bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Powerful tests for multi-marker association analysis using ensemble learning

Badri Padhukasahasram[1]

[1]Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, Michigan, USA, 48202

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Multi-marker approaches are currently gaining a lot of interest in genome wide association studies and can enhance power to detect new associations under certain conditions. Gene and pathway based association tests are increasingly being viewed as useful complements to the more widely used single marker association analysis which have successfully uncovered numerous disease variants. A major drawback of single-marker based methods is that they do not consider pairwise and higher-order interactions between variants. Here, we describe multi-variate methods for gene and pathway based association analyses using phenotype predictions based on machine learning algorithms. Instead of utilizing only a linear or logistic regression model, we propose the use of ensembles of diverse machine learning algorithms for testing multi-variate associations. As the true mathematical relationship between a phenotype and any group of genetic and clinical variables is unknown in advance and may be complex, such a strategy gives us a general and flexible framework to approximate this relationship across different sets of SNPs. We show how phenotype prediction based on our method can be used for constructing tests for SNP set association analysis. We first apply our method to simulated datasets to demonstrate its power and correctness. Then, we apply our method to previously studied asthma-related genes in 2 independent asthma cohorts to conduct association tests.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Genome wide association studies (GWAS) have generated a wealth of information about genes and genetic variants influencing various diseases and traits. The vast majority of GWAS have focused on single-marker analysis and tests for significance were corrected for multiple hypotheses testing to obtain the correct false positive rates. Because the number of markers tested in such studies is large, a SNP needs to have strong effects or the sample size needs to be large enough to cross the stringent genome wide significance thresholds. Furthermore, many complex traits are thought to result from the interplay of multiple genetic and environmental factors, which are not captured by single SNP association tests. Given these limitations of single-marker analysis, many multi-marker approaches for association testing have been proposed and are increasingly being used to complement single SNP analysis.[1-10]

As genes are the basic functional unit of the genome and since genes rarely work in isolation, multi-marker association tests appear to account for the multiplicity that occurs biologically. Therefore, while individual causal variants might show only a marginal signal of association, jointly utilizing all informative SNPs within a gene or a pathway may detect their manifold effects Testing genes and pathways also reduces the burden of multiple testing from millions of individual SNP tests to ~20,000 genes and even fewer pathways. Multi-marker methods may also be less sensitive to differences in allele frequency and linkage disequilibrium between population groups (and, therefore, may produce more replicable results).

To date several gene-based association tests have been proposed.[1;2;4;8-10] Most of these approaches first assign a subset of SNPs to a particular gene based on their location in the

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genome; they then seek to calculate a gene-based *p* value based on the individual SNP association tests. VEGAS is a versatile, gene-based method that combines the chi-square test statistics of individuals SNPs (while accounting for their dependence) to compute gene-level significance.[2] GATES is another gene-based test that uses an extended Simes procedure to integrate the *p* values of individual variants while accounting for pairwise correlations between variants when calculating the effective number of tests.[1] SKAT is a logistic kernel machine based test that can account for non-linear effects when determining the gene-level significance.[9] The methods used for combining *p* values in gene-based tests can be divided into 2 broad categories: best-SNP picking and all SNP aggregating tests. Best-SNP picking tests use only one SNP-based *p* value after accounting for multiple testing adjustment. GATES is an example of such a test. All-SNP aggregating tests like VEGAS-SUM and SKAT attempt to accumulate the effects of all SNPs into a test when determining the overall *p* value. HYST is a recently developed hybrid method that use both these kinds of approaches in its calculations.[8] The initial pathway-based approaches[3] for analyzing GWAS data were developed by adapting ideas from the microarray field where similar methods have been developed for gene expression data.[11] In pathway-based analysis, researchers examine SNPs within predefined sets of genes based on prior biological knowledge or computer predictions. In the recent years, a range of pathway-based analytic methods have been developed.[3;7;8;12-15] These can be broadly classified into 2 types based on whether they utilize SNP-based *p* values or individual-level genotype information to determine significance.

There is considerable room for further improvement in existing gene- and pathway-based methods. For example, many existing approaches use the minimum of the *p* values for variants

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

within a gene to determine the gene-level *p* values. However, this may not be optimal in terms of utilizing the available information and it may be better to determine the joint association of multiple predictive SNPs rather than use individual SNP *p* values. Similarly, multi-marker association tests that jointly utilize all informative genetic variants from a pathway may offer a more powerful test when compared with combining gene-based *p* values. In addition, many existing methods do not account for nonlinear and epistatic effects.

