BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 14 2008, pages 1603–1610
doi:10.1093/bioinformatics/btn239

Genetics and population analysis

EM-random forest and new measures of variable importance for multi-locus quantitative trait linkage analysis

Sophia S. F. Lee[1,2], Lei Sun[1,3], Rafal Kustra[1] and Shelley B. Bull[1,2,*]

[1]Department of Public Health Sciences, University of Toronto, Toronto M5T 3M7, [2]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto M5G 1X5 and [3]Genetics and Genomic Biology, The Hospital for Sick Children Research Institute, Toronto M5G 1L7, Canada

Received on November 15, 2007; revised on May 16, 2008; accepted on May 17, 2008

Advance Access publication May 21, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: We developed an EM-random forest (EMRF) for Haseman–Elston quantitative trait linkage analysis that accounts for marker ambiguity and weighs each sib-pair according to the posterior identical by descent (IBD) distribution. The usual random forest (RF) variable importance (VI) index used to rank markers for variable selection is not optimal when applied to linkage data because of correlation between markers. We define new VI indices that borrow information from linked markers using the correlation structure inherent in IBD linkage data.

Results: Using simulations, we find that the new VI indices in EMRF performed better than the original RF VI index and performed similarly or better than EM-Haseman–Elston regression LOD score for various genetic models. Moreover, tree size and markers subset size evaluated at each node are important considerations in RFs.

Availability: The source code for EMRF written in C is available at www.infornomics.utoronto.ca/downloads/EMRF

Contact: bull@mshri.on.ca

Supplementary information: Supplementary data are available at www.infornomics.utoronto.ca/downloads/EMRF

1 INTRODUCTION

In genomics, proteomics, bioinformatics and other related scientific fields in which thousands of genes are investigated, variable selection is a critical task in identifying a subset of relevant genes for subsequent analysis or with good predictive performance. In quantitative trait linkage studies, variable selection is often accomplished by testing for linkage within a single locus framework, aiming to find genes with marginal effects. But for many diseases the phenotypic variation may be explained primarily by epistatic interactions (Gibson, 1996; Moore, 2003). Although linkage methods can be extended to model multiple loci and complex interactions, it is typically not done for the sake of simplicity. Pair-wise and higher order interactions can be evaluated by testing each of them individually in a model, but the number of interaction terms increases exponentially. Moreover, quantitative traits are often influenced by multiple loci (Falconer, 1989), so the number of parameters including interaction terms in a linear regression model needed to explain the variation can increase substantially requiring many degrees of freedom for estimation. The problem is further exacerbated by the requirement in standard regression models that the number of observations n be larger than the number of parameters p (i.e. $n >> p$). Searching through all possible covariate space in this fashion, especially in high-throughput genome-wide studies where thousands of genes are investigated, is challenging, exhausting and may be impractical. The disadvantage of current single-locus quantitative trait linkage methods is their inability to examine multiple loci and complex interactions and covariate space simultaneously.

A recursive partitioning method, known as trees, can be utilized to tackle these problems, particularly to detect interactions. The tree is a non-parametric predictive model designed to identify important predictors among a large number of covariates by recursively partitioning the covariate space into several more homogeneous non-overlapping subspaces. A well-known caveat of this method is its instability (Breiman, 1996a), but it can be improved by *bagging*, which involves growing a large number of trees using bootstrap samples and averaging the results, thus reducing the variance (Breiman, 1996b). Bagging results in highly correlated trees, but techniques to reduce this correlation can further decrease the variance. This is the motivation for the *random forest* (RF).

The original RF (OrigRF) is claimed to enjoy high prediction accuracy with only a single tuning parameter m_{try} (Breiman, 2001). It also incorporates a variable importance (VI) index to measure the relative importance of markers, making it a useful tool for variable selection. This attractive feature of OrigRF has made it increasingly of interest in genetic studies. It has been utilized in association studies (Bureau *et al.*, 2005; Lunetta *et al.*, 2004), microarray analysis (Shi *et al.*, 2005) and proteomic studies (Izmirlian, 2004). Bureau *et al.* (2003) applied the method to simulated data from Genetic Analysis Workshop 13 (GAW13) to detect linkage to genes governing variation in high-density lipoprotein, triglyceride and glucose levels. To our knowledge, no studies have been reported that systematically examine the performance of OrigRF in quantitative trait linkage analysis.

Genes on the same chromosome are said to be *linked* and they tend to be inherited together. The technique to detect linkage between trait loci and marker loci with *known* positions using family data is called *linkage analysis*. Linkage analysis studies the inheritance of marker alleles among relatives (Ott, 1999). The genetic similarity between two relatives is measured by the number of alleles that they have

*To whom correspondence should be addressed.

inherited of the same allelic form and from the same ancestor. This genetic sharing among relative pairs is called *identical by descent* (IBD). Under the assumption of no linkage, a pair of full siblings with unrelated parents (our focus in this report) can either share 0, 1 or 2 alleles IBD at a given locus with probability 1/4, 1/2, and 1/4, respectively. The underlying principle behind linkage analysis is that relative pairs with similar traits are expected to share more alleles at the trait loci than at unlinked marker loci. Marker loci near to the trait locus will exhibit similar, but attenuated, patterns of excess sharing.

If the IBD sharing of a sib-pair at a marker locus can be unequivocally determined (i.e. complete marker information or fully informative markers), the proportion of alleles shared IBD is either 0, 1/2 or 1 with certainty. In general, IBD sharing cannot be unequivocally determined (i.e. incomplete marker information or ambiguous inheritance) and it has to be estimated probabilistically, conditional on available genotype data, using algorithms such as the Elston–Stewart peeling algorithm (Elston and Stewart, 1971) and the Lander–Green hidden Markov model algorithm (Lander and Green, 1987). For such data, we propose a new *EM-Haseman–Elston regression tree* (EMHE Tree) and following the algorithm of OrigRF, we develop *EM-random forest* (EMRF), which is an ensemble of EMHE trees. We take advantage of the inherent correlation between marker IBD sharing and incorporate it into new VI indices based on the original VI definition and Friedman's definition of partial dependence (Friedman, 2001).

