

# Very Large Scale ReliefF for Genome-Wide Association Analysis

Margaret J. Eppstein and Paul Haake

**Abstract**— The genetic causes of many monogenic diseases have already been discovered. However, most common diseases are actually the result of complex nonlinear interactions between multiple genetic and environmental components. There is thus a pressing need for new computational methods capable of detecting nonlinearly interacting single nucleotide polymorphism (SNPs) that are associated with disease, from amidst up to hundreds of thousands of candidate SNPs. Recently, some progress has been made using feature selection algorithms based on weights from the ReliefF data mining algorithm on sets of up to 1500 SNPs. However, the accuracy of ReliefF does not scale up to the sizes needed for truly large genome-scale SNP association studies. We propose a population-based variant dubbed VLSReliefF, which mitigates this performance drop by stochastically applying ReliefF to SNP subsets, and then assigning each SNP the maximum ReliefF weight it achieved in any subset. A heuristic method is proposed for determining the optimal subset size as a function of heritability, sample size, and order of interactions. The method is validated using a variety of computational experiments on synthetic datasets of up to 100,000 SNPs.

## I. INTRODUCTION

With the advent of the genomic era, characterized by the human genome project [8][28] and the development of methods for rapid and affordable genotyping [26], attention has now turned to large-scale research efforts in identifying human genetic variability, and millions of single nucleotide polymorphisms (SNPs) have now been identified [7][9][11]. Such genomic information carries with it the potential for improved understanding regarding the genetic etiologies of disease, and may revolutionize our abilities to detect, treat, and even prevent disease [5][14][22]. The genetic causes of many monogenic diseases have already been discovered, including those for cystic fibrosis, Alzheimer’s disease, Hirschsprung disease, and phenylketonuria [4][14][22]. However, most common diseases are actually the result of complex nonlinear interactions between multiple genetic and environmental components [15]. Complex diseases include heart disease, obesity, cancer, diabetes, and schizophrenia [4][14][29]. Since complex genetic etiologies cannot be unraveled by simple linkage studies [5], there is a pressing need for new

computational methods capable of detecting nonlinearly interacting SNPs that are associated with disease [5][29].

Detecting which handfuls of SNPs exhibit nonlinear epistatic interactions that predispose for disease is a computationally daunting combinatorial optimization task, because individual SNPs may have little or no detectable ill effects [18]. Solution of this important problem is further exacerbated by low disease heritability, small sample sizes, and a lack of information regarding how many, if any, SNPs interact. New methods, such as multifactor dimensionality reduction [16][17][24] and ROC-based metrics [3] show promise for detecting epistatic interactions, but are only applicable to very small sets of SNPs. Various statistical, data mining, and machine learning strategies have shown some promise for sets of up to a few hundred SNPs [6][12][21], but are not scalable to really large-scale genome-wide association studies (which may contain up to hundreds of thousands of SNPs). When interactions are purely epistatic, standard evolutionary algorithms are no better than random search, since there are no high fitness subsets available to act as “building blocks” for selection to act upon [19][30].

Recently, some progress has been made using feature selection algorithms based on ReliefF weights. The ReliefF data mining algorithm [25] estimates importance weights of individual attributes by comparing similarities between samples across entire attribute sets simultaneously, in case/control studies. Thus, this method can theoretically detect purely epistatic interactions between SNPs, although the accuracy of the weights decreases with increasing number of SNPs and decreasing heritability. Moore and White [20] propose an iterative application of ReliefF, referred to as TuRF (“Tuning ReliefF”), in which they apply ReliefF in order to estimate the importance of each SNP, discard some predefined constant number of SNPs with the lowest weights, and repeat. Consequently, the accuracy of the estimated weights increases with each iteration as they reduce the number of candidate SNPs linearly. McKinney *et al.* [13] combined ReliefF weights with entropy measures in an iterative feature selection algorithm dubbed “Evaporative Cooling”. Moore and White [19] also experimented with using TuRF weights to bias the retention of highly weighted SNPs when evolving binary combinations of SNPs, and found that their results were better than random search. Eppstein *et al.* [3] used a fitness measure based on ReliefF weights in a probabilistic evolutionary approach that reduces the candidate SNP set size logarithmically. However, the power of all of these ReliefF-based approaches are ultimately governed by the accuracy of the weights at the first iteration. Thus, while the ReliefF-based methods

Manuscript received March 28, 2008. The computational resources provided by the Vermont Advanced Computing Center supported in part by NASA (NNX 06AC88G) are gratefully acknowledged.

M. J. Eppstein is with the Computer Science Department and the Complex Systems Center, University of Vermont, Burlington, VT 05405 USA (phone: 802-656-3330; fax:802-656-0696; e-mail: Maggie.Eppstein@uvm.edu).

P. Haake is with the Computer Science Department, University of Vermont, Burlington, VT 05405 USA (e-mail: Paul.Haake@uvm.edu).

described above have been applied to datasets with up to 1500 SNPs, they are not scalable to truly genome-wide association studies which may contain hundreds of thousands of SNPs, since ReliefF computed weights become increasingly random as the number of attributes increases (e.g., see [25], and the results shown in this paper).

In this work, we propose a population based general modification to ReliefF that maintains a high accuracy of weights for very large sets of attributes, such as in genome-wide SNP association studies. Consequently, we dub our modified algorithm VLSReliefF (“Very Large Scale ReliefF”). We compare the performance of ReliefF and VLSReliefF on synthetically generated datasets with purely epistatic SNP interactions, for varying numbers of candidates SNPs, heritabilities, and sample sizes.

