 000 moves were implemented. The temperature was decreased after each run of the Markov chain in equal increments on a $\log(10)$ scale. To speed up the SA algorithm, we kept track of the number of accepted moves. If the number reached 500, we exited the Markov chain and decreased the temperature to its next value. The regularization parameters λ_1 , λ_2 , and α were selected by 10-fold cross-validation (Supplementary materials S1 available at *Biostatistics* online). Our detailed R codes for PRLR are available in a GitHub repository (<https://github.com/kingqwert/R/tree/master/PRLR>).

The performance of selected approaches was evaluated using the following metrics:

- (i) Given the true set of SNPs $\mathcal{S}$, true positive (TP), false positive (FP), positive predictive value (PPV_{snp}) (or precision), and sensitivity (SN_{snp}) (or recall) were evaluated. PPV_{snp} and SN_{snp}

were computed as

$$\text{PPV}_{\text{snp}} = \frac{\# \text{ of SNPs in the proposed approach overlapped by } \mathcal{S}}{\# \text{ of SNPs in the proposed approach}}$$

$$\text{SN}_{\text{snp}} = \frac{\# \text{ of SNPs in the proposed approach overlapped by } \mathcal{S}}{\# \text{ of SNPs in the true set } \mathcal{S}}.$$

A lower value of PPV_{snp} indicates lower prediction accuracy for true underlying SNPs and a lower value of SN_{snp} represents weaker ability to discover true underlying SNPs.

- (ii) Given the true set of all two-SNP interactions in $\mathcal{S}$, which was denoted as $\mathcal{S}_{\text{int}}$, TP, FP, positive predictive value, and sensitivity of two-SNP interactions were evaluated (TP_{int} , FP_{int} , PPV_{int} , and SN_{int}). They were defined as

$$\text{PPV}_{\text{int}} = \frac{\# \text{ of interactions in the proposed approach overlapped by } \mathcal{S}_{\text{int}}}{\# \text{ of interactions in the proposed approach}}$$

$$\text{SN}_{\text{int}} = \frac{\# \text{ of interactions in the proposed approach overlapped by } \mathcal{S}_{\text{int}}}{\# \text{ of interactions in } \mathcal{S}_{\text{int}}}.$$

- (iii) To evaluate $\text{PPV}_{\text{snp(int)}}$ and $\text{SN}_{\text{snp(int)}}$ simultaneously, F-value, the harmonic mean of $\text{PPV}_{\text{snp(int)}}$, and $\text{SN}_{\text{snp(int)}}$, were then calculated.
- (iv) Area under curve (AUC), sensitivity ($\text{SN}_{\text{subject}}$), and specificity ($\text{SP}_{\text{subject}}$), which measure the prediction accuracy for the outcome Y , were evaluated in a newly sampled validation dataset. The validation dataset, the sample size of which was 1000, was randomly sampled in the same manner as the simulation dataset.

3.2. Simulation results

Table 1–3 show the results of the performance measures of our approach (PRLR), Lasso, Logic, and FBLR for the Scenarios 1, 2, 3, 4, and 5. For Scenario 1 in Table 1, PRLR provided slightly worse or similar performance compared with other existing approaches. In this scenario, standard logic regression worked well. For Scenario 2, PRLR provided the best performance especially in terms of TP, SN_{snp} , and SN_{int} . Similar tendency was observed in the results of Scenario 3, 4, and 5. For Scenario 3 in Table 2, PRLR provided the best performance in many settings, especially in terms of SN_{snp} and SN_{int} . For Scenario 4, almost all the performance measures were always better for PRLR than other approaches. For Scenario 5 in Table 3, although the predictive performance of PRLR was comparable to other approaches and it provided relatively large number of FPs compared with standard logic regression, PRLR showed the best performance with respect to SN_{snp} , PPV_{int} , and SN_{int} . In terms of prediction accuracy in the validation dataset (i.e., AUC, $\text{SN}_{\text{subject}}$, and $\text{SP}_{\text{subject}}$), PRLR was the best in Scenario 3 and 4 in most settings and similar to other approaches in Scenarios 2 and 5. The performance of PRLR improved as the number of samples and SNPs in η increased.

As a whole, the results of the simulation studies reveal that our proposed approach outperformed existing approaches in most settings, especially in terms of SN_{snp} , PPV_{int} , and SN_{int} while maintaining the comparable number of FPs (Tables 1–3), indicating that PRLR has improved ability to discover true underlying SNPs, provided that the biological weighted adjacency of SNPs are appropriately specified. Moreover, PRLR appeared to be superior in efficiently capturing SNP–SNP interactions.

Table 1. Performance of logistic regression with Lasso penalty (Lasso), standard logic regression (logic), full Bayesian logic regression (FBLR), and proposed method (PRLR) for Scenario 1 and 2: mean and standard deviation. Bold character indicates the best performance.

