Verma et al. BioData Mining (2018) 11:5
https://doi.org/10.1186/s13040-018-0168-6

BioData Mining

RESEARCH Open Access

Collective feature selection to identify crucial epistatic variants

Shefali S. Verma[1,2,3], Anastasia Lucas[1,3], Xinyuan Zhang[2,3], Yogasudha Veturi[1,3], Scott Dudek[1,3], Binglan Li[2,3], Ruowang Li[3], Ryan Urbanowicz[3], Jason H. Moore[3], Dokyoon Kim[1] and Marylyn D. Ritchie[1,2,3*]

* Correspondence: marylyn@pennmedicine.upenn.edu
[1]Biomedical and Translational Bioinformatics Institute, Geisinger Health System, 100 N Academy Avenue, Danville, PA 17822, USA
[2]Huck Institute of Life Sciences, The Pennsylvania State University, University Park, PA, USA
Full list of author information is available at the end of the article

Abstract

Background: Machine learning methods have gained popularity and practicality in identifying linear and non-linear effects of variants associated with complex disease/traits. Detection of epistatic interactions still remains a challenge due to the large number of features and relatively small sample size as input, thus leading to the so-called "short fat data" problem. The efficiency of machine learning methods can be increased by limiting the number of input features. Thus, it is very important to perform variable selection before searching for epistasis. Many methods have been evaluated and proposed to perform feature selection, but no single method works best in all scenarios. We demonstrate this by conducting two separate simulation analyses to evaluate the proposed collective feature selection approach.

Results: Through our simulation study we propose a collective feature selection approach to select features that are in the "union" of the best performing methods. We explored various parametric, non-parametric, and data mining approaches to perform feature selection. We choose our top performing methods to select the union of the resulting variables based on a user-defined percentage of variants selected from each method to take to downstream analysis. Our simulation analysis shows that non-parametric data mining approaches, such as MDR, may work best under one simulation criteria for the high effect size (penetrance) datasets, while non-parametric methods designed for feature selection, such as Ranger and Gradient boosting, work best under other simulation criteria. Thus, using a collective approach proves to be more beneficial for selecting variables with epistatic effects also in low effect size datasets and different genetic architectures. Following this, we applied our proposed collective feature selection approach to select the top 1% of variables to identify potential interacting variables associated with Body Mass Index (BMI) in ~ 44,000 samples obtained from Geisinger's MyCode Community Health Initiative (on behalf of DiscovEHR collaboration).

Conclusions: In this study, we were able to show that selecting variables using a collective feature selection approach could help in selecting true positive epistatic variables more frequently than applying any single method for feature selection via simulation studies. We were able to demonstrate the effectiveness of collective feature selection along with a comparison of many methods in our simulation analysis. We also applied our method to identify non-linear networks associated with obesity.

Keywords: Feature selection, Epistasis, Non-additive effects, Obesity, Parametric methods, Non-parametric methods

Background

The advancements and cost-effectiveness of genotyping and sequencing technologies have led to the ever increasing "short fat data" problem (where the number of features outnumbers the sample size; p> > n) in applying various machine learning methods to detect epistasis [1, 2]. Gene-gene interactions are considered as crucial components in the origination of the "missing heritability" for testing association of variants with single or multiple disease traits [3, 4]. Various statistical and biological filtering techniques are commonly applied to select variants that are most meaningful in the search for epistatic interactions linked with common and complex diseases [5, 6]. Regression approaches are frequently used to model pairwise interactions but many machine learning approaches such as multi-factor dimensionality reduction (MDR) [7, 8], neural networks [9], support vector machines [10], Bayesian methods [11], among others are contemporary methods now more commonly applied. Most of these methods are limited in the number of features they can handle, and thus dealing with the computational burden poses a challenge in the application of these methods. Beside the computational burden, it is also important to note that the efficiency of most learning-based methods can be improved to a greater extent if the number of input variables can be reduced. In order to do so, many feature selection methods have been proposed in the past and have been applied in the context of detecting statistical epistasis to identify non-linear associations of genetic variants with a disease trait. Hence, feature selection is not a new concept. Several parametric and non-parametric methods such as LASSO [12], Elastic Net [13], Random Forests [14], ReliefF [15], Gradient Boosting [16], etc., have been developed and used frequently to perform feature selection. All methods have some advantages and disadvantages, and thus they do not follow a "one method fits all" criterion.

In this study, we tested an eclectic set of parametric and non-parametric methods on simulated datasets to first pick a few orthogonal methods to use in selecting features that can be used in downstream analysis of epistasis. We compared these methods based on both efficiency and effectiveness. We observed that different methods tend to select variables based on different important aspects. Thus, we suggest a collective feature selection approach. We propose to select the union of features from the top comparable methods. The concept of taking the input from many algorithms, to select variables as a collective opinion is in line with the "no free lunch" theorem of optimization which states that in searching for candidate solutions, no one algorithm can be specialized to all problems [17]. Unknown genetic etiology of complex diseases makes it theoretically impossible for one algorithm to be specialized in identifying all possible combinations of predictors associated with a disease. This concept is also similar to the concept of "Crowd Machine" which has been explained in previous work [18]. Crowd Machine learning refers to combining multiple machine learning methods into a single machine learning method so that the features from all methods can be used effectively. Our proposed method is a variation of this concept. We recommend applying a collective approach using various top performing feature selection methods to identify variants with varying effect sizes (high and low penetrance) and MAF.

Methods

In this section, we will describe the datasets used for simulations and real data analyses as well as the statistical methods applied for conducting feature selection.

Simulation studies

Simulated data experiment 1

We simulated multiple data sets consisting of SNPs (single nucleotide polymorphisms), referred to as variables, using an additive genetic encoding (AA = 0, Aa = 1, aa = 2) with case-control status to test for binary outcome. Our simulation parameters consisted of various combinations of the following epidemiological characteristics:

- Disease penetrance: This refers to the strength of the simulated signal or effect size and thus directly corresponds to the heritability of the phenotype. We have used previously simulated data with the same signal strength, these are listed as 0.1_diagonal, 0.5_diagonal and 0.9_diagonal and have been previously explained in Li et al. [11].
- Number of disease sites: This refers to the number of SNPs that contributes to the total effect in the dataset.
- Minor Allele Frequency (MAF): For many genetic interaction studies, it has been shown that MAF highly influences power to detect true interactions. Therefore, we limited our analysis to only common alleles above MAF 0.4 [19, 20]. For main effect variants, we limited the MAF of the causal SNP to 0.4. For interacting effects, MAF for each of the two interacting SNPs was also set to 0.4.
- Number of Samples: We generated 8 simulation scenarios consisting of balanced datasets with 2000 cases and 2000 controls.
- Number of Variants: We set the number of variants as 100 and 500 to address the computation burden.

We simulated (a) main effect only and (b) interaction effect only datasets using a simulation procedure that has been previously explained [11]. As described, we evaluated feature selection methods on similar datasets. Table 1 lists details of all parameters used in generating main effect and interaction effect datasets. For all the simulation analyses, we generated 10 data sets for each combination of parameters in this experiment since we were interested in obtaining mean accuracy values or scores from the replicates to compare across different methods.

Table 1 Parameters used for generating simulated experiment 1 data

Type of Effect (100 and 500 SNPs)	Dataset name	Causal SNP Model (Penetrance: 0.1, 0.5 and 0.9)
Main Effect	1SNP	G1
	2SNP	G1, G2
	3SNP	G1, G2, G3
	4SNP	G1, G2, G3, G4
Interaction Effect	case1_control0	G1 < –>G2
	case1_control1	G1 < –>G2 G99 < –>G100
	case2_control0	G1 < –>G2 G3 < –>G4
	case2_control2	G1 < –>G2 G3 < –>G4 G97 < –>G98 G99 < –>G100

All datasets consisted of 2000 cases and 2000 controls (4000 samples in total). 'G' here refers to the SNP ID prefix

Using marginal association as an example, the association of a variable with a binary outcome simply indicates differences of allele frequencies between cases and controls. Using a really simple dummy example as in Table 2, we can see that there are more 1 's and 2 's for SNP1 in the controls than cases. Thus, in simulation, we could simulate case and control data separately by specifying different allele frequencies for SNP1. Using frequencies as probabilities, we then used sample with replacement until we reached the desired samples.

The same logic extends to interacting variables. Only in this case, we are specifying different joint allele frequencies between two SNPs. For example, there are more SNP1 = 2 and SNP2 = 2 combination for controls than cases in Table 3. We used the different joint allele frequencies to simulate interaction case and control data.

Therefore, simulated variables with interaction effects were generated in case and control datasets separately. For example, in dataset 1SNP, there is 1 SNP with a main effect and similarly for the interaction effect datasets, such as case2_control2, there are 2 pairs of interaction SNPs simulated in cases (G1 < −>G2 and G3 < −>G4) and 2 interaction pairs in controls (G97 < −>G98 and G99 < −>G100).