Our main goal here is to develop an accurate method for multi-marker association analysis that can incorporate pairwise and higher order interactions. We use phenotype prediction algorithms as a basis for constructing such association tests. Since both the underlying genetic architecture of a trait and the optimal model structure to use for combining the association information across multiple SNPs are not usually known before testing, we propose a machine learning approach for this purpose. The main novelty of our approach is the use of ensembles of diverse learning models to generate phenotype predictions. In this approach, we feed the initial predictions generated from many individual learning algorithms into a second-level learning algorithm which weights their contributions suitably to generate a final prediction.[16-19] Thus, our approach involves blending the results of different learning algorithms by using a "meta-level" learning algorithm. We also use additional variables called "meta-features" (e.g., age, sex, body mass index, SNPs etc) as inputs to guide this blending procedure.[18] In principle, such a combination of models can allow us to better approximate the true underlying relationships between the input variables and phenotype across multiple sets of SNPs. Note that this relationship can be non-linear, complex and variable in nature across different SNP sets.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here, we show how machine learning algorithms can be used to construct powerful tests for multi-marker association analysis. We then show how to construct tests of association in the presence of non-genetic covariates and how to construct a multi-marker test for interactions under this framework. We first apply our method to simulated datasets to demonstrate its power and correctness. Lastly, we apply our method to previously studied asthma-related genes in 2 independent asthma cohorts to conduct gene-based association tests.

METHODS

Approach for predicting phenotypes

Here, we present an overview of our approach to predict phenotypes from genetic and clinical variables through the use of multiple machine learning algorithms. First, we created a list of all genetic variants and clinical covariates that can potentially influence the phenotype of interest. Next, we perform a feature selection step where we identify a subset of variables which are useful for building a predictive model. This can be done in many ways such as using variable importance scores from a random forest algorithm or Pearson's correlation coefficient with the phenotype. Different machine learning algorithms (e.g., random forests, support vector regression, multiple regression, artificial neural networks, and boosted regression trees) are then trained using this subset of informative variables. Subsequently, we use the predictions from these individual models along with the selected features as inputs in a "meta-level" random forest algorithm. Lastly, we assess prediction accuracy by testing the model on an "outside the training set" and through 5-fold cross-validation.

Ensemble learning algorithm for phenotype prediction

Ensemble learning variation 1:

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. *Generate a set of all genetic variables.*
2. *Perform feature selection on the training data in order to identify an informative subset of variables ($f_1, f_2 \ldots f_n$) for phenotype prediction. This can be performed using either pairwise correlation coefficients between variables and phenotype or by using random forest variable importance scores to rank the variables. Then, we can use the top 10%-30% of the variables in a prediction model.*
3. *Train k independent machine learning approaches on the training data using the selected features and generate model predictions $P_1, P_2 \ldots P_k$.*
4. *Use the predictions from step 3, $P_1, P_2 \ldots P_k$ and $f_1, f_2 \ldots f_n$ as inputs and train a "meta-level" learning algorithm using random forests. Note that this is a key step in the algorithm and generates a final prediction by blending many individual predictions in a possibly nonlinear manner. The main goal is to learn the best model to combine individual models from the training data so that we can predict the phenotype as well as possible. The non-linear combination of models along with the meta-features gives us a more general predictive framework which can accommodate different model structures and also allows the overall model to vary across the multi-dimensional parameter space.*
5. *Generate predictions in test data P_{blend1} using the models trained in steps 3 and then 4. Repeat for all cross-validation folds to obtain unbiased phenotype predictions for all samples.*

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Generalization: An ensemble of ensembles

Generalizations of the algorithm described previously are also possible that can potentially further boost the prediction accuracy. In particular, the creation of an ensemble of models (steps 3 and 4 in previous algorithm) can be done in a variety of different ways. For example:

Ensemble learning variation 2: Combining of predictions from individual learning models can be done sequentially using predictions from all previous steps as inputs in the next step i.e. instead of 3 and 4 we can:

i) Train learning algorithm 1 on the training data using the selected features $f_1, f_2 \ldots f_n$ *as inputs and generate model predictions* P_1.

ii) Train learning algorithm 2 on the training data using P_1 and the selected features $f_1, f_2 \ldots f_n$ *as inputs and generate model predictions* P_2.

iii) Training learning algorithm 3 on the training data using P_1, P_2 *and the selected features* $f_1, f_2 \ldots f_n$ *as inputs and generate model predictions* P_3.

iv)……

v)…….

vi) Training learning algorithm k on the training data using $P_1, P_2, \ldots P_{k-1}$ *and the selected features* $f_1, f_2 \ldots f_n$ *as inputs and generate model predictions* P_k.