## II. ALGORITHMS

### A. ReliefF

Pseudo-code for the ReliefF algorithm [25], applied to  $M$  samples, each with  $N$  attributes (missing data can be treated probabilistically, as described in [25]) and a binary class label, is as follows:

**“ReliefF” Algorithm:**  
 Weights  $W_j=0, \forall j \in \{1..N\}$  attributes  
 FOR  $i = 1$  to  $k$  ( $k = \#$  samples)  
   Select an individual sample  $R_i$   
   Hits =  $nn$  nearest neighbors from same class as  $R_i$   
   Misses =  $nn$  nearest neighbors from other class as  $R_i$   
   FOR  $j = 1$  to  $N$  (for all attributes)  
      $H_j =$  proportion of the  $nn$  Hits $_j$  that matched  $R_{ij}$   
      $M_j =$  proportion of the  $nn$  Misses $_j$  that matched  $R_{ij}$   
      $W_j = W_j + (H_j - M_j)/k$  (estimate attribute weights)  
   ENDFOR  
 ENDFOR

As the number of attributes  $N$  gets large, the probability that the “wrong” nearest neighbors will be selected increases, for a given sample size  $k$ . This occurs because, as  $N$  increases, irrelevant attributes are more likely to match by coincidence, and thus to be used in determining who is the nearest neighbor. Thus, ReliefF weights become increasingly more random as  $N$  increases. In all our experiments, we used  $nn = 10$  nearest neighbors, and rather than selecting  $k$  random samples, we deterministically selected each of the  $k=M$  samples in the entire dataset.

### B. VLSReliefF

The principle behind VLSReliefF is simple. Since weights estimated by ReliefF are more accurate when applied to smaller attribute sets, we simply apply ReliefF to a population of randomly selected attribute subsets  $S$ , each of size  $N_s < N$ , with the expectation that at least one subset in the population will contain all interacting attributes that are associated with the class outcome (and will thus have

elevated weights for those attributes). One could conceivably recombine the partial results from this population of subsets in a variety of ways. Here, we simply update the global weight of each attribute (i.e., in the full attribute set) with its maximum local weight computed in any subset, as shown in the following pseudo-code:

**“VLSReliefF” Algorithm:**  
 Global Weights  $W_j = -\infty, \forall j \in \{1..N\}$  attributes  
 FOR  $subset = 1$  to  $nsubsets$   
    $S =$  random subset of  $N_s$  attribute indices  
  
 “ReliefF” is applied to subset  $S$ :  
 Local Weights  $X_j = 0, \forall j \in S$  attributes  
  
 FOR  $i = 1$  to  $k$  ( $k = \#$  samples)  
   Select an individual sample  $R_i$   
   Hits =  $nn$  nearest neighbors from same class as  $R_i$   
   Misses =  $nn$  nearest neighbors from other class as  $R_i$   
  
 FOR  $j = 1$  to  $N$  (for all attributes)  
    $H_j =$  proportion of the  $nn$  Hits $_j$  that matched  $R_{ij}$   
    $M_j =$  proportion of the  $nn$  Misses $_j$  that matched  $R_{ij}$   
    $X_j = X_j + (H_j - M_j)/k$  (estimate local attribute weights)  
 ENDFOR  
  
 ENDFOR  
 $W_{S(j)} = \max(W_{S(j)}, X_{S(j)}), \forall j \in S$  (update global weights)  
 ENDFOR

Suppose that there are  $L$  features that interact nonlinearly to predispose for a given class outcome. The number of subsets that must be randomly generated in order to expect one “positive” subset (i.e., which contains all  $L$  interacting features) is:

$$\frac{(N_s - L)! N!}{(N_s)! (N - L)!} \xrightarrow[L \text{ small}]{} \left( \frac{N}{N_s} \right)^L \quad (1)$$

In our implementation, we thus specify the number of subsets in the population to be as follows:

$$nsubsets = \sigma \cdot \left( \frac{N}{N_s} \right)^L \quad (2)$$

The probability that at least one of the subsets is positive thus asymptotes (for large  $N$  and small  $L$ ) to:

$$1 - \left( 1 - \left( \frac{N_s}{N} \right)^L \right)^\sigma \left( \frac{N}{N_s} \right)^L \quad (3)$$

For  $\sigma = 1, 2, 3, 4$  this value asymptotes to 0.63, 0.86, 0.95, and 0.98, respectively. Consequently, in all results reported here,  $\sigma$  was set to 4, so that the probability of stochastically generating at least one positive subset was 0.98. It should be noted that in a real genomic association study, the number  $L$

of interacting SNPs will not be known *a priori*. However, the sample size of the dataset will determine the maximum  $L$  that is statistically supportable (see, for example, the results to our third set of experiments), and this number can be used to bound the population size of VLSReliefF. When  $N$  is very large,  $L > 2$  is not computationally feasible.

The smaller the subset size  $N_S$  is, the more accurate the computed weights will be. On the other hand, the larger  $N_S$  is, the more computationally efficient the algorithm is, since the number of subsets required decreases proportional to the factor  $N_S^{-L}$ . Thus, a practical implementation must balance these two competing goals by using an  $N_S$  that is as large as possible (for computational efficiency), but for which the weights of the interacting attributes are still detectably higher than for extraneous features.

