binomialRF: Scalable Feature Selection and Screening for Random Forests to Identify Biomarkers and Their Interactions

Samir Rachid Zaim[1-3], Colleen Kenost[1-3], Yves A. Lussier[1-5, *], Helen H. Zhang[6,*]

[1]The Center for Biomedical Informatics & Biostatistics of the University of Arizona Health Sciences, 1230 N. Cherry Ave, Tucson, AZ, 85721, USA
[2]The Department of Medicine, College of Medicine, Tucson, 1501 N. Campbell Ave, Tucson, AZ, 85721, USA
[3]The Graduate Interdisciplinary Program in Statistics, 617 N. Santa Rita Ave. The University of Arizona, Tucson, AZ 85721, USA
[4]The Center for Applied Genetic and Genomic Medicine, 1295 N. Martin, Tucson, AZ 85721, USA
[5]The University of Arizona Cancer Center, 3838 N. Campbell Ave, Tucson, AZ 85721, USA
[6]The Department of Mathematics, College of Sciences, 617 N. Santa Rita Ave. The University of Arizona, Tucson, AZ 85721, USA

* Correspondence
Corresponding Authors
YAL: yves@email.arizona.edu
HHZ: hzhang@math.arizona.edu

Keywords: random forests, feature selection, binomial, variable selection, R package

Abstract

Background: Feature selection is becoming increasingly important in machine learning as users want more interpretation and post-hoc analyses. However, existing feature selection techniques in random forests are computationally intensive and high RAM memory requirements requiring specialized (uncommon) high throughput infrastructures . In bioinformatics, random forest classifiers are widely used as it is a flexible, self-regulating and self-contained machine learning algorithm that is robust to the "*predictors(features)* $P >> subjects\ N$" problem with minimal tuning parameters. The current feature selection options are used extensively for biomarker detection and discovery; however, they are limited to variants of permutation tests or heuristic rankings with arbitrary cutoffs. In this work, we propose a novel paradigm using the binomial framework for feature selection in random forests, *binomialRF*, which is designed to produce significance measures devoid of expensive and uninterpretable permutation tests. Furthermore, it offers a highly flexible framework for efficiently identifying and ranking multi-way interactions via dynamic tree programming.

Methods: We propose a novel and scalable feature selection technique that exploits the tree structure in the random forest, treats each tree as a binomial stochastic process, and determines feature significance by conducting a one-sided binomial exact test to determine if a feature was detected more often than expected by random chance. Since each tree is an independent and identically distributed random sample in a binomial process (from the perspective of choosing a splitting variable in the root node), the test statistic is constructed based on the frequency of a

feature being selected (each tree is a Bernoulli trial for selecting a feature), and then the features are ranked based on the observed test statistics, its resulting nominal p-values, and its multiplicity-adjusted q-values. Furthermore, the binomialRF framework provides a general selection framework to identify 2-way, 3-way, and K-way interactions by generalizing the test statistic to count sub-trees in the random forest using dynamic tree programming.

Results: In simulation studies, the binomialRF algorithm performs competitively with respect to the state of the art in terms of classification accuracy, true model coverage, and controlling for false selection in identifying main effects while attaining substantial computational performance gains (between 30 to 600 times faster in high dimensional settings than the state of the art). In addition, extending the binomialRF using model averaging identified the true model on average with greater accuracy (>20% improvement in reducing false positive feature selection at high dimensions while maximizing true model coverage) and attained greater classification accuracy (between 4-9% improvement across all techniques) without sacrificing computational speed (2nd fastest performance after binomialRF). In addition, the framework easily scales and extends to identifying 2-way and K-way interactions (i) without additional memory requirements (only requires storing original predictor matrix), and (ii) with minimal additional computational complexity cost due to efficient dynamic tree programming interaction searches. The algorithm was validated in a case study to predict bronchospasm-related hospitalization from blood transcriptomes where the binomialRF algorithm correctly identified the previously published relevant physiological pathways, presented comparable classification accuracy in a validation set, and extended previous work in this area by looking at pathway-pathway interaction.

Conclusion: The proposed binomialRF proposes a novel and efficient feature selection method devoid of permutation tests – that scales linearly in the number of trees, with minimal computational complexity, thus outperforming alternate conventional methods from a computational perspective while attaining competitive model selection and classification accuracies and enabling computations on common of cost-effective high throughput infrastructures. Furthermore, the binomialRF model averaging framework greatly improves the accuracy of the feature predictions, controlling for false selection and substantially improving model and classification accuracy. Validated in numerical studies and retrospectively in a clinical trial (case study), the binomialRF paradigm offers a binomial framework to detect feature significance and easily extends to search for K-way interactions in a linear fashion, reducing a known non-polynomial time exploration to linear approximations.

The binomialRF R package is freely available on GitHub and has been submitted to BioConductor, with all associated documentations and help files.

Github: https://github.com/SamirRachidZaim/binomialRF
BioConductor: binomialRF

1 Introduction

Recent advances in machine learning and data science tools have led to a revamped effort for improving clinical decision-making anchored in genomic data analysis and biomarker detection. However, despite these novel advances, random forests (RFs) [1] remain a widely popular

machine learning algorithm choice in genomics given their ability to i) accurately predict phenotypes using genomic data and ii) identify relevant genes and gene products used to predict the phenotype. Literature over the past twenty years has demonstrated [2-9] their wide success in being able to robustly handle the "P >> *N*" issue where there are more predictors or features "*P*" (i.e., genes) than there are human subjects "*N*" while maintaining competitive predictive and gene selection abilities.

However, the translational utility of random forests has not been fully understood as they are often viewed as "black box" algorithms for physicians and geneticists. Therefore, a substantial effort over the past decade has focused around "feature selection" in random forests [5, 6, 10-14] in order to better provide explanatory power of these models and to identify important genes and gene products in classification models. **Table 1** demonstrates the two major classes of feature selection techniques: i) permutation-type measures of importance and ii) heuristic rankings without formal decision boundaries (i.e., no p-values). The second and third columns in **Table 1** indicate where each method falls. These feature selection techniques have been widely used in biomarker discovery by the bioinformatics community as an alternative to classical statistical techniques and have shown promising results in identifying single gene products. However, these techniques do not easily scale computationally nor memory-wise for identifying molecular interactions, limiting their translational utility in precision medicine. To meet this need, we propose the *binomialRF* feature selection algorithm, a wrapper feature selection algorithm that identifies significant genes and gene sets in a memory-efficient and scalable fashion. Building from the 'inclusion frequency"[15] heuristic feature ranking, binomialRF formalizes this concept into a binomial probabilistic framework with measures of p-values and extends to identify gene set interactions of any size. The main algorithm is presented in Section 2 while the extension to identify interactions is presented in Section 3. A section on theoretical computational complexity is presented in Section 4, while applications in numerical analyses and case studies evaluate its utility in Sections 5 and 6. The discussion, limitations, and concluding sections are presented in Sections 7-9.

Table 1: Random forest Feature Selection Techniques. As described, the proposed binomialRF method is the only computationally-efficient feature selection technique (permutation free) that also generates a p-value.

Method	P-value	Computational efficiency	Description
RFE	No	Yes	Iteratively trains Random Forest (RF) models and at each iteration, removes all features with 0 VI and train the subsequent RF with the remaining vars. Repeats until model convergence (i.e., no more features can be eliminated).
VSURF[11]	No	No (permutations)	Two-step feature selection process: Step 1 determines 'important' features by permuting values of predictor and measuring change in error; Step 2 refines model by step-forward introduction until no further error reduction.
PIMP[14]	Yes	No (permutations)	Permutes outcome and determines a feature's importance based on increases in MI or Gini errors. The importance measure is then fit to a distribution to produce p-values for each feature.
Boruta[13]	No	Yes	Creates phony predictors by permuting the values of the shadow vars. Runs RF and collects Z-score on importance of

			all vars. Eliminates features whose Z-score is less than max (Z-score shadow vars). Repeats until convergence.
VarSelRF[5]	No	Yes	Iteratively removes worst .20 (or x-percentage) of all features; retrains RF; selects smallest feature set within 1se of best model.
binomialRF "current manuscript"	Yes	Yes	Performs Binomial feature selection test to identify whether a model selects a predictor more likely than what is expected by random chance.

2 binomialRF: Identifying main effects

2.1 Problem Set-up and Notation

The binomialRF algorithm is a feature selection algorithm designed to develop classifiers in genomics datasets. Datasets are noted as $\boldsymbol{X}$ and are comprised of $\boldsymbol{N}$ subjects (usually < 1,000) and $\boldsymbol{P}$ genes in the genome (usually P > 20,000 expressed genes). Genomics data represent the traditional "high-dimensional" setting where there are many more features than there are samples (e.g., "$P >> N$"). In binary classification settings, the outcome variable $\boldsymbol{Y}$, differentiates the case and control groups (i.e., "healthy" vs. "tumor" tissue samples).

In linear and generalized linear models (hereinafter termed "linear models"), there is a major assumption of linearity (in coefficients) in which either $E[Y|X] = X\beta$ or $E[Y|X] = g(X\beta)$, where the E operator denotes the "expected value" and g is a link function (e.g., logistic or probit for binomial regression), and β is a coefficient vector. Linearity assumptions are often inflexible and many times unjustified, resulting in suboptimal classification performance. Furthermore, linear models offer restrictive results when performing feature selection in the presence of high-dimensional data ("$P >> N$") since linear models saturate at $P = N$, rendering powerful algorithms such as the LASSO[16] and elastic net [17] ineffective as they cannot identify additional features (genes) after saturation. In other words, if a genomics classifier requires a 1,000 gene signature to effectively predict classes, but there are only 100 subjects in the study, the LASSO and elastic net will at most select 100 genes out of the 1,000 gene signature, effectively missing the remaining 900. Similar limitations occur for other penalized regression strategies like the group LASSO[18].

These two major limitations led many bioinformaticians to consider powerful, model-free machine learning algorithms such as random forests (RF)[1] to analyze genomics datasets for developing classifiers and identifying gene-product biomarkers. Trees and random forest require minimal assumptions and can robustly handle high-dimensional data since they do not require any full column rank and other matrix regularity conditions in linear models. RF are collections of randomized decision trees, where for each decision tree, T_z, in the RF, a subset of the data and of the features are selected. This randomization encourages a diverse set of trees and allows each individual tree to make predictions across a variety of features and subjects. The induced diversity in the RF model strengthens the overall classifier and mitigates overfitting. Specifically,

each tree only sees $\boldsymbol{m} < \boldsymbol{P}$[i] features in the root when it determines the first optimal feature for splitting the data into two subgroups. Alternatively, one can re-parameterize the feature subsampling with parameter s ($s \in (0,1)$, which determines what percentage of P is seen by each tree. Letting $F_{j,z}$ denote the random variable measuring whether X_j is selected as the splitting variable for the root at tree T_i,

$$F_{j,z} = \begin{cases} 1\,, & if\ root\ (T_i) = X_j \\ 0\,, & otherwise\,. \end{cases} \tag{1}$$

This results in $F_{j,z}$ being a Bernoulli random variable, $F_{j,z} \sim Bern(p_{root})$, and $F_j = \sum_{z=1}^{V} F_{j,z} \sim Binomial(V, p_{root})$ being a Binomial random variable across all V trees, where $\boldsymbol{p_{root}}$ is the probability of randomly selecting a feature X_j as the optimal splitting variable in the root of tree T_i. Under the null, p_{root} is constant across all trees. **Table 2** details the notation that will be used throughout this study.