Performance measure	Samples	Scenario 1				Scenario 2			
		Lasso	Logic	FBLR	PRLR	Lasso	Logic	FBLR	PRLR
500	TPsnr	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.971 (0.180)	1.590 (0.500)	1.612 (0.660)	1.990 (0.100)
	FPsnr	7.560 (15.080)	1.300 (0.660)	20.000 (0.060)	4.670 (1.006)	7.190 (0.681)	0.642 (1.342)	18.420 (19.340)	3.400 (0.888)
	PPV snr	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.215 (0.209)	0.714 (0.273)	0.080 (0.067)	0.379 (0.067)
	SNsnr	NA	NA	NA	NA	0.983 (0.090)	0.797 (0.251)	0.799 (0.329)	0.995 (0.050)
	TPint	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	NA	0.550 (0.499)	0.278 (0.470)	0.990 (0.100)
	FPint	NA	0.240 (0.770)	10.010 (0.060)	2.030 (1.579)	NA	0.593 (0.536)	9.651 (9.322)	1.758 (1.422)
	PPVint	NA	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	NA	0.481 (0.482)	0.028 (0.048)	0.466 (0.257)
	SNint	NA	NA	NA	NA	NA	0.550 (0.499)	0.278 (0.470)	0.990 (0.100)
	F-value	NA	NA	NA	NA	0.156 (0.111)	0.372 (0.128)	0.073 (0.030)	0.273 (0.036)
	F-value int	0.600 (0.126)	0.572 (0.025)	NA	0.637 (0.023)	NA	0.247 (0.239)	0.027 (0.046)	0.299 (0.110)
1000	AUC	0.487 (0.418)	0.641 (0.397)	0.575 (0.285)	0.588 (0.190)	0.833 (0.043)	0.754 (0.029)	0.852 (0.032)	0.829 (0.021)
	SNsubject	0.521 (0.405)	0.635 (0.392)	0.662 (0.229)	0.620 (0.181)	0.586 (0.045)	0.897 (0.023)	0.554 (0.072)	0.587 (0.064)
	SPsubject	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.566 (0.044)	0.911 (0.058)	0.635 (0.087)	0.966 (0.040)
	TPsnr	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.990 (0.070)	1.571 (0.510)	1.640 (0.670)	2.000 (0.000)
	FPsnr	8.370 (15.250)	1.402 (0.720)	19.976 (0.180)	4.700 (0.823)	9.030 (0.330)	0.761 (1.301)	18.360 (19.330)	3.410 (1.138)
	PPV snr	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.181 (0.179)	0.673 (0.280)	0.082 (0.034)	0.385 (0.074)
	SNsnr	NA	NA	NA	NA	0.997 (0.036)	0.787 (0.253)	0.818 (0.337)	1.000 (0.000)
	TPint	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	NA	0.526 (0.501)	0.416 (0.583)	1.000 (0.000)
	FPint	NA	0.300 (1.120)	9.974 (0.180)	2.040 (1.222)	NA	0.661 (0.554)	9.489 (9.469)	2.030 (2.380)
	PPVint	NA	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	NA	0.443 (0.475)	0.042 (0.058)	0.496 (0.280)
5000	SNint	NA	NA	NA	NA	NA	0.526 (0.501)	0.416 (0.583)	1.000 (0.000)
	F-value	NA	NA	NA	NA	0.138 (0.106)	0.357 (0.130)	0.074 (0.031)	0.276 (0.039)
	F-value int	0.577 (0.098)	0.549 (0.023)	0.747 (0.031)	0.597 (0.017)	NA	0.230 (0.236)	0.040 (0.056)	0.310 (0.121)
	AUC	0.424 (0.416)	0.669 (0.394)	0.508 (0.335)	0.558 (0.185)	0.812 (0.030)	0.752 (0.025)	0.824 (0.024)	0.804 (0.015)
	SNsubject	0.413 (0.415)	0.678 (0.419)	0.586 (0.268)	0.587 (0.187)	0.581 (0.031)	0.890 (0.014)	0.555 (0.061)	0.541 (0.034)
	SPsubject	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.563 (0.031)	0.919 (0.049)	0.582 (0.052)	0.984 (0.014)
	TPsnr	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.990 (0.110)	1.710 (0.461)	1.550 (0.761)	2.000 (0.000)
	FPsnr	6.371 (11.570)	1.571 (0.790)	19.980 (0.140)	4.704 (1.017)	6.792 (0.472)	0.831 (1.290)	18.450 (19.241)	3.700 (1.418)
	PPV snr	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.226 (0.192)	0.674 (0.261)	0.078 (0.038)	0.370 (0.082)
	SNsnr	NA	NA	NA	NA	0.994 (0.056)	0.856 (0.228)	0.775 (0.381)	1.000 (0.000)
5000	TPint	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	NA	0.588 (0.495)	0.450 (0.501)	1.000 (0.000)
	FPint	NA	0.410 (1.021)	9.980 (0.141)	2.143 (1.753)	NA	0.767 (0.574)	9.550 (9.519)	2.760 (3.230)
	PPVint	NA	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	NA	0.434 (0.463)	0.045 (0.050)	0.450 (0.298)
	SNint	NA	NA	NA	NA	NA	0.588 (0.495)	0.450 (0.501)	1.000 (0.000)
	F-value	NA	NA	NA	NA	0.167 (0.110)	0.365 (0.111)	0.070 (0.035)	0.267 (0.044)
	F-value int	0.535 (0.039)	0.527 (0.010)	NA	0.545 (0.007)	NA	0.230 (0.228)	0.042 (0.048)	0.284 (0.133)
	AUC	0.264 (0.305)	0.778 (0.283)	0.409 (0.226)	0.453 (0.166)	0.778 (0.009)	0.750 (0.029)	0.782 (0.019)	0.782 (0.006)
	SNsubject	0.220 (0.313)	0.815 (0.296)	0.486 (0.255)	0.615 (0.163)	0.574 (0.006)	0.886 (0.004)	0.555 (0.054)	0.530 (0.007)
	SPsubject	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.556 (0.025)	0.924 (0.057)	0.535 (0.012)	0.990 (0.002)

TP, true positive; FP, false positive; PPV, positive predictive value; SN, sensitivity; int, two-SNP interactions; AUC, area under curve.