2 METHODS

The classical Haseman–Elston linkage method regresses a response variable Y, defined as the squared difference in trait values, on the proportion of alleles shared IBD, denoted by π. A negative slope is suggestive of linkage. For ambiguous marker inheritance, Haseman and Elston (1972) estimated the IBD proportion by its expected value ($\widehat{\pi}$) and then treated this as fixed in the regression. Alternatively, Kruglyak and Lander (1995) performed the regression utilizing the Expectation-Maximization (EM) algorithm (Dempster *et al*., 1977), which iterates between an E-step to estimate the posterior IBD probabilities and an M-step to estimate the regression coefficients, until convergence, thus avoiding substitution of the expected IBD under the null of no linkage. We denote the former regression by HE and the latter by EMHE. In either case, evidence for linkage at each marker can be assessed by the LOD score, which is proportional to a likelihood ratio test (LRT) statistic comparing the log-likelihood under H_0: null regression slope to that under H_A: negative regression slope.

Our EMRF method is motivated by the EMHE single-locus model which we adapt to tree and tree-ensemble models, such as the OrigRF, developed by Breiman (2001). The OrigRF is a collection of full-sized trees developed using bootstrap samples. At each node of a tree, a random subset of markers of size m_{try} is selected. In regression, the recommended level for m_{try} is one-third of the total number of markers (Liaw and Wiener, 2002). Each marker is evaluated for splitting the bootstrap sample observations into two subsets according to a cutpoint value. The conventional splitting criterion compares the mean-squared error (MSE) of all observations to the sum of the MSEs in the two subsets, and chooses the maximum. The overall prediction of the forest is the average of predictions over all the trees. The out-of-bag (OOB) observations left out of each bootstrap sample are used to estimate the prediction error (PE) of the forest.

2.1 EM-HE regression tree and EM-RF

We propose EMHE Tree and EMRF as tree-based analogues of EMHE regression. Our EMHE Tree employs an EMHE regression at every node. Using posterior probabilities for the IBD sharing values of 0, 1/2 and 1, means that only two candidate cutpoints (between 0 and 1/2 or between 1/2 and 1) have to be evaluated for each marker, similar to the situation with fully informative markers. In contrast, if the expected $\widehat{\pi}$ data are used in the HE regression, multiple candidate cutpoints have to be evaluated for each marker because $\widehat{\pi}$ can take on any values between 0 and 1.

At each split in conventional trees, sib-pairs are assigned either to the left or right child node with 100% certainty. In the final tree, a sib-pair belongs to one unique terminal node. In EMHE Tree, a sib-pair contributes to *both* children nodes at each split with probabilities determined by the split variable, the cutpoint and the associated posterior IBD probabilities estimated in the EMHE regression at the parent node. Probabilities in subsequent splits are conditional on the parent node. In the final tree, each sib-pair contributes to all terminal nodes with different probabilities depending on the tree structure, and the sum of the contributions must be one. Here, it is assumed that markers selected in the path from the root node to a leaf are unlinked. Figure A1 in Appendix A in the Supplementary Material illustrates an example of a tree model and how the probabilities are calculated.

Based on the HE regression LOD-score linkage statistic, we define a splitting rule in the EMHE Tree to select the marker and cutpoint that yields the greatest difference between the parent node and the children nodes in the total linkage statistic. This criterion is equivalent to maximizing the change in the log-likelihood ratio statistic (LLR), defined as the difference in $\log_{10}$ likelihood associated with the selected cutpoint. The rationale for the HE linkage-based log-likelihood split in contrast with the conventional MSE splitting criteria is described in the beginning of Appendix A in the Supplementary Material. An example illustrating the evaluation of two cutpoints is shown in Figure A2 in Appendix A in the Supplementary Material.

After a tree is grown, predictions are assigned to all terminal nodes. In conventional regression trees, the predicted value at a terminal node is given by the mean response at the node. In the EMHE Tree, since an EMHE regression is fitted at each node, the prediction for each IBD state is obtained from the fitted regression and predictions are assigned to each terminal node. Then for a sib-pair and the set of posterior IBD probability weights contributed to each terminal node, the predicted response is given by a weighted average of the predictions across all the terminal nodes. Figure A3 in Appendix A in the Supplementary Material provides an example of how the probabilities contribute to each child node and the overall predicted value is calculated for a sib-pair.

Aside from the tree growing algorithm and tree predictions, EMRF is similar to OrigRF. Each tree is grown on a bootstrap sample and at each node the selection of a cutpoint to split the node is made within a random subset of markers of size m_{try}. The OOB data not used to grow the tree are then used to estimate the PE and measure the importance of markers.

2.2 New VI indices with partial permutation and smoothing

The OrigRF measures the importance of a marker by the increase in PE that occurs when that marker is permuted in the OOB sample while the data for the remaining markers are kept intact. More specifically, for tree b and n_{OOB} OOB samples with observed and predicted responses y_i and $\hat{y}_{ib}$ respectively, the estimated prediction error for tree b is defined,

$$\widehat{PE}(b) = \frac{1}{n_{\text{OOB}}} \sum_{\substack{i=1 \\ i \in \text{OOB}}}^{n} \left(y_i - \hat{y}_{ib}\right)^2$$

Denote the PE after permutation of OOB samples for marker M by $\widehat{PE}_M(b)$. Then the VI index for marker M for tree b is

$$VI\ (M,\ b)\ = \widehat{PE}(b) - \widehat{PE}_M(b).$$

The importance measure for marker M, *VI(M)*, is given by the average of *VI(M, b)* over all B trees.