Table 2: Description of mathematical notation.

Notation	Description
β	coefficient vector
$\boldsymbol{g, k}$	Indices in product term
$\boldsymbol{H_0, H_A}$	"Null" and "Alternative" hypotheses, respectively.
$\boldsymbol{K}$	Integer denoting the interaction depth
$\boldsymbol{L}$	Number of candidate models in binomialRF model averaging
$\boldsymbol{m}$	User parameter to limit the number of features in a tree. **$\boldsymbol{m}$ is constant for all trees**
model_j	Candidate model in the binomialRF model averaging
$\boldsymbol{N}$	Positive integer denoting the number of samples or subjects
O (·)	Big "O" operator, used to denote an algorithm's computational complexity. (O(N^2) for example, denotes an algorithm that is quadratic in N.)
$\boldsymbol{p_{root}(P, m)}$	Probability of randomly selecting a feature X_j as the optimal splitting variable in the root of a tree. It is a function that depends on P and m
$\boldsymbol{p_{K-way}}$	Probability of randomly selecting a set of K features $\{X_j\}_{j=1}^{K}$ as the optimal splitting sequence starting at the root of a tree
$\boldsymbol{P}$	Positive integer indicating the number of measured or expressed genes in a genomics dataset
$\boldsymbol{RF = \{T_1, \ldots, T_V\}}$	Random forest grown with V distinct individual decision trees
$\boldsymbol{s}$	Proportion of features subsampled. $s = \frac{m}{P}$. **$\boldsymbol{s}$ is constant for all trees**
Selected_j	Selected features by model_j in binomialRF model averaging
$\boldsymbol{T_z}$	Decision tree. Specifically, the zth tree in a random forest
$\boldsymbol{V}$	User parameter to determine the number of trees in a random forest
$\boldsymbol{X_{N\times P}}$	Genomics design matrix with N samples and P features. Also known as X
$\boldsymbol{X_j}$	Vector that denotes the jth feature (or gene) in $\boldsymbol{X_{N\times P}}$

[i] The parameter, $\boldsymbol{m}$, is a user-determined input in the random forest algorithm with default values set usually to either $\boldsymbol{m} = \sqrt{\boldsymbol{P}}$ or $\boldsymbol{m} = \frac{\boldsymbol{P}}{\boldsymbol{3}}$.

$\{\boldsymbol{X_j}\}_{j=1}^{K}$	Sequence of K consecutive splitting features starting at the root
$\boldsymbol{X_i} \otimes \boldsymbol{X_j}$	Interaction between the ith and j^{th} feature
$\otimes \boldsymbol{X_i^K}$	Interaction of K different features, for $K \geq 3$
$\boldsymbol{Y_{1\times N}}$	Vector of binary class labels. Alternatively referred to as Y

2.2 High-level overview and background

The proposed binomialRF R package is a software that contains a novel feature selection technique for random forests, which wraps around the existing randomForest R package[19]. Since random forests (RFs) use bootstrap sampling, feature subsetting, and aggregating to create a robust ensemble classifier from individual weak learners, the algorithm essentially averages results from a set of trees to form predictions and provide heuristic feature importance rankings. With a minimal number of tuning parameters and hyperparameters, RFs are essentially a self-regulating and self-sufficient machine learning algorithm. The binomialRF algorithm builds upon the random forest algorithm to exploit its binary-split tree structure to select important features. It treats each tree in the RF as a stochastic binomial process and develops a hypothesis-test-based feature selection criterion for random forests, resulting in a rigorous p-value ranking for feature selection. There are a number of existing feature selection algorithms in random forest algorithms (see **Table 1**) that measure "variable importance" based on permutation tests or heuristic rankings. In this work, we provide an alternative for measuring importance devoid of permutations and heuristic rankings by introducing a binomial hypothesis test into the feature selection process. To the best of our knowledge, this is the first feature selection algorithm in random forests that implements a binomial framework, can provide a relative significance measure without using permutation tests, and naturally scales to screen for interactions in a fast and memory-efficient way. We explain in the subsequent sections the components of the binomialRF algorithm.

The binomialRF R package is freely available on GitHub and has been submitted to BioConductor, with all associated documentations and help files.

Github: https://github.com/SamirRachidZaim/binomialRF
BioConductor: binomialRF

2.3 Optimal splitting variables and decision trees

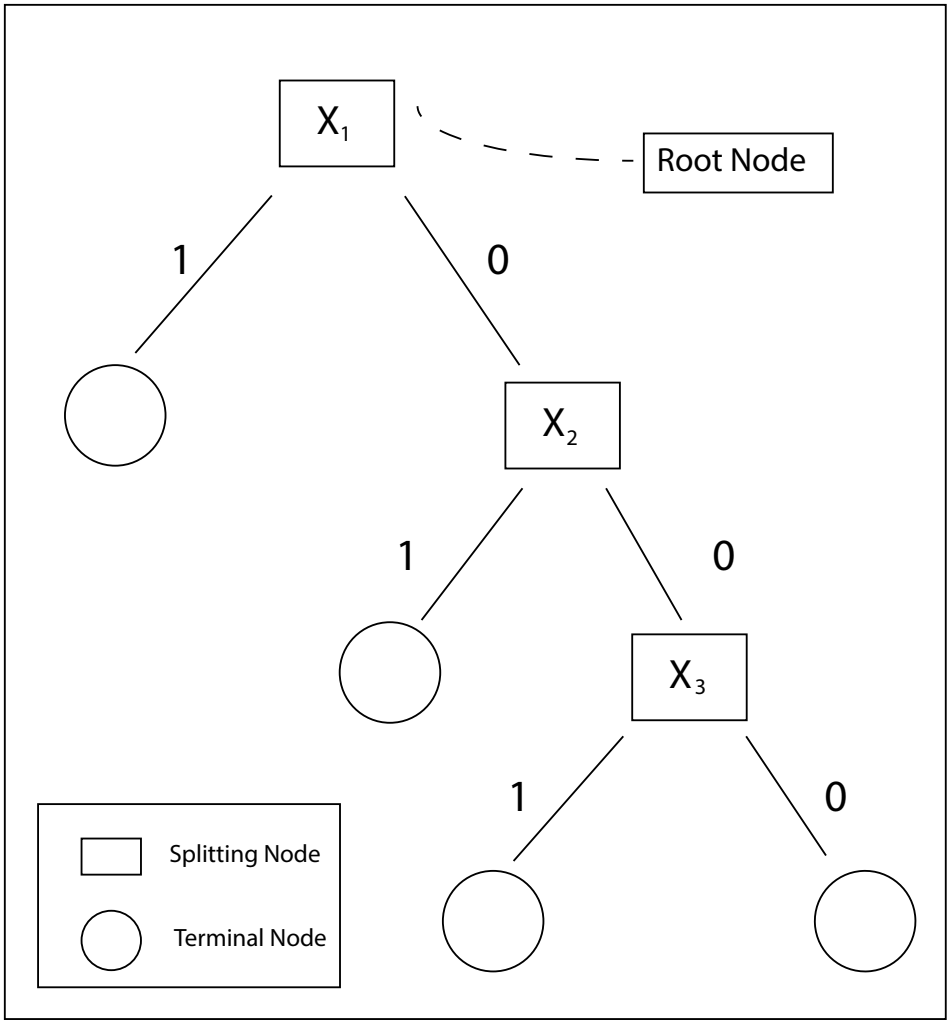


Fig. 1. Decision tree. In a binary split decision tree where X_1 is the optimal splitting feature at the root of the tree, and $\{\boldsymbol{X_j}\}_{j=1}^{3} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}\}$ is the optimal splitting sequence that indicates a potential $X_1 \otimes X_2 \otimes X_3$ 3-way interaction (alternatively denoted as $\otimes X_{j=1}^{3}$).

Consider a decision tree, T_z, in a random forest. At the top-most "root" node, m, features are randomly subsampled from the entire space of P features. These m features are all tested as possible splitting variables for the tree, and the optimal splitting variable, $X_{optimal}$, is selected as the feature X_j that best separates the two classes and provides the best information gain. Formally, this is stated in **Equation 2.**

$$X_{optimal} = argmax_{X_j}(Information\ Gain) \tag{2}$$

Starting from the root, each node either selects its $X_{optimal}$ or becomes a terminal node as seen in **Figure 1**, where in the root, $X_{optimal} = X_1$ and in the subsequent right daughter nodes, $X_{optimal} = X_2, X_3$.

If we focus solely at the root, then under a null hypothesis, each feature has the same probability of being selected as the optimal root splitting feature, denoted by $p_{root} = \Pr(X_{optimal} = X_j)\ \forall\ j \in \{1, \ldots, P\}$. The random variable $F_{j,z}$ (shown in **Equation 1**) is an indicator variable that measures if X_j is selected as the optimal variable for the root at tree T_z. $F_{j,z}$ is a Bernoulli random variable, $F_{j,z} \sim Bern(p_{root})$. Summing across all the trees in the random forest, $\boldsymbol{RF} = (T_1, \cdots, T_N)$, $F_j = \sum_{z=1}^{V} F_{j,z} \sim Binomial(V, p_{root})$ is a Binomial random variable across all V trees, where $\boldsymbol{p_{root}}$ is the probability of randomly selecting a feature X_j as the optimal splitting variable in the root of tree T_z.

2.4 Calculating p_{root} and the binomial exact test for significance

Let T_z represent an individual tree grown in a random forest. Then, $\boldsymbol{RF} = (T_1, \cdots, T_N)$ denotes the random forest (i.e., the collection of independently and identically distributed trees). If at each T_z, $m < P$ features are subsampled to reduce the feature subspace at each tree, then the probability, p_{root}, of X_j being selected by a tree, T_z, is shown in **Equation 3:**

$$p_{root} = 1 - \left(\prod_{g=1}^{m} \frac{P-g}{P-(g-1)} \left(\frac{1}{m}\right) \right) \tag{3}$$

Since features are selected without replacement in the subsampling process, $\prod_{g=1}^{m} \frac{P-g}{P-(g-1)} \left(\frac{1}{m}\right)$ is the probability of not selecting it, and using the complement rule, we get P(A) = 1-P(¬A). Using Equation 3, we can provide a formal measure of significance (i.e., a p-value) regarding if X_j was selected more than expected by chance by determining if its frequency of selection exceeds a statistical threshold. If we are concerned with only conducting a single hypothesis test about predictor X_j using a significance level α, then we conclude that X_j is chosen more often than by random chance if F_j exceeds the critical value $Q_{\alpha,K,p}$, resulting in a one-sided hypothesis test (shown in **Equations 4a-c**). Formally:

binomialRF Feature Selection Hypothesis Test

$H_0: Pr(X_j) \leq p_{root} \qquad H_A: Pr(X_j) > p_{root}$	Hypothesis Test	(4-a)
$F_j = \sum_{i=1}^{V} F_{j,z} \sim Binomial(V, p_{root})$	Test Statistic	(4-b)
$R = \{ X_j : F_j < Q_{\alpha,V,\boldsymbol{p_{root}}} \}$	Rejection Region	(4-c)

Furthermore, since we are conducting simultaneous hypothesis tests when assessing the significance of each feature, we must adjust for multiplicity. Using false discovery rate adjustment procedures, such as Benjamini-Hochberg(BH)[20] or Benjamini- Yekutieli (BY)[21] or Bonferroni family-wise, error rates can all be used depending on the predictor space, though a safe option is always given that it provides a false-discovery-adjusted p-value in the case of dependent predictors.