Note that each algorithm after i) is a meta-level learning algorithm. Then, we generate predictions in test data P_{blend2} *using the models as in training and repeat for all cross-validation folds to obtain unbiased phenotype predictions for all samples.*

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ensemble learning variation 3: Instead of applying an ensemble learning model (variation 1) to all the samples, we can divide the high-dimensional parameter space of variables into different subsets. Then, we can train different ensemble learning models using only samples that fall in these different subsets and finally merge these models to obtain the overall prediction model. Then, we can generate final predictions P_{blend3} *in test data as we did for training data for all cross-validation folds within all subsets to obtain unbiased phenotype predictions for entire sample.*

Lastly, we can train a final learning algorithm that uses P_{blend1}, P_{blend2} *and* P_{blend3} *as inputs to generate the final prediction* P_{final}.

Multi-marker tests of association

Once we have estimated a model using any of the algorithms described in the previous section and predicted phenotypes, we can construct tests of association in the following manner. For continuous traits, we can calculate the Pearson's correlation coefficient between predicted (P_{final}) and observed (P_{actual}) values and obtaining the corresponding *p* values. For case-control studies, we perform a logistic regression using all the genetic variables (SNPs) and P_{final} as explanatory variables. A chi square based likelihood ratio test can then be used to generate *p* values.

Testing multi-marker associations in the presence of covariates

Association testing in the presence of covariates (e.g., age, gender, BMI and smoking status) can be done in the following manner. First, consider both non-genetic covariates and genetic variables together for phenotype prediction according to any of the ensemble learning algorithms described earlier. Let $P_{final\text{-}all}$ be the predicted phenotype values. Then, remove the SNP variables

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and rerun the phenotype prediction algorithm. Let $P_{final\text{-}covariates}$ be the predicted phenotype values. For continuous traits, we first calculate the Pearson's correlation coefficient for both these predicted variables with the true phenotypes (P_{actual}). The strength of association for the genetic variables can then be calculated using the Steiger's Z test for the difference between the 2 calculated correlation coefficients. Let r_{12} and r_{13} denote the Pearson's correlations between the true phenotype (P_{actual}) and $P_{final\text{-}covariates}$, $P_{final\text{-}all}$ respectively. Let r_{23} denote the Pearson's correlation between $P_{final\text{-}covariates}$ and $P_{final\text{-}all}$. The Steiger's test computes p values based on the following test statistic that is assumed to be standard normally distributed:

$$Z = (Z_{12} - Z_{13})\sqrt{N-3}/\sqrt{(2h - 2hr_{23})}$$

Here, Z_{12} and Z_{13} are Fisher's transformations of r_{12} and r_{13}, and

$h = (1 - fr_m^2) / (1 - r_m^2)$ where $f = (1 - r_{23})/(2 - 2r_m^2)$ and $r_m^2 = (r_{12}^2 + r_{13}^2)/2$.

For case-control studies we can use both non-genetic covariates and genetic variables as well as $P_{final\text{-}all}$, $P_{final\text{-}covariates}$ as explanatory variables in a logistic regression model and use a chi square based likelihood ratio test with a model without any genetic variables (i.e. non-genetic covariates, $P_{final\text{-}covariates}$ only) to calculate the p value.

Multi-marker tests for interactions

We can test for interactions between a set of markers in the following manner. First, consider all of the SNPs together in a linear or logistic regression model (for continuous or case-control phenotype) and generate phenotype predictions using cross-validation for all individuals. Let P_{linear} be the predicted phenotype values. Then, generate phenotype predictions for all individuals using any of the ensemble learning algorithms described previously. Let $P_{ensemble}$ denote the

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

predicted phenotype values. For continuous traits, we will use all markers as well as $P_{ensemble}$ and P_{linear} as explanatory variables in a multiple regression model (Model 1) and perform a F test with a model (Model 0) without interactions (i.e. one with all markers and P_{linear} only) to calculate the *p* value. We compare the sum of the squared errors (SSE) of prediction to construct an F statistic with $(1, N - V_{Model1} - 1)$ degrees of freedom. Here:

$F = [SSE_{Model0} - SSE_{Model1}][N - V_{Model1} - 1]/SSE_{Model1}$. N denotes the number of samples and V_{Model1} denotes the total number of explanatory variables in model 1. For case-control studies, we will use all markers as well as $P_{ensemble}$ and P_{linear} as explanatory variables in a logistic regression model and use a chi square based likelihood ratio test with a model without interactions (i.e. one with all markers and P_{linear} only) to calculate the *p* value.