Simulated data experiment 2

We simulated another set of datasets with a slightly different architecture to recapitulate the need to consider the varying architectures of complex disease traits. For this simulation, we used GAMETES software to simulate two different genetic architectures [21]. The two architectures here are reflected by the ease of detection measure (EDM) which makes the model easier (EDM-2) or harder (EDM-1) to detect [22]. We simulated 100 features with MAF 0.2 for all predictive features in each dataset. In this scenario, we simulated datasets with a sample size of 2000 (1000 cases and 1000 controls). For each combination of parameters, we simulated 50 replicates at a heritability value of 0.1, 0.2, or 0.4. There were 3 predictive features and 97 non-predictive features simulated in each dataset. The predictive features are included in such a way that there is a pairwise pure epistatic interaction as well as a third main effect additively combined with the interaction.

Table 2 Dummy example representing the simulation criteria for main effects in simulation experiment #1

Case control status	SNP1
0	1
0	2
0	2
0	0
1	0
1	0
1	0
1	1

Here 0 in column1 refers to controls and 1 refers to cases. In column 2, 0,1 and 2 refers to the genotypes

Table 3 Dummy example representing the simulation criteria for interacting effects in simulation experiment #1

Case control status	SNP1	SNP2
0	2	2
0	2	2
0	2	2
0	0	0
1	0	0
1	0	0
1	0	0
1	1	1

Here 0 in column1 refers to controls and 1 refers to cases. In column 2 and 3, 0,1 and 2 refers to the genotypes

Biological data application

Natural biological data

We applied the proposed collective feature selection to a real dataset obtained from the Geisinger MyCode DiscovEHR collaboration [23, 24]. At the time of these analyses, the DiscovEHR study consisted of 60,000 samples whose genotype data (using Illumina Human Omni Express Exome chip) is linked to their Electronic Health Record (EHR). For our analysis, we extracted unrelated European American samples of age 18 or older. We extracted all available Body Mass Index (BMI) values for all samples who also had genotype data, from the Geisinger EHR. Median BMI was calculated for all samples and used as the basis for the obesity phenotype in the subsequent analyses. Average BMI of DiscovEHR population is 30 [24]. After quality control, 40,449 samples were divided into cases and controls where samples with BMI ranging from 18 to 24.9 (defined as normal range) were considered as obesity controls and samples with BMI > 30 (defined as obese) were considered as obesity cases. We excluded samples in marginal BMI range (25–30) to remove phenotypic heterogeneity and classify samples as normal and obese (extremes of the distribution). To conduct a two-step analysis for feature selection and model testing, we divided the dataset randomly into two parts: variable selection dataset and modelling dataset. Our variable selection dataset was used for feature selection and consisted of 15,201 samples (3917 controls and 11,284 cases) while our modelling dataset, which contained 14,925 samples in total (3767 controls 11,158 cases), was used for downstream analyses. We also performed quality control on genotype data to only include variants with genotyping call rate > 99%, MAF > 20% and HWE *P*-value<1e-07. Lower frequency variants (MAF < 0.2) were excluded from analyses as a first filtration step so as to compare our methodology to the simulated datasets. Additionally, studies also suggest that for variants with MAF < 0.2, the interaction effects do not explain much of genetic variance [19, 20]. To reduce the search space for testing, we LD-pruned the data to only include independent variants. We used an R^2 threshold of 0.2 for LD pruning. After genotype QC, the training dataset consisted of 60,232 variants for feature selection.

Statistical methods

To compare and contrast the different methods that can be used to select features with non-additive effects, we chose a wide range of filter and embedded methods. For filter-

based feature selection methods we tested MultiSURF* and MDR and for embedded methods we tested Random forests, gradient boosting, LASSO and Elastic Net. In this manuscript, we divided methods into parametric and non-parametric methods. We used this terminology throughout the manuscript to classify the methods tested. Figure 1 lists the three categories of methods we tested for feature selection. We will describe in detail how we ran analyses using these methods in this section. Datasets that are imputed or obtained from commercial genotyping chips such as Illumina consist of 500 K to approximately 10 M variants. After quality control and LD pruning to include only the most independent variables, it is common to still be left with over 50,000 variants that can be exhaustively tested for interactions. Feature selection can reduce the number of variants and consequently, the computational burden for downstream analysis. Since we chose different methods to test in our analysis, it is important to note that the format of output from all these methods varies and has limited the way in which we can compare the accuracy of these methods. For example, some methods provide a test-statistic for every model, where others provide a ranked list of variables based on performance. In comparing all methods, we could not choose an arbitrary test statistic threshold for each method as that could create bias in selecting variables based on different test statistics as followed in each method (for example MDR and uses balanced accuracy for ranking models, LASSO and elastic net uses lambda for ranking variables, Ranger and gradient boosting uses variable importance measure based on prediction accuracy for ranking variables, etc.). Therefore, to compare all different feature selection approaches, we employed a ranking based method to extract the highest ranked features based on a user-defined percentage after running each algorithm. Ranking here refers to the score or the accuracy estimates from each algorithm separately. For all our simulation tests, we showed results for selecting variables at several different user-defined thresholds: 2% or 3% (based on number of effect SNPs in the two sets of simulated datasets), 5% and 10%, thus we select the top 2/3%, 5%, or 10% of the

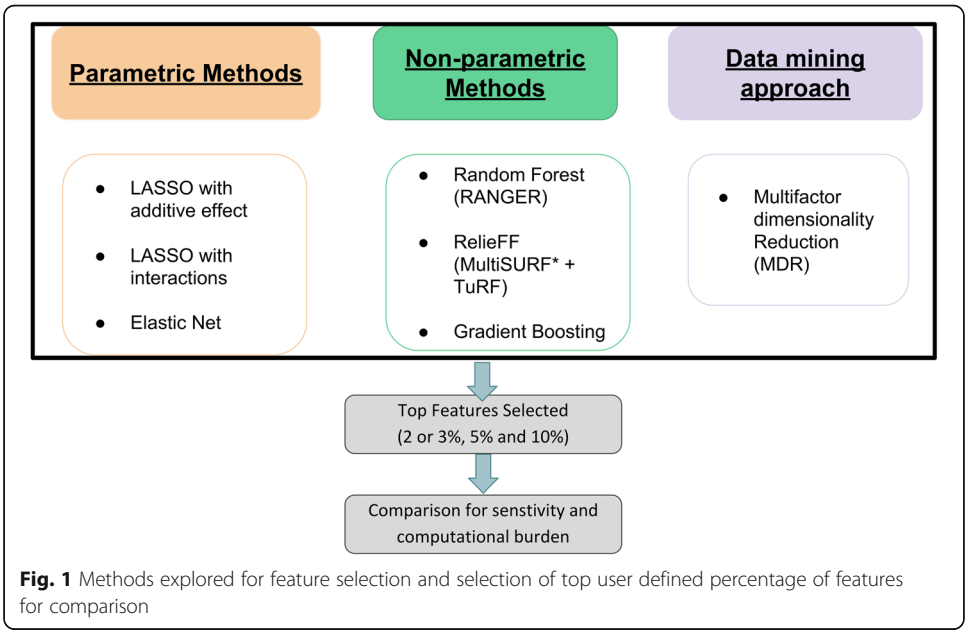

Fig. 1 Methods explored for feature selection and selection of top user defined percentage of features for comparison

variables in both simulated data experiments to investigate whether the methods perform better or worse, i.e. selection of true positives in comparison to false positives (see Fig. 2). Next, we ran all of the methods on all replicates of datasets generated from the combinations of parameters as explained in the simulated data section to reduce the error and increase robustness of the models selected. We then averaged across the replicated runs to compare the results.

Parametric methods

LASSO and Elastic Net regularizations are widely accepted methods for feature selection [6]. Least Absolute Shrinkage and Selection Operator, or LASSO [12, 25], is a shrinkage and variable selection method with imposed *L1* regularization on the regression coefficients. Since the main goal of this analysis is to detect features that exhibit interaction effects, we ran LASSO regression to include both additive effects of SNPs in the models and exhaustive pairwise interactions of all SNPs in the model. Below are the equations representing LASSO regularization for single SNP and interactions:

Penalized estimates for the model with additive effects alone can be derived as the solution to the following optimization problem:

$$(\mu, \hat{\beta}) = min\left\{\sum_{i}^{n} \left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 + \lambda J(\beta)\right\}$$

where $\sum_i^n (y_i - \mu - \sum_{j=1}^p x_{ij}\,\beta_j)^2$ is the residual sum of squares, $\lambda >= 0$ is the regularization parameter and $J(\beta)$ is the penalty function.

