Methods for Identifying SNP Interactions: A Review on Variations of Logic Regression, Random Forest and Bayesian Logistic Regression

Carla Chia-Ming Chen, Holger Schwender, Jonathan Keith, Robin Nunkesser, Kerrie Mengersen, and Paula Macrossan

Abstract—Due to advancements in computational ability, enhanced technology and a reduction in the price of genotyping, more data are being generated for understanding genetic associations with diseases and disorders. However, with the availability of large data sets comes the inherent challenges of new methods of statistical analysis and modeling. Considering a complex phenotype may be the effect of a combination of multiple loci, various statistical methods have been developed for identifying genetic epistasis effects. Among these methods, logic regression (LR) is an intriguing approach incorporating tree-like structures. Various methods have built on the original LR to improve different aspects of the model. In this study, we review four variations of LR, namely Logic Feature Selection, Monte Carlo Logic Regression, Genetic Programming for Association Studies, and Modified Logic Regression-Gene Expression Programming, and investigate the performance of each method using simulated and real genotype data. We contrast these with another tree-like approach, namely Random Forests, and a Bayesian logistic regression with stochastic search variable selection.

Index Terms—Logic regressions, Genetic Programming for Association Studies, Modified Logic Regression-Gene Expression Programming, Random Forest, Bayesian logistic regression with stochastic search algorithm, candidate gene search.

1 Introduction

Single nucleotide polymorphism (SNP) is the most common genetic variation among individuals and it was estimated that the human genome has approximately 10 million SNPs [25]. With the recent mapping of the human genome [41] came the availability of high throughput laboratory procedures for the identification of SNPs. Strong correlation among blocked SNPs, i.e., linkage disequilibrium, allows scientists to study the association between genetic and phenotypic variation using a subset of SNPs. Genome Wide Association Studies (GWAs) attempt the mapping of SNPs to phenotypic variation among individuals. Such procedures require a sound statistical methodology and associated computational capability to cope with the analysis of a large data set. Most studies are focused on single locus analysis, which directly tests the association between individual SNP and phenotypic variant. The most commonly implemented statistical approach for these studies is a SNP-by-SNP testing algorithm. This procedure requires an additional statistical correction for the Type 1 error associated with multiple testings. Rice et al. [35] provide summaries of commonly used correction methods, including Bonferroni correction, permutation test and false discovery rate, and discuss the benefits and drawbacks of each of these.

Although the SNP-by-SNP approaches are relatively fast and capable of incorporating covariates (e.g., [43]), the major limitation of such approaches is the difficulty of detecting possible gene epistasis effects [17], which is often suggested as the reason for lack of success in genetic studies of complex diseases [9]. Although "epistasis" is commonly defined as the interaction of different genes, there is some confusion on the definition of epistasis in the literature owing to the existence of different types of interaction [9]. Cordell [8] and Phillips [33] provide thorough reviews on different types of epistasis. In this study, we are focused on using statistical methods to identify gene interaction, this is the "statistical epistasis" according to [33].

Various statistical methods that have been developed for searching for epistasis effects in complex diseases include Bayesian epistasis association mapping (BEAM, [44]), multifactor dimensionality reduction [36], Polymorphism Interaction Analysis [29], logic regression [37], Bayesian model selection [11], and a two stage approach

- *C.C.-M. Chen, K. Mengersen, and P. Macrossan are with the Discipline of Mathematical Sciences, Queensland University of Technology, Gardens Point, Brisbane, Queensland 4001, Australia. E-mail: gcewels@gmail.com, k.mengersen@qut.edu.au, plennon@activ8.net.au.*
- *H. Schwender is with Department of Biostatistics, Johns Hopkins University, Baltimore, MD, and also with the Department of Statistics, TU Dortmund University, Dortmund 44221, Germany. E-mail: holger.schwender@udo.edu.*
- *J. Keith is with the School of Mathematical Sciences, Monash University, Clayton VIC 3800, Australia. E-mail: jonathan.keith@monash.edu.*
- *R. Nunkesser is with Department of Computer Science, TU Dortmund University, Germany and Collaborative Research Centre 475, TU Dortmund University, Dortmund 44221, Germany. E-mail: Robin.Nunkesser@tudortmund.de.*

Manuscript received 9 May 2010; revised 29 Oct. 2010; accepted 28 Dec. 2010; published online 7 Mar. 2011.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2010-05-0118. Digital Object Identifier no. 10.1109/TCBB.2011.46.

that first selects SNPs with strong marginal effects, then identifies interactions among the SNPs [30]. A.G. Heidema et al. [16] provide an overview and evaluation of the performance of five widely applied methods in detecting interaction effects. One of these, logic regression (LR, [37]), is a unique method that has the structure of a generalized regression method but with a Boolean combination of variables as predictors. LR is motivated and developed for a plausible but difficult association pattern between SNPs and phenotype, which often involves using words like "AND," "OR," and "NOT." For example, an individual may have a higher chance of having a specific trait when *"the homozygous variant genotype is at SNP S_1* ***AND*** *the homozygous reference genotype is at SNP S_2* ***OR*** *both SNP S_3* ***AND*** *S_4 are* ***NOT*** *of the homozygous reference genotype"*

LR has been widely applied in the analysis of SNP data for various phenotypes including sporadic breast cancer [12], [39], trachoma [2], bladder cancer [1], renin-angiotensin [22], and myocardial infraction [22]. Schwender [38] indicates that LR is more preferred when compared with other tree-based approaches, such as Random Forests (RF, [4]) and Classification and Regression Trees (CART, [5]).

Although LR was initially developed for prediction, its capability has been extended through algorithms such as logic Feature Selection (LogicFS, [39]), Monte Carlo Logic regression (MCLR, [23]), and Full Bayesian logic regression (FBLR, [12]).

Another extension to the original LR involves variations in the searching algorithm. Kooperberg [23] pointed out two drawbacks with the simulated annealing algorithm implemented in original LR. First, it identifies a single best model which potentially neglects competing models. Second, simulated annealing is not geared for the identification of SNPs in linkage disequilibrium (LD). Although the latter limitation has not yet been resolved, the former limitation can potentially be resolved by using different searching algorithms. Methods such as Reversible Jump MCMC [15]), Genetic Programming for Association Studies (GPAS, [32]) and Gene Expression Programming (MLR-GEP, [27]) have a framework similar to logic regression but implement different searching algorithms.

The aim of this paper is to summarize these variations of LR for a case-control study and compare the performance of the methods using simple simulated examples. Due to the fact that LR is a tree-based algorithm, we also consider Random Forests [4] in this paper. Furthermore, we compare the methods with a Bayesian logistic model. Therefore, the methods included in this study include LogicFS, MCLR, GPAS, MLR-GEP, RF, and Bayesian logistic regression.

2 Methods

2.1 Logic Regression

Before introducing LR, it is important to note how SNPs may be coded in LR. Let allele "A" be a disease allele; that is, having allele "A" increases the probability of expressing a certain phenotype. Typically the SNP is coded as 0, 1, 2 which corresponds to genotypes "aa," "aA," and "AA." Alternatively, the SNPs may be coded as a binary variable, which represents the dominant and recessive effect, for instance, genotype "Aa" or "AA" at SNP S may be coded as $S_{i,1}$ and genotype "AA" as $S_{i,2}$.

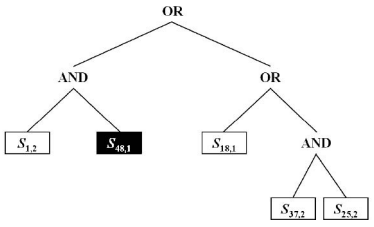

Fig. 1. An example of a logic tree of LR.