Power and Type 1 Error rates of gene-based association test for data simulated under multiplicative and additive models

We tested the performance of the proposed gene-based test by simulating genotype data for 30 biallelic SNPs assuming Hardy Weinberg equilibrium. We assumed 3 different scenarios of linkage disequilibrium (LD) structures for the 30 SNPs: i) SNPs are within blocks with high LD (r = 0.9 or 0.8 within blocks) ii) SNPs are within blocks in moderate LD (r = 0.5 or 0.4) iii) SNPs are completely independent of one another and in linkage equilibrium. The choice of simulation settings are similar to what was used previously for comparisons in [(1)]. For each LD scenario, we considered 3 different gene sizes with the first 3, first 10 and all 30 SNPs with 1, 2 and 6 disease SNPs respectively. For each gene size, we tested 3 models i) a null model with no disease loci. ii) additive model where one SNP in each LD block has a minor allele that increases the risk additively by 0.14. iii) multiplicative model where one SNP in each LD block has a minor allele

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that increases the risk by a factor of 1.14. The baseline risk for individuals with non-risk alleles was calculated using risk ratios and allele frequencies and the population disease prevalence was 0.1. We used a sample of 1,500 cases and 1,500 controls drawn from a simulated population of 100,000 individuals for each scenario. For more details about these LD patterns, please refer to [(1)]. Type1 error rates and statistical power was obtained as the fraction of 1,000 and 500 simulated case-control datasets respectively, for which the gene-based association test generated significant *p* values (i.e. $p <= 0.05$).

Power and Type 1 Error rates of gene-based test for models with interactions

The simulations in the previous section assumed that the effect of various disease susceptibility SNPs are independent of one another and they increase the risk additively or multiplicatively. To explore the effect of pairwise and higher order interactions between genetic variants, we also compared the performance of methods for data simulated under models with interactions. We simulated a quantitative trait for many different models with one or more interactions among variants in addition to main effects. In addition, we also considered scenarios where there is pure epistasis i.e. where the effect of a group of SNPs is simply due to their interactions and there are no main effects. We simulated samples of 3000 individuals and genes with 5 or 10 SNPs assuming linkage equilibrium. The phenotype was drawn from a complex distribution involving a sum of a standard normal variable and products of SNPs. Power and Type 1 Error rates were estimated based on 100 and 500 simulated datasets respectively. We calculated the fraction of simulated datasets for which the gene-based method generated a significant *p* value ($p <= 0.05$) and compared it with a gene-based test with linear regression as well as with GATES[(1)]. For a gene-based test with linear regression, *p* values were obtained by using an F test statistic.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Power and Type 1 Error rates for multi-marker test for interactions

For all the models simulated in the previous section, we also constructed a multi-marker test for interactions as described previously and estimated the power of such a test. We simulated samples of 3000 individuals and genes with 5 or 10 SNPs assuming linkage equilibrium. The phenotype was drawn from a complex distribution involving a sum of a standard normal variable and interactions terms involving SNPs. Power and Type 1 Error rates were estimated based on 1000 simulated datasets. For each model with interactions, we calculated the fraction of simulated datasets for which the multi-marker test of interactions generated a significant p value ($p <= 0.05$). p values were based on an F test statistic with two parameters as described previously.

Datasets

We apply the methods developed in this paper to 2 different datasets from independent studies. These datasets are briefly described below:

i) The Study for Asthma Phenotypes and Pharmacogenomic Interactions by Race-ethnicity (SAPPHIRE) is an ongoing NIH-funded project that seeks to identify the genomic determinants of asthma controller medication response in a population based sample of asthmatic individuals. In particular, this cohort includes individuals with asthma who visit the Henry Ford Health System (HFHS) and Henry Ford Medical Group (HFMG), who consent to participation, and who undergo a detailed enrollment evaluation. The health system serves the primary and specialty medical needs of people in southeastern Michigan, including Detroit and its surrounding metropolitan area. The enlisted SAPPHIRE patients meet the following criteria: age 12-56 years, a prior clinical diagnosis of asthma, and no recorded diagnosis of chronic obstructive pulmonary

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

disease or congestive heart failure. In this cohort, we had genome wide data from 586,952 SNPs for testing associations in the primary GWAS data in a sample of 1,401 African American individuals. This includes 1,073 asthma cases and 328 healthy controls (For more details about dataset and quality control refer to [20]).

ii) GALA (Gene-environment studies of asthma in Hispanic/Latino children) II study is a case control study in a cohort of Latino/Hispanic children between 8-22 years in age which aims to i) assess interactions between ancestry, environment and asthma ii) examine candidate gene-environment interactions with asthma and related phenotypes and iii) Determine whether migration and acculturation are associated with asthma and severe asthma. We have genome wide genotype data from 747,075 markers and various clinical covariates for a collection of 3,772 individuals from different regions of the United States. This dataset includes 1,891 asthma cases and 1,881 controls.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

Multiplicative and Additive models-Comparisons

Tables 1-3 shows comparisons for the performance of various methods for disease case-control datasets simulated under additive and multiplicative models. We can see that the performance of the newly proposed method based on an ensemble of machine learning algorithms is competitive with other approaches and the Type 1 error rates produced by all methods are close to expectations. Power for many existing approaches are similar to one another for the parameters investigated here and for the machine learning and logistic regression methods, power is not sensitive to the strength of linkage disequilibrium.