This VI index defined in OrigRF has some drawbacks in the context of linkage analysis. The first concerns the permutation of a single marker, which is intended to break its relationship with the response so that PE will increase if the marker is important. However, due to the correlation between neighbouring markers in linkage studies, permuting data for a single marker while linked markers are still present in the tree may not affect the PE greatly even if the permuted marker is linked with the trait and hence important. Thus, the inherent correlation between IBD sharing at markers that are close together on a chromosome thwarts the intended effect of the single marker permutation for the importance index calculation. We define partial permutation as a permutation that is limited to a random subset of samples, where the subset size is determined by the degree of permutation discussed below. We propose *multi-marker partial permutation* in which all the linked markers that have been chosen as split variables in the tree are jointly and partially permuted according to the distance between markers, which influences the correlation between them. We define the degree of permutation by $\gamma = (1-2\theta)^2$, which is the correlation between sib-pair IBD sharing where θ is the recombination fraction between two marker loci. In this partial permutation procedure, only γn_{OOB} samples for a marker are permuted. The shorter the distance between the two markers, the larger the proportion of data that is permuted. Figure A4 in Appendix A in the Supplementary Material illustrates the relationship between the correlation γ and recombination fraction θ between two markers, M and M'.

The second drawback of the original definition of VI concerns potential bias in the measure. The fact that a marker M' is not chosen as a split variable in a tree does not necessarily imply that it is not important. It is possible that marker M' is not in the set of m_{try} markers to be evaluated as a split variable, or if it is in this set, it is possible that it is linked to another marker M, that may be more important and therefore chosen as a split variable instead. In the original VI definition, the importance measure of M' would be set to zero for tree T, i.e. $VI(M', T) = 0$. Then averaging these measures over all trees would produce downward bias in VI for a possibly important marker. Intuitively, if markers linked to marker M' are chosen as split variables, then $VI(M', T)$ should reflect a value similar to its linked markers.

We propose to *smooth* or interpolate the VI measure for marker M' by a weighted fraction of the values from its linked markers. For marker M' and its neighbouring linked markers A and B, which are split variables in tree T, define

$$VI(M', T) = \frac{\gamma_{AM'}^2 \, VI(A, T) + \gamma_{BM'}^2 \, VI(B, T)}{\gamma_{AM'} + \gamma_{BM'}} \quad (1)$$

where $\gamma_{MM'}$ is the sib-pair IBD correlation between markers M and M'.

We define *VIPP* as a new definition of VI that adopts the multi-marker partial permutation and smoothing procedures using the OOB sib-pair data. For marker M and tree T, *VIPP(M, T)* is obtained using the multi-marker permuted data if marker M is a split variable or the smoothed importance measure if marker M is not a split variable. Then *VIPP(M)* is the average of *VIPP(M,T)* across all trees.

2.3 New VI measures based on partial dependence

We define another VI index, denoted by *VIPD* that is based on Friedman's (2001) definition of *partial dependence*. Partial dependence provides a marginal summary of a marker by averaging over markers not of interest from the model. It is the prediction of the response under the different values of the marker of interest, averaged over the changes in values of the other markers. Partial dependence is calculated by a weighted traversal of the tree. In the sib-pair EMHE Tree, if the split variable is the marker of interest, then the weights contributed to the children nodes are updated by the probability the sib-pair belongs to the corresponding nodes. However, if the split variable is not of interest, the weights are updated by the sum of the weights of all sib-pairs contributing to the respective children nodes. The final summary of partial dependence of the model on the marker of interest is given by a weighted average of the predictions over the visited terminal nodes.

The partial dependence for marker M and tree T for each OOB sib-pair is computed and treated as a prediction from tree T in what follows. We define *VIPD (M, T)* using the LOD score, which is a likelihood ratio statistic in $\log_{10}$ scale, for testing linkage. Here, the likelihoods under the alternative and null hypotheses are calculated using the OOB data and permuted OOB data respectively for marker M. The smoothing strategy discussed earlier in the situation where marker M is not a split variable in tree T is also applied in this definition of importance index. Averaging *VIPD (M, T)* across all trees provides the importance measure *VIPD (M)* for marker M based on the partial dependence VI definition. Multi-marker partial permutation is not effective here because the sib-pair IBD probabilities only for the marker of interest M contribute to a change in the partial dependence prediction. For markers not of interest, including the linked markers, the sum of the weights of all sib-pairs contributing to the respective children nodes is used to update the weights during the tree traversal; therefore permuting the IBD data for the linked markers has no effect on the measure.

3 APPLICATION OF EMRF TO GAW13 FRAMINGHAM LINKAGE DATA

We applied HE and EMHE regressions, OrigRF and EMRF to localize putative genes influencing systolic blood pressure (SBP) in sibships from the Framingham Heart Study (FHS) (Dawber *et al.*, 1951). Levy *et al.* (2000) previously reported significant genome-wide linkage to SBP in the FHS pedigrees on chromosome 17 (with a LOD score of 4.7) using 398 micro-satellite markers. Using a related longitudinal SBP phenotype, Briollais *et al.* (2003) confirmed the chromosome 17 finding (LOD 3.5) and also found linkage to a marker on chromosome 8 (LOD 3.6). When Wu *et al.* (2006) repeated this analysis in nuclear families extracted from the FHS pedigrees, the chromosome 8 marker LOD score increased to 4.6, but the LOD for the chromosome 17 marker dropped below 3.0. These LOD scores were all obtained with variance component methods.

Following Briollais *et al.* (2003), we adopted a two-step approach using a multi-level model in the first stage and linkage methods in the second stage. Residuals from a multi-level model of the relationship between SBP and age, accounting for the correlation between multiple observations per individual and adjusted for covariates including sex, body mass index, treatment and cohort, were used as the phenotype in subsequent linkage analysis. Based on our analysis of the same 398 markers in 1263 sib-pairs, multiple chromosomal regions showed evidence of linkage according to the Haseman–Elston regressions, but did not meet genome-wide significance (i.e. LOD score <3.0). The LOD scores from HE and EMHE regressions were similar. The EMHE regression LOD score and the OrigRF and EMRF VI indices are displayed in Figure 1. A region on chromosome 9 was common to EMHE and OrigRF, whereas a region on chromosome 6 was common to the EMRF VIs. The ranking of the top 30 markers based on these indices are shown in Figures C3 through C7 in Appendix C in the Supplementary Material. The ATC6A06 marker reported by Levy *et al.* (2000) and a second nearby marker appear on the HE or EMHE lists, but not on the EMRF lists. The chromosome 8 marker GATA72C10 reported by Briollais *et al.* (2003) and Wu *et al.* (2006), however, appears on the VIPD list.