In contrast to ReliefF weights, the global VLSReliefF weights are relatively insensitive to the overall number of SNP attributes  $N$ , since  $N_S$  is constant and the local weights are always computed for subsets of size  $N_S$ . However, the accuracy of VLSReliefF weights is not completely independent of  $N$ , because as  $nsubsets$  increases (with increasing  $N$ ), the use of the “max” operator for combining local weights means that the global weights of irrelevant attributes will also tend to increase slowly with  $N$ , due to stochastic effects.

### C. *iVLSReliefF*

If one is trying to identify which SNPs interact nonlinearly (nonlinear feature selection) in a genome-wide association study, the VLSReliefF algorithm could be applied iteratively, discarding lowest ranked SNPs after each iteration, in place of ReliefF in the proposed TuRF [20] or Evaporative Cooling [13] algorithms. In the current proof-of-concept study, we adopt a simple iterative scheme very much like TuRF, which we refer to as *iVLSReliefF*, in which we simply continue iteratively applying VLSReliefF, retaining only those features that are above a specified percentile rank (PR) of VLSReliefF weights at the end of each iteration, until the feature set has been sufficiently reduced. This basic iterative scheme is outlined in the pseudo-code below.

#### **“iVLSReliefF” Algorithm:**

LOOP

    Compute global Weights from VLSReliefF

    Rank order all attributes by Weight

    Discard all attributes below a given Percentile Rank

UNTIL number of attributes is sufficiently small

Because the VLSReliefF weights are much more accurate for large  $N$  than are ReliefF weights, one can do more aggressive feature pruning at each iteration that suggested in the TuRF [20] or Evaporative Cooling [13] algorithms, and still achieve higher power on larger SNP sets and lower heritabilities. In the experiments reported here, we retained

all SNPs with  $\geq$  the median rank (i.e., at or above the 50<sup>th</sup> percentile). Iteration terminates when the retained set of SNPs is sufficiently reduced in size, at which point more accurate methods of estimating penetrance of particular SNP combinations can be employed (e.g., see [3][16]).

## III. EXPERIMENTS

Three types of computational tests were performed: 1) We compared the accuracy of the weights computed by individual runs of VLSReliefF (with  $N_S = 50$ ) to those of ReliefF, on sets of up to 5000 SNPs containing two purely epistatically interacting SNPs, 2) we applied *iVLSReliefF* (with  $N_S = 50$ ), discarding any SNPs below the 50<sup>th</sup> PR in each iteration, ranked according to their VLSReliefF weights, for genome-scale tests to identify two purely epistatically interacting SNPs from out of sets of 100,000 SNPs and a sample size of only  $M = 1600$ , and 3) we performed experiments towards assessing the optimal subset size  $N_S$  as a function of heritability and sample size. Further details of the experimental designs of each of these 3 types of experiments are described below.

### A. *Comparing VLSReliefF to ReliefF*

We compare VLSReliefF to the full ReliefF algorithm, using several sets of synthetic datasets with purely epistatic associations, in order to validate the method under the most difficult conditions. Each SNP value was encoded as a ternary number, encoding the 3 possible diploid values of 2 possible alleles at each locus. We stochastically generated 30 distinct penetrance tables for each heritability  $h^2 \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$  with  $L = 2$  purely epistatically interacting SNP loci (referred to hereafter as the “true” SNPs). Thus, there were a total of 270 distinct penetrance tables employed in this set of experiments. Interactions between the 2 true SNPs were forced to be purely epistatic, in that there were no class association effects attributable to the individual true SNPs. The penetrance tables were created by a stochastic method designed to achieve the specific desired (narrow sense) heritabilities that meet the criteria described in [2]. It should be noted that some penetrance tables create problem sets that are simply inherently more difficult than others of the same heritability, which is why we used 30 distinct tables at each heritability. Using these penetrance tables, we generated data values for the 2 true SNPs for  $M = 1600$  samples in each experiment (50% cases and 50% controls) and embedded them within a total of 100,000 SNPs, where the values for the additional SNP values were randomly generated and exhibited no true association with the disease class. Experiments were run selecting  $N \in \{100, 200, 500, 1000, 5000\}$  of the 100,000 SNPs, where each set included the 2 true SNPs along with  $N-2$  other extraneous SNPs, for each of the penetrance tables. For these initial proof-of-concept tests, we used  $N_S = 50$  for all experiments. Results of these tests were assessed as follows. For each combination of

experimental parameters, weights of all  $N$  SNPs were estimated using both ReliefF and VLSReliefF. SNPs were then rank ordered by weight, and we determined the percentile rank (PR) of the *lowest* weighted of the 2 true SNPs, for each of the 30 repetitions at a given heritability. The percent of the 30 repetitions that ranked both true SNPs above a designed PR cutoff is shown as the “power” of the method at the given PR cutoff.

### B. Genome-Scale Application of *iVLSReliefF*

We applied the iterative algorithm *iVLSReliefF*, with  $N_S = 50$ , to two genome-scale datasets with  $M = 1600$  samples for  $N = 100,000$  SNPs containing  $L = 2$  purely epistatically interacting SNPs, with heritability  $h^2 = 0.15$ . At the end of each iteration, we retained only those SNPs with  $\geq$  the median percentile rank, according to the VLSReliefF weights (i.e., we used a 50<sup>th</sup> PR cutoff). We iterated 10 times, so that the feature set was reduced from 100,000 SNPs to approximately 100 SNPs. We also computed the ReliefF weights of the current feature set inside each iteration, for comparison, but these were not used as the basis for deciding which SNPs to retain.