Models with epistatic effects

In Table 4, we compare the power of our approach for models with pairwise and higher-order interactions between SNP variants using a simulated quantitative trait. We compare the ensemble learning approach with a gene-based test constructed using multiple linear regression as well as with the extended Simes procedure as implemented by GATES. In all situations, our simulations indicate that the machine learning approach which can model interactive effects is uniformly more powerful than the other 2 approaches. Table 4 also shows that the estimated gain in power can be substantial. Among the other 2 methods, multiple linear regression performed second best while the GATES method which only integrates the *p* values from single marker tests had the lowest power. In Table 5, we show the Power and Type 1 Error rates for a multi-marker test for interactions for the same models as in Table 4. These results clearly demonstrate the ability of our approach to detect the presence of interactions by considering the difference between ensemble learning and linear model based predictions.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Application to real datasets

Lastly, we applied the proposed gene-based association test to an empirical dataset consist of 328 healthy non-asthmatic individuals and 1,073 individuals with asthma. These individuals are of African American descent and are part of the Study of Pharmacogenomic Interactions by Race Ethnicity (SAPPHIRE) cohort. We tested 9 previously studied asthma-related genes[21-23] in this cohort, to see if these are also associated with asthma status in the African American population. When constructing this gene-based test, we adjusted for age, gender and principal components 1-10 as covariates.

In addition to the SAPPHIRE cohort, we also applied the gene-based association tests to the same set of 9 genes in 3,772 Hispanic/Latino children (1891 cases and 1881 controls) from the GALAII study. Again, we adjusted for age, gender and principal components 1-10 as covariates when constructing a gene-based association test. Tables 6 and 7 show the results of our ensemble learning gene-based association tests in the African American and Latino groups respectively. We also show comparison of results with the GATES and logistic regression methods. At a Bonferroni adjusted significance threshold of 0.0027 (= 0.05/18), we can see that the ensemble learning gene-based test finds more hits than the other 2 approaches.

.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

We have introduced a new method for assessing the significance of association of a set of SNPs with a particular phenotype. This method uses diverse machine learning algorithms to construct predictive models for the phenotype using SNPs from a gene or a pathway, and, subsequently uses such predictions to construct tests of association. Machine learning algorithms represent powerful tools for inferring the relationship between multiple variables and a response variable of interest and can account for complicated interactions between the predictor variables. Although the use of machine learning for prediction is not a new idea, the construction of ensembles using diverse machine learning approaches yields a more general predictive model which can accommodate different kinds of model structures. Because the "true" multi-variate relationship between a set of variables and response is not known in advance and may be variable, such ensembles of models allow us to better approximate this relationship across different sets of SNPs by learning from the data. The use of ensemble-based ML predictions leads to novel multi-marker tests of association. We expect these tests to be useful for gene- and pathway-based association analysis. The method can applied to any arbitrary set of SNP variables (e.g. from a region of interest or SNPs from a functional class) and for admixed populations (e.g. African Americans, Latinos etc) it should be straightforward to use both SNP variables and local ancestry variables as inputs to construct similar tests.

There are 3 key advantages of using our gene-based approach compared to existing approaches. The first is our method is flexible does not have to assume a particular genetic effects model (i.e. additive, recessive, dominant etc) for a SNP. When constructing our tests, we can include 3 variables for each SNP where the variants are encoded according to these 3 models (i.e. additive,

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

recessive, dominant etc). Thus, we can test genes/pathways/SNP-sets where the genetic effects are heterogeneous in nature across SNPs. Another advantage is the ability to include any number of covariates (note that covariates can also be other SNPs from related genes) in the gene-based association tests and accounting for the interactions between them. This allows us to model higher-level multi-variable interactions (e.g. SNP x SNP x COVARIATE, SNP x COVARIATE x COVARIATE), which are not considered using single marker and other gene-based association tests. Lastly, creating an ensemble of diverse multivariate models as well as the incorporation of meta-features, makes our method less restrictive than other methods allowing us to approximate a wider range of models accurately. All these novel aspects can boost the power of the method and may allow us to discover new genetic associations missed by existing approaches.