LR was initially developed for classification and regression, which aims to find Boolean combinations that enhance the prediction of the model. The LR thus comprises Boolean combinators such as AND- and OR-, and variables, i.e., SNPs, in a logic expression, L. Using the same example as in the Introduction, L is then

$$L = \left(S_{1,2} \wedge S_{2,1}^{C}\right) \vee (S_{3,1} \wedge S_{4,1}), \tag{1}$$

where $\wedge$ and $\vee$ denote the AND and the OR operator, respectively, and ^{C} denotes the complement of a boolean variable.

Logic expressions can be structured into a tree representation which is referred to as a logic tree. The terminology of the logic tree is very similar to that used in CART, although the trees of LR and CART are different structurally, as discussed later in this paper. A node is a point on the tree structure where a split occurs. In LR, a node represents one of the Boolean operators (AND-) and (OR-), and each leaf corresponds to one of the variables (SNPs). Fig. 1 is an example of a logic tree of LR, and with a logic expression given by

$$L = \left(S_{1,2} \wedge S_{48,1}^{c}\right) \vee (S_{18,1} \vee (S_{37,2} \wedge S_{25,2})). \tag{2}$$

Here, the leaves include the dominant effect of SNP 18; and the recessive effect of SNP 1, 37, and 25. SNP 48 is highlighted in dark shade, representing the complement of SNP 48 (i.e., NOT ($\mathrm{SNP}_{48,1}$)).

When the number of SNPs increases, searching among all possible logic trees/expressions becomes unmanageable. This motivates the implementation of a stochastic searching algorithm. The simulated annealing algorithm proposed by Ruczinski [37] and Kirkpatrick [21] starts with a logic tree, L_1, consisting of randomly selected variables. At each iteration s, a new logic expression, L_{new} is proposed by randomly selecting one of six possible moves: alternate a leaf, alternate an operator, grow a branch, prune a branch, split a leaf, or delete a leaf. Each move is assigned with a prespecified probability, and not all moves are permissible at an iteration. For instance, when the maximum size of the tree is reached, moves which result in adding a leaf/leaves are prohibited. The acceptance of L_{new} depends upon the acceptance probability, given by

$$a(MCR_s, MCR_{new}, T) = \min\left\{1, \frac{\exp(MCR_s - MCR_{new})}{T}\right\}, \tag{3}$$

where MCR_s is the misclassification rate of the tree s and T denotes the "temperature," which decreases with the

duration of the annealing process. Thus, the acceptance rate of a new logic tree is much higher at the beginning of the process (when T is large) and eventually becomes almost zero at the end of the search.

For more complicated problems, multiple trees can be combined using a generalized linear model

$$g(y) = \beta_0 + \sum_{q=1}^{Q} \beta_q L_q, \tag{4}$$

where $g(.)$ is a link function, β_0 is the intercept, $\beta_q, q = 1, \ldots, Q$, is the coefficient of the tree L_q, and Q is the maximum number of trees allowed. Using such a format increases the versatility of LR for the analysis of different types of phenotypes [37] and can be easily modified for more complicated models such as the Cox proportional hazards model.

2.2 Monte Carlo Logic Regression

Kooperberg and Ruczinski [23] proposed that instead of selecting a single optimal model, it is preferable to identify various competing models and combinations of covariates that are potentially associated with the phenotype. Their method incorporates Bayesian model selection techniques using Markov Chain Monte Carlo to explore a large number of models. Therefore, the model is called Monte Carlo Logic Regression (MCLR).

The main difference between MCLR and LR is in the use of priors and the searching algorithm. MCLR requires specification of a prior on the model size. The model size is defined as $\sum_{q=1}^{Q} |L_q|$, where $|L_q|$ is number of terminal nodes of the tree q. Because the model parameters of (4) are not essential for detecting the SNP interaction, Kooperberg and Ruczinski [23] adapted the maximum likelihood approaches for parameter estimation instead of using a fully Bayesian approach.

Compared with LR, the searching algorithm of MCLR is more complicated as it uses Reversible Jump MCMC (RJMCMC, [15]. At each iteration, a logic tree is selected at random and modified using the same moves as the LR. Once a new model is selected, the acceptance of the new model will depend upon the prior, posterior, and likelihood ratio as described in [15].

Like other MCMC methods, a large number of iterations are required to ensure the convergence of a MCMC chain. The importance of SNPs and SNP interactions is determined from the post burn-in samples, i.e., samples after the chain has converged. For instance, the importance of a two-way SNP interaction is defined as the frequency of the pair of SNPs found in the same logic tree over all post burn-in models. The same paradigm is used for finding the interactions of three variables.

2.3 Logic Feature Selection

LogicFS is more closely related to LR in that it follows the same paradigm as the LR and uses simulated annealing as the searching algorithm. However, instead of seeing them as two separate methods, LogicFS improves the variable selection of LR by repetitively fitting logic regression models to different bootstrap samples. This is achieved by employing bagging [3] with the base learner LR.

LogicFS draws a bootstrap sample from the original samples, i.e., n samples are randomly drawn with replacement from the original samples, and then applies logic regression to the bootstrap sample. This process is repeated several times (typically 50-100 times). LogicFS also improves the interpretation of the logic expression by transforming the expression into a disjunctive normal form (DNF). This makes the SNP interactions directly identifiable. For example, assume a standard logic expression

$$L = \left(S_{1,1} \wedge S_{2,1}^c\right) \vee \left(S_{3,2} \vee S_{4,2}\right) \wedge S_{5,1}^c, \tag{5}$$

of an original LR. L is then transformed into a DNF, which becomes

$$L = \left(S_{1,1} \wedge S_{2,1}^c\right) \vee \left(S_{3,2} \wedge S_{5,1}^c\right) \vee \left(S_{4,2} \wedge S_{5,1}^c\right). \tag{6}$$

Compared with (5), the identification of interactions is much easier in (6). The two-way SNP interactions are SNPs connected by "AND" operators, which are $S_{1,1}$ AND $S_{2,1}^c$, $S_{3,2}$ AND $S_{5,1}^c$, and $S_{4,2}$ AND $S_{5,1}^c$. This representation can then be used to estimate the importance of any interactions based on its predictability, which is essential for distinguishing a "real" influential interaction from noise. Moreover, transforming the logic expression into a DNF pools the AND-combination and makes some variables redundant. For example, if both $S_{1,1} \wedge S_{2,1} \wedge S_{3,1}$ and $S_{1,1} \wedge S_{2,1} \wedge S_{3,1}^c$ are in the logic expression, LogicFS shortens the logic expression by removing $S_{3,1}$ and the expression becomes $S_{1,1} \wedge S_{2,1}$.

The importance of each interaction is estimated using the out-of-bag (OOB) approach, which is similar to that used in Random Forests. During each iteration, about 60-65 percent of the subjects are drawn to become the bootstrap samples for the construction of a logic tree. The remaining subjects which are not included in the construction are called OOB samples. In the case-control study, the importance of an interaction P is estimated as the value of the variable importance measure (VIM) which is the average difference in the misclassification rate of OOB samples with and without the interaction P in the logic regression model over all iterations of LogicFS, i.e.,

$$\text{VIM}_{\text{single}} = \frac{1}{b}\left(\sum_{b:P \in I_b} (N_b - N_b^-) + \sum_{b:P \notin I_b} (N_b^+ - N_b^-)\right), \tag{7}$$

where I_b is a set of all interactions identified in the bth iteration, $b = 1, \ldots, B$, N_b is the number of OOB samples that are correctly classified with P in the model and N_b^- is the number of OOB samples that are correctly classified without P in the logic expression. Similarly, N_b^+ is the number of OOB samples that are correctly classified when P is added to the logic expression when P was not originally included in the expression.