### C. Estimating the Optimal Subset Size $N_S$

The optimal subset size  $N_S$  will vary a function of heritability and sample size. In order to begin to empirically assess this, we ran the following set of experiments. We tested 16 logarithmically spaced SNP set sizes, ranging from  $N = 10$  to  $N = 10,000$ , each containing  $L \in \{2,3\}$  true SNPs which interact with pure epistasis to predispose for a disease. As in our other experiments, there were no marginal effects of any individual SNPs, and the 3-way interactions exhibited no 2-way interactions. Each of the above experiments was repeated for 100 stochastically generated penetrance tables at each combination of the 9 heritabilities  $h^2 \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$  and  $L \in \{2,3\}$ . Thus, there were a total of 1800 distinct penetrance tables employed for this set of experiments. We generated  $M \in \{1600, 3200, 6400\}$  samples for each penetrance table and each value of the 16 values of  $N$ , and then assessed the minimum, over all 100 repetitions at each combination of parameters, of the PR of the lowest ranked true SNP in each set. For each combination of  $L \in \{2,3\}$  and  $M \in \{1600, 3200, 6400\}$ , we computed the 50<sup>th</sup> percentile contour lines, over the domain of varying  $N$  and  $h^2$ . We then fit power-law curves to these contours. These curves serve as indicators as to what maximum set size  $N$ , for a given  $h^2$ , ReliefF will reliably rank  $L$  epistatically interacting SNPs at or above the 50<sup>th</sup> percentile. Based on the regression statistics, we devised a combined heuristic for optimizing  $N_S$  as a function of sample size ( $M$ ), heritability ( $h^2$ ), and order of interactions ( $L$ ).

## IV. RESULTS

### A. Results of comparing VLSReliefF to ReliefF

Figure 1 illustrates the results for purely epistatic 2-way SNP interactions, for both VLSReliefF (Figure 1a-d) and

ReliefF (Figure 1e-h), applied to datasets with heritabilities ranging from 0.4 down to 0.1, sample sizes of  $M = 1600$ , and  $N$  ranging from 100 to 5000 SNPs. The small circles denote the PRs of the *lowest* ranked of the 2 true SNPs, in each of the 30 penetrance tables at each combination of experimental parameters (in cases where fewer than 30 circles at a given value of  $N$  are evident in a given column of data points, this is because many of the PRs were identical and lie on top of each other). The dash-dot lines denote the power of the methods at ranking both the true SNPs at or above the 95<sup>th</sup> percentile of the ranks, and the solid line denote the power of the methods at ranking both the true SNPs at or above the 50<sup>th</sup> percentile of the ranks. While ReliefF performs well at high heritabilities and with low numbers of SNPs, the power of this method drops markedly with decreasing heritability and increasing  $N$  (Figure 1e-h). Indeed, at  $N = 5000$  and  $h^2 = 0.1$ , the distribution of the PRs of the ReliefF weights for the two true SNPs from all 30 repetitions was not significantly different from that of a uniform distribution of PRs between 0 and 100 (Kolmogorov-Smirnov test,  $p = 0.27$ ).

In contrast, these experiments demonstrate that VLSReliefF maintains very high power for up to at least the 5000 SNP sets tested, placing all true SNPs at or above the 95<sup>th</sup> percentile at least 93% of the time for all heritabilities above 0.1. At heritability 0.1, the power of VLSReliefF begins to drop off on this relatively small sample of  $M = 1600$ . However even with 5000 SNPs it placed both true SNPs at or above the 95<sup>th</sup> percentile 53% of the time, and at or above the 50<sup>th</sup> percentile 93% of the time. Even at  $N = 5000$  and  $h^2 = 0.1$ , the median of the lowest VLSReliefF PR of the true SNPs in each of the 30 datasets was still 96.4 PR, and the distribution of these PRs was highly significantly different from uniform (Kolmogorov-Smirnov test,  $p = 1.5 \times 10^{-24}$ ).

The difference in performance between the two methods is more easily visualized by the three dimensional surface plots shown in Figure 2. Here, we plot the surface of the power of the two methods in placing both correct SNPs at or above the 95<sup>th</sup> percentile of ranked weights over heritabilities ranging from extremely low (0.01) to relatively high (0.4), and we have truncated the lower limit of the vertical axis at the level that would be expected if the weights were entirely random. There is a rapid performance drop of ReliefF with both decreasing heritability  $h^2$  and increasing numbers of SNPs  $N$  (Figure 2b). On the large  $N$  and low  $h^2$ , the percent of ReliefF weights at or above the 95<sup>th</sup> PR was no better than that of random weights (Figure 2b). In contrast, VLSReliefF (Figure 2a) maintained very high power for all  $h^2 > 0.1$ , and was nearly independent of  $N$  over the ranges tested. At very low heritabilities (0.01 to 0.1), the power of VLSReliefF dropped precipitously with these datasets, because 1600 samples is simply insufficient for ReliefF to reliably score these very low heritability signals in subsets of size  $N_S = 50$ . While the accuracy of ReliefF is clearly not scalable to very large genome-wide