For LASSO Regularization with additive effects only, the penalty function is the L1 norm and can be expressed as follows:

$$J(\beta) = \sum_{j=1}^{p} |\,\beta_j\,|$$

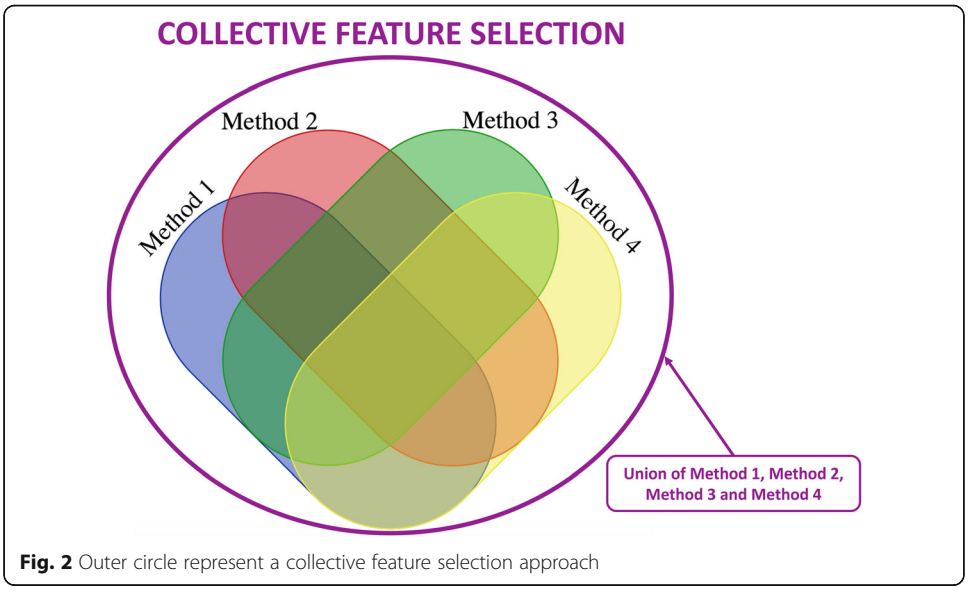

Fig. 2 Outer circle represent a collective feature selection approach

Likewise, penalized estimates for the regression model with additive and SNP-SNP interaction effects can be derived as the solution to the following optimization problem:

$$\widehat{(\mu, \beta, \hat{\gamma})} = \min\left\{\sum_{i}^{n}\left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\,\beta_j - \sum_{j=1}^{p}\sum_{k<j}^{p} x_{ij}\,x_{ik}\gamma_{jk}\right)^2 + \lambda J(\beta,\gamma)\right\}$$

Again, for LASSO Regularization with additive effects and interactions together, the penalty function can be expressed as follows:

$$J(\beta,\gamma) = \sum_{j=1}^{p} |\,\beta_j\,| + \sum_{j=1}^{\mathrm{p}}\sum_{k<j}^{p} |\,\gamma_{jk}\,|$$

LASSO combines variable selection and shrinkage of variables, but it has a drawback when the number of predictors is greater than the number of samples ($p > \mathrm{n}$), in which case it tends to select at most n predictors. Also, when predictors are correlated, LASSO is outperformed by ridge regression. Thus, we modeled the data with ridge regression in a preliminary part of our analysis but did not include those results in this manuscript since they were similar to those from LASSO. Next, we explored another penalized regression method, the Elastic Net, which works well in selecting a group of correlated variables and does not limit the selection of the number of variables. Elastic Net uses a weighted average of the *L1* and *L2* norms for its penalty function.

Similar to the LASSO penalty function, the elastic net penalty function [Zou et al.; 2005] for the model with additive effects and interactions can be expressed as follows:

$$J(\beta,\gamma) = \alpha\left(\sum_{j=1}^{p}\left|\beta_j\right| + \sum_{j=1}^{p}\sum_{k<j}|\beta_{jk}|\right) + (1-\alpha)\left(\sum_{j=1}^{p}\left|\beta_j\right|^2 + \sum_{j=1}^{p}\sum_{k<j}\left|\beta_{jk}\right|^2\right]$$

Both these penalized regression methods (LASSO and elastic net) require optimization of λ. Elastic net involves another tuning parameter called α, which is commonly set to 0.5. In order to help tune these parameters, we performed 5-fold cross validations for these two methods and chose the most optimal regularization parameter for feature selection.

Non-parametric methods

Even though parametric methods are simple and easy to understand, they do not always fit the complex nature of biology. Thus, exploring some non-parametric methods is also necessary. Non-parametric methods do not make assumptions about the distribution of variables and underlying genetic architecture. These methods usually work best for "big data" problems. We tested two decision-tree based methods, including Random forests and Gradient Boosting, and we also tested a non-heuristic ReliefF algorithm variation called Multiple Threshold Spatially Uniform ReliefF (MultiSURF*) [26].

For our random forests implementation, we used the RANGER R package [27]. We tuned random forests to get better results, setting number of trees as 1000 for main effect datasets and 4500 for datasets with interaction effects. The other parameter that we tuned is the number of variables that split each node; we used 35 for main effects, 70 for interaction effects in datasets with 100 SNPs, and 200 for interaction effects in datasets with 500 SNPs. We also used gradient boosting implementation in the GBM R

package. For gradient boosting, we set the number of trees as 800 for main effect datasets and 15,000 trees for interaction effects datasets. We set the bag fraction as 0.5 and shrinkage as 0.01, which have been suggested to result in the best performance based on the best practices from R package manual (https://cran.r-project.org/web/packages/gbm/gbm.pdf). TuRF refers to Tuned ReliefF and it performs feature selection recursively. It is suggested to use TuRF along with ReliefF algorithms to get better performance when using a large number of variables [5, 15, 28]. Thus, it is important to test the number of variables that will be thrown out at every iteration. We tested discarding 1%, 5% and 10% of least predictive variables at each iteration to determine the appropriate threshold for MultiSURF* + TuRF in order to identify more true positives in a computationally feasible amount of time.

Non-parametric data mining approach for feature selection

Multifactor Dimensionality Reduction (MDR) has been traditionally applied to several association studies including gene-gene and gene-environment interaction studies [7, 8]. Many different versions of MDR have also been proposed for different data types [29–31]. Using MDR as filtering method has also been previously tested and compared with other methods [32]. We utilized parallel MDR (pMDR) {https://ritchielab.psu.edu/software/mdr-download} in similar way to Oki NO et al. [33], where we ran all main effect models and two-way interactions without cross validations and then ranked the variables based on their training accuracy.

Collective feature selection

A plethora of machine learning and feature selection methods have been proposed and tested in various studies [6, 12, 13, 16, 32]. In this manuscript, we aimed to compare a few of these methods; however, picking one method can be convenient but not always pertinent. Thus, in our analysis we proposed to select a few orthogonal feature selection methods from what were tested and then use the union of all variables selected from these methods for any downstream analysis. Figure 2 depicts the concept of our collective feature selection approach in selecting variables in a dummy set of 4 methods listed as Method 1 to 4. We applied this approach to simulated data experiments 1 and 2 but are only showing results from experiment 2 where we selected top 3%, 5% and 10% features from each of the 4 methods. Results from experiment 1 are similar. We compared these methods in terms of number of overlapping features, number of true positives, and number of false positives detected from each. To represent the overlapping features and features selected for EDM-1 and EDM-2 models, we merged the datasets with 3 different heritability values (0.1,0.2 and 0.4) together.

Using this approach, we also propose a pipeline as shown in Fig. 3 for performing analysis using feature selection as an essential step before applying machine learning methods, such as neural networks, support vector machines, Bayesian approaches, etc., in downstream analyses. Figure 3 represents a three-step pipeline, beginning with testing several feature selection methods in simulated datasets in Step 1, as covered in this manuscript. Steps 2 and 3 involve applying the selected methods to a real dataset. We propose to apply top performing methods from Step 1 on our natural biological dataset in Step 2. In Step 2, we select variables based on our collective approach in our real training dataset. Finally, in Step 3, we propose to extract collectively selected variables

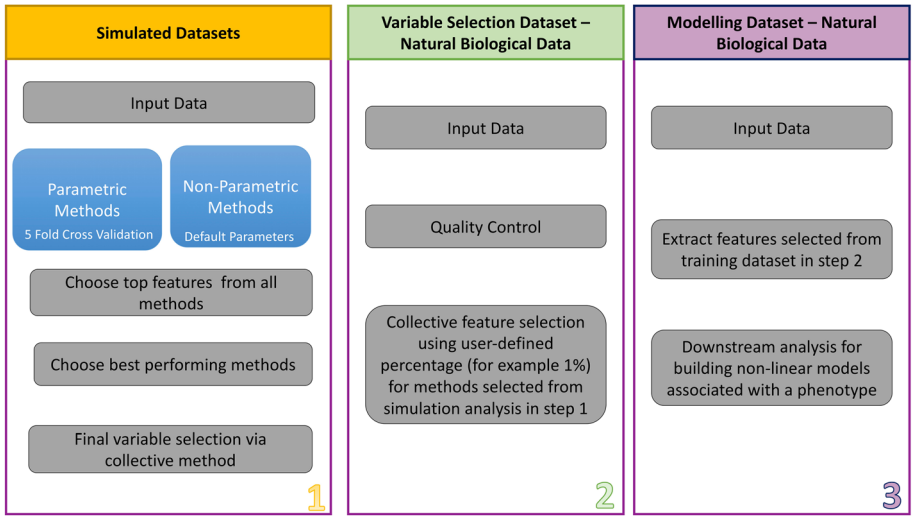

Fig. 3 Pipeline of feature selection procedure and downstream analysis in both simulated and natural biological data

from the variable selection subset of our natural biological dataset to then use for downstream analysis.