2.5 The binomialRF algorithm

Combining methods of Sections 2.3 and 2.4, the binomialRF feature selection algorithm constructs a formal hypothesis test to determine whether X_j is an important feature or not. It first calculates the probability of selecting X_j as $X_{optimal}$ a tree and the test statistic $F_j = \sum_{T_i \in T} F_{j,z}$. It conducts a hypothesis test comparing the observed test statistics F_j for all features X_j, and its expected value, yielding a p-value for each feature. Finally, it returns a feature selection ranking with p-values and FDR-corrected q-values denoting variable importance. In other words, the binomialRF ranks and assigns p-values to features based on how frequently they were the optimal splitting variable in the tree. By restricting the search to the root of each tree, it allows for modeling it via a binomial hypothesis testing framework. This procedure is illustrated in **Figure 2** and formalized in **Algorithm 1** (Appendix A1).

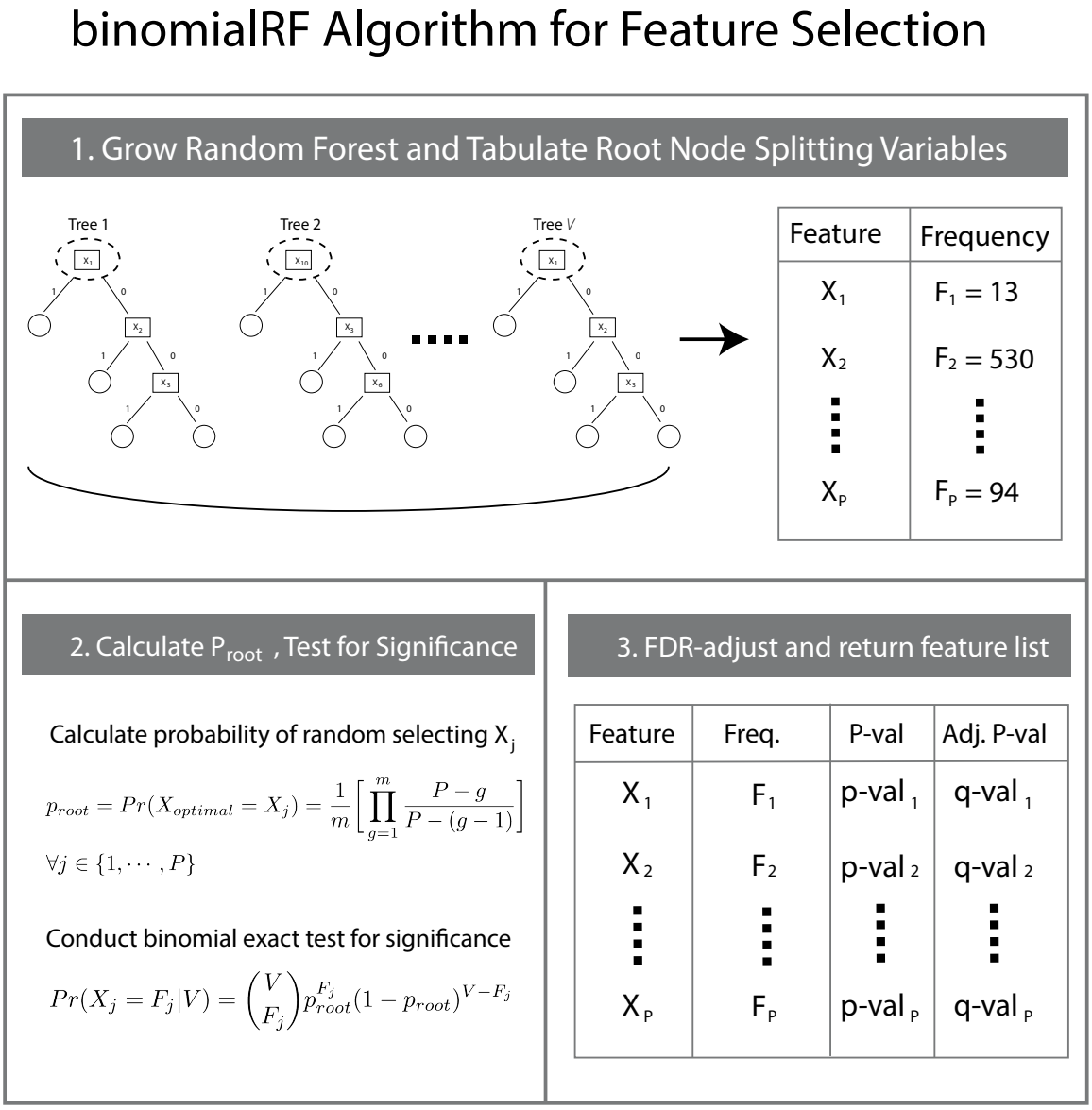


Fig. 2. The binomialRF algorithm. The binomialRF algorithm is a feature selection technique in random forests (RF) that treats each tree as a stochastic binomial process and determines whether a feature is selected more often than by random chance as the optimal splitting variable, using a top-bottom sampling without replacement scheme. The main effects algorithm identifies whether the optimal splitting variables at the root of each tree are selected at random or whether certain features are selected with significantly higher frequencies, and the interaction-screening extension is detailed in Section 3. **Legend:** $T_z = z^{th}$ tree in random forest; X_j = feature j; F_j = observed frequency of selecting X_j; Pr = probability; P = number of features; V= number of trees; m = user parameter to limit the number of features; g = index of the product.

2.6 Cross-validated binomialRF

Since the binomial exact test is contingent on a test statistic measuring the likelihood of selecting a feature and if there is a single dominant feature, it will render all remaining 'important' features useless as they will always be selected as the splitting variable. Letting s represent the percent_feature parameter, it is used to determine how many features are considered at each node. Features may be collinear or some dominant features may overshadow other important features. If s is too small, there will be minimal competition among features, allowing noisy ones to be deemed important, and if s is too large, dominant features will always overshadow important but small-signal features. Therefore, it is important to test a number of possible values of $s \in (0,1)$ and optimize it via cross-validation. An extension to the binomialRF has been written to conduct a [default] 5-cross-validation[ii] to tune and identify the optimal s hyper-parameter.

2.7 binomialRF Model Averaging: A tool for final model selection

Model averaging is an alternative way to perform model selection by combining different models based on their performance and feature selection. In particular, assume that there are L different candidate models, $\text{model}_1, \text{model}_2, \cdots, \text{model}_L$, that may contain the full true model structure, a subset of the truly important terms, or even none of the significant variables. We can implement binomialRF under each model, model_a, and identify important terms under each model; denote the selected set by Selected_1, under model_1. Then one can define an importance metric called "Proportion Selected" – see **Equation 5** – to measure how often a given feature X_j was selected. Formally, the proportion selected metric is defined as

$$\text{Proportion Selected } (X_j) = \sum_{i=1}^{L} \text{I}[\text{X}_\text{j} \in \text{Selected}_i] \tag{5}$$

where I is the indicator function measuring if X_j was selected by model_i. If Proportion Selected (X_j) is equal to 1, then that feature is determined important by every single possible candidate model; however, if it equals 0, then every possible candidate model rejected it.

In **Figure 3**, we use a simple toy example to illustrate the power of model averaging under binomialRF. As shown in **Figure 3A,** consider a dataset where the design matrix X contains 10 predictors, of which the first 5 are related to the binary class label, Y, and the last 5 are noise.

[ii] 5-fold is the default but a user-modifiable parameter in the algorithm.

Suppose that you do not know what the true features are, so you consider a set of 10 possible candidate models (**Figure 3B)** to choose which is the most likely set of 'significant' features. Each of the proposed candidate models ($\text{model}_1, \text{model}_2, \cdots, \text{model}_{L=10}$) is ran as shown by the top bars (in black and white in **Figure 3B**), and determine their selections (Selected_1, $\text{Selected}_2, \cdots, \text{Selected}_{L=10}$) as shown by the bottom bars (in green and purple in **Figure 3B**). Then, to validate whether a feature is truly important, a model average is obtained by calculating the Proportion Selected (X_j) metric for all possible features and rank them by their Proportion Selected value (**Figure 3C**). Using Proportion Selected > 0.5 as a cutoff, we would conclude that $X_1 X_2, X_3, X_4, X_5$ are all significant while the remaining features can be discarded as noise.

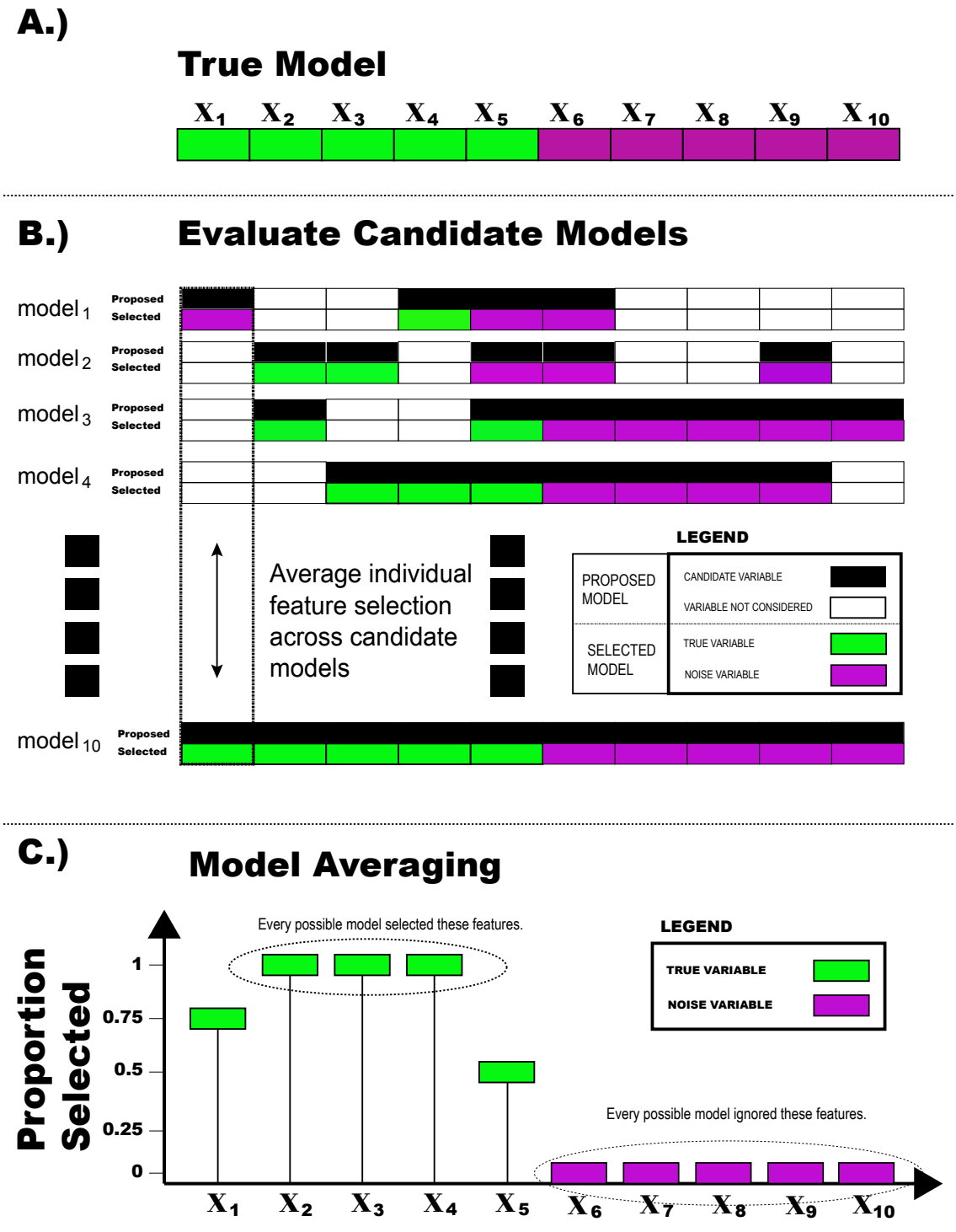


Fig. 3. binomialRF model averaging visualization example. (A) shows the true simulated model with P= 10 predictors. The first five of these predictors are "true" variables while the remaining 5 are pure noise. In (B), we try 10 different candidate models, each of which includes

a random subset of predictors, while the last model includes all predictors. Each of the candidate models, (like model_1 for example) sees a set of features (black = included, white = excluded) and then determines whether they are significant (green) or not (purple). (C) Each of the 10 model's recommendations are averaged into a final model consensus by showing how often each of the features was deemed significant. True features will be closer to 1 (i.e., selected 100% of the times across all models), while noisy features will tend closer to 0. The power and robustness of model averaging lies in that individual candidate models may make occasional errors due to random chance and noise (like model_1 for example ignoring feature X_1), but the noise tends to get averaged out via model averaging, as majority voting will overcome individual mistakes.