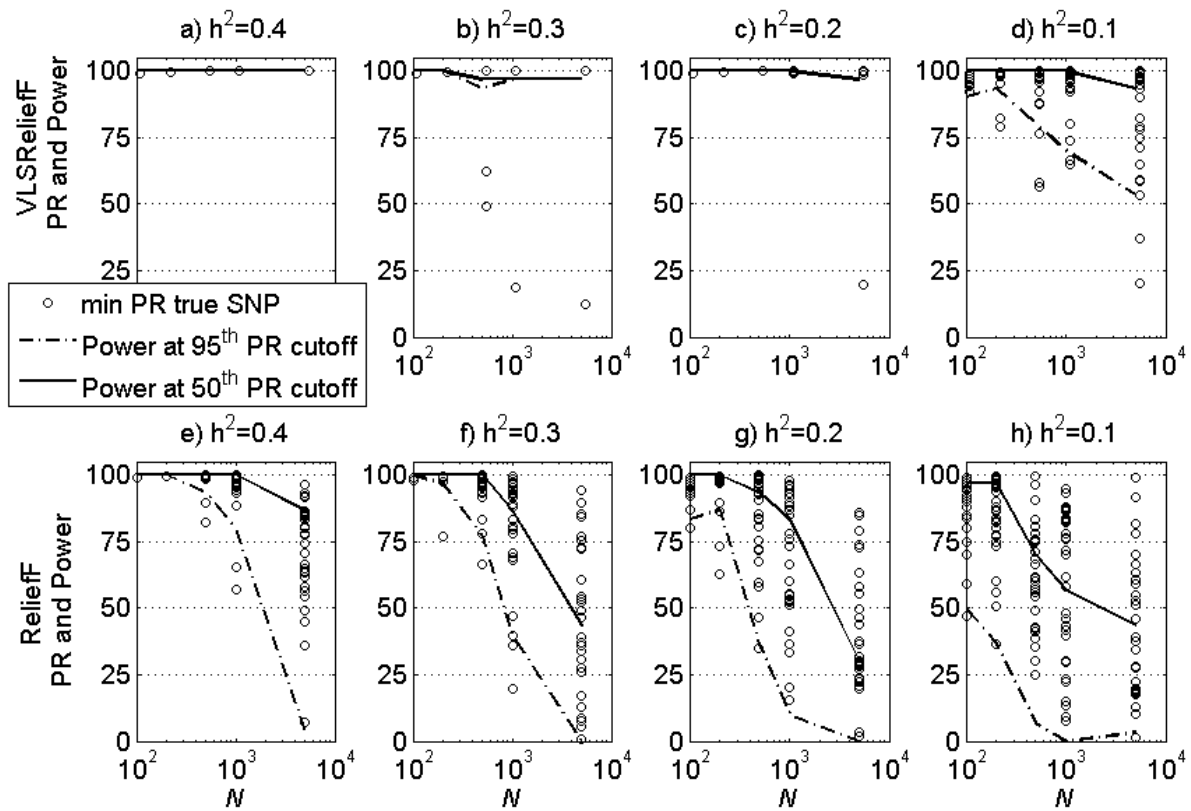


Figure 1. Results in detecting 2 epistatically interacting SNPs, for both VLSRelief (Figure 1a-d) and ReliefF (Figure 1e-h), applied to datasets with sample sizes of  $M = 1600$  and varying heritability and numbers of SNPs  $N$ . The small circles denote the PR (Percentile Rank) of the *lowest* ranked of the 2 true SNPs, for 30 distinct penetrance tables at each combination of  $h^2$  and  $N$ . The lines denote the power of the methods at ranking both the true SNPs at or above the 95<sup>th</sup> PR (dash-dot lines) and the 50<sup>th</sup> PR (solid lines). The legend applies to all subplots, and is placed so that it covers no data points.

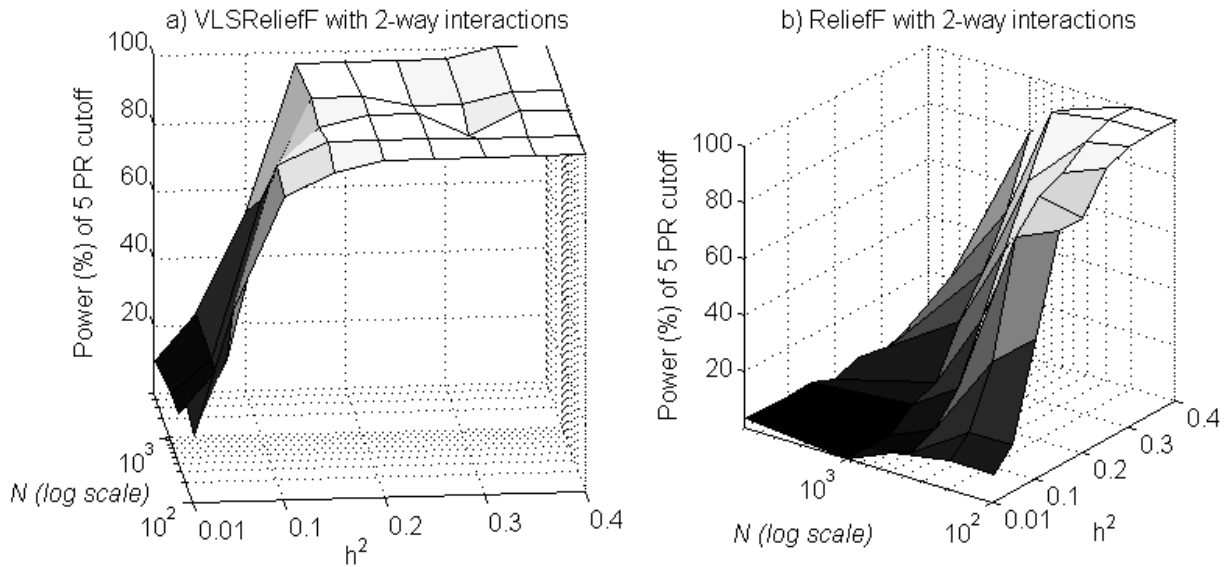


Figure 2. Three-dimensional surface plots of the power (over 30 independent datasets) of a) VLSRelief and b) ReliefF, in their ability to compute weights two purely epistatically interacting SNPs in the 95<sup>th</sup> PR, for  $N$  up to 5000 and  $h^2$  down to 0.01. Each plot is rotated to most clearly elucidate the shape of the surface.

association studies, these results are encouraging that the accuracy of VLSReliefF may be.