Feature selection and downstream analyses

We applied this proposed approach to test for SNPs that are associated with obesity among samples from the MyCode DiscovEHR study [23]. On quality controlled data, we selected features using MDR, MultiSURF* and TuRF, and Ranger collectively, and then performed downstream analyses using Analysis Tool for Heritable and Environmental Network Associations (ATHENA) [9, 34]. We choose to apply Grammatical Evolution Neural Networks (GENN) implemented in ATHENA for this analysis to select non-linear epistatic interactions between SNPs selected from the feature selection strategy described above. Grammatical evolution methods are alternatives to classical genetic programming approaches in machine learning methods. This approach has been widely accepted and its effectiveness has been explained in previous studies [35–38]. We used the following parameter criteria to identify networks associated with BMI case control outcome:

1. Five-fold cross validation

   Modelling data as described in Step 3, which included 14,925 samples and features selected via collective approach, were divided into 5 equal parts.

2. Process

   The first iteration begins with selecting a training set to generate random population (popsize 10,000), dividing into sub-populations, and then preforming an analysis on 30 nodes. The grammar for GENN is then used to evaluate the training set using Area Under Curve (AUC) fitness criteria. This step is then repeated 20 times (numsteps) after which migration takes place to select the

best solution from all 30 nodes. This process is repeated 4 more times, once for each remaining cross-validation fold to perform 5-fold cross validation as explained in step #1.

3. Results

   Training and testing AUC for each network model associated with outcome is reported from all cross validations.

Results

Simulation studies results

Optimizing TuRF iterations

We aimed to use TuRF along with MultiSURF to help increase its efficiency. We tested 3 different thresholds, 1%, 5%, and 10%, to iteratively remove that percent of lowest ranking variables at each iteration. Figure 4 below shows the comparison of results. Here sensitivity is defined as the proportion of true positives selected where a sensitivity of 1 means 100% of true positives were selected from the simulated dataset.

It is interesting to note here that MultiSURF* without TuRF performs better for strong effect models (0.5 and 0.9 penetrance) for both 100 SNPs and 500 SNPs datasets. This could be due to the fact that TuRF works better for larger datasets with many variables whereas 500 variables is still considered relatively "small" and can be handled by MultiSURF alone. In this case, TuRF does not help but instead makes it worse. The poor performance of MultiSURF and TuRF could be explained by the algorithm accidentally discarding the important variable or variables in the first iteration.

Distribution of accuracy from MDR

We ranked all MDR generated models based on their training accuracy to select top user-defined percentages of models as explained in model selection. Figure 5 shows the distribution of median training accuracy for all models that were selected from MDR feature selection. It is to be noted that the accuracies for the selected features vary greatly based on the strength of signal. For example, training accuracies in 0.1 effect signal datasets are close to 55% for all models whereas accuracies for 0.9 effect signal

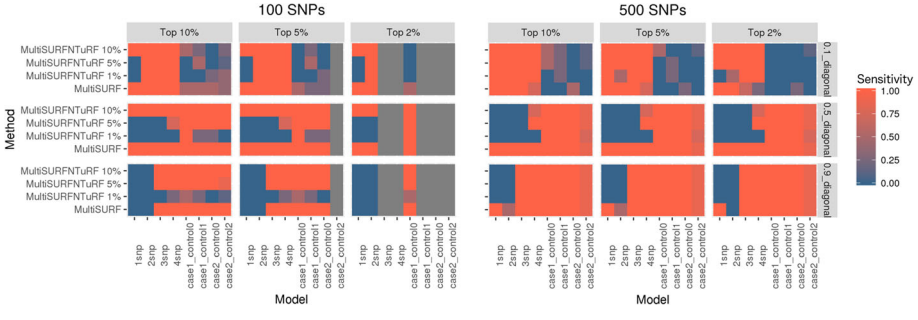

Fig. 4 Comparison of results for TuRF parameters when using it with MultiSURF* as well as MultiSURF* without TuRF implementation. The plot on left is for 100 variables and the right plot is for 500 variables. The x-axis lists main effect and interaction datasets while the y-axis lists all methods tested. These plots are faceted by percentage of top variables selected and the strength of signal. The color gradient refers to the sensitivity (percentage of true positives), ranging blue to orange, or 0 to1

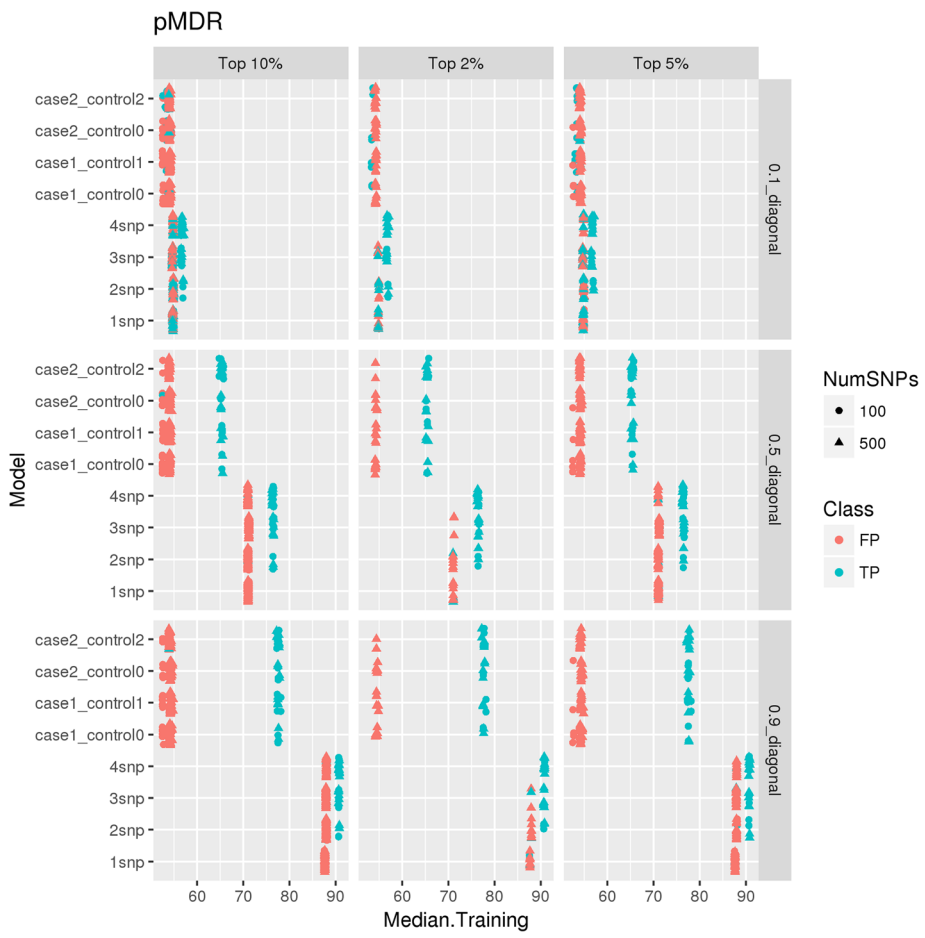

Fig. 5 Distribution of median training accuracy from pMDR analyses. X-axis is median training accuracy values of the model, Y-axis lists all main effect and interaction simulated datasets. These plots are faceted by effect size and top percentage of models selected. 100 SNP data is shown in circles and 500 SNP in triangles. The two colors represent actual true and false positives in results

datasets are closer to 80%. Notably, in many two-way interaction models, we observed that false positives are paired with true positives. Figure 5 represents overall accuracy of the model. Since both false and true positives exist in model, the accuracies reported are also higher for false positive (in red).