2.8 Evaluations in Numerical Studies

2.8.1 Logistic Dataset Generation

To understand the strengths and limitations of the binomialRF feature selection algorithm and to compare its performance with the state of the art, we conduct a variety of numerical studies. These simulation scenarios generate logistically-distributed data to mimic binary classification settings. The gold standard is generated by first creating a coefficient vector β whose first five elements are non-zero and the remaining are zero. Then, $X_{N\times P}$, random multi-variate standard normal matrix with N samples and P features, is generated to mimic a standardized and centered genomics matrix, and finally, X, undergoes a logistic transformation from which a Bernoulli random variable is generated to mimic the binary class vector, Y.

2.8.2 Modifying Signal-to-Noise Ratio

To understand the strengths and limitations of the binomialRF feature selection algorithm and to compare its performance with the state of the art, we conduct a variety of numerical studies that compare different signal-to-noise ratio settings. These simulation scenarios generate logistic data as the gold standard with a coefficient vector β whose first five elements are non-zero and the remaining are zero. The signal-to-noise ratio is altered in two ways to determine how robustly each technique is able to handle noise. First, the magnitude of the non-zero coefficients are increased from 1 to 3 to evaluate how much better each technique is able to determine the true variables from noise. Later, we increase the number of features, P, in the design matrix, $X_{N\times P}$, 10-fold from 10 to 100 to 1,000. Each time we fix the number of true features to 5, making the remaining $P-5$ noise features, which enables us to evaluate how well each technique is able to select the 5 true features in the presence of 5, 95, and 995 noisy features, respectively. The numerical studies are presented in Section 5. Results for $\beta = [1_5\ 0_{P-5}]^T$ are shown in **Figure 6** while the results for $\beta = [3_5\ 0_{P-5}]^T$ are available upon request and are omitted as they present minimal additional information. In the simulation study, we seeded a small number of true features relative to the number of noisy features and then used model averaging to perform feature selection. We ran the model averaging algorithm twice, using the following decision rules.

- Proportion Selected $(X_j) \geq 0.5$,
- Proportion Selected $(X_j) \geq 0.9$.

This second cutoff was chosen due to empirical results suggesting that as the number of candidate models increases, the Proportion Selected (X_j) approaches 1. That is, the limit of its selection proportion will approach 1 as the number of candidate models L goes to infinity (i.e., $\text{Limit}_{L\to\infty}$ Proportion Selected $\left(X_j\right) \to 1$). We report simulation results for Proportion Selected $(X_j) \geq 0.9$.

2.9 Evaluations in Clinical Studies

2.9.1 Overview of the Asthma Clinical Validation Study

To determine the utility of the binomialRF feature selection algorithm in translational bioinformatics, we conducted a validation study mirroring a prior study that focused on the translational impact of random forest classifiers.

Specifically, the study by Gardeux et al [26] determined whether

1.) a classifier predicting symptomatic subjects among healthy adults who were inoculated with human rhinovirus (HRV) (i.e., the common cold) using their blood transcriptome data before and during infection could forecast which pediatric asthmatic patients will experience recurrent exacerbations using their transcriptomes derived from *ex vivo* incubation of their blood with and without HRV.
2.) we ccould develop a fully-specified random forest classifier (i.e., set of features) to make these predictions using dynamic genomic information (i.e., gene expression data) before and after HRV exposure.

The study examined a few different machine learning techniques and developed a random forest classifier that identified key pathways to predict asthma exacerbation.

2.9.2 Asthma Clinical Validation Datasets

The two datasets in this clinical validation are described in **Table 3** and contain microarray data from two different studies. The first study, conducted by researchers at Duke [22] in 2009, looked at understanding dynamic changes in gene expression data in healthy adults as a response to human rhinovirus (HRV) infection and measured the rhinovirus symptoms (about 50% were asymptomatic yet shedding the virus and the remainder were symptomatic and was used as a "training set" to develop the classifier) [23]. The second dataset was a tightly-controlled clinical trial conducted at the University of Arizona within severe asthmatic patients to determine whether the *ex vivo* HRV incubation of the peripheral blood mononucleocytes would be associated to ulterior asthmatic exacerbation. This dataset was designed to predict asthmatic exacerbation, a phenotype that was established during a 1-year follow-up and was defined as

- No Exacerbation: patients with no hospitalizations and/or emergency room visits; or
- Recurrent Exacerbation: patients with hospitalizations and/or emergency room visits.

The clinical trial was designed in a tightly-controlled fashion in which all patients' demographic and clinical profiles were verified and shown equally distributed between the two groups [23]

(no obvious confounders), and all patients received the same maximal treatment to mitigate exacerbations.

Table 3: Asthma Validation Study Datasets.

	Training Data [22]	Testing Data [23]
Description	*In vivo* HRV stimulus-response in 19 healthy adults	*Ex vivo* HRV stimulus-response of peripheral blood mononucleocytes of 23 severe asthmatic children (followed one year with maximal treatment)
Data and Platform	Microarray, Affy. Human Gene U133A 2.0	Microarray, Affy. Human Gene 1.0ST33297
Outcome Prediction	10 symptomatic, 9 asymptomatic	12 recurrent exacerbations, 11 no asthma exacerbation
Data Collection	19 microarrays before HRV, 19 microarrays after HRV	23 microarrays before HRV, 23 microarrays after HRV
GEO Access Tag	GSE17156	GSE68479

Both datasets contained microarray data at the gene-product and were transformed using the *N-of-1-pathways*[24] framework to determine whether a gene-ontology biological process pathway was dysregulated in each patient. Thus, the final design matrix $X_{N\times P}$ contained ~ 20 subjects in each group and approximately 3,000 pathways. The goal of our Asthma case study is to determine whether our feature selection technique could confirm the clinical findings (i.e., reproduce the predicted pathways in the study), attain similar prediction performance, propose new pathway discoveries, and extend their study by proposing pathway-pathway interactions.

3 binomialRF: identifying interactions

3.1 Current Available Interactions Screening and Selection Techniques

In classical linear models when detecting 2-way interactions, interactions are included in a multiplicative fashion and treated as a separate feature in the model with its own linear coefficient term. Here, we denote $X_i \otimes X_j$ as an interaction between X_i and X_j. The main regularity condition imposed in interactions in linear models is strong heredity. Strong heredity is the requirement that if $X_i \otimes X_j$ is an interaction in the model, then both X_i and X_j must be included in the model. Similarly, under weak heredity, at least one of the two features must be included individually in the model if the interaction is included. The argument between allowing weak vs. strong heredity revolves around which properties a model can attain under certain regularity conditions as well as feasibility and utility [25, 26]. However, under tree-based models, strong heredity hierarchy is automatically induced as a natural consequence of the binary split tree's structure. This reduces computational inefficiencies in forcing strong heredity as well as avoids irregularities present under weak heredity. Therefore, it makes the interaction search more computationally feasible and statistically rigorous.

3.2 binomialRF: Identifying 2-way Interactions

To generalize the binomialRF algorithm to search for 2-way interactions, we generalize **Equation 3** by adding another product term to denote the second feature in the interaction set to calculate p_{2-way}.

$$p_{2-way} = \frac{1}{2}\left[\left(1-\left(\prod_{g=1}^{m}\frac{P-g}{P-(g-1)}\left(\frac{1}{m}\right)\right)\right)\left(1-\left(\prod_{g=1}^{m}\frac{(P-1)-g}{(P-1)-(g-1)}\left(\frac{1}{m}\right)\right)\right)\right] \tag{6}$$

As we are interested in selecting interactions and if X_j is selected at the root node, then it is no longer available for selection subsequently. Thus, we replace P with $(P-1)$, and we include a 1/2 normalizing constant since the interaction can happen two different ways (either via the left or right child nodes). **Figure 4A** provides the visual representation of how to generalize binomialRF to identify a 2-way interaction by looking at pairs of features starting at the root node.

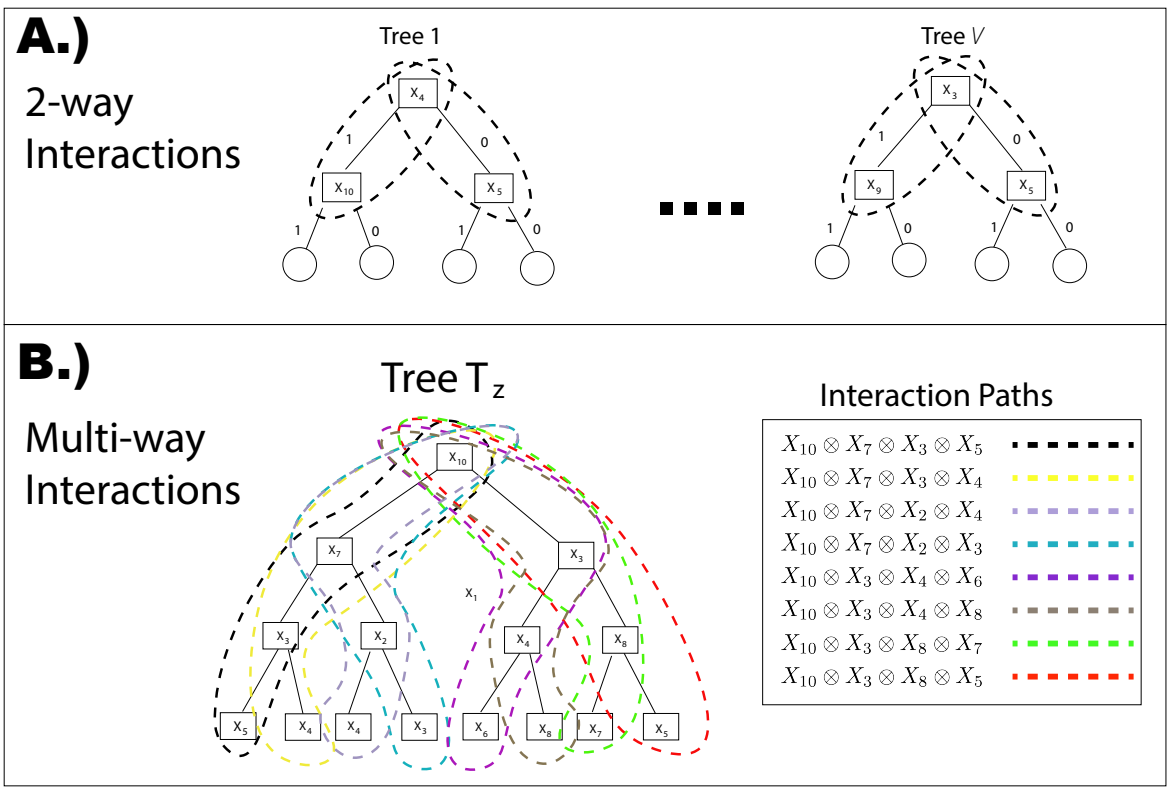


Fig. 4. Calculating Interactions. (A) 2-way Interactions. To illustrate how the binomialRF algorithm extends to identify 2-way interactions, the test statistic is generalized to evaluate the frequency, F_{ij} of each $X_i \otimes X_j$ pair of interactions that occur in the random forest. The probability of the interaction occurring by random chance is recalculated and normalized by a factor of a half, and then a binomial exact test is conducted. **(B) *K*-way interactions, *K* = 4.** Here, we illustrate the tree traversal process to identify all 4-way interactions, $\bigotimes \boldsymbol{X}_{i=1}^{4}$, with each color denoting a possible interaction path. The legend in the right shows how each interaction path results in a set of 4-way feature interactions. In general, for any user-desired

K, the k.binomialRF algorithm traverses the tree via dynamic tree programming to identify all possible paths from the K -terminal nodes to the root, where $\boldsymbol{K}$ -terminal nodes are all nodes K-steps away from the root node.