Table 2. Performance of logistic regression with Lasso penalty (Lasso), standard logic regression (logic), full Bayesian logic regression (FBLR), and proposed method (PRLR) for Scenario 3 and 4: mean and standard deviation. Bold character indicates the best performance.

Performance measure	Samples	Scenario 3					Scenario 4				
		Lasso	Logic	FBLR	PRLR	Lasso	Logic	FBLR	PRLR		
500	TPsn	1.212 (0.464)	0.952 (0.476)	1.112 (0.444)	2.430 (0.832)	0.189 (0.469)	0.070 (0.259)	0.371 (0.567)	1.510 (0.759)		
	FPsn	2.587 (0.897)	0.379 (0.787)	18.745 (19.738)	3.505 (1.190)	10.311 (5.544)	1.116 (0.902)	19.155 (19.683)	1.106 (0.346)		
	PPV sn	0.319 (0.341)	0.715 (0.377)	0.056 (0.022)	0.414 (0.135)	0.018 (0.078)	0.059 (0.223)	0.019 (0.028)	0.521 (0.236)		
	SNsn	0.303 (0.116)	0.238 (0.119)	0.278 (0.111)	0.608 (0.208)	0.027 (0.067)	0.010 (0.037)	0.053 (0.081)	0.216 (0.108)		
	TPint	NA	0.052 (0.192)	0.056 (0.188)	2.080 (1.680)	NA	0.001 (0.001)	0.001 (0.028)	0.660 (0.536)		
	FPint	NA	0.724 (0.588)	6.166 (6.526)	2.344 (2.105)	NA	6.999 (6.999)	6.999 (9.305)	0.219 (0.548)		
	PPVint	NA	0.067 (0.246)	0.009 (0.028)	0.493 (0.327)	NA	0.000 (0.000)	0.000 (0.003)	0.583 (0.471)		
	SNint	NA	0.013 (0.048)	0.014 (0.047)	0.347 (0.280)	NA	0.000 (0.000)	0.000 (0.004)	0.073 (0.060)		
	F-value	0.111 (0.065)	0.172 (0.082)	0.046 (0.018)	0.244 (0.078)	0.008 (0.022)	0.009 (0.031)	0.014 (0.021)	0.151 (0.073)		
	F-value int	NA	0.011 (0.039)	0.005 (0.017)	0.188 (0.133)	NA	0.000 (0.002)	0.000 (0.002)	0.064 (0.051)		
1000	AUC	0.598 (0.021)	0.595 (0.024)	0.602 (0.023)	0.690 (0.028)	0.507 (0.016)	0.550 (0.017)	0.522 (0.020)	0.817 (0.059)		
	SNsubject	0.120 (0.103)	0.971 (0.027)	0.220 (0.062)	0.245 (0.083)	0.768 (0.327)	0.645 (0.302)	0.438 (0.328)	0.996 (0.005)		
	SPsubject	0.953 (0.023)	0.200 (0.084)	0.954 (0.024)	0.980 (0.029)	0.240 (0.331)	0.211 (0.328)	0.795 (0.246)	0.240 (0.234)		
	TPsn	1.504 (0.580)	1.148 (0.884)	1.316 (0.564)	3.480 (0.717)	0.511 (0.742)	0.245 (0.434)	0.749 (0.714)	2.080 (0.580)		
	FPsn	2.920 (1.178)	1.409 (1.492)	18.623 (19.579)	3.540 (1.086)	6.302 (3.226)	0.968 (0.708)	18.962 (19.119)	1.102 (0.393)		
	PPV sn	0.340 (0.330)	0.449 (0.372)	0.066 (0.028)	0.502 (0.105)	0.075 (0.187)	0.202 (0.380)	0.038 (0.036)	0.646 (0.136)		
	SNsn	0.376 (0.145)	0.287 (0.221)	0.329 (0.141)	0.870 (0.179)	0.073 (0.106)	0.035 (0.062)	0.107 (0.102)	0.297 (0.083)		
	TPint	NA	0.316 (0.524)	0.068 (0.208)	4.500 (1.856)	NA	0.000 (0.028)	0.007 (0.077)	1.230 (0.863)		
	FPint	NA	1.029 (0.834)	6.732 (6.502)	2.646 (2.205)	NA	0.000 (2.517)	6.993 (7.623)	0.202 (0.568)		
	PPVint	NA	0.235 (0.386)	0.010 (0.031)	0.652 (0.247)	NA	0.000 (0.011)	0.001 (0.010)	0.826 (0.337)		
5000	SNint	NA	0.079 (0.131)	0.017 (0.052)	0.750 (0.309)	NA	0.000 (0.004)	0.001 (0.011)	0.137 (0.096)		
	F-value	0.130 (0.071)	0.170 (0.132)	0.055 (0.024)	0.316 (0.061)	0.026 (0.041)	0.009 (0.031)	0.014 (0.021)	0.202 (0.049)		
	F-value int	NA	0.054 (0.087)	0.006 (0.019)	0.338 (0.126)	NA	0.000 (0.002)	0.000 (0.002)	0.112 (0.067)		
	AUC	0.604 (0.019)	0.584 (0.028)	0.610 (0.020)	0.660 (0.023)	0.515 (0.021)	0.550 (0.019)	0.532 (0.022)	0.779 (0.049)		
	SNsubject	0.197 (0.082)	0.955 (0.022)	0.175 (0.096)	0.214 (0.044)	0.830 (0.249)	0.629 (0.311)	0.454 (0.334)	0.995 (0.005)		
	SPsubject	0.974 (0.027)	0.234 (0.038)	0.946 (0.018)	0.992 (0.012)	0.184 (0.260)	0.164 (0.286)	0.769 (0.278)	0.263 (0.211)		
	TPsn	2.428 (0.504)	1.516 (0.824)	1.956 (0.564)	3.977 (0.151)	2.842 (0.931)	1.113 (0.931)	2.142 (0.910)	3.500 (0.966)		
	FPsn	6.243 (1.614)	2.516 (2.697)	18.003 (19.579)	3.114 (1.125)	6.891 (2.678)	1.422 (1.414)	17.877 (18.873)	0.796 (0.810)		
	PPV sn	0.280 (0.238)	0.376 (0.234)	0.098 (0.028)	0.574 (0.092)	0.292 (0.258)	0.439 (0.397)	0.107 (0.046)	0.820 (0.149)		
	SNsn	0.607 (0.126)	0.379 (0.206)	0.489 (0.141)	0.994 (0.038)	0.406 (0.133)	0.159 (0.133)	0.306 (0.130)	0.500 (0.138)		
5000	TPint	NA	0.452 (0.684)	0.184 (0.348)	5.932 (0.452)	NA	0.280 (0.469)	0.336 (0.434)	3.148 (1.156)		
	FPint	NA	2.921 (2.290)	6.387 (6.344)	2.250 (2.589)	NA	0.827 (0.667)	7.478 (7.457)	0.426 (1.700)		
	PPVint	NA	0.134 (0.230)	0.028 (0.052)	0.782 (0.204)	NA	0.253 (0.413)	0.043 (0.055)	0.956 (0.166)		
	SNint	NA	0.113 (0.171)	0.046 (0.087)	0.989 (0.075)	NA	0.040 (0.067)	0.048 (0.062)	0.350 (0.128)		
	F-value	0.164 (0.087)	0.186 (0.105)	0.081 (0.024)	0.362 (0.039)	0.133 (0.059)	0.115 (0.096)	0.079 (0.034)	0.308 (0.067)		
	F-value int	NA	0.056 (0.085)	0.017 (0.032)	0.430 (0.074)	NA	0.033 (0.053)	0.023 (0.029)	0.248 (0.067)		
	AUC	0.620 (0.020)	0.589 (0.037)	0.622 (0.019)	0.626 (0.009)	0.565 (0.022)	0.531 (0.014)	0.609 (0.012)	0.753 (0.014)		
	SNsubject	0.244 (0.022)	0.942 (0.009)	0.164 (0.112)	0.203 (0.010)	0.836 (0.137)	0.235 (0.183)	0.722 (0.353)	0.991 (0.002)		
	SPsubject	0.974 (0.032)	0.228 (0.045)	0.951 (0.022)	0.998 (0.003)	0.211 (0.173)	0.419 (0.352)	0.464 (0.256)	0.457 (0.073)		