2.4 Genetic Programming for Association Studies (GPAS)

Genetic Programming for Association Studies (GPAS, [32]) is, as the name suggests, a genetic programming (GP, [24]) approach for genome-wide association studies. Unlike all methods discussed so far, GPAS does not require the fitting

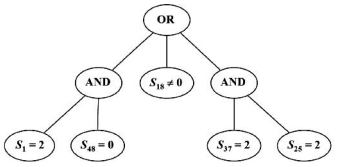

Fig. 2. An example of an individual in the GPAS algorithm. There are five literals and two monomials. $S_1 = 2$ indicated SNP 1 is AA (or aa, depending on user's preference), and it is called a literal. An example of a monomial is $S_1 = 2$ AND $S_{48} = 0$.

of (4), but directly searches for logic expressions in DNF using the GP method.

Fig. 2 is an example of an individual (tree) in GPAS. Although there are some similarities between Figs. 1 and 2, these are essentially quite different. First, in contrast to other methods, variables in GPAS can be polytomous. Thus, SNPs can be coded as 0, 1, and 2, and consequently, when applied to GWAs, it is not necessary to recode the genotypes.

Because GPAS is based on the concept of genetic programming, the terminology used in this approach is more aligned with biological evolutionary terminology than that of LR. For example, the logic tree of the LR is referred to as an "individual" in GPAS and the combination of many individuals becomes a "population." Moreover, the "literal" of GPAS is the same as a leaf of a tree in LR, and a "monomial" refers to a case where two or more literals are connected with an AND-operator, which is similar to the interaction of two SNPs. For example, there are five literals and two monomials in Fig. 2. For the consistency of this paper, we converted the GPAS terminology into comparable terms of the LR.

Like other searching methods, GPAS is also an iterative approach. The algorithm starts with a random population of two individuals, each consisting of randomly selected SNPs. A new set of individuals is generated as candidates for the next iteration (or so-called generation). These candidates are generated in three different ways. First, all individuals of the current generation automatically become candidates for the next generation. Second, two individuals are randomly selected from the population and a "crossover" is performed by randomly selecting a part of an individual (namely monomial) and attaching the selected part to the other individual to form a new individual. Third, five different moves (mutation or alteration) are applied to randomly selected individuals. The moves (mutation) in GPAS include inserting a literal (adding a SNP), deleting a literal (removing a SNP), replacing a literal with another literal, inserting a new monomial (adding a new "AND" combination) and deleting a monomial (deleting a SNP_x AND SNP_y). These additions and deletions are performed at random, meaning that the locations of deletion/insertion are chosen at random and items to be inserted are also chosen at random.

After having generated a pool of candidates, a set of individuals is then selected from the pool to form the next generation. The selection criterion used in GPAS is called "fitness," which aims to balance the number of correct classifications (NCR) of both cases and controls and to also penalize the size of the classifier, s. The fitness of the GPAS tree in the ith iteration of GPAS is expressed as a set of objectives

$$\text{fitness}_i = (\text{NCR}_i^{\text{Cases}}, \text{NCR}_i^{\text{Controls}}, s_i). \quad (8)$$

An individual is said to be *dominant* to others if at least one of the objectives is superior and none of the objectives is inferior. Only the dominant individuals are then selected for the next generation. This selection process is called *domination selection* [32]. The iteration repeats until either the number of generations reaches the predetermined number of generations, or the desired fitness level is achieved.

The size of an individual is restricted in GPAS, although it is possible to have more monomials in an individual. Nunkesser et al. [32] limited the individual to only one monomial.

2.5 Modified Logic Regression—Gene Expression Programming (MLR-GEP)

Although MLR-GEP [27] is based on LR, it is actually more closely aligned with GPAS. Since MLR-GEP has the aim of identifying SNP interactions, the model parameters of (4) are considered to be less relevant and are thus ignored. The advantage of this approach is it increases the computational efficiency, thereby making it more capable of accommodating the computational burden of GWAs. Using the same notation as earlier, the MLR model becomes:

$$g(y) = \sum_{i=1}^{K} L_i, \quad (9)$$

where $g(.)$ is a link function. For a case-control study, the most commonly used link is logit. The stochastic searching algorithm used in MLR-GEP is the Gene Expression Programming (GEP, [10]), which is a hybrid of genetic algorithms (GA, [18]) and genetic programming (GP, [24]).

The terminologies of GPAS and MLR-GEP are interchangeable with a key difference in the definition of an "individual." In GA, individuals are linear strings with fixed length, whereas in GP, individuals are nonlinear objects with different sizes and shapes. GEP combines the features of individuals of GP and GA, leading to individuals of GEP encoded as strings with fixed length, which can be later expressed as nonlinear objects with different shapes and sizes. Therefore, GEP has the advantages of both GA and GP, with the ease of manipulation of GA and the functional complexity of GP.

The linear string in GEP is referred to as a "gene," and a gene is composed of "nodes" representing either functions (i.e., Boolean—AND, OR, and NOT) or "terminals" (i.e., SNPs). A number of genes can be linked by functions to form a "chromosome." The structure of the GEP gene is divided into a "head" and a "tail" (Fig. 3). The head contains both functions and terminals, whereas the tail contains only terminals. The first head node of each gene, or "root" node, must be a function. The tail length is a fixed function of the head length and the maximum function arity (number of function arguments). The structure of the GEP gene and the translation system from a fixed length string to an expression tree guarantees that all modifications arising from evolution of the individuals result in syntactically correct expression trees (ETs). Despite the fixed length of the GEP genes, they have the potential to code for ETs of widely differing shapes and sizes. The number and length of GEP genes is peculiar to the problem at hand.

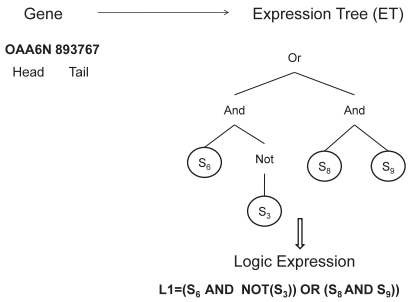

Fig. 3. An example of an individual in MLR-GEP, showing the translation of single string to an object of shape and size. The length of the gene is fixed, therefore node 767 at the end of the gene tail is redundant.

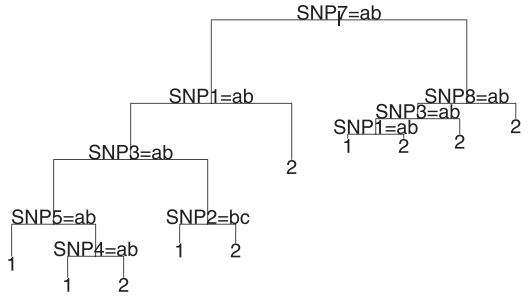

Fig. 4. An example of a classification tree in RF, where 1 and 2 are the disease status. This tree contains 10 terminal nodes and 9 binary splits. Code a, b, and c represent genotype aa, aA, and AA.

The moves (also called mutations or genetic operations) of MLR-GEP can take place at genes and chromosomes, and include mutation, transposition, insertion of sequence, root insertion of sequence, and recombination. Mutation is a change occurring in a single node of a gene and can occur at both the head and tail of a gene. When it occurs in the gene head (other than at the root node) it may produce either a function or terminal, whereas tail mutation must result in a terminal. Transposable elements of GEP are fragments of the genome that can relocate to another place in the chromosome. Insertion Sequence (IS) elements are short fragments with a function or terminal in the first position that may transpose to the head of genes except the root. Root Insertion Sequence (RIS) elements are short fragments with a function in the first position, and which transpose to the root of genes. In addition, an entire gene may transpose to the beginning of the chromosome (gene transposition). Recombination in GEP is similar to crossover in GPAS. It may take one of three forms. In all cases, two parent chromosomes are randomly chosen and paired to exchange "genetic" material. During one-point recombination, two parent chromosomes crossover at a randomly chosen point to form two daughter chromosomes. During two-point recombination, two parent chromosomes exchange the fragment contained between two randomly chosen points to form two daughter chromosomes. In gene recombination, an entire gene is exchanged during crossover. "Elitism," or the survival and cloning of the best individual chromosome in each generation into the next generation, is practiced.