Next, we update the hypothesis test in (4) and modify it to identify 2-way interactions for all possible $X_i \otimes X_j$ pairs.

binomialRF 2-way Interaction Selection Hypothesis Test

$H_0: Pr(X_i \otimes X_j) \leq p_{2-way} \quad H_A: Pr(X_i \otimes X_j) > p_{2-way}$	Hypothesis Test	(7-a)
$F_{X_i \otimes X_j} = \sum_{z=1}^{V} F_{X_i \otimes X_j, z} \sim Binomial(V, p_0)$	Test Statistic	(7-b)
$R = \left\{X_i \otimes X_j : F_{ij} < Q_{\alpha, V, p_{2-way}}\right\}$	Rejection Region	(7-c)

3.3 binomialRF: Generalizing to identify *K*-way interactions

To generalize **Equation 6** into multi-way interactions and calculate p_{K-way}, we first note that for any multi-way interaction of size K in a binary split tree results in 2^{K-1} terminal nodes. Therefore, there are 2^{K-1} possible ways of obtaining the K-way interaction (**Figure 4B**). Thus, the normalizing constant in **Equation 6** is replaced to 2^{K-1} in **Equation 8.** The two products in **Equation 6** are now expanded to become K different product terms (each representing the probability of selecting individual features in the interaction set), and $(P-2)$ is replaced with $(P-k)$ to account for sampling without replacement, which yields **Equation 8**.

$$p_{K-way} = \frac{1}{2^{K-1}} \prod_{k=1}^{K} \left(1 - \left(\prod_{g=1}^{m} \frac{(P-k)-g}{(P-k)-(g-1)} \left(\frac{1}{m}\right) \right) \right) \tag{8}$$

Next, we update the hypothesis test in **Equation 7** and modify it to identify 2-way interactions for all possible $\bigotimes X_{i=1}^{K}$ sets.

$H_0: Pr\left(\bigotimes \boldsymbol{X}_{i=1}^{K}\right) \leq p_{k-way} \quad H_A: Pr\left(\bigotimes \boldsymbol{X}_{i=1}^{K}\right) > p_{k-way}$	Hypothesis Test	(9-a)
$F_{\bigotimes X_{i=1}^{K}} = \sum_{z=1}^{V} F_{\bigotimes X_{i=1}^{K}, z} \sim Binomial(V, p_{K-way})$	Test Statistic	(9-b)
$R = \left\{\bigotimes \boldsymbol{X}_{i=1}^{K} : F_{\bigotimes X_{i=1}^{K}} < Q_{\alpha, V, p_{k-way}}\right\}$	Rejection Region	(9-c)

3.4 Using dynamic tree programming to search for interactions

Let any node at the Kth level of a tree be a "K -terminal" node. Binary split trees have exactly two child nodes for every non-terminal node; therefore, to climb up from every K -terminal node to the root, we can calculate the path recursively back to the root. This is done by traveling up a node's parent node until we reach the root. The climbToRoot algorithm is provided in the Appendix (A2), and a pseudo-algorithm is provided below to illustrate the main concepts.

Pseudo-Algorithm

- Identify K-terminal (child) node
- For each K-terminal (child) node
    - Initialize each interaction path as the child node
    - While (child node ≠ root node)
        - Determine if child node is left or right daughter node
        - New.parent.node ← Identify parent node
        - Append new.parent.node to interaction path
        - Child.node ← new.parent.node
    - Return (interaction path)
- Return (all 2^{K-1} K-set interaction paths)

3.5 Evaluations in Numerical Studies

To understand the strengths and limitations of the binomialRF feature interaction selection algorithm, we conducted various small-scale numerical studies in which an interaction or set of interactions were seeded and evaluated how well the binomialRF algorithm detected the interaction. In contrast to the main effects numerical studies that examined signal-to-noise ratio across different dimensions, this focus was to evaluate whether binomialRF could detect the interaction structures in absence of explicitly mapping them into the design matrix.

3.6 Evaluations in Clinical Studies

Biological and genomics analyses that omit interactions will assume biology occurs in isolation; thus, to fully validate the binomialRF algorithm, we extend the clinical study from traditional biomarker discovery to biomarker interactions. The prior study by Gardeux[23] examined a pathway-level random forest classifier, and in this component of the evaluation, we consider pathway-pathway interactions.

As described in Section 2.9, the clinical trial datasets each contained approximately 3,000 pathways, yielding approximately $\binom{3000}{2} = 4{,}498{,}500$ possible 2-way interaction combinations. A brute-force approximation would require storing a matrix 1,500 times larger in order to give every 2-way combination an equal chance of attaining significance. In our case study, we show how to use binomialRF as an interaction screening process to drastically reduce the computation time.

4 Complexity Analysis

The algorithm binomialRF has a 2-way computational efficiency as compared to traditional feature selection methods in random forests since a) it attains a minimal computation complexity, and b) it requires minimal memory storage during runtime. Section 4.1 describes the memory required at runtime, while Section 4.2 notes the theoretical computational complexity and conducts some studies to show computational gain over the state of the art.

4.1 Memory Storage Requirements

To illustrate the magnitude of memory gained by binomialRF, we use a simple case with 10 variables to show how much more memory is required to calculate 2-way interactions. As seen in **Table 4,** to calculate 2-way interactions, binomialRF only requires an $n \times 10$ matrix whereas any other technique would require an $n \times 55$, effectively 5 times more RAM during runtime. To calculate 2-way interactions in a moderately larger dataset with 1000 variables, it would require approximately 500 times more memory. **Table 4** illustrates the relative memory requirements for calculating 2-way and 3-way interactions when there are 10, 100, and 1000 predictors in the design matrix, X.

Table 4: Required Memory Storage for 2-way and 3-way computations. One advantage of the binomialRF algorithm is that it can screen for sets of gene interactions in a memory efficient manner by only requiring a constant-sized matrix

Dimension P	**K = Interaction Size**	**Dim(X)**	**Dim ($\bigotimes X_{i=1}^{K}$)**	**Memory Gain = $\frac{\text{Dim}(\bigotimes X_{i=1}^{K})}{Dim(X)}$**
10	2	N x 10	N x 55	~ 5
	3		N x 175	~ 17
100	2	N x 100	N x 5050	~ 50
	3		N x 166750	~ 1,700
1000	2	N x 1000	N x 500500	~ 500
	3		N x 166667500	~ 170,000

Thus, the memory storage gains are not trivial for even simple 2-way interactions, let alone K-way interactions. Note that in linear models, efficient solution paths for $\bigotimes X_{i=1}^{K}$ only exist for $K \in \{1,2\}$ (LASSO[16] for K=1 and RAMP[27] for K=2). For K>2, no algorithm guarantees computational efficiency. In RF-based feature selection techniques examined in this paper, no efficient solutions exist to scalably identify interactions.

4.2 Computational Complexity

The computational complexity of detecting interaction is on the order of $O(V2^{K-1})$ where V is the number of trees grown in the forest, and K (usually small) is the depth of the interaction search in a binary split tree. For example, in order to calculate 2-way interactions, the algorithm complexity requires only twice as many operations as for main effects, rather than $\binom{P}{2}$ more operations (one for each explicitly-mapped interaction). As seen in the clinical study validation, when P=3,000 pathways, looking at 2-way interaction screening only requires ~6,000 calculations, rather than $\binom{3000}{2} = 4{,}498{,}500$ brute-force calculations. For a permutation-based algorithm, this would mean in an additional 4,498,500 permutation-tests to determine interaction significance. This substantial computational gain occurs because a decision tree's binary split limits the search space drastically, at most growing by powers of 2. This is still exponential growth, however, since interactions are usually limited to measuring 2-way or 3-way interactions, the 2^{K-1} term is for all practical purposes a constant, thus empowering the binomialRF interaction search to provide a quasi-linear approximation to a non-polynomial time problem.

5 Numerical results

In this section, each method is evaluated based on its computation speed (runtime), classification accuracy, and model selection. To evaluate computation time, each method's runtime is reported in total seconds with the range of performances displayed in a boxplot; to evaluate classification accuracy, the standard 0-1 unweighted classification loss function is used; and to evaluate model selection, false selection rates and discovery rates will be used to determine how well the techniques recover the 'true' model.

5.1 Simulation study: Main effects

We generate a simple simulated logistic dataset by generating a multivariate standard normal feature matrix composed of P features and 100 data points. The true model consists of a β vector where the first five coefficients are non-zero and the last P-5 are zero. Finally, the binomial outcomes are generated using a logistic transformation to calculate the probabilities for the Bernoulli random variable. We simulate various signal-to-noise ratio settings in two different ways. First, we let the five nonzero coefficients be either all 1s or all 3s to determine if the algorithm is robust to decreasing the magnitude of the coefficients in the logistic model. Second, we add noise by increasing the number of irrelevant features relative to the five nonzero 'true' variables. We increase the dimension of P 10-fold while keeping constant the number of nonzero coefficients to increase the relative noise in the predictor matrix. Formally, the simulation structure is illustrated below:

$$\beta = [\text{ nonzero coefficients } \in \{1_5, 3_5\} \qquad 0_{P-5}], \quad \forall\, P \in \{10, 100, 1000\} \tag{10}$$
$$X_{100\times P} \sim MVN(0_{10}, I_{10x10}))$$
$$Y \sim Binomial\left(n = 100, \qquad prob = \frac{1}{1+e^{-X^T\beta}}\right)$$

5.1.1 Computation time

In order to compare the computation time of each model as accurately as possible, we strictly measure the time it takes for the algorithm to produce a feature ranking or p-value and omit all other portions of the algorithm using the base system.time R function as seen below:

```
1. ## binomialRF time profile
2. binomialRf.time = system.time(binom.rf <-binomialRF(X,y, …))
```

In order to measure scalability in the predictor space, 500 random forest objects are grown with 500 trees, measuring each algorithm's run time. We repeat this experiment expanding the dimension of $X_{N\times P}$ 10-fold each time and measure the runtime as the predictor matrix increases dimension. The runtimes are graphically summarized in powers of 10 (i.e., $\log_{10}$) in **Figure 5** as the larger runtimes otherwise dominate the boxplots and do not allow for visual differentiation between all techniques. The tests were all run on a 2017 MacBook Pro laptop with 16GB RAM, Intel Core i5, with 4 cores using R version 3.4.0 (64-bit).