TP, true positive; FP, false positive; PPV, positive predictive value; SN, sensitivity; int, two-SNP interactions; AUC, area under curve.

Table 3. Performance of logistic regression with Lasso penalty (Lasso), standard logic regression (logic), full Bayesian logic regression (FBLR), and proposed method (PRLR) for Scenario 5: mean and standard deviation. Bold character indicates the best performance.

Performance measure	Samples	Scenario 5			
		Lasso	Logic	FBLR	PRLR
TPsnp	500	1.080 (0.620)	0.850 (0.410)	1.160 (0.540)	1.340 (0.650)
FPsnp		3.026 (1.330)	0.352 (0.639)	18.840 (19.460)	7.600 (7.280)
PPVsnp		0.263 (0.318)	0.707 (0.391)	0.058 (0.027)	0.149 (0.081)
SNsnp		0.108 (0.062)	0.085 (0.041)	0.116 (0.054)	0.134 (0.065)
TPint		NA	0.000 (0.030)	0.010 (0.110)	0.270 (0.420)
FPint		NA	0.000 (2.697)	4.990 (8.352)	7.910 (4.560)
PPVint		NA	0.000 (0.011)	0.002 (0.013)	0.032 (0.083)
SNint		NA	0.000 (0.003)	0.001 (0.011)	0.027 (0.042)
F-value		0.056 (0.031)	0.075 (0.035)	0.039 (0.018)	0.069 (0.033)
F-value int		NA	0.000 (0.002)	0.001 (0.006)	0.013 (0.021)
AUC		0.587 (0.038)	0.611 (0.023)	0.599 (0.036)	0.589 (0.028)
SNsubject		0.986 (0.030)	0.931 (0.113)	0.136 (0.212)	0.993 (0.021)
SPsubject		0.024 (0.051)	0.098 (0.170)	0.971 (0.060)	0.013 (0.039)
TPsnp	1000	1.410 (0.690)	1.030 (0.280)	1.440 (0.630)	1.730 (0.670)
FPsnp		3.010 (1.480)	0.192 (0.765)	18.56 (19.058)	7.100 (6.860)
PPVsnp		0.319 (0.318)	0.843 (0.268)	0.072 (0.032)	0.195 (0.088)
SNsnp		0.141 (0.069)	0.103 (0.028)	0.144 (0.063)	0.173 (0.067)
TPint		NA	0.020 (0.110)	0.050 (0.200)	0.520 (0.480)
FPint		NA	1.091 (0.717)	8.283 (8.133)	7.130 (3.500)
PPVint		NA	0.018 (0.133)	0.006 (0.024)	0.067 (0.119)
SNint		NA	0.002 (0.011)	0.005 (0.020)	0.052 (0.048)
F-value		0.071 (0.030)	0.091 (0.022)	0.048 (0.021)	0.091 (0.035)
F-value int		NA	0.001 (0.010)	0.003 (0.011)	0.026 (0.025)
AUC		0.611 (0.018)	0.610 (0.022)	0.617 (0.019)	0.605 (0.012)
SNsubject		0.990 (0.023)	0.929 (0.115)	0.137 (0.214)	0.983 (0.035)
SPsubject		0.020 (0.044)	0.091 (0.159)	0.939 (0.085)	0.031 (0.064)
TPsnp	5000	4.070 (1.050)	1.830 (0.790)	2.770 (1.010)	4.890 (0.890)
FPsnp		15.687 (4.95)	0.811 (1.992)	17.302 (18.794)	5.000 (6.530)
PPVsnp		0.206 (0.175)	0.693 (0.284)	0.138 (0.051)	0.490 (0.119)
SNsnp		0.407 (0.105)	0.183 (0.079)	0.277 (0.101)	0.489 (0.089)
TPint		NA	0.280 (0.400)	0.190 (0.380)	3.180 (0.840)
FPint		NA	0.620 (0.475)	8.071 (7.881)	7.020 (4.040)
PPVint		NA	0.311 (0.457)	0.023 (0.046)	0.311 (0.171)
SNint		NA	0.028 (0.040)	0.019 (0.038)	0.318 (0.084)
F-value		0.115 (0.048)	0.142 (0.057)	0.092 (0.034)	0.243 (0.047)
F-value int		NA	0.025 (0.036)	0.010 (0.021)	0.150 (0.049)
AUC		0.639 (0.019)	0.622 (0.034)	0.636 (0.018)	0.623 (0.006)
SNsubject		0.984 (0.016)	0.891 (0.108)	0.231 (0.208)	0.962 (0.055)
SPsubject		0.043 (0.041)	0.044 (0.061)	0.907 (0.103)	0.088 (0.099)