Extensions of these methods towards the case of multiple correlated phenotypes should also be straightforward. If instead of a single phenotype, we are interested in many phenotypes that are correlated with one another in some manner, we can construct a joint association test for all of them in the following manner. First, we will apply the ensemble learning based gene-based association test to each phenotype individually and obtain their corresponding p values. Subsequently, we can obtain an overall *p* value from these individual *p* values using the TATES multi-trait association method[24], which is analogous to the extended Simes procedure of GATES developed for testing multi-marker associations.

We applied our method to both simulated and empirical datasets to demonstrate its power and utility. For models without interactions between variables, the ensemble learning approach works as well as many currently existing methods available for testing gene-based association tests. In

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

particular, power is close to a likelihood ratio test based on a logistic regression model alone. In contrast, for models dominated by interactions, the ensemble learning approach can be considerably more powerful than the alternate approaches considered here. Thus, for genes or pathways or phenotypes where epistatic effects are important, our approach is more likely to detect associations than other approaches that don't consider such effects. By using a collection of diverse multi-variate models with interactions, the method developed here can complement the existing set of multi-marker association tests and can find novel correlations in existing GWAS datasets that are not visible through alternate approaches.

Computation can be a possible limiting factor when applying ensemble learning algorithm based associations tests to thousands of genes in genome wide datasets. For genome wide data, we suggest using a multi-stage approach to obtain results within a reasonable time. We can start by testing with a computationally efficient method like GATES to identify a smaller subset of likely candidate genes (e.g. top 100 or genes with *p* values below a threshold). Then, the machine learning based multi-marker association test (and generalizations described before) can then be applied to these high priority genes to obtained highly refined gene-based *p* values. We note that it should be trivial to split large datasets into gene regions and parallelize such tests when many computer nodes are available.

In summary, ensemble learning algorithms provide a general and flexible framework for conducting association analysis. We have shown that phenotype predictions made by such algorithms can be used for many common tasks encountered in association analysis such as testing of multi-marker associations, adjusting for multiple non-genetic (and possibly genetic) covariates and testing for interactions at a gene-level. Because machine learning is a highly

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

developed area of study, prediction of response from many input variables is a well-studied problem and numerous well-established algorithms are already available which can be readily incorporated as components in an ensemble learning framework to maximize prediction accuracy and construct powerful tests of association.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

(1) Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet 2011; 88(3):283-93.

(2) Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet 2010; 87(1):139-45.

(3) Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genome wide association studies. Am J Hum Genet 2007; 81(6):1278-83.

(4) Li M, Wang K, Grant SF, Hakonarson H, and Li C. ATOM: a powerful gene-based association test by combining optimally weighted markers. Bioinformatics 2009; 25(4):497-503.

(5) Huang H, Chanda P, Alonso A, Bader JS, Arking DE. Gene-based tests of association. Plos Genetics 2011; 7(7):e1002177.

(6) Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP et al. Pathway analysis using random forests classification and regression. Bioinformatics 2006; 22(16):2028-36.

(7) Chung RH, Chen YE. A two-stage random forest-based pathway analysis method. PLoS One 2012; 7(5):e36662.

(8) Li MX, Kwan JS, Sham PC. HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. Am J Hum Genet 2012; 91(3):478-88.

(9) Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 2010; 86(6):929-42.

(10) Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011; 89(1):82-93.

(11) Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005; 102(43):15545-50.

(12) Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N et al. Pathway analysis by adaptive combination of P-values. Genet Epidemiol 2009; 33(8):700-9.

(13) Guo YF, Li J, Chen Y, Zhang LS, Deng HW. A new permutation strategy of pathway-based approach for genome-wide association study. BMC Bioinformatics 2009; 10:429.

(14) Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M. Genome-wide gene and pathway analysis. Eur J Hum Genet 2010; 18(9):1045-53.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(15) Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U et al. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet 2010; 86(6):860-71.

(16) Bell RM, Koren Y, Volinksky C. The BellKor solution to the Netflix Prize. 2007.
Ref Type: Internet Communication

(17) Toscher A, Jahrer M, Bell RM. The BigChaos Solution to the Netflix Grand Prize. 2009.
Ref Type: Internet Communication

(18) Sill J, Takacs G, Mackey L, Lin D. Feature-Weighted Linear Stacking. Arxiv:0911 0460 2009.

(19) Breiman L. Stacked Regression. Machine Learning 1996; 24.

(20) Padhukasahasram B, Yang JJ, Levin AM, Yang M, Burchard EG, Kumar R et al. Gene-based association identifies SPATA13-AS1 as a pharmacogenomic predictor of inhaled short-acting beta-agonist response in multiple population groups. Pharmacogenomics J 2014.