Like GPAS, GEP individuals are selected according to their fitness. In contrast to GPAS, the fitness here is defined as the ability of the solution to predict the case/control status of each datum. This is the same as the correct classification. For any GEP individual i, the fitness is

$$\text{fitness}_i = \sum_{j=1}^{J} (c_{ij} = T_j), \tag{10}$$

where J is the number of subjects j in the data set, T_j is the case/control status for the subject j, and c_{ij} is the predicted case/control status under GEP individual i for subject j.

Like GPAS, MLR-GEP starts with randomly generated individuals (not limited to two), then evaluates the fitness of all individuals. Each individual is altered with one of the moves described earlier. The fitness of altered individuals is evaluated. Individuals with reasonable fitness then evolve into the next generation. Like GPAS, the process continues until a predetermined number of generations is achieved, or until a desired fitness is achieved. Finally, the interactions of SNPs are identified from the surviving expressions where SNPs are connected by Boolean operator "AND."

2.6 Random Forests

Random Forests (RF, [4]) is a method which involves a collection of numerous classification or regression trees (CART, [5]). CART is a simple statistical tool applying recursive binary partitioning of the feature space. CART is well known for its efficiency in coping with large data sets. However, as the data become noisier, and less information is contained in each variable, the predictive ability of CART diminishes. RF overcomes this problem by introducing random elements into the model by which subsets of variables are chosen at random and bootstrap samples are selected with replacement for tree growing.

Although the Boolean operators are not physically present in the actual CART structure, the CART tree can be translated into a combination of SNPs, AND- and OR-operators. For example, Fig. 4 is an example of a classification tree. Following the far right path of this figure, it is equivalent to "when an individual has genotype AA at SNP 7 and genotype AA at SNP 8, this individual is more likely to have the phenotype." Moreover, in contrast to LR, CART trees aim to predict both affected and nonaffected individuals. Because variables of RF can have more than two levels, the coding of SNP can remain in the original genotype forms, i.e., "aa," "aA," and "AA."

A binary split is denoted as a node, and is defined as a parent or a child. For instance, in Fig. 4, SNPs 1 and 8 are the children nodes of SNP7. A leaf is where the splitting terminates (also called terminal nodes). The training data sets is first split into two subsets using the criteria which resulted in the lowest misclassification rate, i.e., genotype "aa" at SNP7 in the example tree shown in Fig. 4. The binary splitting continues until the child nodes have a reasonable level of homogeneity, or the sample sizes (n) of the child nodes are smaller than a prespecified value. In the standard CART, the trees are required to be pruned/shrunk to avoid overfitting; however, this is not required in RF.

The error rate of RF depends on the correlation between any two trees in the forest and the strength of individual trees. Higher correlation between trees in the forest results in a higher error rate, and greater strength of trees reduces the error rate. These two indicators are affected by the size of the subset of variables used in tree building. Reducing

the size of the subset also reduces both correlation and strength. The optimal size of the subset is not directly estimated from the data, but determined by users [4].

The prediction error of RF is estimated using OOB samples, which are the same as described in LogidFS. At each bootstrap iteration, the prediction of OOB samples is estimated from the tree grown in that iteration. The OOB error is the average of the ratio of the number of times that OOB cases are misclassified to the number of times the respective case is an OOB sample, across the entire forest.

RF provides a variable importance ranking via the variable predictive importance, which is estimated also using the OOB cases. The importance of variable j is estimated as the average difference between the correct classification rate of OOB cases, and the correct classification rate of OOB cases with the value of the variable of interest (j, in our example) replaced with a randomly permuted value over all trees.

Variables, j and k, say, are defined here as interacting if, when one variable is used for a split, the other variable is systematically more or less likely to be used for another split. The measurement used for the interaction importance ranking is the gini index. The gini value is calculated and ranked for each tree and each pair of variables within the tree. The absolute difference between the rank of the tree and the rank of a pair of variables is the gini measure for that pair of variables, which is then averaged across the forest.

2.7 Bayesian Logistic Regression with Stochastic Search Variable Selection (BV)

The last method included in this paper is different from methods described so far. This model does not have a tree-like structure, but is instead based on logistic regression of a dichotomous phenotype [28] in conjunction with a stochastic search algorithm for variable selection. Stochastic search variable selection (SSVS) using MCMC [13], [14] is a commonly used model for variable selection in the Bayesian framework. The earliest implementation of this model for genetic research was for the identification of multiple quantitative trait loci for complex traits [42]. Similar methods have also been applied to SNP data [7], [11]. BV is different from the Bayesian epistasis association mapping (BEAM) proposed by Zhang and Liu [44], which detects epistasis effects by applying a Bayesian method to partition the markers into three groups: markers unlinked to the disease risk, markers contributing independently to the disease risk, and markers jointly influence the disease risk, and then confirms the association using a frequentist approach. In contrast, BV assumes both independent and epistasis SNP effects can be modeled in a linear framework. Letting Y_i denote the phenotype of individual i and q_i be the probability of individual i having the phenotype, the typical logistic model is

$$\log\left(\frac{q_i}{1-q_i}\right) = \mu + \sum_{s=1}^{n_s} \nu_s x_{is} + \varepsilon_i, \tag{11}$$

where μ is the population mean, x_{is} is the genotype of SNP s for individual i, ν_s is the coefficient of x_{is}, and n_s is the total number of SNPs. Instead of using SSVS proposed in [13], we implement a variation of SSVS, which is more closely aligned with the one discussed in [7]. Let z_s be a latent indicator variable, where $z_s = 0$ indicates that SNP s is not in the model, conversely, $z_s = 1$ indicates that SNP s is included in the model. Assuming that genotypes are diallelic, $x_s \in \{0, 1, 2\}$, the model then becomes

$$\log\left(\frac{q_i}{1-q_i}\right) = \mu + \sum_{s=1}^{n_s} z_s \sum_{l=0}^{l=2} \nu_{s_l} g_{is_l} + \varepsilon_i, \tag{12}$$

where g_{is_l} is an indicator variable taking the value of 0 or 1 depending on whether individual i has genotype l at SNP s. The parameter ν_{sl} is the contribution of genotype l at SNP s to the expression of the phenotype and ε_i is the residual. This single model can be easily built upon to incorporate two-way interaction effects, so that

$$\begin{aligned} Y_i = \mu &+ \sum_{s=1}^{n_s} z_s \sum_{l=0}^{2} \nu_{s_l} g_{is_l} \\ &+ \sum_{j=1}^{n_s} \sum_{k=1, j\neq k}^{n_s} \eta_{jk} \sum_{l_j=0}^{2} \sum_{l_k=0}^{2} \gamma_{jl_jkl_k} g_{ijl_jkl_k} + \varepsilon_i, \end{aligned} \tag{13}$$

where η_{jk} is an indicator variable, with $\eta_{jk} = 1$ if the SNP $j \times k$ is included in the model, else 0. The parameter $\gamma_{jl_jkl_k}$ is the contribution due to the interaction between genotype l of SNP j and genotype l of SNP k. Similarly, $g_{ijl_jkl_k}$ is an indicator variable taking the value of 0 or 1 depending on whether individual i has genotype l at SNP l_j and genotype l_k at SNP k.