The runtimes are shown in the boxplots in **Figure 5**, ranked left to right by median runtime. The binomialRF is consistently the fastest tree-based feature selection algorithm while the binomialRF model averaging is on average the second fastest algorithm. Note, the model averaging algorithm considers 10 different candidate models to perform feature selection, and

the runtime reported measures the time it took to average the results of all 10 candidate models. We increased the number of features from 10 to 1000 to mimic a high dimensional space (where $P >> N$) and asses which techniques scale well in high dimension. In the high-dimensional setting of $P = 1{,}000$, the binomialRF's mean runtime was 0.96 seconds while other techniques required on average between 30 to 600 seconds per run to analyze 1,000 features. A tabular summary of the runtimes was omitted to remove redundancy and is available upon request.

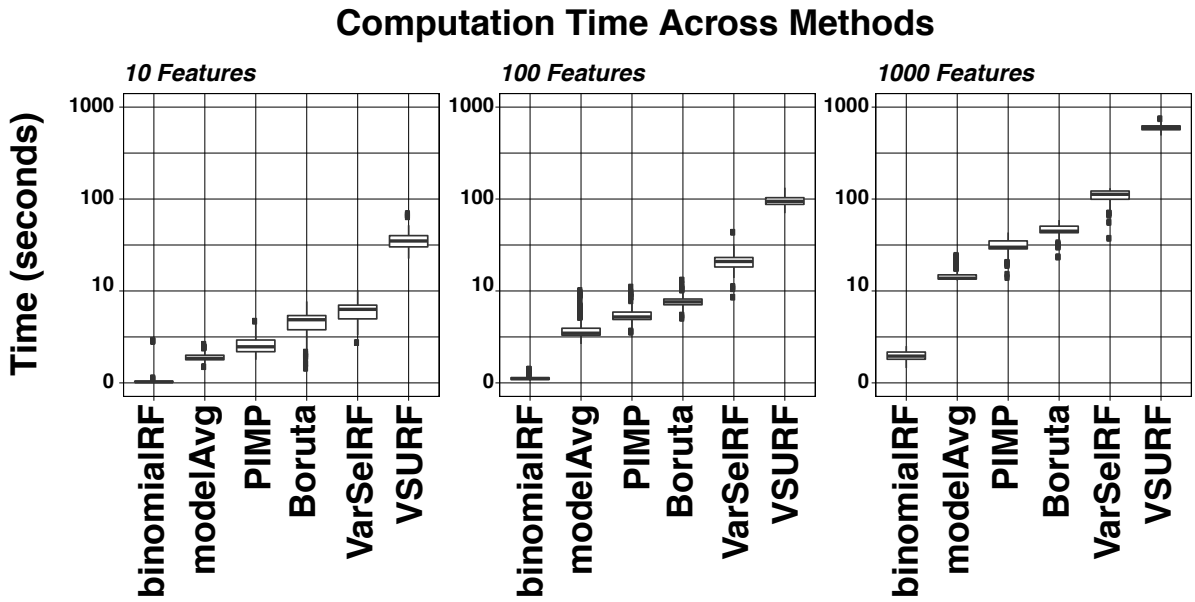


Figure 5: Time profiling across techniques. The simulation scenarios are detailed in Section 3.1, where the length of the coefficient vector, β, varies, but the first five elements are nonzero and P-5 are zero. Rather than measure accuracy and model identifiability criteria across each of the techniques in this set of simulation runs, we only measure the computational runtime of each technique after the main randomForest function call in R. The runtimes are reported in log base 10 (adding +1 to all values to avoid 'negative' runtimes), and all simulations were conducted on a 2017 MacBook Pro with 3.1 GHz Intel Core i5 and 16 GB of RAM. All simulations resulted in binomialRF obtaining the fastest runtimes followed by binomialRF model averaging (denoted 'modelAvg' in graph).

5.1.2 Misclassification test error, model size, false selection rate and coverage

To assess the feature selection techniques, we evaluate them based on the accuracy of the induced final model, average model size, false selection rate (FSR), and true variable coverage. The test error is measured via a standard 0/1 loss function, and the FSR and Coverage formulas are defined below:

$$\text{FSR} = \frac{U}{U+I+1} \qquad \text{Coverage} = \frac{I_{Selected}}{I_{Total}} \tag{11}$$

where U = uninformative variables in model, I = informative variables in model, $I_{selected}$ is the number of "true" features each algorithm detected, and $I_{Total} = 5$ is the number of true

predictors. The +1 is added in the denominator of the FSR formula to ensure non-degenerative cases when no variables are selected (or alternatively viewed as an intercept in a linear model). Results are shown in **Figure 6.** Across the majority of the simulations, the final binomialRF-induced model, on average, results in a lower test error and attains the highest true variable coverage. However, the Boruta algorithm consistently attains the best FSR. In high dimension, the binomialRF trade-off results in a higher FSR in order to attain better coverage.

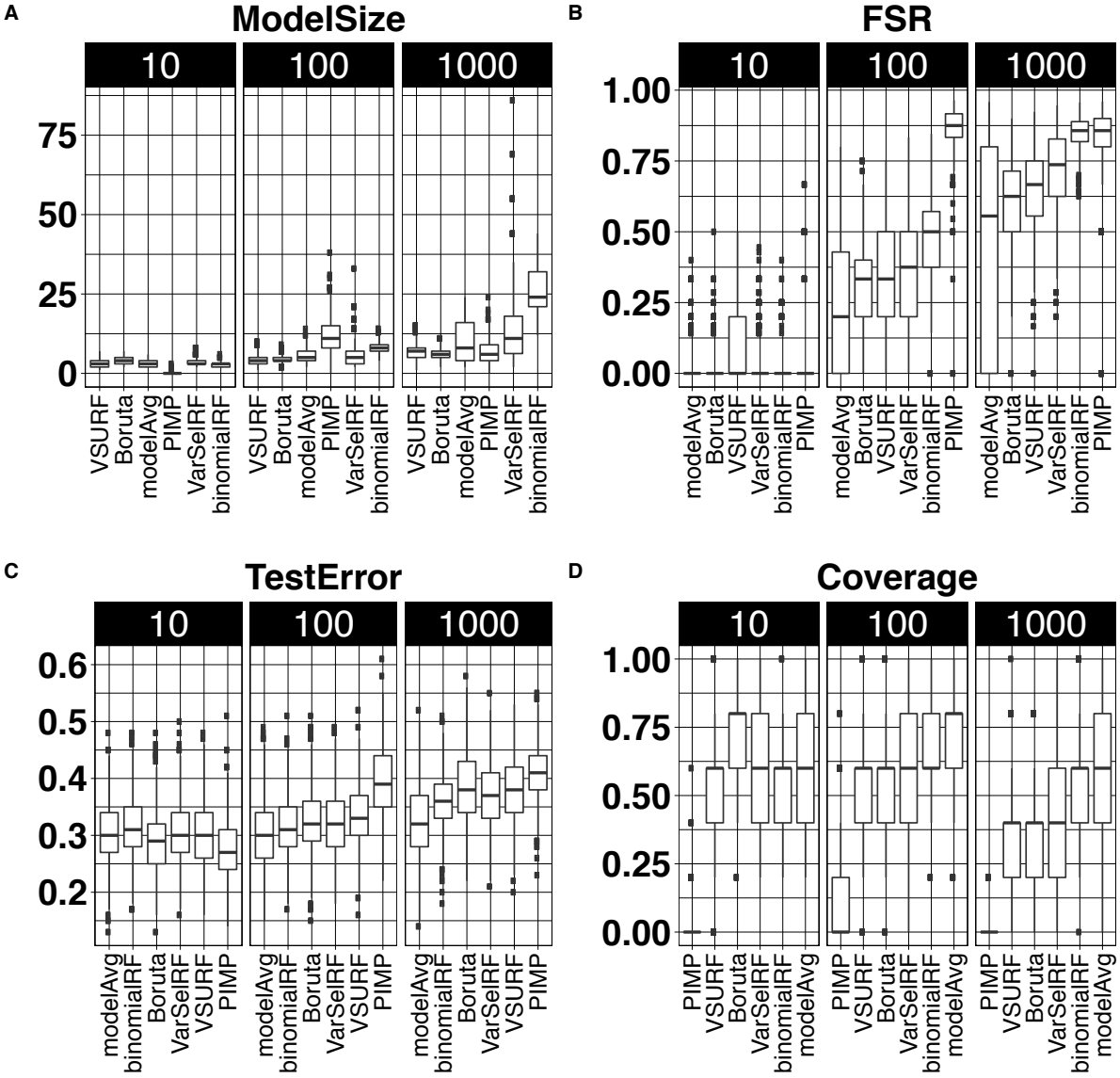


Figure 6: Main Effects Simulation Study. Each technique is ranked using model size, false selection rate, true coverage and test error as the evaluation criteria. In each panel, the model size, false selection rate (FSR), coverage rate, and testing errors are reported and each technique. (A) Each technique selects their 'model', based on the features they selected, and

then the test error is shown, repeating this process for P=10, 100, and 1000 features. (B) and (C) measure how often each technique selects true and phony variables, where lower values (B) and higher values in Panel C indicate greater model selection accuracy. Finally, (D) records the model size to put into context the values in (B) and (C). Across these simulations, the binomialRF model averaging selection, on average, resulted in the lowest test error and optimal model selection (i.e., maximized the true variable coverage controlled for false selection even in high noise settings).

Table 5: Simulation Results for *P* = 1000. Averaged results (with 1 sd) are reported for the high-dimensional (*P*=1000) where there is the largest signal-to-noise ratio as there are 5 true features and 995 noisy features. As shown, the binomialRF algorithm attains high coverage but at the expense of high false selection in high dimensional settings whereas Boruta trades lower coverage for sparser, more accurate models. The binomialRF model averaging selection, however, is able to better control for false selection and maximize coverage, resulting in a better classification accuracy in the test set.