Application of all methods on simulated datasets

We tested all chosen methods on the two experiments of simulated datasets with different ranges for effect sizes and various additive main effect and interaction effect models as explained in the data section. To compare results, we are using the degree of effectiveness described as “Sensitivity” where a sensitivity of 1 is equivalent to 100% of true positives being selected in the top features. Figure 6 represents the results for all methods tested using simulated data experiment 1 and Fig. 7 represents results from all methods tested using simulated data experiment 2. It seems evident from these plots that MDR used as a feature selection tool helps to select true positives every time for models tested in simulated data experiment 1 whereas Ranger and Gradient Boosting perform best in terms of selecting true positives for data experiment 2. In the first set

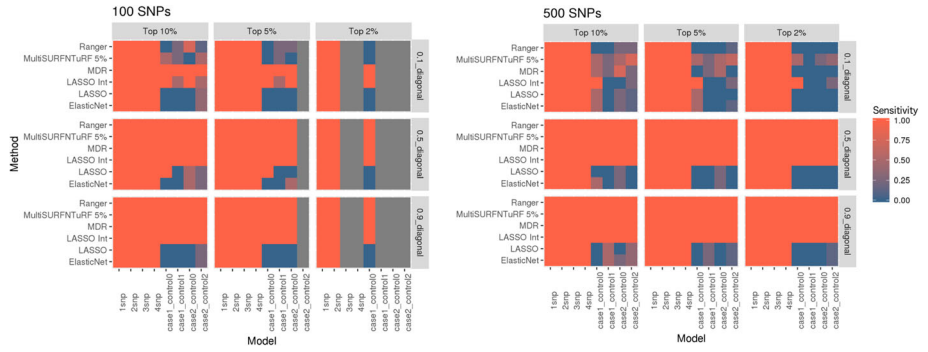

Fig. 6 Comparison of results from all methods tested on simulated dataset 1. These heat maps show the sensitivity of results for all methods (on y-axis) and all simulated models (on x-axis) for both 100 SNPs and 500 SNPs datasets in combination with different effect sizes and selection percentage of top features

of simulations, we see that nonparametric methods do not perform as well in detecting interacting effects. Additionally, we see that most methods do not perform well for the weakest signal tested (0.1 penetrance). MultiSURF and TuRF seem to perform well for interaction effects but do not perform well for main effects. In the second set of simulations, LASSO (without explicitly adding interactions in the model) and elastic net fail to find true positives in both EDM-1 and EDM-2 models, while MDR fails to identify true positives for EDM-2 model. MultiSURF alone and MultiSURF with TuRF both struggle in finding true positives from EDM-2. MDR and ReliefF algorithms work well for EDM-1 model architecture. Lastly, LASSO with interaction models can identify interactions similar to best performing methods in both simulated sets. When we compared these methods for their efficiency (computation burden and memory requirements) as shown in Fig. 8 and Table 4, we observed that the parametric methods (especially LASSO with interactions) take more computation time than most non-parametric methods and the data mining approach. Additionally, LASSO generates an pxp matrix (SNPxSNP) for all exhaustive pairs of SNPs and also requires more memory than other methods to perform computation. Methods like LASSO and Ranger (R package) were also not computationally feasible to run on large genome-wide datasets including over 50,000 SNPs. Thus, a pre-filtration of SNPs based on criteria like LD pruning, MAF filter would be necessary.

We also estimated the time it would take for most of these methods to run when the number of samples and variants are increased. From our analysis, we estimated that MDR scales linearly with number of samples and quadratically with number of features whereas MultiSURF* scales quadratically with number of samples and linearly with number of features. [15]. Thus, the computational burden of MDR increases more when number of samples is increased. Gradient boosting seems to only perform well with larger effects sizes in terms of detecting true positives [16] and fewer variants. For larger numbers of variants and samples, the best way to perform analyses using gradient boosting is to create subsets of SNPs with low intercorrelations and then aggregate results (email conversation with Dr. Gitta Lubke). Since we do not want to make any pre-assumption about the nature of interactions and only test subset of SNPs in a smaller region of genome, we decided to not use Gradient Boosting in such manner.

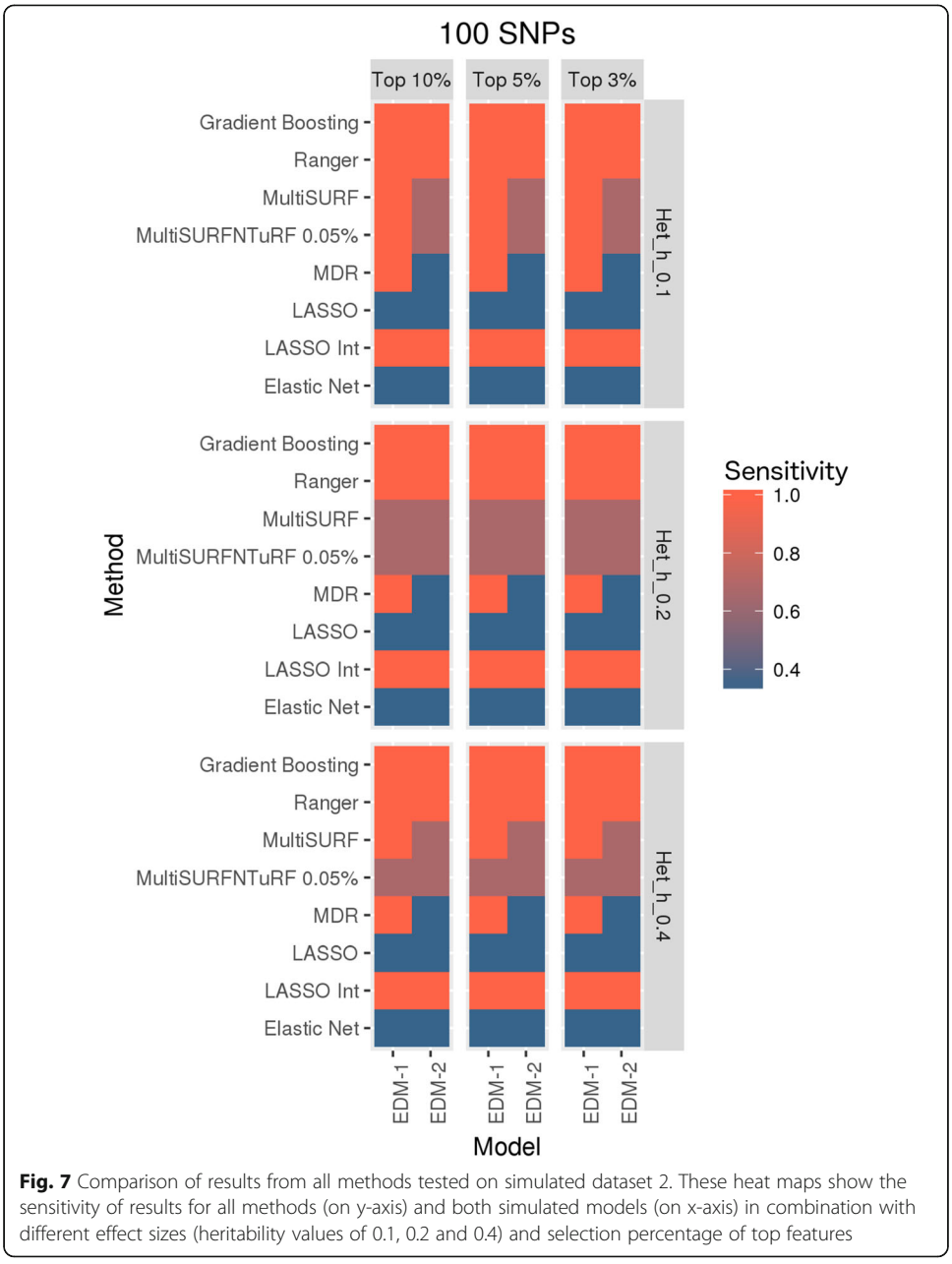

Fig. 7 Comparison of results from all methods tested on simulated dataset 2. These heat maps show the sensitivity of results for all methods (on y-axis) and both simulated models (on x-axis) in combination with different effect sizes (heritability values of 0.1, 0.2 and 0.4) and selection percentage of top features

Collective feature selection on simulated dataset

We applied collective feature selection on simulated experiment data 2 to obtain the number of features that will be selected from top performing methods. Figure 9a and b show the overlap among top features selected from MDR, Ranger, Gradient Boosting, and MultiSURF* and TuRF on EDM-1 and EDM-2 model architectures. Based on information known about merged results from simulated datasets, we expected to obtain 9 true positives (3 from each heritability parameter) in each set of top features selected by every method. However, we again observe that each method does not pick all true positives as shown in 3rd panel of Fig. 9.