The importance of SNP s is measured as the number of times that SNP s is included in the iterations after burn-in over the total number of post burn-in iterations. The importance measure is thus confined between 0 and 1. The importance of SNP interactions is also estimated following the same paradigm.

In the following examples, we used noninformative priors for all parameters, as follows:

$$\begin{aligned} &\varepsilon \sim \text{Normal}(0, \tau^{-1}); \quad &&\tau \sim \text{InverseGamma}(0.05, 0.05); \\ &z \sim \text{Bernuolli}(p_z); \quad &&\mu, \nu, \gamma \sim \text{Normal}(0, 1); \\ &\eta \sim \text{Bernuolli}(p_\eta); \quad &&p_z, p_\eta \sim \text{Uniform}(0, 1). \end{aligned} \tag{14}$$

Model parameters were estimated using a Gibbs sampling algorithm. With the exception of z and η, all parameters have nonstandard conditional distributions, so a slice sampler [31] was used. The estimation of z and η was based on a combination of Gibbs and Metropolis-Hasting algorithms [6]. At each MCMC iteration, the value of z and η depend on the ratio of the conditional posterior probability of the model including and excluding a SNP. For example, if the condition posterior probability of the model with SNP i is larger than the model without SNP i if the ratio exceeds a random value drawn uniformly between 0 and 1, then z_i is assigned with value 1, else 0.

Ten independent chains were generated with 100,000 iterations each. The first half of the iterations of each of the chains were treated as the burn-in and the variable importance measures were derived from the last 50,000 samples, that is the number of times the SNP or the SNP interaction is included in the model at each of the remaining 50,000 iterations. The convergence of MCMC chains was assessed by comparing the model likelihoods of different simulation sequences, all of which started from different points.

TABLE 1
The Four Conjunctions $P_1, \ldots, P_4$ Used in the First Simulation

Conjunction	Interaction	Number of Cases (Controls)	Proportion of Data
P_1	$S_{1,2}$	100(0)	10%
P_2	$S^c_{2,1}$ and $S_{3,2}$	150(0)	15%
P_3	$S_{4,2}$ and $S_{5,2}$ and $S_{6,2}$	100 (0)	10%
P_4	$S_{7,2}$ and $S_{8,2}$	150(0)	15%
No	None	0 (500)	50%

These represent SNP interactions responsible for the presence of the phenotype. The number of cases simulated for each conjunction and the proportion of the observations described by each of these conjunctions are summarized in the third and fourth column. The last row indicates the number of controls included in the data set, which made up half of the total population.

3 Data

We use two data sets to evaluate the performance of the six methods described in the previous sections. These comprised a simulated data sets and a real data set obtained from the GENICA study [20].

3.1 Simulated Data

For each of these fifty data sets, 500 cases and 500 controls are generated so that for each case exactly one of the conjunctions $P_1, \ldots, P_4$, summarized in Table 1, is true, and none of these conjunctions is true for any of the controls. Thus, employing the logic expression

$$L = P_1 \vee P_2 \vee P_3 \vee P_4,$$

as classification rule leads to a correct classification of all 500 cases and 500 controls in each of the 50 data sets. Apart from the values of the informative SNPs, i.e., the SNPs forming $P_1, \ldots, P_4$, the genotypes of the noninformative SNPs are randomly drawn with a minor allele frequency randomly selected in the range from 0.2 to 0.4.

Similar methods of simulation were also implemented by Schwender [40] and Nunkesser et al. [32].

3.2 Real Data: GENICA

The GENICA study is an age-matched and population-based case-control study that has been carried out by the Interdisciplinary Study Group on Gene ENvironment Interaction and Breast CAncer in Germany, a joint initiative of researchers dedicated to the identification of genetic and environmental factors associated with sporadic breast cancer. Further details on the GENICA study, such as data collection and cleaning, are in [20].

In this paper, we focus on a subset of the genotype data from the GENICA study. More precisely, data of 1,234 women (609 cases and 625 controls) and 39 SNPs belonging to the estrogen, the DNA repair, or the control of cell cycle pathways are considered in the analyses.

Because a few of the women show a large number of missing genotypes, all observations with more than three missing values are removed from the analysis leading to a total of 1,199 women (including 592 cases and 607 controls). The remaining missing genotypes are imputed by a weighted k nearest neighbors approach described in [40] and implemented in the R package *Scrime* [34].

4 Results

Table 2 provides a parallel comparison of features of all the methods included in this study. The comparison is mainly focused on the difference in structure of the methods, genetic implementation, alterations allowed from one state to another and tree structures. Among all methods, even though the structure of RF and BV does not directly utilize boolean operators, the tree of RF can potentially be interpreted as a combination of "OR," "AND," and SNPs, while the addition (+) of BV is similar to "AND."

To prevent a local maximum, all methods required adaptation of some form. For LogicFS, GPA, and MLR-GEP, this is achieved by repeating the analysis a number of times. For methods utilizing a form of MCMC (MCLR and BV), this is done by using multiple chains. RF avoids a local maximum by generating multiple trees in the forest and basing inferences on the results of the forest. We present here the results after these types of repetition, i.e., after applying each of the approaches once to each of the fifty simulated data sets, and fifty times to the GENICA data sets with different starting points of the search.

In the simulated data sets, although the methods compared here are somewhat different, except for the RF, all other methods are able to identify at least some of the prespecified SNPs. Among six methods compared in this paper, only LogicFS, MCLR, RF, and BV provide rankings for the variable importance. Of all these methods, LogicFS most successfully identifies all four SNP interactions in each of the fifty data sets with relatively large importance (usually, shown in the Top 4 rankings). For MCLR, only one of the four interactions is always detected, namely P_2. The other interactions, P_1, P_3, and P_4, are identified in 90, 50, and 80 percent of the fifty samples, respectively. RF, on the other hand, did not identify any of these conjunctions in its interaction rankings. However, when considering individual SNPs separately, SNPs involved in the interactions all appeared with high rankings.

After 50,000 iterations, BV is able to identify two-way interactions, namely P_2 and P_4, in all fifty data sets. Because the BV model we used here is designed for detecting only the main and/or two-way interaction effects, it is not possible to identify the three-way interaction (P_3 of Table 1). However, the effects of the three-way interaction can be identified by BV as subsets of three-way interactions, i.e., S_4 AND S_5, S_4 AND S_6, and S_5 AND S_6. The conjunction, P_1 on the other hand, is often identified as a part of an interaction effect rather than a solitary effect.

Similar to the results of LogicFS, GPAS detects all four SNP interactions explaining the cases in each of the fifty data sets. However many nonrelated SNPs are also identified.

MLR-GEP is limited in identifying many conjunctions. Of all interactions listed in Table 1, the only conjunction consistently identified is when the SNP is a main effect, namely P_1. The conjunction with the second highest chance of detection is P_4 with an average of over 50 percent; however, the chance of detecting this interaction varies from 10 to 100 percent. The other two conjunctions, P_2 and P_3, on the other hand were not found under the MLR-GEP approach.