### B. Results of Genome-Wide Application of iVLSReliefF

Based on the results above, we tested the ability of the iterative nonlinear feature selection algorithm iVLSReliefF on genome-scale datasets with  $M = 1600$  samples and  $N = 100,000$  SNPs containing a 2-way purely epistatic interaction that predisposed for disease with heritability  $h^2 = 0.15$ . As indicated by Figure 2a, this heritability is at the lower end of where we anticipate that iVLSReliefF will be able to reliably detect the true SNPs, given a sample size of only 1600. However, it was not clear if the strong results shown in Figure 2a would scale out to very large genome-scale SNP sets with  $N = 100,000$ . At the time of this writing, we have completed two full tests of iVLSReliefF as described above, using two penetrance tables, one of which had appeared to be of moderate difficulty and the other which had appeared to be relatively hard in the tests described above. In Figure 3 we illustrate the PR of the lowest ranked of the true SNPs at each iteration (where iterations proceed right to left, as features are eliminated) for the two datasets. For comparison, we also show the corresponding PRs of the lowest ranked of the true SNPs, using ReliefF weights applied to the same remaining feature set. Clearly, VLSReliefF far outperformed ReliefF on these two datasets, and even with a relatively aggressive 50<sup>th</sup> percentile rank cutoff, iterative feature selection with iVLSReliefF always retained the 2 true epistatically interacting SNPs in these two datasets, with the lesser weighted of these at the 98<sup>th</sup> PR in

the final iteration.

### C. Results of Optimal Estimation of Subset Size $N_S$

In the experiments described above, we used a subset size of  $N_S = 50$ . Ideally, however, the subset size  $N_S$  used in VLSReliefF should be set as large as possible, in order to minimize computational costs, but small enough so that interacting SNPs that are nonlinearly associated with disease class are ranked highly by ReliefF. Thus, our third set of experiments was designed to begin to assess how to determine an appropriate  $N_S$  as a function of heritability, sample size, and order of interactions. In Figure 4 we show the empirically determined 50<sup>th</sup> PR contours (dashed lines) as a function of heritability  $h^2$  and number of SNPs  $N$ , for 2-way and 3-way interactions at 3 different sample sizes. All these contour curves exhibited power-law behaviors ( $R^2 > 0.95$ ). We then heuristically combined the parameters for the best-fit curves to devise an ad-hoc formula for choosing an approximately optimal subset size  $N_S$ , as a function of sample size  $M$ , heritability  $h^2$ , and the anticipated order of interactions  $L$ , as follows:

$$N_S \approx 3.75 \cdot (L-1)^{-4} \cdot M^{(L-1) \cdot 0.25} \cdot h^{4 \cdot (L-1)^{-1}} \quad (4)$$

(where  $h^4$  is the square of the heritability  $h^2$ ). The heuristic curves yielded by equation (4) are also shown on Figure 4 (solid lines). Choosing an  $N_S$  by equation (4) would thus mean that, with very high probability,  $L$  purely epistatically interacting SNPs would be ranked in the top 50<sup>th</sup> percentile rank of VLSReliefF weights, facilitating iterative feature selection. Further work is clearly needed to see how