Model	Coverage	FSR	Model Size	Test Error
Boruta	0.33 (0.19)	0.61 (0.15)	**6 (1.96)**	0.38 (0.06)
PIMP	0 (0.03)	0.8 (0.19)	6.7 (3.9)	0.41 (0.05)
VSURF	0.33 (0.19)	0.64 (0.16)	6.87 (2.48)	0.38 (0.06)
VarSelRF	0.45 (0.22)	0.71 (0.15)	13.44 (10.19)	0.37 (0.06)
binomialRF	0.57 (0.19)	0.85 (0.05)	25.97 (7.11)	0.36 (0.05)
binomialRF model averaging	**0.6 (0.21)**	**0.41 (0.39)**	9.93 (6.6)	**0.32 (0.07)**

5.2 **Simulation study:** 2-way Interactions

To determine whether binomialRF can successfully detect interactions, we ran two different simulation studies. The first study considers the following model:

$$X \sim MVN(0_{10}, I_{10x10})) \qquad (12)$$
$$z = 1 + X_1 + X_3 + (X_1 \otimes X_3)$$
$$Y \sim Binomial\left(n = 100, \quad prob = \frac{1}{1 + e^{-z}}\right)$$

500 different simulation runs were conducted, each growing 500 trees in the forest. The averaged results for the top 10 interactions (ranked in order of significance) are shown in **Table 6-A**. As shown in **Table 6**, the binomialRF is clearly detecting the signal in the $X_1 \otimes X_3$ interaction at a higher rate than random combinations of main effect signals and noise (i.e., $X_1 \otimes X_8$). Thus, in the simple 2-way interaction, the binomialRF model clearly detects the signal. The second interaction simulation study considers the following slightly more complicated model where two sets of 2-way interactions are now present:

$$X \sim MVN(0_{10}, I_{10x10})) \qquad (13)$$

$$z = 1 + X_1 + X_3 + (X_1 \otimes X_3) + X_2 + X_4 + (X_2 \otimes X_4)$$
$$Y \sim Binomial\left(n = 100, \quad prob = \frac{1}{1 + e^{-z}}\right)$$

Similarly, as before, 500 trees were grown, and 500 different simulation runs were conducted. The averaged results for the top 10 interactions (ranked in order of significance) are shown in **Table 6-B.**

Table 6: Interaction Simulation Studies. Panel A.) Simulation Study 1 (equation 12). The only interaction that was seeded in the simulated data was the $X_1 \otimes X_3$ interaction, and our feature selection algorithm is detecting it as the only true significant interaction and disregarding everything else. **Panel B.) Simulation Study 2 (equation 13).** The binomialRF is detecting the signal in the $X_1 \otimes X_3$ and $X_2 \otimes X_4$ interactions (which were the true interactions seeded in the simulation) at a higher rate than random combinations of main effect signals and noise, which are the true interactions seeded in the data. However, since $X_1, X_2, X_3,$ and X_4 have such a dominant signal, binomialRF is also suggesting we screen for all possible 2-way interactions between them (in yellow). Thus, since in real-life datasets, the true solution is never known, the binomialRF algorithm will detect a signal where it can and screen for potentially meaningful interactions where they arise, and the binomialRF model averaging framework can then be applied to parse through the noise.

A.) Simulation Study 1
Interaction seeded: $X_1 \otimes X_3$

Interaction	Frequency	Adj.Pval
$\boldsymbol{X_1 \otimes X_3}$	**130.1**	**4.41E-06**
$X_1 \otimes X_{10}$	34.3	0.074
$X_3 \otimes X_9$	31.7	0.091
$X_1 \otimes X_4$	32.6	0.093
$X_1 \otimes X_5$	32.5	0.094
$X_1 \otimes X_7$	32.4	0.095
$X_2 \otimes X_3$	30.3	0.096
$X_3 \otimes X_4$	31.9	0.010
$X_1 \otimes X_8$	32.5	0.10
$X_1 \otimes X_6$	35.8	0.10

B.) Simulation Study 2
Interactions seeded:
$X_1 \otimes X_3$ and $X_2 \otimes X_4$

Interaction	Frequency	Adj.Pval
$\boldsymbol{X_2 \otimes X_4}$	**72.68**	**0.008**
$\boldsymbol{X_1 \otimes X_3}$	**70.46**	**0.02**
$X_3 \otimes X_4$	59.87	0.03
$X_1 \otimes X_2$	56.14	0.03
$X_1 \otimes X_4$	60.32	0.04
$X_2 \otimes X_3$	57.25	0.06
$X_4 \otimes X_6$	18.95	0.28
$X_4 \otimes X_9$	19.55	0.31
$X_4 \otimes X_7$	21.27	0.31
$X_1 \otimes X_8$	18.32	0.31

6 Asthma Validation study

6.1 Asthma Validation study: Identifying Pathways

The study by Gardeux et al [23] was designed to assess whether it was possible to develop a fully-specified classifier from healthy adults infected with HRV that predicted asthma exacerbation in pediatric asthmatic patients who were infected with HRV. They used a variety of

machine learning techniques and restricted their classifier to consider 10, 20, and 30 pathways at a time. The optimal classifier developed was a fully-specified random forest classifier that attained approximately 73 percent accuracy in the pediatric asthmatic validation-set cohort. Since our goal is to not develop the optimal classifier but rather provide more interpretability and assess interactions, we conducted the asthma clinical validation study in a slightly different fashion following the same overarching concepts. As shown in **Table 3**, the healthy adult [training] dataset was used to identify meaningful pathways, and the pediatric asthmatic [validation] cohort was used to confirm their utility. We first ran the binomialRF algorithm on the healthy adults considering all 3014 pathways and validated its selected pathways in the asthmatic patients (shown in Appendix A.3). The binomialRF algorithm identified 67 significant pathways at FDR < 5%. 19 out of the 20 pathways identified by [23] were confirmed by binomialRF while a number of other candidate pathways identified by binomialRF may hold predictive power and may be physiologically related to the pathway classes identified in [23]. For example, "GO:0006342" is a "chromatin silencing" pathway and this pathway is physiologically related to Class "V-Chromatin Organization." Another such example is "GO:0001763", a "morphogenesis of a branching structure" pathway that falls under the "Class III - Morphogenesis" pathway class. Thus, binomialRF is able to confirm known physiologically discoveries as well as propose biologically-feasible novel candidate pathways for predicting HRV-induced asthma exacerbation. A binomialRF model averaging analysis was also conducted – the pathway selection results from the model averaging are located in Appendix **A4. binomialRF Model Averaging in Asthma Validation Study**. A total of 10 candidate models were proposed by randomly selecting between 200 and 3,000 pathways per model. Each pathway was then ranked based on its "Proportion Selected" value, and the pathways identified in more than half the models (i.e., Proportion Selected >0.5) were selected into the final model.

6.2 Asthma Validation study: Classification Accuracy

To validate these results in the classifier, we also compared the classification accuracy on the pediatric asthma patients using the pathways selected by the binomialRF model against that of a random forest trained on all 3,014 pathways. We then compared the binomialRF model averaging (see **Appendix A4**) to the results obtained by the naïve random forest and the binomialRF model. The optimal binomialRF model averaging selection resulted in a classifier with a 67% and 70% mean and median classification accuracy, respectively. It is worth noting that although the validation accuracy did not surpass the 73% obtained by Gardeux [23], the binomialRF is a feature selection framework for which classification accuracy is an important outcome but not the only objective.

6.3 Asthma Validation study: Identifying Pathway-Pathway Interaction

To extend the Asthma Validation study from Gardeux et al [23], we also screen for pathway-pathway interactions. To do dimension reduction and reduce the search space for interactions, the binomialRF algorithm was first ran (prediction set and testing set) with 67 pathways selected by a previous run of binomialRF for single factors. Using these 67 pathways, we then screened for all possible 2-way interactions using the 2-way binomialRF algorithm and identified 107 pathway-pathway interactions (interactions are listed in **Appendix A5,** ranked in order of significance)**.** These results indicate all the significant interactions that occur between 'main effects' (or pathways deemed significant on their own). As a first step, exploring interactions

between ‘main effects’ allows for a bioinformatician to explore “expected” interactions, which are interactions between the main pathways at play. Restricting the interaction space to the 67 pathways identified by the first run of binomialRF greatly reduces the computational burden, but it is worth noting that this type of search leads to a confirmation bias as we only allow the algorithm to search for interactions in pathways in which we expect interactions, rather than allow for the algorithm to also search for unexpected interactions. Therefore, a proper interaction search would be more robust under a model averaging setting in order to determine which interactions are consistently important versus those that are pure noise.

7 Discussion

7.1 Feature selection in random forests

In general, machine learning algorithms are judged on three main concepts: interpretability, scalability, and accuracy. The binomialRF feature selection algorithm provides a relatively simple interpretation, attains competitive model identifiability and accuracy performance, and provides a framework that easily scales in high dimensional main effect and interaction settings. Furthermore, it offers a distinct advantage compared to other feature selection algorithms in trees, as it easily generalizes to search for interactions using the same binomial exact test. The heuristic rankings based on the random forest importance measure (like RFE and VarSelRF) clearly cannot search for interactions since they are restricted to univariate analyses, and iterations can only eliminate non-important univariate relationships, thus providing zero information on interactions. The permutation-based tests (like the Permutation-Importance algorithm) do not generalize (easily at least) to search for interactions since the literature demonstrates that it is challenging enough as is to estimate empirical null distributions of Z-scores of importance metrics (Vita, Altmann, etc..), let alone estimating joint Z-score distributions, or how to construct scalable and efficient algorithms to estimate them.

P-value based rankings offer statistical rigor in machine learning-based models as they allow for more robust measures of gene selection rather than arbitrary cutoffs (i.e., “we selected the top 10%”), while permutation-type tests shuffle predictor values across nodes in a tree which are expensive operations and disregard node hierarchy in trees. In the former, techniques with heuristic stopping criteria are subject to arbitrary user-imposed decisions which lack scientific rigor (i.e., why is selecting the top 10% of genes more scientifically valid than the top 20%?). In the latter, feature selection techniques that disregard the innate hierarchy in tree-based structures are losing valuable signal. In the construction of a decision tree, the predictor chosen as the “splitting variable” in each node is the optimal predictor choice. It intuitively follows that features selected higher up the hierarchy of the tree have more weight as they are deemed to be “optimal” in a larger pool of samples, and the root node is theoretically the best feature in each tree as it is the “first best” feature selected. Therefore, the hierarchy must be respected.

The power of model averaging results from the same principle that allows random forests to be powerful classifiers: an ensemble of weak learners is stronger than an individual strong classifier. A prediction made from a majority vote or consensus of weak decision trees (i.e., a random forest) is more robust than a well-pruned decision tree as the ensemble minimizes the effects of mistakes done by individual trees as long as the majority make the right choice. If we extend this to feature selection, then the same occurs with binomialRF. An individual run of binomialRF

might be sensitive to noise; however, if we rely instead on the consensus of multiple iterations of binomialRF and add some randomization (both at the feature and sample-size level), then in order for a feature to be selected into the final model, it has to be selected across the majority of the candidate models. Therefore, this makes the final model selection more robust and stable. The binomialRF model averaging runs across the simulation studies resulted in the best classification rate in the test sets while controlling for false selection and maximizing true feature coverage. These results confirm the intuition that extending ensemble techniques from classifier development into feature and model selection will improve the latter. It is also worth noting that while model averaging is independent of the feature selection technique (and could have been applied to all other methods), we will explore whether the binomialRF (and largely the binomial distribution) framework offers theoretical advantages to propose any theoretical asymptotic results (see limitations and future studies).

7.2 Pathway-Pathway Interactions in the Asthma Validation Study

The first interaction listed in the supplemental table **(A5)** is GO:0016570 $\otimes$ GO:0009581. GO:0016570 is a histone protein modification pathway while GO:0009581 is a pathway indicating a response to external stimulus. Their interaction indicates that differential expression of pathways associated to histone protein regulation is interacting with response to an external stimulus (likely the HRV-inoculation stimulus). In fact, a few pathway interactions down the list we see GO:0016570 $\otimes$ GO:0009615, which suggests that differential histone modification in response to a virus is predictive of recurrent asthma exacerbations as well as of healthy subjects' symptoms to HRV (plausibly, the human rhinovirus infection). Indeed, histone modification has been linked to the development of asthma[28, 29] and HRV infection has also been shown to cause DNA methylation changes in epithelial cells of healthy and asthmatic subjects[30]. These two pathway-pathway interactions indicate that histone modifications are potentially highly susceptible to environmental stimuli, suggesting an epigenetic component to asthmatic children's response to therapy. The previous "genome by environment" classifier by Gardeux et al as well as the epigenetic literature in asthma corroborate the existence of these "genome by environment" interactions [23, 31-33], illustrating the utility of looking for pathway-pathway interactions beyond single pathway response effect that they reported. The pathway-pathway interaction screening can then be used to corroborate known biological phenomenon as well as potentially shed light on previously unknown interacting mechanisms.