When applying the methods to the GENICA data, except for RF, all other methods identify a probable association of the interaction of ERCC2_18880 and ERCC2_6465 with sporadic breast cancer. These two SNPs are from the Excision Repairs Cross-Complementing group 2 region (ERCC2, formerly XPD). LogicFS, MCLR, GPAS, and MLR-GEP all indicate that having the homozygous reference genotype at ERCC2_6465 and either heterozygous or

TABLE 2
Parallel Comparisons of Features, Genetic Implementation, Alteration (Move) and Tree Structures of LR, LogicFS, MCLR, GPAS, MLR-GEP, RF, and BV

Methods	LR	logicFS	MCLR	GPAS	MLR-GEP	RF	BV
Features							
Model based	y	y	y	n	n[1]	n	y
Iterative searching Algorithm							
Require (y/n)	y	y	y	y	y	n	y
Algorithm	Simulated Annealing	Simulated Annealing	RJMCMC	Genetic Programming	Gene Expression Programming	NA	MCMC (Gibbs+MH)
Iterative/Evoluationary[2]	I	I	I	E	E		I
Quantify Interactions	n	y	y	n	n	y	y
Use Boolean	y	y	y	y	y	n[3]	n[3]
Boolean Operators	AND, OR	AND,OR	AND, OR	AND, OR	AND, OR	OR, AND[3]	AND[3]
Genetic Implementation							
SNP Coding	R/D[4]	R/D	R/D	A/F	A/F	A/F	A/F
LD	y[5]	y[5]	y	y[5]	y[5]	y	y
Max SNPs	1000	1000	1000	GWAs[6]	at least 23,000	*	at least 23,000[7]

[1] *Although it is based on LR, the parameters are ignored.* [2] *Iterative (I) indicates a state depends immediate previous state only, Evolutionary (E) indicates a state depends previous states.* [3] *Strictly, RF and BV do not have Boolean operators, however, the trees of RF can be interpreted as combination of OR and AND. Similarly, the additive of BV model is like AND operator.* [4] *RD: Recessive/Dominance; A/F: Allele Frequency.* [5] *Although LD is not directly considered in the method, LD can be detected via runs with different starting points.* [6] *Nunkesser et al. [32] stated that GPAS is able to analyze the GWA data, however it is yet to be verified.* [7] *Considering the additive effect only.* *Unclear.

homozygous genotype at ERCC2_18880 is likely to increase the chance of breast cancer. This result is also supported by BV with more detail. According to the results of BV, the highest chance of developing sporadic breast cancer is when individuals show the homozygous genotype at ERCC2_18880 and homozygous reference genotype at ERCC2_6465 with an odds ratio of 4.17 (CI: 2.63-6.67), followed by individuals with heterozygous genotype at ERCC2_18880 and homozygous reference genotype at ERCC2_6465 with an odds ratio of 2.37 (CI: 1.01-5.58).

Another interesting finding which is identified only by the BV approach is the functionality of ERCC2_6465. The results of BV show that ERCC2_6465 is potentially associated with the sporadic breast cancer in two different ways, by acting as a solitary additive effect or by interacting with SNPs other than ERCC2_18880.

5 Discussion

In this study, we review different variations of logic regression, Random Forest, and Bayesian logistic regression with stochastic search variable selection, for their ability to identify SNP interactions. The methods are then discussed and compared using simulated and real data sets.

In the simulated evaluation, because the data are simulated with the conditions closely aligned with logic regression, i.e., using Boolean expression,"AND," "OR," "NOT," it is not surprising that the overall results are more favorable for logic tree-based approaches. GPAS and LogicFS both identified all expected SNPs interaction of the simulation data. In contrast, BV is a regression type approach which does not use Boolean operators and the level of interactions between variables is required to be specified prior to analysis (i.e., the current coding of BV was only designed to detect up to two-way interactions). However, considering all these potential constraints, BV showed better results in detecting the conjunctions compared with RF and MLR-GEP.

Among the different methods, the results of the RF analysis of the simulation data are the most unexpected. Although the RF is a tree-based method, it did not identify any conjunctions listed in Table 1. However, when considering SNPs at an individual level, these SNPs involved in the interactions were all successfully identified by RF with relatively high importance measures. The same pattern was also found in the results of the analysis of the GENICA data: even though RF did not find the interaction of ERCC2_18880 and ERCC2_6465 to be important, these two SNPs were the top two ranking SNPs when SNPs were considered individually.

These findings reflect the problem with the definition/measurement of interaction importance that is currently implemented in the RF code. The program we used for carrying out the analysis is not the *randomForest* package of R, but the Fortran code available from the author's website.[1] In this version of RF, the importance of a pair of variables is defined as the absolute difference between the ranking of the pair and the ranking of the tree which is averaged across the forest. Although developers of this code stated that

1. http://www.stat.berkeley.edu/breiman/RandomForests/.

TABLE 2
(Cont.) Parallel Comparisons of Features, Genetic Implementation, Alteration (Move) and Tree Structures of LR, logicFS, MCLR, GPAS, MLR-GEP, RF and BV (Continued)

Methods	LR	logicFS	MCLR	GPAS	MLR-GEP	RF	BV
Tree Structure							
Have Tree Structure	y	y	y	y	y	y	n
Boolean[1]	y	y	y	y	y	n	
Operators	y	y	y	y	y	n	
Node	B	B	B	B	B	S	
Terminal Node	S	S	S	S	S	P	
Binary/Mutiple Split	Binary	Binary	Binary	Multiple	Binary	Binary	
Fitness Measure	MCR	MCR	MCR	Multiple NCR	NCR	OOB MCR	
Moving between States	Acceptance Prob	Acceptance Prob	RJMCMC	Fitness	Fitness	NA	
Alteration							
Allow Alteration[2]	y	y	y	y	y	n	n[9]
No. Alterations	6	6	6	7	5	2	
Method of Alteration							
Change SNP[3]	√	√	√	√	√		
Change Boolean[4]	√	√	√		√		
Grow Branch[5]	√	√	√	√	√	√	
Prune Branch[6]	√	√	√	√	√		
Split leaf	√	√	√	√	√	√	
Delete leaf	√	√	√	√			
Crossover[7]				√	√		
Insert new split at Root node				√			
Require Pre-setting[8]	√	√	√		√		

[1] *Tree structure.* [2] *Changes made to the tree of current state.* [3] *Change SNP with another SNP.* [4] *Change Boolean with another Boolean.* [5] *Adding a part to existing tree.* [6] *Deleting a part of existing tree.* [7] *Exchange parts between two trees.* [8] *Strictly, the model does not have these alterations. However, some alterations are equivalent to the addition and deletion embedded in BV.* [9] *Need to assign the probability to each alterations prior to analysis. B-Boolean operators, S-SNPs, P-Prediction (case or controls).*

"caution" is required for the interpretation of the interaction effects, the results confirm the problem of using such criteria. This criterion is only useful for detecting the interaction of a pair of SNPs, say A and B, when these two SNPs are often selected jointly in the random selection of the potential predictors used for tree growing. Furthermore, this criterion is easily obscured due to the nature of recursive partitioning embedded in CART. For example, using the dummy example of Fig. 4, at the root node, the training samples are split into two subgroups, one group with genotype aa and aA at SNP 7, while the other group has the complement genotype at the same SNP. The further splitting of these two subgroups depends only on the structure embedded within each subgroup, i.e., the splitting which resulted in the most reduction of the impurity measure within that subgroup. Therefore, unless the interaction of SNP A and B is prominent in the subsets, the importance of these two SNP interactions is likely to be overlooked using current criteria.

Although the interaction cannot be identified directly under current settings, the interaction effects are captured by the solitary variable importance measured using the permutation methods and OOB samples. The assertion is confirmed in [26]. Therefore, with some improvements, RF can be a useful tool for identifying SNP interactions. For instance, Jiang et al. [19] suggest the use of a sliding window sequential forward feature selection in conjunction with statistical testing to find epistasis effects.