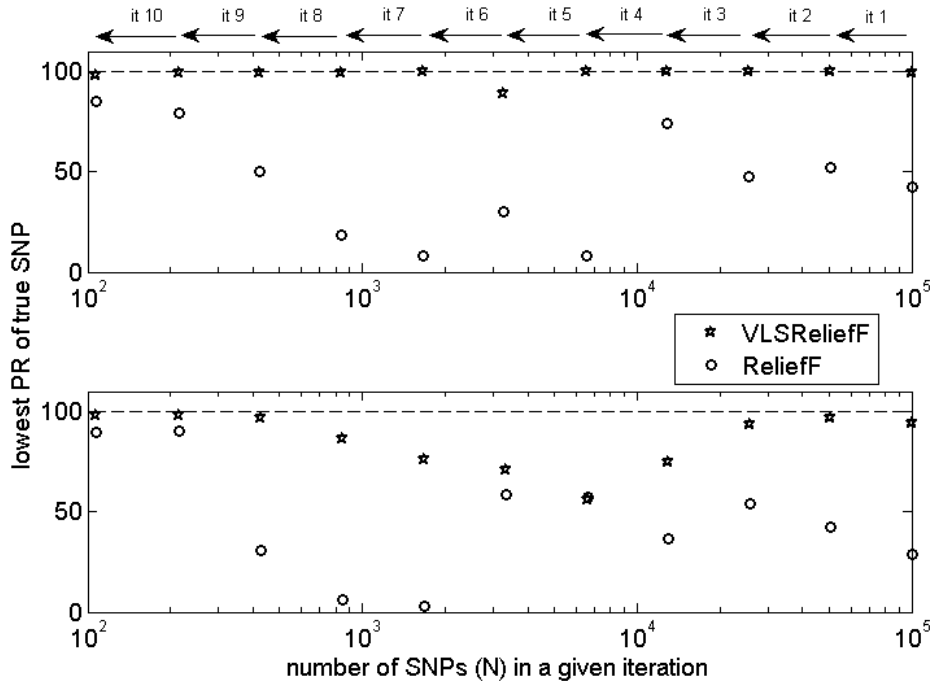


Figure 3. Percentile Rank (PR) of the lowest ranked of the 2 true SNPs at each iteration (where iterations of iVLSReliefF proceed right to left, as features are eliminated from the initial genome-scale set size of  $N = 100,000$  SNPs) for two datasets with  $M = 1600$  sample datasets with 2-way epistatic interactions that predispose for disease with heritability  $h^2 = 0.15$ . Based on prior tests on 30 penetrances tables at  $h^2 = 0.15$ , we selected the penetrance table used for the top plot as an example of moderate difficulty and the one used for the bottom plot as an example of a relatively hard table.

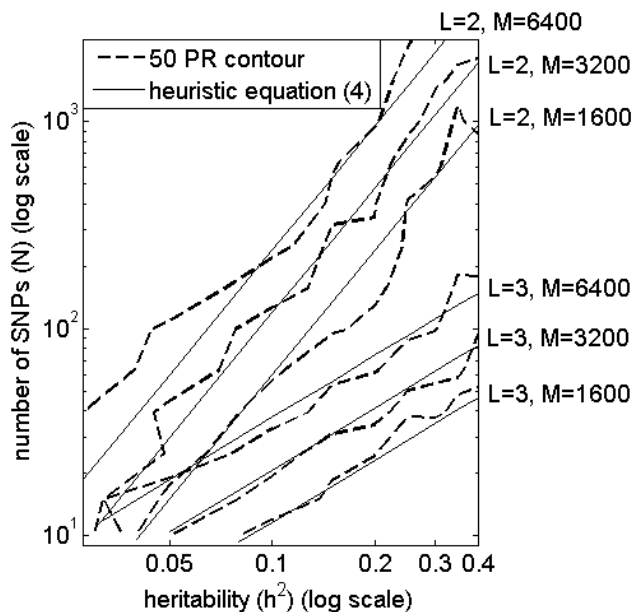


Figure 4. The 50<sup>th</sup> Percentile Rank contour lines from empirical tests, along with the heuristic approximation curve (equation 4).

generally equation (4) applies, but these results are encouraging that such a heuristic rule-of-thumb may be found. Figure 4 also highlights the fact that purely epistatic 3-way interactions are much more difficult to reliably identify with ReliefF than are purely epistatic 2-way interactions, and require much larger sample sizes and, if embedded in VLSReliefF, much smaller  $N_S$ , in order to yield reliable results.

## V. DISCUSSION AND CONCLUSIONS

Our results confirm that ReliefF is quite accurate on relatively small feature sets and high heritabilities. However, the accuracy of ReliefF does not scale up to the sizes needed for truly large genome-scale SNP association studies, where ReliefF weights are statistically no better than random. VLSReliefF mitigates this performance drop by stochastically applying ReliefF to SNP subsets of manageable size, and then assigning each SNP the maximum ReliefF weight it achieved in any subset. An iterative application of VLSReliefF, in which only SNPs in the top half of the percentile ranks are retained between iterations, was shown to successfully retain both of two purely epistatically interacting SNPs with heritability 0.15, in two datasets originally containing 100,000 SNPs for each of only 1600 samples. Future studies are needed to more fully characterize the robustness of iVLSReliefF on genome-scale datasets with different heritabilities and higher-order interactions.

Although the performance of VLSReliefF appears to scale well up to truly genome-scale association studies, the computational complexity unfortunately still scales exponentially, when equation (2) is used to specify  $\theta((N/N_S)^L)$  evaluations of ReliefF, each of which has time complexity bounded by  $O(M^2 N_S)$  [25]. This time complexity is reduced proportional to the factor  $N_S^L$ , so it is

important to choose  $N_S$  as large as possible that will still yield accurate weights. Our results here indicate that it may be possible to choose an appropriate  $N_S$  heuristically, as a function of sample size and heritability. Due to the exponential scaling of equation (2), and the difficulty ReliefF has in detecting 3-way pure epistasis except in relatively small SNP sets (Figure 4), VLSReliefF, like exhaustive 3-way search, is not computationally practical for true genome-wide association studies searching for pure epistasis between  $L > 2$  SNPs. However, it certainly extends the reliability of attribute weights into very large scale attribute sets where ReliefF weights are not distinguishable from random. Additionally, unlike exhaustive pair-wise search, VLSReliefF assigns importance scores to all  $N$  SNPs in the dataset. Thus, it may still provide insights into higher-order interactions involved in the genetic etiology of complex disease, even if one uses  $L = 2$  in determining  $N_S$  by equation (4) and the number of subsets by equation (2), especially if there are also some detectable marginal or 2-way epistatic effects among the higher-order interacting SNPs. We are continuing to characterize and optimize the performance of the algorithm, and look forward to applying it to SNP feature selection in large-scale genomic association studies in the near future.