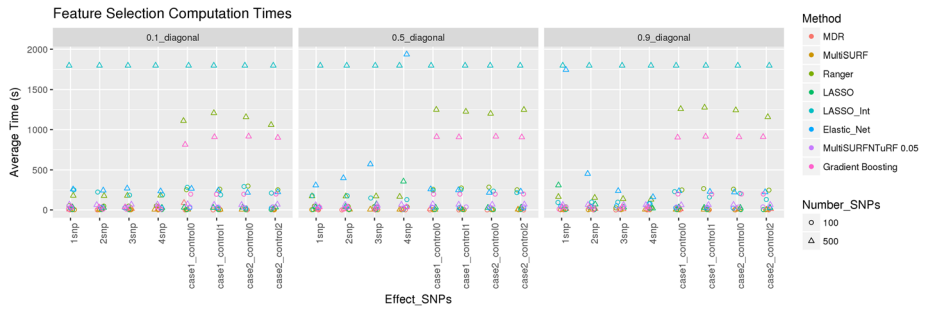

Fig. 8 Plot showing time in seconds (on y-axis) taken for running all feature selection methods. All simulated models are presented in x-axis. Color represents each method. Circles are for 100 SNPs datasets and triangles for 500 SNPs datasets

Therefore, the practice of applying a collective approach seems advantageous. Figure 10 shows the number of features selected in each model by top 3, 5, and 10% model selection criteria. One point to note is that by choosing collective feature selection, we picked all 9 true positives every time whereas by picking one method alone, we risk the chance of picking the "best" method based on one scenario and applying it to a dataset where it is unable to detect all of the true positives.

Biological data application

Collective feature selection

The first step in identifying non-linear models associated with obesity (defined here based on BMI values) is to perform feature selection. We selected 3 methods (MultiSURF and TuRF, MDR and RANGER) for feature selection as described in the methods section. As shown in Table 4, Ranger R package was not computationally feasible (in terms of memory) to run on > 50,000 SNPs; we performed feature selection via random forests by combining Ranger with GenABEL R package to load GWAS data. The computational time for collective feature selection is the combination of the time it took to run each method which is 13 days for Ranger + 1 day for MultiSURF

Table 4 Computational time and memory requirements for all feature selection methods, compared in terms of number of SNPs

Method	Computational time based on number of SNPs (in seconds)				Memory requirements based on number of SNPs			
	100	500	50,000	100,000	100	500	50,000	100,000
LASSO	4.65	58.49	NA	NA	10gb	10gb	NA	NA
LASSO with interactions	186.9	1800	NA	NA	20gb	20gb	NA	NA
Elastic Net	31.44	401.11	NA	NA	10gb	10gb	NA	NA
Ranger	151.31	681.83	NA	NA	8gb	8gb	NA	NA
Gradient Boosting	103.22	466.95	NA	NA	8gb	8gb	NA	NA
MDR	0.25	15.03	6102	89,777	1gb	1gb	10gb	30gb
MultiSURF	2.48	5.13	NA	NA	18gb	39gb	NA	NA
MultiSURF + TuRF 0.05	36.72	65.72	4420	8321	18gb	39gb	28gb	28gb

Note that "NA" here stands for where the model could not be tested due to computational infeasibility while keeping all parameters for simulated datasets same

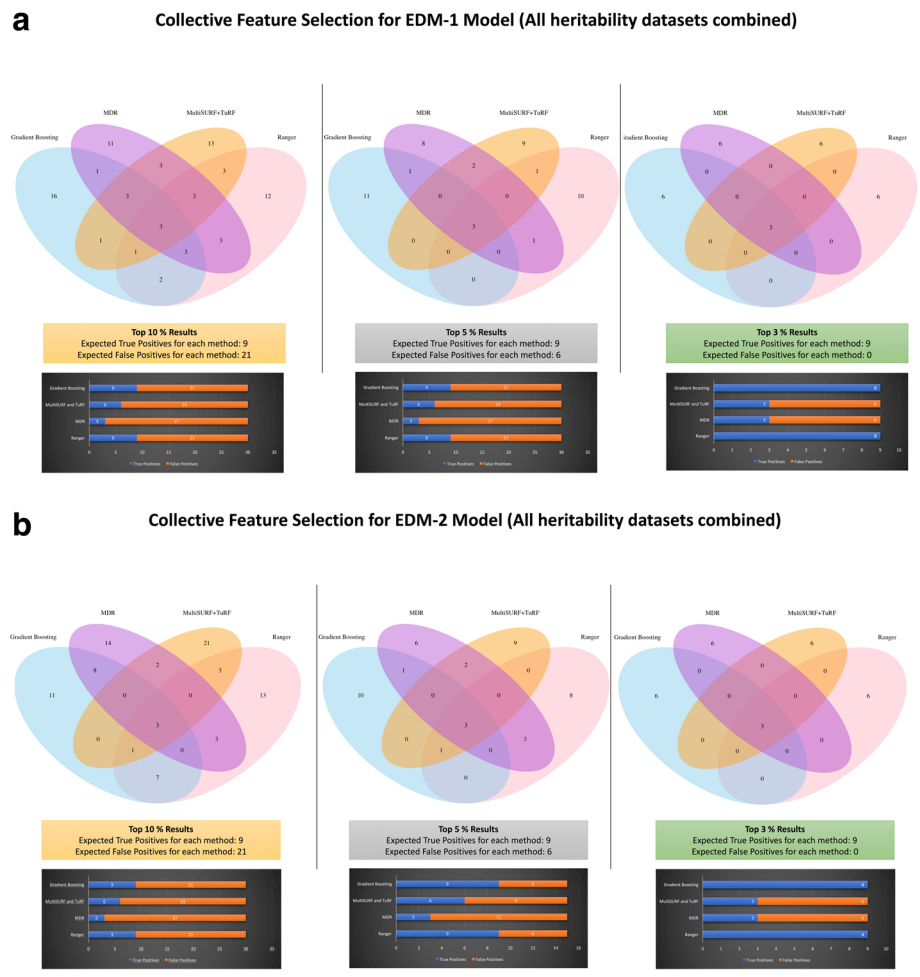

Fig. 9 Venn Diagrams shown here represent the overlap among the top features selected by all methods while bar charts below each Venn diagram show the number of true positives and false positives selected by each method. Plot (**a**) illustrates results for EDM-1 datasets and Plot (**b**) contains results for EDM-2 datasets

and TuRF + 3.5 days for MDR = 17.5 days. Input data consisted of 60,032 SNPs and after feature selection, we selected the top 1% results from each method. This resulted in 1758 variables selected using collective feature selection (note that intersection of methods only selects 2 genes which do not include well known SNPs linked to obesity such as variants in *FTO* and *MYO16)*. The overlap of these variables among the different methods is shown in Fig. 11.

ATHENA results

Five different networks were obtained as a result of applying GENN to identify non-additive interactions associated with BMI outcome. The training and testing area under curve (AUC) for the 5 models are presented in Table 5.

We choose the best network from this analysis, which is shown in Fig. 12. Figure 12 also represents the selection of variants by each feature selection method. In this analysis, we did not adjust for any confounding effects of age, sex, or principal components (PCs) on BMI, but for the variants selected in top models from ATHENA, we ran

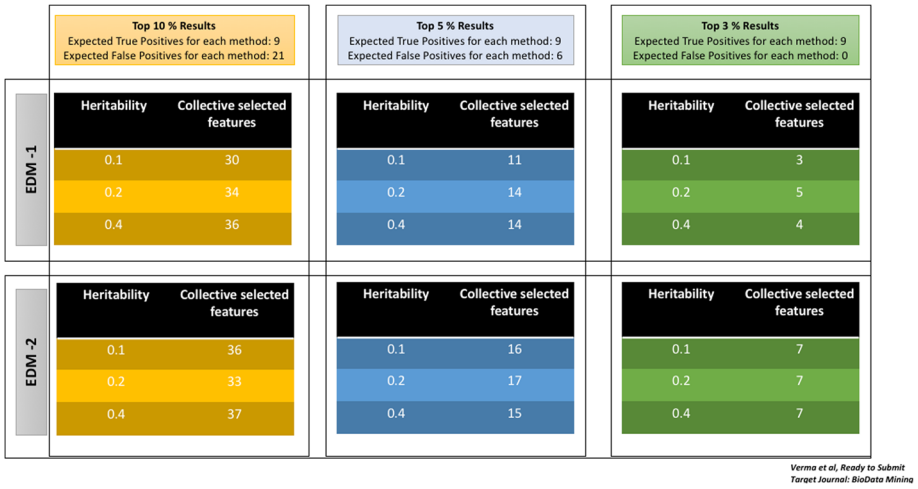

Fig. 10 Number of features selected by collective approach

regression using PLATO [39] to see if the effect sizes and *P*-values for these variants change drastically when BMI, the dependent variable, is adjusted by covariates (age, sex and first 4 PCs). Therefore, to identify if there is significant effect of co-variates on SNPs, we tested these variants by running logistic regression with and without adjusting for covariates. Table 6 lists the *p*-values and betas from regression analyses.