TP, true positive; FP, false positive; PPV, positive predictive value; SN, sensitivity; int, two-SNP interactions; AUC, area under curve.

4. APPLICATION STUDY: SNP–SNP INTERACTIONS INFLUENCING BONE MINERAL DENSITY IN CHILDHOOD LEUKEMIA SURVIVORS

4.1. Application settings

Survivors of childhood acute lymphoblastic leukemia (ALL) are frequently at risk of deficits in bone mineral density (BMD) because of ALL treatments received during childhood cancer, which include radiation therapy, antimetabolites, and glucocorticoids (Kaste and others, 2006; Wasilewski-Masker and others, 2008). Cranial radiation therapy decreases BMD due to injury to the hypothalamic–pituitary axis,

which controls sex- or growth-hormone secretions (Wasilewski-Masker and others, 2008). Antimetabolites and glucocorticoids decrease BMD because they impair the activities of osteoblast and osteoclast cells (Wasilewski-Masker and others, 2008). Although it is well known that survivors of childhood ALL may experience low BMD later in their adult lives, studies of the effects of single SNPs and SNP–SNP interactions to explain the substantial variations in BMD among survivors remain limited (Im and others, 2018).

We analyzed data from the St. Jude Lifetime (SJLIFE) Cohort Study, which includes BMD measurements and SNP genotype information. Detailed cohort profiles can be found in Hudson and others (2011). In SJLIFE, BMD was measured using quantitative computed tomography (QCT), and a BMD Z-score was calculated using the age- and sex-matched general population as a reference. Since BMD Z-scores less than -1 are clinically considered as an indicator for low BMD/osteopenia (i.e., one standard deviation below the mean BMD for an age- and sex-matched reference population), the outcome variable of interest was created by binarizing the BMD Z-score with a threshold of -1 (Kumar and others, 2011). Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA) was used to genotype DNA samples in SJLIFE. Genotyping quality control (QC) thresholds for exclusion were $< 95\%$ per sample call rate across all SNPs, $< 95\%$ SNP call rate across samples, $MAF < 1\%$, and Hardy–Weinberg equilibrium p -value $< 10^{-6}$. Following Im and others (2018), the ChromHMM chromatin annotation method was used to map SNPs to enhancer or promoter regions (Ernst and others, 2011) and we retained SNPs mapped to enhancer or promoter regions in any of nine ChromHMM-annotated cell types. Following QC and the ChromHMM enhancer/promoter filter, this analysis focuses on 5814 SNPs in a region of approximately 130 Mbp on chromosome 12 (Case 1). For each of these SNPs, a binary variable was created to represent the recessive minor allele genetic model.

Since all included SNPs were mapped to regulatory regions (i.e., promoter or enhancer), the smoothing parameter γ in Figure 2 (lower) to regulate the effect of the network matrix N on the biological weighted adjacency matrix Σ was set to 0. A total of 835 participants were divided into training and validation datasets to assess the prediction accuracy of the constructed models ($N = 735$ for training, $N = 100$ for validation). The number of participants with low BMD were 225 and 21 for the training and validation datasets, respectively. For PRLR, five Markov chains were implemented in parallel ($M = 5$), and the regularization parameters λ_1 , λ_2 , and α were optimized by using 10-fold cross-validation. Other settings, including the parameter settings for Lasso, Logic, and FBLR, were the same as those used in the simulation study.