(21) Li X, Howard TD, Zheng SL, Haselkorn T, Peters SP, Meyers DA et al. Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. J Allergy Clin Immunol 2010; 125(2):328-35.

(22) Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010; 363(13):1211-21.

(23) Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet 2011; 43(9):887-92.

(24) van der SS, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet 2013; 9(1):e1003235.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Comparison of empirical power and Type 1 error rates of gene-based association tests for simulated datasets assuming linkage equilibrium.

	#SNP (#DSL)	Logistic Regression	Fisher	Vegas-Sum	Original -Simes	Vegas-Max	GATES	Machine-Learning Ensemble
Linkage Equilibrium								
Type 1 Error	3(0)	4.66	4.67	4.70	4.61	4.62	4.61	4.90
Type 1 Error	10(0)	5.10	5.00	5.04	5.06	5.07	5.06	4.80
Type 1 Error	30(0)	5.26	4.96	4.97	4.97	5.04	4.97	5.60
Power Additive	3(1)	43.71	41.79	42.67	45.28	45.22	45.28	56.00
Power Additive	10(2)	56.88	53.32	54.56	54.76	54.00	54.76	57.60
Power Additive	30(6)	65.32	61.5	63.28	47.18	45.62	47.18	69.00
Power Multiplicative	3(1)	46.61	44.72	45.54	48.39	48.3	48.39	53.00
Power Multiplicative	10(2)	69.00	65.25	66.88	67.00	66.26	67.00	69.00
Power Multiplicative	30(6)	93.45	91.44	92.28	82.21	80.18	82.21	94.60

DSL denotes the number of disease susceptibility markers.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2. Comparison of empirical power and Type 1 error rates of gene-based association tests in simulated datasets for moderate Linkage Disequilibrium.

	#SNP (#DSL)	Logistic Regression	Fisher	Vegas-Sum	Original -Simes	Vegas-Max	GATES	Machine-Learning Ensemble
Linkage Disequilibrium								
Type 1 Error	3(0)	4.86	7.17	4.91	4.54	4.81	4.98	5.20
Type 1 Error	10(0)	4.88	9.8	4.83	4.55	4.92	5.00	5.60
Type 1 Error	30(0)	5.63	11.09	5.03	4.97	5.29	5.56	5.70
Power Additive	3(1)	44.59	55.8	49.36	49.71	50.51	51.23	55.20
Power Additive	10(2)	56.25	72.38	61.36	58.39	59.12	60.72	63.80
Power Additive	30(6)	65.47	83.04	71.96	53.29	52.24	55.65	68.00
Power Multiplicative	3(1)	46.52	57.5	50.98	51.19	52.00	52.65	53.40
Power Multiplicative	10(2)	68.42	81.73	72.48	70.66	70.9	72.4	70.20
Power Multiplicative	30(6)	93.68	98.04	95.59	86.07	84.34	87.52	94.70

DSL denotes the number of disease susceptibility markers.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3. Comparison of empirical power and Type 1 error rates of gene-based association tests on simulated datasets for strong Linkage Disequilibrium.

	#SNP (#DSL)	Logistic Regression	Fisher	Vegas-Sum	Original -Simes	Vegas-Max	GATES	Machine-Learning Ensemble
Linkage Disequilibrium								
Type 1 Error	3(0)	4.96	11.49	5.23	3.88	5.22	5.35	6.00
Type 1 Error	10(0)	5.33	15.68	4.84	3.37	4.88	5.34	5.70
Type 1 Error	30(0)	5.57	17.9	4.89	3.38	4.89	5.64	5.90
Power Additive	3(1)	45.03	72.29	58.81	53.88	58.2	60.43	61.00
Power Additive	10(2)	57.20	89.82	75.74	66.39	71.71	74.3	59.00
Power Additive	30(6)	65.56	96.04	86.3	62.84	66.80	72.75	65.80
Power Multiplicative	3(1)	47.13	74.28	60.88	56.28	60.74	62.77	65.00
Power Multiplicative	10(2)	68.45	94.41	84.89	77.14	80.59	83.00	74.40
Power Multiplicative	30(6)	93.4	99.92	99.2	91.42	92.24	95.38	94.00

DSL denotes the number of disease susceptibility markers.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 4. Comparison of empirical power and type 1 error rates of gene-based association tests for a quantitative trait simulated under a model with interactions.