The detection of false informative SNPs is commonly observed across all methods; however, it is difficult to compare the false positive and false negative rates of these methods. GPAS and MLR-GEP identify a set of SNPs showing possible association without giving a quantitative measure, such as variable importance ranking, to show the degree of association between a SNP and disease. In this study, the set of possible models according to GPAS is exponentially large, and without the variable importance ranking, it is more difficult to identify the false informative SNPs in the real data. Despite the fact that the ranking of variable (interaction) importance is available in other methods, an appropriate threshold point for these measures is still not well understood. This is because a threshold point may potentially depend on the underlying genetic model and the ratio of the causal and noise SNPs, which is often impossible to know prior to the analysis [26]. Therefore, instead of basing conclusions on the results of a single method, a more sensible approach is to analyze data with different methods and to compare the results.

Further investigation on how to integrate the results of different methods would be beneficial.

Methods incorporating tree-based structures are more robust in identifying the higher order interactions (e.g., three or more way interactions). In the tree-based methods, higher order interactions are directly identified from a tree or a collection of trees. In contrast, to find higher order interactions using regression models, the order of interactions needs to be specified a priori. Moreover, as the number of terms increases in a regression model, the parameter space increases exponentially and consequently reduces the computational feasibility which is especially difficult in a genome-wide association study.

BV, on the other hand, gives better results for understanding the allele effects on the expression of the phenotype. This information is available from the magnitude of the coefficients of the different terms. For example, the coefficient of g_{sl} gives the relative measure of the effect of the genotype l of the SNP s. BV also provides a quantified measure of the risk of having the phenotype for different genotype combinations at causal loci.

Among all methods, GPAS and MLR-GEP are the only methods capable of coping with the intensity and computational power required for the analysis of large data sets. This is because these algorithms are based on a machine learning algorithm (i.e., GP and GEP). LogicFS and MCLR, on the other hand, are limited to a maximum of 1000 SNPs in the written code. It is noted that BV has been used for finding individual SNP additive effects (but not for two-way or higher interactions) for up to 23,000 SNPs. Unless more effective programming or a fast searching algorithm is adopted, most of the methods described here are only suitable for candidate gene search or fine mapping.

The major drawbacks of GPAS and MLR-GEP are in the accuracy and specificity of the identification of important interactions. Both of these methods implemented a machine learning algorithm, and although fast, the results are less reliable. This problem is especially noticeable in MLR-GEP. The performance of MLR-GEP can be improved in various ways, such as paying greater attention to the parameter setting in the evolutionary process, incorporating model parameters and use of more sophisticated fitness measures [27].

The most relevant genetic questions for such models concern their ability to detect genetic heterogeneity and linkage disequilibrium (LD) SNPs, and the effect of LD SNPs on the model. Of all methods, LogicFS is expected to be less capable of identifying any of these effects given that it is highly related to logic regression and has therefore inherited the same shortcomings identified in [23]. However, this problem can arguably be solved by applying LogicFS for several repetitions to several subsets of the data sets thereby identifying a large number of different models.

All other methods potentially have strategies for detecting genetic heterogeneity. Bayesian methods (MCLR and BV) identify heterogeneity from a collection of multiple models [23] and/or the use of various Markov chains [6]. In GPAS and MLR-GEP, by repeating the analysis with different starting populations, the heterogeneities are potentially identifiable from a collection of tree structures. In these two methods, trees are connected by the "OR" operator and the subtree therefore represents different possible genetic pathways. Similarly, in RF, genetic heterogeneity can be determined from trees nested within the full tree.

When LD SNPs are in the data sets, Bayesian approaches again have the advantage of multiple chains. When two SNPs are highly correlated, if one SNP is selected in the model, although the chance of the other SNP being selected is very small, it does have an equal chance of being selected in the model. When the number of chains (or models) is large enough, the LD SNPs are identified. In RF, LD SNPs are identified as surrogate variables. However, as noted by Lunetta et al. [26], correlated SNPs can diminish the variable importance ranking.

Although some of the methods included in this study have the same foundations, they are manifestly different in various facets. Each method has its advantages, and conversely some limitations. Even so, the methods included in this study, in general, are superior in identifying SNPs in which the effect of the SNP is highlighted by the presence of other SNPs. For instance, although the results of the analysis are not included here, we tested the SNP effect of the GENICA data using SNP-by-SNP Fisher's exact test and found the p-value of ERCC2_18880 is far from significant (p-value $= 0.106$, prior to power adjustment).

None of the methods included in this study, exhibits distinct superiority over another. In conclusion, the GPAS and MLR-GEP may be preferred for searching through large dimensional spaces; LogicFS, MCLR, RF, and BV may be preferred for candidate gene/region searches, and BV may be preferred for providing detail on the allele effects.

REFERENCES

[1] A.S. Andrew, M.R. Karagas, H.H. Nelson, S. Guarrera, S. Polidoro, S. Gamberini, C. Sacerdote, J.H. Moore, K.T. Kelsey, and E. Demidenko, "DNA Repair Polymorphisms Modify Bladder Cancer Risk: A Multi-Factor Analytic Strategy," *Human Heredity*, vol. 65, no. 2, pp. 105-118, 2008.

[2] B. Atik, T.A. Skwor, R.P. Kandel, B. Sharma, H.K. Adhikari, L. Steiner, H. Erlich, and D. Dean, "Identification of Novel Single Nucleotide Polymorphisms in Inflammatory Genes as Risk Factors Associated with Trachomatous Trichiasis," *PLoS ONE*, vol. 3, no. 10, pp. e3600, 2008.

[3] L. Breiman, "Bagging Predictors," *J. Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees.* Chapman and Hall CRC, 1984.

[6] C.C.-M. Chen, K. Mengersen, and J.M. Keith (In prep), "Bayesian Method for Genome-Wide Association Studies: Review and Illustration,"

[7] D.V. Conti and W.J. Gauderman, "Snps, Haplotypes, and Model Selection in a Candidate Gene Region: The Simple Analysis for Multilocus Data," *Genetic Epidemiology*, vol. 27, no. 4, pp. 429-441, 2004.

[8] H.J. Cordell, "Epistasis: What it Means, What it Doesn't Mean, and Statistical Methods to Detect it in Humans," *Human Molecular Genetics*, vol. 11, no. 20, pp. 2463-2468, 2002.

[9] H.J. Cordell, "Detecting Gene-Gene Interactions that Underlie Human Diseases," *Nature Rev. Genetics*, vol. 10, no. 6, pp. 392-404, 2009.

[10] C. Ferreira, "Gene Expression Programming: A New Adaptive Algorithm for Solving Problems," *Arxiv Preprint cs.AI/0102027*, 2001.

[11] B.L. Fridley, "Bayesian Variable and Model Selection Methods for Genetic Association Studies," *Genetic Epidemiology*, vol. 33, no. 1, pp. 27-37, 2009.

[12] A. Fritsch and K. Ickstadt, "Comparing Logic Regression Based Methods for Identifying SNP Interactions," *Lecture Notes in Computer Science,* vol. 4414, pp. 90-103, 2007.

[13] E.I. George and R.E. McCulloch, "Variable Selection via Gibbs Sampling," *J. Am. Statistical Assoc.,* vol. 88, no. 423, pp. 881-889, 1993.

[14] E.I. George and R.E. McCulloch, "Approaches for Bayesian Variable Selection," *Statistica Sinica,* vol. 7, pp. 339-374, 1997.

[15] P.J. Green, "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika,* vol. 82, no. 4, pp. 711-732, 1995.

[16] A.G. Heidema, J.M.A. Boer, N. Nagelkerke, and E.C.M. Mariman, "The Challenge for Genetic Epidemiologists: How to Analyze Large Numbers of SNPS in Relation to Complex Diseases," *BMC Genetics,* vol. 7, no. 23, 2006, doi: 10.1186/147/2156-7-23.

[17] J. Hoh, A. Wille, and J. Ott, "Trimming, Weighting, and Grouping SNPS in Human Case-Control Association Studies," *Genome Research,* vol. 11, no. 12, pp. 2115-2119, 2001.