Lastly, to check the performance of our approach under the high-dimensional case, we applied PRLR and the existing approaches to a larger dataset without limiting SNPs to those mapped to the enhancer/promoter regions: this analysis included a total of 376 777 SNPs (Case 2). The smoothing parameter γ to incorporate the network matrix N was set to 0.1 as illustrated in Figure 2 (lower). In terms of N , if two SNPs belonged to enhancer/promoter regions, the corresponding element of N was set to 1; if one SNP belonged to an enhancer/promoter region and another SNP belonged to a non-enhancer/promoter region, the element of N was set to 0.5; otherwise, the element of N was set to 0. Other settings used in Case 2 were identical to Case 1.

4.2. Application results

Consistent with the results of the simulation study, PRLR detected more single SNPs and SNP–SNP interactions associated with low BMD compared with existing approaches. In Case 1, the regularization parameters were set to $\lambda_1 = 2$, $\lambda_2 = 5$, and $\alpha = 2$ after 10-fold cross-validation. The best model selected by PRLR was $\text{logit}(P(Y = 1)) = -1.133 - 0.352L_1 + 0.805L_2 - 0.727L_3$, where $Y = 1$ indicates low BMD and L_1 , L_2 , and L_3 are the trees in Supplementary materials S2 available at *Biostatistics* online. The estimated standard errors of the coefficients were 0.204 for the intercept ($p < 0.001$), 0.249 for the

first term ($p = 0.157$), 0.213 for the second term ($p = 0.002$), and 0.174 for the last term ($p < 0.001$). The result was robust to changes in the random number seed, the number of Markov chains, and the regularization parameter σ in the restricted weighted adjacency matrix Σ_{gg} .

Table 4 shows the results of PRLR and existing approaches for Cases 1 and 2 in terms of the selected single SNP, SNP-pair interactions, and AUC as the prediction performance measure. The SNP-pair interactions were created by eliminating the operator from the Boolean expressions. In Case 1, PRLR discovered 12 single SNPs and 21 SNP-SNP interactions, which included the findings of previous work (Zheng and others, 2016). Among the 12 single SNPs selected by PRLR, 3 SNPs (rs4766415, rs2114870, and rs1355472) were shared with other conventional approaches. This suggests that our approach may discover informative SNPs more often than alternative approaches, while keeping some relevant SNPs in common. In terms of prediction performance, PRLR outperforms existing approaches: PRLR produced AUC = 67.24, while other approaches produced 52.05 for Lasso, 59.31 for Logic, and 57.74 for FBLR, respectively, in the validation dataset. Additionally, we checked the performance of PRLR in the higher dimensional case. In Case 2, the regularization parameters were set to $\lambda_1 = 0.01$, $\lambda_2 = 1$, and $\alpha = 5$ after 10-fold cross-validation, and similar results as Case 1 were observed. PRLR discovered 9 SNPs and 10 SNP-SNP interactions, some of which were shared with other conventional approaches (rs11052372 and rs7313912); PRLR's prediction performance, measured by AUC, was superior (AUCs of PRLR, Lasso, Logic, and FBLR are 62.69, 54.81, 53.10, and 52.38, respectively.) The best model selected by our approach was $\text{logit}(P(Y = 1)) = -2.987 - 0.678L_1 + 2.164L_2 + 0.487L_3$, where L_1, L_2 and L_3 are shown in Supplemental materials S2 available at *Biostatistics* online. The estimated standard errors of the coefficients were 0.746 for the intercept ($p < 0.001$), 0.172 for the first term ($p < 0.001$), 0.735 for the second term ($p = 0.003$), and 0.179 for the last term ($p = 0.007$).

5. DISCUSSION

The logic regression framework provides an important analytic methodology for genetic association studies when the primary research interest is to study interactions between genetic variants. However, its potential effectiveness in practice is diminished by the computational burden of the method and the problem of being trapped in local solutions. To tackle these problems, this article proposed a new approach that entails running several Markov chains for a logic regression model in parallel, while avoiding the inefficient overlap of estimation paths by introducing two ideas: (i) biologically-induced distances between two SNPs to define the similarity of estimation paths and trees, which are derived from the physical positions of SNPs in the genome and biological gene pathways; and (ii) two repulsive forces between and within estimation paths. The repulsive forces were designed to encourage searches in dissimilar areas when the trajectories of different Markov chains of different estimation paths/trees come close to each other during SA iterations. Due to this repulsive system, PRLR can effectively search the interaction space and avoid being trapped in local solutions. The performance of PRLR was evaluated using simulation and epidemiological data. The simulation results indicate that PRLR was able to discover more true SNP interactions than conventional approaches and improve the prediction accuracy in many settings. Our application study shows that while PRLR proposed several SNPs in common with other approaches, it also unveiled additional interactions, which partially replicate the results of previous work (Zheng and others, 2016).