An approximate null distribution for the genome-wide maximal VI statistic and P-values were obtained for all markers, using 1000 permutation samples as suggested by Churchill and Doerge (1994) for genome-wide 5% significance level. The peak on chromosome 6 was not statistically significant based on the permutation test (Figs C1 and C2 in Appendix C in the Supplementary Material). The null distribution of the genome-wide maximal VI statistic for the

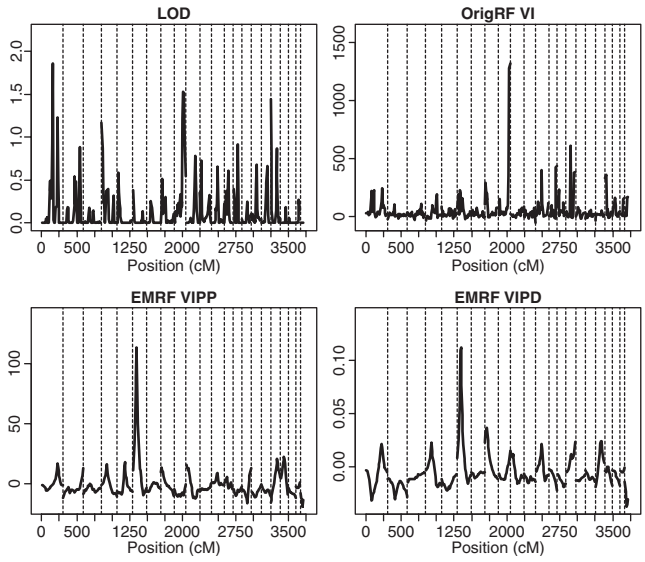

Fig. 1. GAW13 FHS: EMHE regression LOD score, OrigRF VI index and EMRF VI indices from genome-wide linkage analysis (dotted vertical lines separate the chromosomes).

permutation test is evaluated in Figure C8 for OrigRF and Figures C9 and C10 for EMRF. Because case studies of a dataset in which true linkages are not known with certainty are limited, we further examined the behaviour of the VI indices via simulation studies, as described in the next section.

4 SIMULATION STUDIES

4.1 Design

We examined various genetic models in two sets of simulation studies. The response was simulated at the *genotype level* in Simulation Study I and at the *IBD level* in Simulation Study II. In all simulations, the new definitions of VI were applied to EMRF and the original definition was applied to OrigRF.

In Simulation Study I, the trait value for each *individual* was generated according to a joint phenotypic distribution for two quantitative trait loci (QTLs) under three genetic models, and the squared difference in sib-pair trait values was calculated as the dependent variable for subsequent linkage analysis. Overall heritability (h^2) for the genetic models was set at either 0.8 or 0.4. The three models examined when $h^2 = 0.8$ were: Model 1 in which two QTLs contribute equal additive variance components ($\sigma^2_{a1} = \sigma^2_{a2} = 0.4$); Model 2 which involved only additive–additive interaction variance component ($\sigma^2_{aa} = 0.8$); and Model 3 which involved all four additive and dominance interaction variance components with unequal contributions from the two QTLs ($\sigma^2_{aa} = 0.1, \sigma^2_{ad} = 0.55, \sigma^2_{da} = 0.1, \sigma^2_{dd} = 0.05$). The variance components for models with $h^2 = 0.4$ are half of these variance components. We show results only for the models with $h^2 = 0.8$ in this report; results for models with $h^2 = 0.4$ are available in the Supplementary Material. The phenotypic values for the nine genotypes underlying the genetic models with $h^2 = 0.8$ are shown in Table 1. The joint distribution plots of the expected means and marginal means for these models are shown in Figure D1 in Appendix D in the Supplementary Material.

Table 1. Simulation Study I: expected phenotypic values for Models 1–3

	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
Model 1	12.24	11.12	10.00	11.12	10.00	8.88	10.00	8.88	7.76
Model 2	16.15	11.68	7.20	11.68	10.00	8.32	7.20	8.32	9.44
Model 3	5.16	10.42	10.98	8.82	12.72	10.31	12.48	10.12	11.38

In this set of simulations, we compared the performance of the EMRF VI and OrigRF VI indices. With a sample size of 1000, we examined the effect of tree size on EMRF by growing either 15 node trees or full sized trees, and the effect of marker subset size m_{try}, which was set either at the recommended level (a third of the total number of markers, rounded down to the nearest integer) or bagging level (all markers) on EMRF and OrigRF. We also examined the effect of sample size on OrigRF and EMRF in comparison with EMHE regression LOD score.

In Simulation Study II, the dependent variable (squared difference in trait value) for the linkage analysis was directly generated at the IBD level for each *sib-pair* according to a joint IBD distribution between two fully informative QTLs with expected means shown in Table 2. Then to introduce random noise, the square of a normally distributed random variable with mean 0 and variance 2 or 3 was added to the response. We show the results only for the model with noise variance of 2 in this report. The joint distribution plot of the expected means and marginal means are shown in Figure D2 in Appendix D in the Supplementary Material.

For all simulation studies, we sampled 100 replicates under each genetic model with two QTLs and 500 replicates with no QTL. The two QTLs were unlinked and placed in the middle of two chromosomes with respective lengths of 100 cM and 150 cM. The QTL minor allele frequency was 0.2. All markers were biallelic with allele frequency of 0.5 and were simulated under incomplete inheritance. Markers were distributed evenly across the chromosomes at every 5 cM resulting in 52 markers. Three sizes (250, 500 and 1000) of nuclear families were simulated, resulting in three sample sizes of independent sib-pairs in the linkage analysis for examining the effect of sample size.