[18] J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* Univ. of Michigan Press, 1975.

[19] R. Jiang, W. Tang, X. Wu, and W. Fu, "A Random Forest Approach to the Detection of Epistatic Interactions in Case-Control Studies," *BMC Bioinformatics,* vol. 10, Suppl 1, pp. S65, 2009.

[20] C. Justenhoven, U. Hamann, B. Pesch, V. Harth, S. Rabstein, C. Baisch, C. Vollmert, T. Illig, Y.-D. Ko, T. Bruning, and H. Brauch, "For the Interdisciplinary Study Group on Gene Environment Interactions, and N. Breast Cancer in Germany," *Cancer Epidemiol Biomarkers Prev,* vol. 13, no. 12, pp. 2059-2064, 2004.

[21] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, no. 4598, pp. 671-680, 1983.

[22] C. Kooperberg, J.C. Bis, K.D. Marciante, S.R. Heckbert, T. Lumley, and B.M. Psaty, "Logic Regression for Analysis of the Association between Genetic Variation in the Renin-Angiotensin System and Myocardial Infarction or Stroke," *Am. J. Epidemiology,* vol. 165, no. 3, pp. 334-343, 2007.

[23] C. Kooperberg and I. Ruczinski, "Identifying Interacting SNPS Using Monte Carlo Logic Regression," *Genetic Epidemiology,* vol. 28, no. 2, pp. 157-170, 2005.

[24] J.R. Koza and J.P. Rice, *Genetic Programming.* Springer, 1992.

[25] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, and W. FitzHugh, "Initial Sequencing and Analysis of the Human Genome," *Nature,* vol. 409, no. 6822, pp. 860-921, 2001.

[26] K. Lunetta, L.B. Hayward, J. Segal, and P. Van Eerdewegh, "Screening Large-Scale Association Study Data: Exploiting Interactions Using Random Forests," *BMC Genetics,* vol. 5, no. 1, p. 32, 2004.

[27] P. Macrossan, C.C.-M. Chen, and K.L. Mengersen (In prep)., "Using Gene Expression Programming with Modified Logic Regression for the Investigation of SNP Interactions in Large Dimensional Data," *In Prep.,*

[28] P. McCullagh and J.A. Nelder, *Generalized Linear Models.* Chapman and Hall, 1983.

[29] L.E. Mechanic, B.T. Luke, J.E. Goodman, S.J. Chanock, and C.C. Harris, "Polymorphism Interaction Analysis (PIA): A Method for Investigating Complex Gene-Gene Interactions," *BMC Bioinformatics,* vol. 9, no. 1, p. 146, 2008.

[30] Y. Meng, Q. Yang, K.T. Cuenco, L.A. Cupples, A.L. DeStefano, and K.L. Lunette, "Two-Stage Approach for Identifying Single-Nucleotide Polymorphisms Associated with Rheumatoid Arthritis Using Random Forests and Bayesian Networks," *BMC,* vol. 1, Suppl 1, pp. S56, 2007.

[31] R.M. Neal, "Markov Chain Monte Carlo Methods Based on "Slicing" the Density Function," Technical Report No. 9722, Dept. of Statistics, Univ. of Toronto, 1997.

[32] R. Nunkesser, T. Bernholt, H. Schwender, K. Ickstadt, and I. Wegener, "Detecting High-Order Interactions of Single Nucleotide Polymorphisms Using Genetic Programming," *Bioinformatics,* vol. 23, no. 24, pp. 3280-3288, 2007.

[33] P.C. Phillips, "Epistasisthe Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems," *Nature Rev. Genetics,* vol. 9, no. 11, pp. 855-867, 2008.

[34] "R Development Core Team 2008," R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[35] T.K. Rice, N.J. Schork, and D.C. Rao, "Methods for Handling Multiple Testing," *Genetic Dissection of Complex Traits,* D.C. Rao and C.C. Gu, eds., Academic Press, 2008.

[36] M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, and J.H. Moore, "Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer," *Am. J. Human Genetics,* vol. 69, no. 1, pp. 138-147, 2001.

[37] I. Ruczinski, C. Kooperberg, and M. Leblanc, "Logic Regression," *J. Computational and Graphical Statistics,* vol. 12, no. 3, pp. 475-511, 2003.

[38] H. Schwender, "Statistical Analysis of Genotype and Gene Expression Data," PhD thesis, Dept. of Statistics, TU Dortmund Univ., 2007.

[39] H. Schwender and K. Ickstadt, "Identification of SNP Interactions Using Logic Regression," *Biostatistics,* vol. 8, no. 1, pp. 187-198, 2008.

[40] H. Schwender and K. Ickstadt, "Imputing Missing Genotypes with K Nearest Neighbors," technical report, Collaborative Research Center 475, Dept. of Statistics, Univ. of Dortmund, 2008.

[41] "The International HapMap Consortium the International Hapmap Project," *Nature,* vol. 426, pp. 789-796, 2003.

[42] N. Yi, V. George, and D.B. Allison, "Stochastic Search Variable Selection for Identifying Multiple Quantitative Trait Loci," *Genetics,* vol. 164, pp. 1129-1138, 2003.

[43] L.J. Zhao, X.G. Liu, Y.Z. Liu, Y.J. Liu, C.J. Papasian, B.Y. Sha, F. Pan, Y.F. Guo, L. Wang, and H. Yan, "Genome-Wide Association Study for Femoral Neck Bone Geometry," *J. Bone and Mineral Research,* vol. 0, pp. 1-34, 2009.

[44] Y. Zhang and J.S. Liu, "Bayesian Inference of Epistatic Interactions in Case-Control Studies," *Nature Genetics,* vol. 39, no. 9, pp. 1167-1173, 2007.

Carla Chia-Min Chen recieved the PhD degree in statistics at the Queensland University of Technology, Brisbane, Australia. Her research interests include Bayesian statistical methods, particularly in their application to genetics and the statistical aspects of genome-wide association studies.

Holger Schwender received the PhD degree in statistics at the TU Dortmund University in Dortmund, Germany. Afterward, he was a PostDoc at the Department of Biostatistics, Johns Hopkins University, Baltimore. Currently, he is a PostDoc at the Department of Statistics, TU Dortmund University. His research interests include statistical analysis of genetic data, and in particular, genetic epidemiology and statistical genetics.

Jonathan Keith received the PhD degree from the Julius Kruttschnitt Mineral Research Centre, Unviersty of Queensland. Currently, he is a senior lecturer at Monash University. His research interests include bioinformatics, computational biology, genetic epidemiology, comparative genomics, Bayesian methods, changepoint models, phylogenetics and stylometry.

Robin Nunkesser received the PhD degree from Dortmund University of Technology in the development of robust regression algorithms and a genetic programming algorithm for association studies (GPAS). Currently, he is a software engineer and consultant in an IT service company. His research interests include computational statistics and evolutionary algorithms an software engineering.

Kerrie Mengersen received the PhD degree in mathematical statistics from University of New England. Currently, she is a professor in the School of Mathematical Sciences, Queensland University of Technology. Her specific methodological interests are in Bayesian statistics, mixture models, hierarchical modeling, and metaanalysis. Her research interests include biometrics, biostatistics, environmetrics, genetic statistics, and controls. She is an accredited member of Statistical society of Australia (2001), an elected fellow of the Royal Statistical Society (2004), and an elected fellow of the Institute for Mathematical Sciences (2005).

Paula Macrossan received the PhD degree from University of New England in the development of prediction algorithms in animal breeding. Currently, she is a private consultant in animal breeding and genetics, working from a farm in rural New South Wales that is her home. Her research interests include quantitative genetics, animal genetics, and computation(al) sciences.