Previous studies have incorporated external biological knowledge of SNPs in analysis (Afreixo and others, 2009). Using Σ , PRLR can be considered as a new approach to integrate the biological expertise into model building for logic regression; such approaches can detect powerful and biologically meaningful models for SNP associations. Employing the biological weighted adjacency matrix Σ to define the similarity of estimation paths and trees is crucial for our approach to avoid being trapped in local solutions in the search space. Therefore, σ and γ in Σ should be carefully selected based on prior knowledge, possible

Table 4. Performances of our method and existing methods for low bone mineral density

Performance measure	Models			
	Lasso	Logic	FBLR	PRLR
Case 1: 5814 SNPs				
# of selected single SNPs	3	6	20	12
# of selected SNP-SNP interactions	NA	1	10	21
AUC	52.05	59.31	57.74	67.24
Selected single SNPs	rs1355472, rs4766415, rs11055896	rs2114870, rs2733840, rs11055896, rs1355472, rs6538109, rs12318643	rs1355472, rs11055896, rs7978809, rs4766415, rs7953247, rs830379, rs10772010, rs11835013, rs177079, rs3847835, rs11053915, rs12321085, rs2362845, rs17395410, rs16915479, rs2277363, rs10858943, rs12303892, rs3858655, rs730560	rs7301596, rs4765826, rs6489301, rs7132285, rs4766415, rs7301596, rs6489301, rs4765826, rs7132285, rs4765826, rs4766415, rs6489301, rs7132285, rs4766415, rs7132285, rs1355472, rs12303496, rs16914640, rs1560007, rs2114870, rs1560007, rs10772079, rs1560007, rs11053515, rs1560007, rs2114870, rs16914640, rs11053515, rs16914640, rs10772079, rs2114870, rs11053515, rs2114870, rs11053515, rs10772079
Selected SNP-SNP interactions	NA	(rs6538109, rs11055896), (rs12318643, rs2733840)	(rs1833211, rs4758658), (rs4766415, rs11055896), (rs12321085, rs7137797), (rs2171576, rs10861405), (rs11055896, rs7295425), (rs2114870, rs1833472), (rs11045272, rs12369386), (rs11048654, rs11179003), (rs1355472, rs3811170), (rs763581, rs830379)	
Case 2: 376 777 SNPs				
# of selected single SNPs	11	3	20	9
# of selected SNP-SNP interactions	NA	3	6	10
AUC	54.81	53.10	52.38	62.69
Selected single SNPs	rs11055299, rs11055896, rs16910664, rs4762758, rs7134807, rs33162, rs11182500, rs454387, rs704110, rs821858, rs7313912	rs11052372, rs2120575, rs1420742	rs10735033, rs16910664, rs4931087, rs7313912, rs11055896, rs11615703, rs454387, rs7138297, rs7305668, rs1861691, rs10748071, rs9971886, rs4762758, rs2125093, rs10842386, rs7978809, rs7953247, rs1355472, rs2728817, rs25643	rs11052372, rs4763813, rs12368539, rs1112795, rs7313912, rs2216291, rs2270380, rs741404, rs12425885
Selected SNP-SNP interactions	NA	(rs2120575, rs11052372), (rs1420742, rs11052372), (rs1420742, rs2120575)	(rs11053518, rs7977857), (rs12317113, rs4379922), (rs1059845, rs17548327), (rs10877835, rs10773227), (rs735393, rs11832465), (rs7313831, rs7969337)	(rs11112795, rs12368539), (rs7313912, rs12368539), (rs7313912, rs11112795), (rs4763813, rs11052372), (rs2270380, rs2216291), (rs741404, rs2216291), (rs12425885, rs2216291), (rs741404, rs2270380), (rs12425885, rs2270380), (rs12425885, rs741404)

Bold characters indicate the selected SNPs by PRLR that are shared with the existing methods.

trial and error, or reliable databases. However, requiring the correct Σ is not realistic in some practical cases. If the settings of σ and γ are wrong, the efficiency of our algorithm in avoiding overlap of estimation paths would be degraded. If there are no good candidates for σ and γ , we recommend starting at $\sigma = 1$ and $\gamma = 0$, and changing them as sensitivity parameters. Note that in practice the physical chromosomal locations (and linkage disequilibrium structures) of SNPs for $C_{g,g'}(s, s')$ can be obtained from publicly-available databases such as the 1000 Genomes Project (<http://www.internationalgenome.org>), and also SNP-gene mapping for N can be obtained from databases such as FANTOM (<http://fantom.gsc.riken.jp>) which can be integrated with the aforementioned pathway information from databases such as Gene Ontology and KEGG in forming N . Also, since σ only changes the degree of the adjacency (i.e., the values of $C_{g,g'}(s, s', \sigma)$) and λs can absorb the impact of changes in σ , σ can be set more flexibly than γ . Our framework for model building can also be applied without using biological knowledge. For example, the proposed approach can be used by replacing the measure of the biological adjacency with the similarity of geometric tree-structures: it is possible to replace the terms ψ_1 and ψ_2 in (2.2) with the similarities of the tree-structure measured by a tree edit distance between two labeled trees (Zhang and Shasha, 1989; Yang and others, 2005), and the repulsive forces can be defined based on the tree edit distance. In this sense, our approach can be regarded as a new model-building framework for logic regression to efficiently construct the best model by using repulsive forces. However, this requires further discussions about appropriate formulations for repulsive forces or Σ to represent the biological distance between variants of interest.

Ruczinski and others (2003) discussed several approaches to include continuous predictors in logic regression models. Generally, there are three ideas: (i) to discretize the continuous variables into binary variables; (ii) to use the residuals terms when the outcome variable is continuous; and (iii) to update trees and the continuous covariates separately rather than to propose moves to both the trees and the continuous variables jointly. The first idea is simple. We can binarize a continuous covariate by defining a threshold value. With the binarized variables, our method can be used in the same manner. The second idea is a two-step regression procedure and can be applied when the outcome variable is continuous. Firstly, we run a linear regression on a set of the continuous variables, and then the second regression uses the residuals of that regression as the dependent variable to regress on trees as the independent variables. The third idea is to use the technique proposed in Clark and others (2007). They proposed an iterative algorithm: the first step is to update the trees and their regression coefficients, while fixing the regression coefficients of the continuous covariates, and the second step is to update the regression coefficients of the continuous covariates while fixing the trees and their coefficients.