Obesity is a worldwide epidemic and it predisposes to many other metabolic traits and diseases [40]. In our network, we observed a well-known hit for a variant in the *FTO* gene which has been identified by many GWAS analyses. It is to be noted that *FTO* variant was not selected by every feature selection method and similar is the case

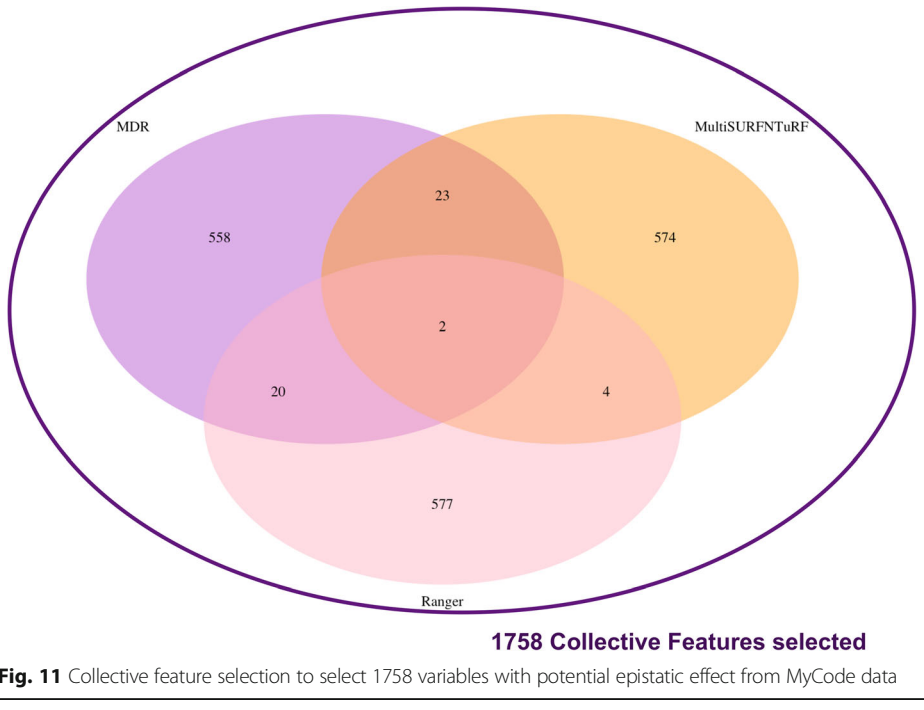

Fig. 11 Collective feature selection to select 1758 variables with potential epistatic effect from MyCode data

Table 5 Training and testing AUC for models selected by ATHENA

Cross Validation	Training AUC	Testing AUC
CV1	0.552115	0.537071
CV2	0.546601	0.540414
CV3	0.543943	0.547598
CV4	0.549398	0.538175
CV5	0.555795	0.541373

for other variants that are reported in Fig. 12. The reported model suggests interaction of *FTO* [40, 41] variants with variants in *TNC*, *MYO16*, and *TTBK1* genes. Notably, these three genes have known associations with other phenotypes influenced by BMI, such as *TNC* with Alzheimer's and schizophrenia [42, 43]. *TTBK1* is also known to be associated with Alzheimer's disease [44–46] while *MYO16* has been found to be associated with pulse pressure [47]. It is also interesting to note that variants in genes *MYO16* and *TNC* are not significant when tested for independent main effect (as reported in Table 6) but they are included in the interaction model as suggested by ATHENA (Fig. 12) which suggests that these variants might work in combination to affect the etiology of obesity but would not be identified otherwise in an additive model.

Discussion

Epistatic features of genes are necessary to consider when investigating the genetic etiology of disease traits. Gene-gene interactions are believed to account for hidden genetic variability [48]. Testing exhaustive pairwise or higher order interactions among all genetic variants poses various challenges including computational burden and correction for multiple hypotheses. Along with these challenges that affect efficiency, it is also important to note that adding more variables to test also reduces the effectiveness of the predictions. Thus, performing feature selection before modelling is

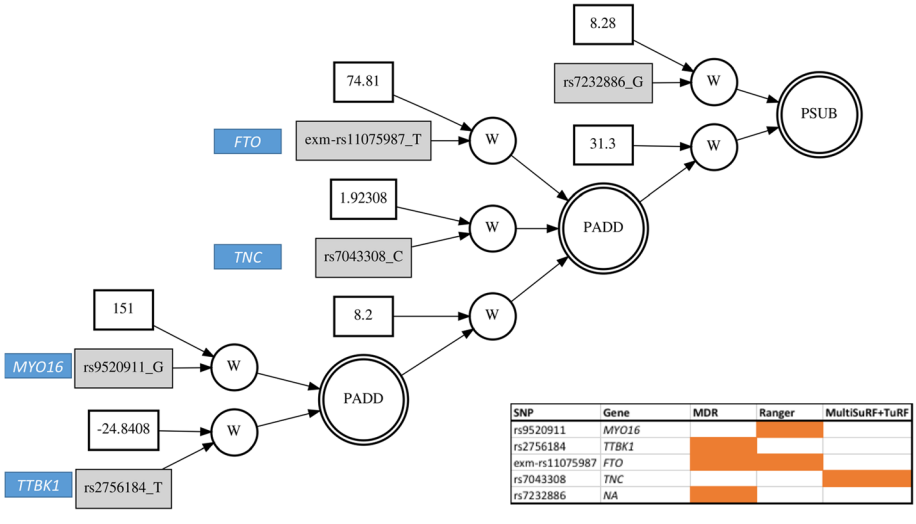

Fig. 12 Best GENN model selected from ATHENA. The SNPs are annotated to gene names. On the bottom right is the list of variants and genes in the model and which feature selection method selected the variant are colored in the table to represent the presence (in orange) and absence (in white) of variant in each feature selection method

Table 6 *P*-values and betas from regression analyses on 5 SNPs in the network selected by ATHENA

SNP	Gene name	No covariates		With covariates	
		p-val	beta	p-val	beta
exm-rs11075987	*FTO*	6.76E-11	0.173939	8.02E-09	0.1690
rs7232886	*N/A*	9.36E-09	−0.15448	6.24E-06	−0.1336
rs2756184	*TTBK1*	0.014104	−0.06635	0.018812	−0.0699
rs9520911	*MYO16*	0.045223	0.053358	0.051185	0.0573
rs7043308	*TNC*	0.768626	− 0.00958	0.745697	− 0.0116

necessary. In our study, we tested parametric, nonparametric, and data mining approaches for feature selection and compared them based on the top models selected as well as the computational time. Through our simulation experiments, we observed that every method is trained to pick variants based on different underlying models that could have potential epistatic effects on disease traits which is reflected by the selection of different false positives from each method on our simulated datasets. Similarly, every method that we tested does not pick all main effect variables every time. This is evident from the non-selection of *FTO* variant by MultiSURF+TuRF and non-selection of variants in genes *MYO16* and *TTBK1* by MDR and Ranger respectively. One possible explanation for selection of different features from different algorithm corresponds to the "no free lunch" theorem [17] and the understanding that no particular feature selection method is specifically designed to pick all epistatic effects. We recommend selecting a user-defined percentage based on combination of sample size, number of variables and trait complexity to obtain the union of features from all methods, referred to here as collective feature selection, to potentially increase power to detect more biologically pertinent associations. It is likely that using a collective approach could result in adding more noise to the analysis, but our analysis suggests that applying different feature selection strategies yield such majorly dissimilar results that the payoff is greater than the cost. In future studies, we aim to test the collective feature selection approach on other natural biological datasets. Our simulation analysis showed that applying non-parametric approaches, like MDR, random forest, gradient boosting and ReliefF results in selecting more true positives epistatic effects in a computationally feasible amount of time than using parametric approaches. But using one method does not always yield all true positives. Thus, we propose collective feature selection utilizing non-parametric methods as a powerful approach for epistatic discovery analysis.

One of the limitations of this study is that we tested our analyses for binary outcome in both simulated and natural datasets. Future work would include the application of these methods to quantitative phenotypes. Additionally, in our simulation analyses we were not able to identify any patterns among the SNPs that were selected across methods. One possible reason could be because of the way that noise is simulated in our dataset, we selected all variants at similar MAF. More studies including simulations of different sets of MAF could also help validate this approach further. In addition, the inclusion of other types of underlying models of epistasis would be useful to further discern which orthogonal or complementary methods perform best in a collective feature selection strategy.

Conclusions

Although our current study is limited in terms of the simulations we performed, they clearly indicate that different methods select varying features depending on the genetic architecture of the trait. Thus, using a collective approach by selecting union of results from different methods rather than selecting an intersection could help preserve features with non-additive effects during feature selection. We applied our approach to select features that were later tested in an independent dataset to identify networks using GENN. Our model was able to select known signals as well as potential interacting effects of known signals with other variants that could be influencing the risk of obesity.

Abbreviations

ATHENA: Analysis Tool for Heritable and Environment Network Analysis; BMI: Body Mass Index; CFS: Collective Feature Selection; GENN: Grammatical Evolution Neural Networks; GWAS: Genome wide Association Study; MAF: Minor Allele Frequency; MDR: Multifactor Dimensionality Reduction; SNP: Single Nucleotide Polymorphism; SURF: Spatially Uniform ReliefF; TuRF: Tuned ReliefF

Acknowledgements

We acknowledge discussions at EDGE conference and inspirations from Dr. Jason Moore and Dr. James Malley.