Multipoint estimates of marker IBD sharing probabilities were estimated in GENEHUNTER version 2.1r_5 (Kruglyak *et al.*, 1996). GENEHUNTER was also used to perform EMHE regression in Simulation Study I. In Simulation Study II, we programmed an EM algorithm in C to perform the regression. All random forests consisted of 1000 trees, which were empirically determined to be sufficient for the PE to converge. The randomForest package version 4.5-16 (Liaw and Wiener, 2002) in R (R Development Core Team, 2008) was implemented to perform the OrigRF using the default settings. We developed a C program to perform the EMRF, setting the minimum number of observations per node to 10 and tolerance level for convergence in the EM algorithm to 10^{-5}.

4.2 Measure of performance in simulation study

Receiver operating characteristic (ROC) curves and the area under the curve (AUC) were used as the measures of performance to compare the ability of the LOD score and random forests VI indices in detecting linkage to QTLs. A ROC curve displays the relationship between the false positive rate (*fpr*) and true positive rate (*tpr*) of detection for a linkage method. Given m_0 replicates simulated under the null with no QTL and m_1 replicates simulated with QTLs under

Table 2. Simulation Study II: expected squared difference in sib-pair trait values

Proportion of alleles shared IBD		QTL_2		
		0	1/2	1
QTL_1	0	8	1	10
	1/2	4	6.5	3
	1	4	6	4

an alternative model, we calculate the number of replicates that are declared significant for a linkage method among the null and alternative replicates, denoted by V and S, respectively, for each threshold value. The ROC curve displays the set of *fpr* and *tpr* pairs given by the ratios of V to m_0 and S to m_1, respectively, at each possible threshold value. The AUC provides a summary measure to compare the accuracy between different linkage methods.

We evaluated test statistics (i.e. VI index or LOD score) from the two susceptibility markers at the two QTLs in each of the m_1 replicates simulated with QTLs. For each test statistic, we assessed three facets of discrimination ability, (1) ability to detect linkage to each QTL separately; (2) ability to detect linkage to at least one QTL and (3) ability to detect linkage to all QTLs. To detect linkage to each QTL separately, the test statistics at the two susceptibility markers were evaluated separately. The maximum of the two test statistics for each replicate was evaluated to detect linkage to at least one QTL and the minimum was evaluated to detect linkage to both QTLs. For each of the discrimination abilities, one value of the test statistic was obtained from each of the m_1 replicates and combined with the genome-wide maximum test statistic from each of the m_0 null model replicates. Then for every possible threshold values, the associated pair of *fpr* and *tpr* was calculated from the combined set. The set of *fpr* and *tpr* pairs and AUCs for the three discrimination abilities were obtained for each linkage method using the ROCR package version 1.0-1 (Sing *et al.*, 2005) in R.

4.3 Simulation results

We denote by *Reg* the forest with 15 nodes and $m_{try} = 17$, *Full* the forest with full sized trees and $m_{try} = 17$, *Bag* the forest with 15 nodes and $m_{try} = 52$, and *BagFull* the forest with full sized trees and $m_{try} = 52$. Full sized trees were grown for OrigRF as defined by Breiman (2001), thus only two settings, *Full* and *BagFull*, were compared. The AUCs to compare the abilities of all VI indices to detect at least one QTL in Simulation Study I, for the three models with $h^2 = 0.8$ and sample size of 1000 are shown in Figure 2. The plots displaying their abilities to detect each QTL separately and both QTLs simultaneously are shown in Figures E1–4 in Appendix E in the Supplementary Material.

Overall, the OrigRF and EMRF VI indices performed well at the recommended level of m_{try} compared to the bagging level (Fig. 2, *Reg* versus *Bag* and *Full* versus *BagFull*). Growing small sized EMRF trees with 15 nodes at the recommended m_{try} level (*Reg*) yielded improved AUCs for both VI indices. The OrigRF VI performed reasonably well in the presence of interacting genes (Models 2 and 3) but performed poorly in Model 1 with only main effects.

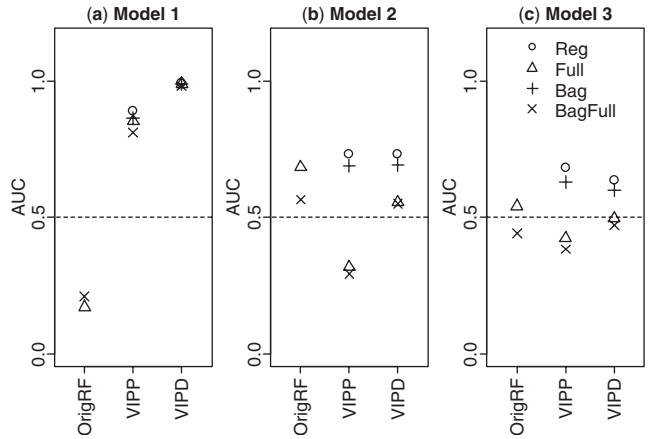

Fig. 2. Simulation Study I: AUCs for VI indices from OrigRF and EMRF in detecting at least one QTL for Models 1–3 with $h^2 = 0.8$ and sample size of 1000. *OrigRF*: Original Random Forest; *VIPP*: EMRF with VI based on Multi-marker Partial Permutation and smoothing using neighbouring markers; *VIPD*: EMRF with VI based on Partial Dependence and smoothing using neighbouring markers; *Reg*: 15 nodes, $m_{try} = 17$; *Full*: full sized trees, $m_{try} = 17$; *Bag*: 15 nodes, $m_{try} = 52$ and *BagFull*: full sized trees, $m_{try} = 52$.

We also evaluated EMRF VI indices defined using combinations of permutation and smoothing procedures. The permutation procedure included: (1) single marker permutation or (2) multi-marker partial permutation where markers are permuted either (i) jointly or (ii) independently. The smoothing procedure included (1) no smoothing, (2) smoothing using *neighbouring* markers that are chosen as split variables or (3) smoothing using *all* markers that are chosen as split variables. Partial dependence definitions of VI indices based on one of the three smoothing procedures were also evaluated. In total, 12 VI indices were compared and they are listed in Tables B1 and B2 in Appendix B in the Supplementary Materials.