## ACKNOWLEDGMENTS

We thank Jason H. Moore and the Computational Genetics Laboratory at Dartmouth College for the use of their synthetic data generator, and Joshua L. Payne for his previous assistance in modifications to this data generator.

## REFERENCES

- [1] H.H. Barrett and K.J. Myers, Foundations of Image Science. John Wiley & Sons, Inc., New Jersey, 2004.
- [2] R. Culverhouse, B.K. Suarez, J. Lin, and T. Reich, "A perspective on epistasis: Limits of models displaying no main effect," *Am. J. Hum. Genet.*, vol. 70, pp. 461-471, 2002.
- [3] M.J. Eppstein, J.L. Payne, B.C. White and J.H. Moore, "Genomic mining for complex disease traits with 'Random Chemistry'", *Genetic Programming and Evolvable Machines (special issue on Medical Applications)*, 8:395-411, 2007. (DOI 10.1007/s10710-007-9039-5)
- [4] A.M. Glazier, J.H. Nadeau, and T.J. Aitman, "Finding genes that underlie complex traits," *Science*, vol. 298, pp. 2345-2349, 2002.
- [5] J.N. Hirschhorn, and M.J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, pp. 95-108, 2005.
- [6] J. Hoh, A. Wille, and J. Ott, "Trimming, weighting, and grouping SNPs in human case-control association studies," *Gen. Res.*, vol. 11, pp. 2115-2119, 2001.
- [7] International HapMap Consortium, "The international HapMap project," *Nature*, vol. 426, pp. 789-796, 2003.
- [8] International human genome sequencing consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921, 2001.
- [9] International SNP map working group, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, pp. 928-933, 2001.
- [10] S. Kauffman, At Home in the Universe: The Search for the Laws of Self-Organization and Complexity. Oxford Univ. Press, USA, 1996.
- [11] L. Kruglyak and D.A. Nickerson, "Variation is the spice of life," *Nature Genet.*, vol. 27, pp. 234-236, 2001.
- [12] P.R. Lucek and J. Ott, "Neural network analysis of complex traits," *Gen. Epidemiol.*, vol. 14, pp. 1101-1106, 1997.

- [13] B.A. McKinney, D.M. Reif, B.C. White, J.E. Crowe, Jr. and J.H. Moore, "Evaporative cooling feature selection for genotypic data involving interactions", *Bioinf*, vol. 23, pp. 2113-2120, 2007.
- [14] K.R. Merikangas, N.C.P. Low, and J. Hardy, "Understanding sources of complexity in chronic diseases – the importance of integration of genetics and epidemiology," *Int. J. Epid.*, vol. 35, pp. 590-592, 2006.
- [15] J.H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Hum. Hered.*, vol. 56, pp. 73-82, 2003.
- [16] J.H. Moore, "Computational analysis of gene-gene interactions in common human diseases using multifactor dimensionality reduction," *Expert. Rev. Mol. Diagn.*, vol. 4, pp. 795-803, 2004.
- [17] J.H. Moore, J.C. Gilbert, C.T. Tsai, F.T. Chiang, T. Holden, N. Barney, and B.C. White, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *J Theor Biol.*, vol. 241, pp. 252-261, 2006.
- [18] J.H. Moore and M.D. Ritchie, "The challenges of whole-genome approaches to common diseases," *JAMA*, vol. 291, pp.1642-1643, 2002.
- [19] J.H. Moore and B.C. White, "Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge," In *Genetic Programming Theory and Practice IV*, R.L. Riolo, T. Soule, and B. Worzel (eds.), Ann Arbor, 2006.
- [20] J.H. Moore and B.C. White, "Tuning ReliefF for genome-wide genetic analysis," In *Lecture Notes in Computer Science*, Rajapakse, J.C. et al. (Eds), 4447, 166-175, Springer, New York, 2007.
- [21] J. Ott and J. Hoh, "Statistical multilocus methods for disequilibrium analysis in complex traits," *Hum. Mut.*, vol. 17, pp. 285-288, 2001.
- [22] L. Peltonen and V.A. McKusick, "Dissecting Human Disease in the Postgenomic Era," *Science*, vol. 291, pp. 1224-1229, 2001.
- [23] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabasi, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551-1555, 2002.
- [24] 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," *Amer. J. of Hum. Gen.*, vol. 69, pp. 138-147, 2001.
- [25] M. Robnik-Sikonja, and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learning*, vol. 53, pp. 23-69, 2003.
- [26] A.C. Syvanen, "Assessing genetic variation: genotyping single nucleotide polymorphisms," *Nature Rev. Genet.*, vol. 2, pp. 930-942, 2001.
- [27] T.A. Thornton-Wells, J.H. Moore, and J.L. Haines, "Genetics, statistics and human disease: analytical retooling for complexity," *Trends Genet.*, vol. 20, pp. 640-647, 2004.
- [28] J.C. Venter et al., "The sequence of the human genome," *Science*, vol. 291, pp. 1304-1351, 2001.
- [29] W.Y. Wang, B.J. Barratt, D.G. Clayton, J.A. Todd, "Genome-wide association studies: theoretical and practical concerns," *Nat. Rev. Genet.*, vol. 6, pp. 109-18, 2005.
- [30] B.C. White, J.C. Gilbert, D.M. Reif, J.H. Moore, "A statistical comparison of grammatical evolution strategies in the domain of human genetics," in *Proc. of the IEEE Congress on Evol. Computing*, D. Corne et al. (eds.) IEEE Press, Edinburgh, UK, 2005, pp. 676-682.