Value	Phenotype distribution	#SNP (#DSL)	Linear Regression	GATES	Machine learning Ensemble
Type 1 error	*P ~ N(0,1)*	5(0)	5.66	4.00	5.00
Type 1 error	*P ~ N(0,1)*	10(0)	5.30	6.00	5.66
Power	*P~N(0,1)+0.20*snp1*snp2*snp9*snp10*	10(4)	40.0	36.0	44.0
Power	*P~N(0,1)+0.002*snp1 +0.002*snp2 +0.12*snp1*snp2 + 0.18*snp3*snp4*	5(4)	84.0	66.0	87.0
Power	*P~N(0,1)+0.25*snp1*snp2*snp3*	5(3)	38.0	34.0	56.0
Power	*P ~N(0,1)+0.3*snp1*snp2*snp3*	5(3)	49.0	36.0	64.0
Power	*P~N(0,1)+0.35*snp2*snp3*snp4*	5(3)	64.0	55.0	87.0
Power	*P~N(0,1)+0.65*snp1*snp2*snp3*snp8*snp9* snp10*	10(6)	58.0	47.0	85.0
Power	*P~N(0,1)+0.002*snp1 +0.002*snp2 + [0.2*(1+snp1)/(1+snp2)] + 0.3*snp4*snp5*	5(4)	66.0	50.0	80.0
Power	*P~N(0,1)+0.002*snp1 +0.002*snp2 +0.3*snp1*snp2 + 0.2*snp3*snp4*	5(4)	61.0	49.0	96.0

DSL denotes the number of trait associated SNPs.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 5. Empirical Power and Type 1 error rate of a gene-based test of interactions for a simulated quantitative trait.

Value	Phenotype distribution	#SNP (#DSL)	Machine learning Ensemble
Type 1 error	*P ~ N(0,1)*	5(0)	6.10
Type 1 error	*P ~ N(0,1)*	10(0)	5.10
Power	*P~N(0,1)+0.20*snp1*snp2*snp9*snp10*	10(4)	55.5
Power	*P~N(0,1)+0.002*snp1 +0.002*snp2 +0.12*snp1*snp2 + 0.18*snp3*snp4*	5(4)	52.5
Power	*P~N(0,1)+0.25*snp1*snp2*snp3*	5(3)	64.6
Power	*P ~N(0,1)+0.3*snp1*snp2*snp3*	5(3)	78.2
Power	*P~N(0,1)+0.35*snp2*snp3*snp4*	5(3)	87.5
Power	*P~N(0,1)+0.65*snp1*snp2*snp3*snp8*snp9*snp10*	10(6)	87.4
Power	*P~N(0,1)+0.002*snp1 +0.002*snp2 + [0.2*(1+snp1)/(1+snp2)] + 0.3*snp4*snp5*	5(4)	53.5
Power	*P~N(0,1)+0.002*snp1 +0.002*snp2 +0.3*snp1*snp2 + 0.2*snp3*snp4*	5(4)	83.6

DSL denotes the number of trait associated SNPs.

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 6. Gene-based *p* values for previously reported asthma-related genes in 1,401 African American individuals from the SAPPHIRE cohort.

Chromosome	Gene	Length in base pairs	Number of SNPs tested	Gene-based *p* value from Ensemble Learning	Gene-based *p* value from Logistic Regression	Gene-based *p* value from GATES
1	PYHIN1	45513	13	0.198	0.230	0.130
2	IL1RL1	7466	6	0.982	0.982	0.832
5	TSLP	6333	5	0.063	0.064	0.533
9	IL33	42198	12	0.408	0.180	0.130
17	GSDMB	14056	15	0.401	0.401	0.870
5	IL13	2937	3	0.156	0.164	0.387
15	SMAD3	57175	23	0.323	0.323	0.359
5	SLC22A5	25906	15	0.0095	0.162	0.0076
5	RAD50	87698	34	0.010	0.010	0.367

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 7. Gene-based *p* values for previously reported asthma-related genes in 3,772 Latino individuals from the GALA study.

Chromosome	Gene	Length in base pairs	Number of SNPs tested	Gene-based *p* value from Ensemble Learning	Gene-based *p* value from Logistic Regression	Gene-based *p* value from GATES
1	PYHIN1	45513	15	0.320	0.320	0.530
2	IL1RL1	7466	16	0.038	0.038	0.046
5	TSLP	6333	7	0.270	0.270	0.250
9	IL33	42198	14	0.0014	0.095	0.069
17	GSDMB	14056	13	2.33E-09	4.20E-08	6.24E-11
5	IL13	2937	10	0.280	0.280	0.100
15	SMAD3	57175	28	0.464	0.464	0.063
5	SLC22A5	25906	12	0.838	0.838	0.956
5	RAD50	87698	33	0.217	0.217	0.050

bioRxiv preprint doi: https://doi.org/10.1101/005405; this version posted May 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.