Our simulations found similar performances for joint and independent multi-marker partial permutation and for smoothing using either neighbouring or all linked markers. The best performance was obtained with both multi-marker permutation and smoothing. Thus, we focused on *VIPP* and *VIPD* in this report.

We also examined the performance of EMRF VI indices under the *Reg* setting and OrigRF VI under the *Full* setting across sample size, in comparison with the EMHE regression LOD score. The AUCs from these statistics are displayed in Figure 3 and Figure E5 in the Appendix E in the Supplementary Material. As expected, the AUCs increased with sample size for all indices. The LOD score yielded larger AUCs at all sample sizes, even in Models 2 and 3 where the underlying genetic model was governed by epistatic QTLs. Among the EMRF VI indices, *VIPP* performed similarly or better in Models 2 and 3, while *VIPD* performed better in Model 1. The OrigRF VI displayed the poorest performance.

The EMHE regression LOD score performed well when genotype data (individual level) are mapped to IBD level (sib-pair level), because the interaction effect at the genotype level translates into marginal effect at the IBD level (Fig. D1 in Appendix D in the Supplementary Material). More specifically, the expected marginal squared difference in trait values at both QTLs show negative association with the proportion of alleles shared IBD, making it ideal for the single-locus EMHE regression to detect linkage to the QTLs.

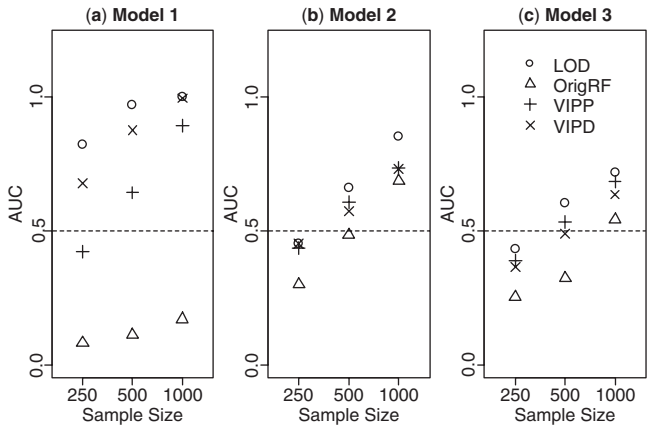

Fig. 3. Simulation Study I: AUCs across sample size for VI indices from OrigRF and EMRF, and LOD score in detecting at least one QTL for Models 1–3 with $h^2 = 0.8$.

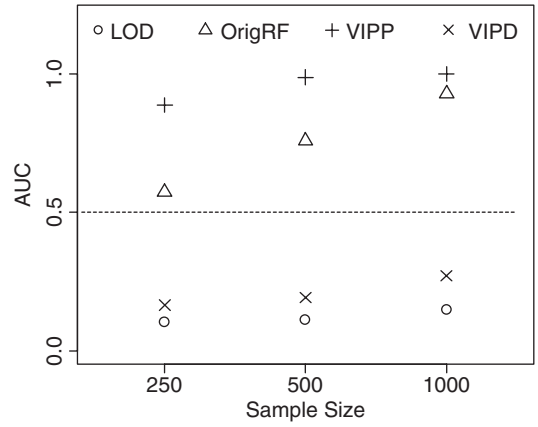

Fig. 4. Simulation Study II: Comparison of AUCs for model simulated at IBD level.

In practice, the RF would be applied to the whole genome in which some chromosomes contain QTLs and others do not. Thus, additional simulations were performed to examine the performance of the VI indices from OrigRF and EMRF under a scenario in which two chromosomes have a QTL each and a third chromosome has no QTL, with similar settings as Models 1 through 3 in Simulation Study I. In this study, the third chromosome was generated with a length of 100 cM. Three chromosomes were also generated for replicates with no QTL. The marker subset size m_{try} was set to 24, which is the recommended level using one-third of the total number of available markers. AUCs from these models were compared with those in the model with two chromosomes from Simulation Study I. The performance of the LOD score, OrigRF and EMRF VI indices deteriorated as expected with the inclusion of the third 'null' chromosome (Fig. E7 in Appendix E in the Supplementary Material). In general, relatively good performance of the EMRF VI indices compared to the OrigRF VI and LOD score were obtained.

In Simulation Study II, no main effects were simulated at the IBD level, thus by design the EMHE regression LOD score performed very poorly (Fig. 4 and Fig. E6 in Appendix E in the Supplementary Material). *VIPD* which performed reasonably well compared to the OrigRF VI in Simulation Study I now showed very poor performance. Since *VIPD* is based on a summary measure capturing the marginal effects of each marker in the model, it performed better in the presence of strong marginal effects at the IBD level, similar to single-locus EMHE regression LOD score. However, since it is based on tree-based models and accommodates interaction present in the model, it performed slightly better than the LOD score in this simulation study, but poorer than the LOD score in Simulation Study I where interaction effects were weaker than marginal effects at the IBD level.

The OrigRF VI performed well in this simulation study since tree-based models are designed for multiplicative models. However, it did not perform as well as *VIPP*, which was usually the best.

5 DISCUSSION AND CONCLUSION

We examined the performance of VI indices from OrigRF and EMRF and compared them to the EMHE regression LOD score under various genetic models, parameter settings and sample sizes using Monte Carlo simulation studies. Our simulations found better performance for all VI indices when m_{try} was set at the recommended level, in agreement with Liaw and Wiener (2002). Our simulations also showed that reducing tree size can have a large positive impact on the performance of VI indices for EMRF. Thus, in agreement with Segal *et al.* (2004) who investigated the effect of tree size on OrigRF, we also recommend reducing the tree size in EMRF. Moreover, growing small trees in RF can save a substantial amount of computational time.