A limitation of this study is that avoiding over-fitting when estimating a complex model using small samples remains a challenge since it is difficult to distinguish the current best model from others that fit the data almost as well in small samples (Ruczinski and others, 2003; Kooperberg and Ruczinski, 2005; Fritsch and Ickstadt, 2007). From this perspective, we consider the set of parameter estimates with the minimum value of the objective function (2.4) as the best set among M parallel runs, but we could also consider alternative approaches to construct the best set of parameter estimates from M sets of parameter estimates such as the use of information criteria or model averaging methods (Clark and others, 2007). In particular, using the average of M sets of parameter estimates in this context has the potential to mitigate the impact of over-fitting (Claeskens and others, 2008).

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

REFERENCES

- AFREIXO, V., BASTOS, C. A. C., PINHO, ARMANDO J., GARCIA, S. P. AND FERREIRA, P. J. S. G. (2009). Genome analysis with inter-nucleotide distances. *Bioinformatics* **25**, 3064–3070.
- ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T. AND OTHERS (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29.
- CLAESKENS, G., HJORT, N. L. (2008). *Model Selection and Model Averaging*, Volume 330. Cambridge: Cambridge University Press.
- CLARK, T. G., DE IORIO, M. AND GRIFFITHS, R. C. (2007). Bayesian logistic regression using a perfect phylogeny. *Biostatistics* **8**, 32–52.
- ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R., COYNE, M. AND OTHERS (2011). Systematic analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43.
- FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1.
- FRITSCH, A. AND ICKSTADT, K. (2007). Comparing logic regression based methods for identifying snp interactions. In: *Bioinformatics Research and Development*. Berlin, Heidelberg: Springer, p. 90–103.
- HUDSON, M. M., NESS, K. K., NOLAN, V. G., ARMSTRONG, G. T., GREEN, D. M., MORRIS, E. B., SPUNT, S. L., METZGER, M. L., KRULL, K. R., KLOSKY, J. L. AND OTHERS (2011). Prospective medical assessment of adults surviving childhood cancer: study design, cohort characteristics, and feasibility of the st. jude lifetime cohort study. *Pediatric Blood & Cancer* **56**, 825–836.
- IM, C., NESS, K. K., KASTE, S. C., CHEMAITILLY, W., MOON, W., SAPKOTA, Y., BROOKE, R. J., HUDSON, M. M., ROBISON, L. L., YASUI, Y. AND OTHERS (2018). Genome-wide search for higher order epistasis as modifiers of treatment effects on bone mineral density in childhood cancer survivors. *European Journal of Human Genetics* **26**, 275.
- KANEHISA, M., GOTO, S., SATO, Y., KAWASHIMA, M., FURUMICHI, M. AND TANABE, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* **42**, D199–D205.
- KASTE, S. C., RAI, S. N., FLEMING, K., MCCAMMON, E. A., TYLAVSKY, F. A., DANISH, R. K., ROSE, S. R., SITTER, C. D., PUI, C.-H. AND HUDSON, M. M. (2006). Changes in bone mineral density in survivors of childhood acute lymphoblastic leukemia. *Pediatric Blood & Cancer* **46**, 77–87.
- KOOPERBERG, C. AND RUCZINSKI, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology* **28**, 157–170.
- KUMAR, J., SWANBERG, M., MCGUIGAN, F., CALLREUS, M., GERDHEM, P. AND ÅKESSON, K. (2011). Lrp4 association to bone properties and fracture and interaction with genes in the wnt-and bmp signaling pathways. *Bone* **49**, 343–348.
- MANSOUR, Y. (1997). Pessimistic decision tree pruning based on tree size. In: *Machine Learning-International Workshop then Conference*. Morgan Kaufmann Publishers, Inc.
- PHILLIPS, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9**, 855.
- RUCZINSKI, C. K. I., LEBLANC, M. L. AND HSU, L. (2001). Sequence analysis using logic regression. *Genetic Epidemiology* **21**, S626–S631.
- RUCZINSKI, I., KOOPERBERG, C. AND LEBLANC, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics* **12**, 475–511.

- SCHWENDER, H. AND ICKSTADT, K. (2007). Identification of SNP interactions using logic regression. *Biostatistics* **9**, 187–198.
- THE 1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.
- TIBSHIRANI, R. AND HASTIE, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559–567.
- WASILEWSKI-MASKER, K., KASTE, S. C., HUDSON, M. M., ESIASHVILI, N., MATTANO, L. A. AND MEACHAM, L. R. (2008). Bone mineral density deficits in survivors of childhood cancer: long-term follow-up guidelines and review of the literature. *Pediatrics* **121**, e705–e713.
- YANG, R., KALNIS, P. AND TUNG, A. K. H. (2005). Similarity evaluation on tree-structured data. In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. Baltimore, Maryland: ACM, pp. 754–765.
- ZHANG, K. AND SHASHA, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing* **18**, 1245–1262.
- ZHENG, Y., WANG, C., ZHANG, H., SHAO, C., GAO, L.-H., LI, S.-S., YU, W.-J., HE, J.-W., FU, W.-Z., HU, Y.-Q. AND OTHERS (2016). Polymorphisms in wnt signaling pathway genes are associated with peak bone mineral density, lean mass, and fat mass in chinese male nuclear families. *Osteoporosis International* **27**, 1805–1815.

[Received March 12, 2018; revised March 5, 2019; accepted for publication March 24, 2019]