Funding

This work has been performed using funds from the Pharmacogenomics of Statin Therapy (POST) grant.

Availability of data and materials

Additional information for reproducing the results described in the article is available from authors upon request. Availability of natural biological data from DiscovEHR cohort may subject to data user agreement.

Authors' contributions

SSV, AML and XYZ performed analyses on simulated and real data. Simulated study data was provided by RL and RU. SSV, YV, DK, MDR were involved in designing the study and appropriate pipelines. SD developed software for running ATHENA and parallel MDR analyses. SSV, DK, YV, MDR participated in drafting and finalizing the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Biomedical and Translational Bioinformatics Institute, Geisinger Health System, 100 N Academy Avenue, Danville, PA 17822, USA. [2]Huck Institute of Life Sciences, The Pennsylvania State University, University Park, PA, USA. [3]Institute for Biomedical Informatics, University of Pennsylvania, Perelman School of Medicine, Richards Building, 3700 Hamilton Walk, Philadelphia, PA 19104, USA.

Received: 29 November 2017 Accepted: 4 April 2018
Published online: 19 April 2018

References

1. Clarke B, Chu J-H. Generic feature selection with short fat data. J Indian Soc Agric Stat. 2014;68:145–62.
2. Steen KV. Travelling the world of gene-gene interactions. Brief Bioinform. 2012;13:1–19.
3. Maher B. Personal genomes: the case of the missing heritability. Nature. 2008;456:18–21.
4. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet. 2002;11:2463–8.
5. Sun X, Lu Q, Mukherjee S, Mukheerjee S, Crane PK, Elston R, et al. Analysis pipeline for the epistasis search - statistical versus biological filtering. Front Genet. 2014;5:106.
6. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinforma. 2015;2015:198363.
7. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001;69:138–47.
8. De R, Verma SS, Drenos F, Holzinger ER, Holmes MV, Hall MA, et al. Identifying gene-gene interactions that are highly associated with body mass index using quantitative multifactor dimensionality reduction (QMDR). BioData Min. 2015;8:41.
9. Holzinger ER, Dudek SM, Frase AT, Krauss RM, Medina MW, Ritchie MD. ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. Pac Symp Biocomput. 2013:385–96.

10. Chen S-H, Sun J, Dimitrov L, Turner AR, Adams TS, Meyers DA, et al. A support vector machine approach for detecting gene-gene interaction. Genet Epidemiol. 2008;32:152–67.
11. Li R, Dudek SM, Kim D, Hall MA, Bradford Y, Peissig PL, et al. Identification of genetic interaction networks via an evolutionary algorithm evolved Bayesian network. BioData Min. 2016;9:18.
12. Ghosh D, Chinnaiyan AM. Classification and selection of biomarkers in genomic data using LASSO. J Biomed Biotechnol. 2005;2005:147–54.
13. Zou H, Hastie T. Regularization and Variable Selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67:301–20.
14. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7:3.
15. Greene CS, Penrod NM, Kiralis J, Moore JH. Spatially uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Min. 2009;2:5.
16. Lubke G, Laurin C, Walters R, Eriksson N, Hysi P, Spector T, et al. Gradient Boosting as a SNP Filter: an Evaluation Using Simulated and Hair Morphology Data. J Data Min Genomics Proteomics [Internet]. 2013:4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3882018/
17. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput. 1997;1:67–82.
18. Battogtokh B, Mojirsheibani M, Malley J. The optimal crowd learning machine. BioData Min [Internet]. 2017 [cited 2017 Nov 27];10. Available from: http://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0135-7
19. Wan X, Yang C, Yang Q, Zhao H, Yu W. The complete compositional epistasis detection in genome-wide association studies. BMC Genet. 2013;14:7.
20. Gyenesei A, Moody J, Semple CAM, Haley CS, Wei W-H. High-throughput analysis of epistasis in genome-wide association studies with BiForce. Bioinformatics. 2012;28:1957–64.
21. Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Min [Internet]. 2012 [cited 2017 Nov 27];5. Available from: http://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-5-16
22. Urbanowicz RJ, Kiralis J, Fisher JM, Moore JH. Predicting the difficulty of pure, strict, epistatic models: metrics for simulated model selection. BioData Min [Internet]. 2012 [cited 2017 Nov 27];5. Available from: http://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-5-15
23. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med [Internet]. 2016 [cited 2016 Jun 17]; Available from: http://www.nature.com/gim/journal/vaop/ncurrent/full/gim2015187a.html
24. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016;354
25. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. BMC Bioinformatics. 2007;8:60.
26. Granizo-Mackenzie D, Moore JH. Multiple Threshold Spatially Uniform ReliefF for the Genetic Analysis of Complex Human Diseases. SpringerLink [Internet]. Springer, Berlin, Heidelberg; 2013 [cited 2017 Sep 22]. p. 1–10. Available from: https://link.springer.com/chapter/10.1007/978-3-642-37189-9_1
27. Wright MN, Ziegler A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. ArXiv150804409 Stat [Internet]. 2015 [cited 2017 Sep 22]; Available from: http://arxiv.org/abs/1508.04409
28. Moore JH. Bioinformatics. J Cell Physiol. 2007;213:365–9.
29. Yu W, Lee S, Park T. A unified model based multifactor dimensionality reduction framework for detecting gene-gene interactions. Bioinforma Oxf Engl. 2016;32:i605–10.
30. Lee S, Kwon M-S, Oh JM, Park T. Gene-gene interaction analysis for the survival phenotype based on the cox model. Bioinforma Oxf Engl. 2012;28:i582–8.
31. Yang C-H, Lin Y-D, Yang C-S, Chuang L-Y. An efficiency analysis of high-order combinations of gene–gene interactions using multifactor-dimensionality reduction. BMC Genomics [Internet]. 2015 [cited 2017 Sep 22];16. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4487567/
32. Multifactor dimensionality reduction as a filter-based approach for genome wide association studies. - PubMed - NCBI [Internet]. [cited 2017 Sep 22]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22303374
33. Oki NO, Motsinger-Reif AA. Multifactor dimensionality reduction as a filter-based approach for genome wide association studies. Front Genet. 2011;2:80.
34. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. ATHENA: the analysis tool for heritable and environmental network associations. Bioinforma Oxf Engl. 2014;30:698–705.
35. Kim D, Li R, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. Knowledge-driven genomic interactions: an application in ovarian cancer. BioData Min. 2014;7:20.
36. Turner SD, Dudek SM, Ritchie MD. ATHENA: a knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait loci. BioData Min. 2010;3:5.
37. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol. 2008;32:325–40.
38. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics. 2003;4:28.
39. Hall MA, Wallace J, Lucas A, Kim D, Basile AO, Verma SS, et al. PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies. Nat Commun [Internet]. 2017 [cited 2017 Nov 3];8. Available from: http://www.nature.com/articles/s41467-017-00802-2
40. Cronin RM, Field JR, Bradford Y, Shaffer CM, Carroll RJ, Mosley JD, et al. Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. Appl Genet Epidemiol. 2014;5:250.

41. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197–206.
42. Chan MK, Krebs M-O, Cox D, Guest PC, Yolken RH, Rahmoune H, et al. Development of a blood-based molecular biomarker test for identification of schizophrenia before disease onset. Transl Psychiatry. 2015;5:e601.
43. Mi Z, Halfter W, Abrahamson EE, Klunk WE, Mathis CA, Mufson EJ, et al. Tenascin-C is associated with cored amyloid-β plaques in Alzheimer disease and pathology burdened cognitively normal elderly. J Neuropathol Exp Neurol. 2016;75:868–76.
44. Lund H, Cowburn RF, Gustafsson E, Strömberg K, Svensson A, Dahllund L, et al. Tau-tubulin kinase 1 expression, phosphorylation and co-localization with phospho-Ser422 tau in the Alzheimer's disease brain. Brain Pathol Zurich Switz. 2013;23:378–89.
45. Yu N-N, Yu J-T, Xiao J-T, Zhang H-W, Lu R-C, Jiang H, et al. Tau-tubulin kinase-1 gene variants are associated with Alzheimer's disease in Han Chinese. Neurosci Lett. 2011;491:83–6.
46. Vázquez-Higuera JL, Martínez-García A, Sánchez-Juan P, Rodríguez-Rodríguez E, Mateo I, Pozueta A, et al. Genetic variations in tau-tubulin kinase-1 are linked to Alzheimer's disease in a Spanish case-control cohort. Neurobiol Aging. 2011;32:550.e5–9.
47. Basson J, Sung YJ, Schwander K, Kume R, Simino J, de las Fuentes L, et al. Gene–education interactions identify novel blood pressure loci in the Framingham heart study. Am J Hypertens. 2014;27:431–44.
48. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009;10:392–404.