The proposed indices, *VIPP,* which is based on the original definition of VI with multi-marker partial permutation and smoothing, and *VIPD*, which is based on Friedman's definition of partial dependence with smoothing, in combination with growing small trees with m_{try} set at the recommended level, exhibited better performance than the OrigRF VI. This suggests that application of these new VI definitions to OrigRF may improve its discrimination abilities in detecting linkage to QTLs for a random sample of sib-pairs. This idea requires further investigation. For selected samples however, EMRF may perform better than OrigRF, based on results obtained by Dolan *et al.* (1999a) for the variance–components model. They obtained less biased estimates and the expected null distribution of log-likelihood ratio statistics using the IBD distribution in the likelihood as compared to using $\widehat{\pi}$ in selected sib-pairs analysis [while assigning expected values (1/4, 1/2, 1/4) to the missing IBD probabilities for the unselected sib-pairs].

In Simulation Study II where pure interaction was simulated at the IBD level, the RFs consistently showed better performance, as expected. More specifically, the EMRF VI index involving smoothing and multi-marker permutation (*VIPP*) showed the best performance, followed by the OrigRF VI. The *VIPD* that showed good performance in Simulation Study I performed poorly in these examples, almost as poor as the LOD score. These results suggest that *VIPD* may be more suitable when there are strong marginal effects at the IBD level.

Generally for EMRF, VI indices showed similar performances with smoothing using neighbouring markers or all markers. VI indices where multiple markers were permuted jointly or independently also performed similarly. Improved performance was observed with the inclusion of multi-marker partial permutation or

smoothing for all indices. The greatest improvement was obtained when both procedures were applied to the original definition of VI index and when smoothing was applied to the partial dependence VI definition.

To assess whether gains associated with EMRF in comparison to OrigRF were due to differences in the splitting criteria or to modifications to the VI measures, we conducted additional simulations under Model 1 with $h^2 = 0.8$, comparing the proposed HE linkage-based splitting rule to the conventional MSE splitting rule with constant mean squared difference in trait values within node (denoted by OrigRF-MSE). In both tree methods, membership to the children nodes was determined using the EM algorithm. We defined a linkage region to be within 5 cM from either side of the true QTL. In one set of simulations, we compared the marker selected at the root node in the two RFs using 100 replicates with one tree per forest where only the root node was split; all the observations were used at the root node (i.e. no bootstrapping), and all 52 markers were assessed. In another simulation under Model 1 with parameter settings as described in Simulation Study I, we compared the mean VI indices across replicates and the power of the two RFs for each VI index in detecting the linkage region. Power is defined as the number of times out of the 100 replicates that a marker with maximum VI is within the linkage region. In the simulation studies, the HE linkage-based splitting rule in EMRF performed better with higher power than the MSE splitting rule (Fig. E8–10 in Appendix E in the Supplementary Material) for all VI indices, suggesting that both the modifications are useful.

In the simulation studies, a 5 cM density marker map was used and IBD sharing was measured at every marker. If higher density maps are available, we suggest that IBD sharing should still be measured at every marker. However, higher density will increase the IBD information, thus the advantage of the EMRF with respect to incomplete data may diminish. On the other hand, higher density will increase the correlation between markers, thus improving the advantage of the partial permutation importance index.

Higher density maps would increase the computation time to execute the EMRF since a larger number of markers would have to be evaluated at each node (m_{try}), following the recommendation to set it to one third the total number of markers. If m_{try} is fixed, then the computation time would be similar between two datasets with the same number of sib-pairs and number of trees in the RF, but different number of markers. Using a 2.8 GHz 32-bit Intel Xeon Processor to execute a EMRF with 1000 trees, a simulated dataset with 52 markers and $m_{try} = 17$ consumed approximately 26, 24 and 17 min for sample size 1000, 500 and 250, respectively. The computation time increased by roughly 50% for a dataset simulated with 21 more markers (i.e. 3 chromosomes model) and $m_{try} = 24$ (approximately 15, 10 and 8 min, respectively for sample size of 1000, 500 and 250). The EMRF analysis took 38 h to analyse one set of FHS data with 1263 sib-pairs and 398 markers. Assuming computation time to be quadratic with the number of markers, a dataset with 1000 markers, 1000 sib-pairs and 1000 trees is predicted to require about 278 h to analyse with EMRF.

One of the assumptions of the EMRF is that markers selected in the path from the root node to a leaf are unlinked. However, linked markers can be found in the same path. Further investigation is required to evaluate how frequently linked markers appear in the same path and the impact of the violation of this assumption.

Our EMRF was developed for sib-pair data rather than extended pedigrees. Use of extended pedigrees has been shown to provide greater information, and thus, greater power in linkage analysis (Dolan *et al.*, 1999b; Schork, 1993; Williams and Blangero, 1999; Williams *et al.*, 1997). This limitation could be addressed in a modified RF by employing regression methods that model the variance–covariance structure among families. One such method is the two-level Haseman–Elston regression that has been shown to be asymptotically equivalent to the variance component model for linkage studies (Wang and Elston, 2005). Another method uses a generalized estimating equation model in which the Haseman–Elston regression and variance component methods are special cases utilizing different working covariance matrices (Chen *et al.*, 2004). Likelihood from the former model or quasi-likelihood from the latter model could then be incorporated in the splitting rule at each node for each modified regression tree that makes up the RF.

If our proposed definitions of VI indices were incorporated in the OrigRF, we could apply the OrigRF for association studies using marker genotypes as opposed to marker IBD data for linkage. Then instead of employing correlation between IBD sharing among sib-pairs as the degree of permutation on linked marker data or as weights in the smoothing procedure, measures of linkage disequilibrium between markers that range in values between 0 and 1 could be utilized for association.

A modified RF could be developed for association studies using genotype data to model variation in a quantitative trait. Similar to an EMHE Tree, posterior genotype or haplotype probabilities, instead of posterior IBD probabilities, could be assigned as weights to each individual. Appropriate splitting criterion could be defined; the simplest being the MSE used in conventional regression trees. An ensemble of these probabilistic trees would then make up the modified RF.

The RF methods described in this report are exploratory tools. They can be used as an initial step in a genome scan linkage analysis, followed by application of other linkage and/or association method on regions identified by the RFs. In this case, a less stringent genome-wide significance level can be used in the RF to select relevant markers for subsequent research. Another strategy is to perform standard linkage methods first with a specific genome-wide significance level to adjust for multiple testing, and then perform exploratory analysis using the RF.

One of the strengths of RF methods is their ability to assess all available markers simultaneously. This feature allows the model to explore complex interactions without the need to individually model each interaction separately as is done in regression models. Another advantage of the RF methods is that independence between trees allows individual trees to be grown in parallel and then combined, so that parallel processing can be used to ameliorate the disadvantage of computational intensity.

In summary, we develop EMRF as a tool for mapping complex traits that allows simultaneous assessment of all available markers. We propose new definitions of VI indices as a measure of relative importance and find them to be an improvement over the original definition of VI. We compare and evaluate the methods in simulated data under various genetic models and conditions, and apply the RFs to the GAW13 FHS data. Like other methods, EMRF has some limitations, but we find the method to be feasible and useful for exploratory multi-locus quantitative trait linkage analysis.

ACKNOWLEDGEMENTS

Conditions for use of the Framingham Heart Study data were approved by the Research Ethics Board of Mount Sinai Hospital, Toronto Canada. We gratefully acknowledge the contribution of Boston University staff (Framingham Heart Study) and the Investigators within the Framingham Heart Study who have collected and analysed these data. We gratefully acknowledge the two anonymous reviewers for their detailed and insightful comments and suggestions which greatly improved the quality of the article.

Funding: S.S.F.L. was supported by a Natural Sciences and Engineering Research Council Doctorate Research Grant. Research support was also provided by the Canadian Institutes of Health Research (Grants NPG-64871, MOP-84287 to L.S. and S.B.B. and Senior Investigator Award MSS-55118 to S.B.B.) and the Network of Centres of Excellence in the Mathematics of Information Technology and Complex Systems. The Framingham Heart Study is conducted and supported by the National Heart, Lung and Blood Institute (NHLBI, Contract No. N01-HC-25195) in collaboration with Boston University School of Medicine. This research was not conducted in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University or NHLBI.

Conflict of Interest: none declared.

REFERENCES

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

Breiman,L. (1996b) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Briollais,L. *et al.* (2003) Multilevel modeling for the analysis of longitudinal blood pressure data in the Framingham heart study pedigrees. *BMC Genet.*, **4**, S19.

Bureau,A. *et al.* (2003) Mapping complex traits using random forests. *BMC Genet.*, **4**, S64.

Bureau,A. *et al.* (2005) Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.,* **28**, 171–182.

Chen,W.M. *et al.* (2004) Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. *Genet. Epid.*, **26**, 265–272.

Churchill,G.A. and Doerge,R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics,* **138**, 963–971.

Dawber,T.R. *et al.* (1951) Epidemiological approaches to heart disease: the Framingham study. *Am. J. Public Health*, **41**, 279.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.

Dolan,C.V. *et al.* (1999a) A simulation study of the effects of assignment of prior identity-by-descent probabilities to unselected sib pairs, in covariance-structure modeling of a quantitative-trait locus. *Am. J. Hum. Genet.*, **64**, 268–280.

Dolan,C.V. *et al.* (1999b) A note on the power provided by sibships of sizes 2, 3, and 4 in genetic covariance modeling of a codominant QTL. *Behav. Genet.*, **29**, 163–170.

Elston,R.C. and Stewart,J. (1971) A general model for the genetic analysis of pedigree data. *Hum. Hered.,* **21**, 523–542.

Falconer,D.S. (1989) *Introduction to Quantitative Genetics.* 3rd edn. Longmans Green/John Wiley & Sons, Harlow, Essex, UK/New York.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.

Gibson,G. (1996) Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor. Popul. Biol.*, **49**, 58–89.

Haseman,J.K. and Elston,R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 3–19.

Izmirlian,G. (2004) Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Ann. N. Y. Acad. Sci.*, **1020**, 154–174.

Kruglyak,L. and Lander,E.S. (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.*, **57**, 439–454.

Kruglyak,L. *et al.* (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.

Lander,E.S. and Green,P. (1987) Construction of multilocus genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA,* **84**, 2363–2367.

Levy,D. *et al.* (2000) Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. *Hypertension*, **36**, 477–483.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.

Lunetta,K.L. *et al.* (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.*, **10**, 32.

Moore,J.H. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.*, **56**, 73–82.

Ott,J. (1999) *Analysis of Human Genetic Linkage*. 3rd edn. Johns Hopkins University Press, Baltimore, MD.

R Development Core Team (2008) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.

Schork,N.J. (1993) Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Am. J. Hum. Genet.*, **53**, 1306–1319.

Segal,M.R. *et al*. (2004) Relating HIV-1 sequence variation to replication capacity via trees and forests. *Stat. Appl. Genet. Mol. Biol.*, **3**, 2.

Shi,T. *et al.* (2005) Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Mod. Pathol.*, **18**, 547–557.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics,* **21**, 3940–3941.

Wang,T. and Elston,R.C. (2005) Two-level Haseman-Elston regression for general pedigree data analysis. *Genet. Epidemiol.*, **29**, 12–22.

Williams,J.T. and Blangero,J. (1999) Power of variance component linkage analysis to detect quantitative trait loci. *Ann. Hum. Genet.*, **63**, 545–563.

Williams,J.T. *et al.* (1997) Statistical properties of a variance components method for quantitative trait linkage analysis in nuclear families and extended pedigrees. *Genet. Epidemiol.*, **14**, 1065–1070.

Wu,L.Y. *et al.* (2006) Locus-specific heritability estimation via the bootstrap in linkage scans for quantitative trait loci. *Hum. Hered.*, **62**, 84–96.