8 Limitations and Future Studies

8.1 Pushing towards more theoretical guarantees in machine learning

The binomialRF technique framework provides a novel paradigm shift that can be extended into multiple directions. On one hand, binomialRF can be extended into a Bayesian framework by placing priors (the current implementation enforces an equal-weighted discrete uniform prior) on the likelihood of selecting a feature and determining significance using posterior probabilities on a beta-binomial process. On the other hand, the binomialRF algorithm can be extended into a binomialRF model averaging framework where candidate models comprised of feature subsets can be assessed and 'averaged' across. Similar to Bayesian Model Averaging (BMA)[34] and Sparsity Orientated Importance Learning (SOIL)[35], the binomialRF can weight candidate model based on its utility, however, in the model-free case, a likelihood-induced weighting is not

possible so we can alternatively weigh by out-of-bag (or validation) error. Since the model averaging data is composed of binomial random variable test statistics, future work will explore whether any asymptotic results occur if we let the number of candidate models go to infinity. At the moment, model averaging still requires arbitrary cutoffs (for our simulations studies experiments we used Proportion Selected > 0.9 as our cutoffs) to make the final model selection, with empirical results suggesting it helps reduce false selection rate without sacrificing true feature coverage. However, these results are still empirical and offer no theoretical guarantees. We need stronger theoretical results to inform which cutoffs to use when considering a specific number of candidate models. Ideally, it would allow us to guarantee model selection results (as is the case with SOIL), however at the moment this is not guaranteed beyond a few empirical studies and will thus be explored in future studies.

8.2 Improving interpretative power and translational utility: Incorporating pathways and ontologies in feature selection

On the other hand, as pathway-based biomarker studies gain more traction in the genomics realm [24, 36, 37], the machine learning community needs to continue developing domain-specific methods that can cater to the bioinformatics and genomics research community. These techniques must be further explored in order to improve the translational power and interpretation of machine learning results in bioinformatics. Too often we develop power predictive "black box" algorithms that lack explanation or interpretive power that is required to translate this information into knowledge. Therefore, future work must be prioritized in this direction. One possible direction to consider is ontology-enriched binomialRF models. Ontologies offer well-curated knowledge graphs that represent complex interplay of biological networks. Incorporating this information beforehand into the feature selection method, and later sending the results back into the ontology-domain for visualization can yield more interesting network-level analyses. Since feature selection in the binomialRF are composed of binomial distribution test statistics, there are numerous statistical possibilities with which one can enrich gene-based binomialRF predictions into aggregated pathway-level features. For example, one way would be to extend gene detection to ontology-based pathway-level analyses via over-representation tests from binomial test statistics. Another would incorporate gene ontology hierarchies between pathways to eliminate redundant signal and incorporate ontologies into smarter pathway detection. Model averaging can also be conducted by incorporating knowledge graphs to make the 'candidate' models more ontologically meaningful by looking at clusters of genes or pathways or identifying which elements dominate the signal in a biological process. We will explore these in future studies.

9 Conclusion

As the biomarker discovery process moves away from identifying single-gene products and moves towards interactions and pathways (say from gene ontologies like GO), the statistical machine learning community will need to continue to develop corresponding interpretable and scalable techniques. The binomialRF algorithm provides an early step in this direction in order to match the technical and computational requirements for these novel large-scale genomics analyses, as well as to extend to other 'omics.

10 List of acronyms

- RF = random forest
- BH = Benjamini Hochberg adjustment
- BY = Benjamini Yekutieli adjustment
- BMA = Bayesian Model Averaging
- SOIL = Sparsity Oriented Importance Learning
- HRV = Human Rhinovirus
- FSR = False Selection Rate
- LASSO = Least Absolute Shrinkage and Selection Operator
- RAMP = regularization algorithm under marginality principle
- DTP = Dynamic Tree Programming
- GO = Gene Ontology
- GO-BP= Gene Ontology Biological Processes

11 Conflict of Interest

The authors declare no conflict of interest.

12 Author Contributions

SRZ conducted all the analyses in R; HHZ contributed to the statistical framework and analysis; all authors contributed to the evaluation and interpretation of the study; SRZ contributed to the figures and tables; SRZ, HHZ, CK, and YAL contributed to the writing of the manuscript; all authors read and approved the final manuscript.

13 Funding

This work was supported in part by The University of Arizona Health Sciences Center for Biomedical Informatics and Biostatistics, the BIO5 Institute, and the NIH (U01AI122275, NCI P30CA023074, 1UG3OD023171). This article did not receive sponsorship for publication.

References

1. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
2. Chen, X. and H. Ishwaran, *Random forests for genomic data analysis.* Genomics, 2012. **99**(6): p. 323-329.
3. Bienkowska, J.R., et al., *Convergent Random Forest predictor: methodology for predicting drug response from genome-scale data applied to anti-TNF response.* Genomics, 2009. **94**(6): p. 423-432.
4. Boulesteix, A.L., et al., *Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012. **2**(6): p. 493-507.
5. Diaz-Uriarte, R., *GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest.* BMC bioinformatics, 2007. **8**(1): p. 328.
6. Díaz-Uriarte, R. and S.A. De Andres, *Gene selection and classification of microarray data using random forest.* BMC bioinformatics, 2006. **7**(1): p. 3.
7. Goldstein, B.A., et al., *An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings.* BMC genetics, 2010. **11**(1): p. 49.
8. Izmirlian, G., *Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial.* Annals of the New York Academy of Sciences, 2004. **1020**(1): p. 154-174.
9. Jiang, P., et al., *MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.* Nucleic acids research, 2007. **35**(suppl_2): p. W339-W344.
10. Archer, K.J. and R.V. Kimes, *Empirical characterization of random forest variable importance measures.* Computational Statistics & Data Analysis, 2008. **52**(4): p. 2249-2260.

11. Genuer, R., J.-M. Poggi, and C. Tuleau-Malot, *VSURF: an R package for variable selection using random forests.* The R Journal, 2015. **7**(2): p. 19-33.
12. Szymczak, S., et al., *r2VIM: A new variable selection method for random forests in genome-wide association studies.* BioData mining, 2016. **9**(1): p. 7.
13. Kursa, M.B. and W.R. Rudnicki, *Feature selection with the Boruta package.* J Stat Softw, 2010. **36**(11): p. 1-13.
14. Altmann, A., et al., *Permutation importance: a corrected feature importance measure.* Bioinformatics, 2010. **26**(10): p. 1340-1347.
15. Chipman, H.A., E.I. George, and R.E. McCulloch, *BART: Bayesian additive regression trees.* The Annals of Applied Statistics, 2010. **4**(1): p. 266-298.
16. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.
17. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the royal statistical society: series B (statistical methodology), 2005. **67**(2): p. 301-320.
18. Meier, L., S. Van De Geer, and P. Bühlmann, *The group lasso for logistic regression.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008. **70**(1): p. 53-71.
19. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.
20. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
21. Benjamini, Y. and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency.* The annals of statistics, 2001. **29**(4): p. 1165-1188.
22. Zaas, A.K., et al., *Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans.* Cell host & microbe, 2009. **6**(3): p. 207-217.
23. Gardeux, V., et al., *A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations.* Journal of the American Medical Informatics Association, 2017. **24**(6): p. 1116-1126.
24. Gardeux, V., et al., *'N-of-1-pathways' unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine.* Journal of the American Medical Informatics Association, 2014. **21**(6): p. 1015-1025.
25. Nelder, J.A., *The selection of terms in response-surface models—how strong is the weak-heredity principle?* The American Statistician, 1998. **52**(4): p. 315-318.
26. Choi, N.H., W. Li, and J. Zhu, *Variable selection with the strong heredity constraint and its oracle property.* Journal of the American Statistical Association, 2010. **105**(489): p. 354-364.
27. Hao, N., Y. Feng, and H.H. Zhang, *Model selection for high-dimensional quadratic regression via regularization.* Journal of the American Statistical Association, 2018. **113**(522): p. 615-625.
28. Martino, D. and S. Prescott, *Epigenetics and prenatal influences on asthma and allergic airways disease.* Chest, 2011. **139**(3): p. 640-647.

29. Yang, I.V. and D.A. Schwartz, *Epigenetic mechanisms and the development of asthma.* Journal of allergy and clinical immunology, 2012. **130**(6): p. 1243-1255.
30. McErlean, P., et al., *Human rhinovirus infection causes different DNA methylation changes in nasal epithelial cells from healthy and asthmatic subjects.* BMC medical genomics, 2014. **7**(1): p. 37.
31. Kidd, C.D., et al., *Histone modifications and asthma. the interface of the epigenetic and genetic landscapes.* American journal of respiratory cell and molecular biology, 2016. **54**(1): p. 3-12.
32. Miller, R.L. and S.-m. Ho, *Environmental epigenetics and asthma: current concepts and call for studies.* American journal of respiratory and critical care medicine, 2008. **177**(6): p. 567-573.
33. Bégin, P. and K.C. Nadeau, *Epigenetic regulation of asthma and allergic disease.* Allergy, Asthma & Clinical Immunology, 2014. **10**(1): p. 27.
34. Hoeting, J.A., et al., *Bayesian model averaging: a tutorial.* Statistical science, 1999: p. 382-401.
35. Ye, C., Y. Yang, and Y. Yang, *Sparsity Oriented Importance Learning for High-Dimensional Linear Regression.* Journal of the American Statistical Association, 2018. **113**(524): p. 1797-1812.
36. Zaim, S.R., et al., *Evaluating single-subject study methods for personal transcriptomic interpretations to advance precision medicine.* bioRxiv, 2018: p. 428581.
37. Zaim, S.R., et al. *Emergence of pathway-level composite biomarkers from converging gene set signals of heterogeneous transcriptomic responses*. in *Pac. Symp. Biocomput*. 2018. World Scientific.

14 Appendix

A1: binomialRF Feature Selection Algorithm

Let $F_{j,z}$ denote the random variable measuring whether X_j is selected as the splitting variable for the root at the root of a tree T_i,

$$F_{j,z} = \begin{cases} 1\,, & if\ root\ (T_i) = X_j \\ 0\,, & otherwise\,. \end{cases} \tag{1}$$

This results in $F_{j,z}$ being a Bernoulli random variable, $F_{j,z} \sim Bern(p_{root})$, and $F_j = \sum_{z=1}^{V} F_{j,z} \sim Binomial(V, p_{root})$ is a Binomial random variable across all V trees, where $\boldsymbol{p_{root}}$ is the probability of randomly selecting a feature X_j as the optimal splitting variable in the root of tree T_i. $\boldsymbol{p_{root}}$ is given by

$$p_{root} = 1 - \left(\prod_{g=1}^{m} \frac{P-g}{P-(g-1)} \left(\frac{1}{m}\right) \right)$$

Under the null , p_{root} is constant across all trees. The binomialRF feature selection algorithm below illustrates the process of identifying main effects for the binomialRF algorithm.

Algorithm 1 binomialRF feature selection

calculate $p_{root} = \frac{1}{m}\left[1 - \prod_{g=1}^{m} \frac{P-g}{P-(g-1)}\right]$

for z=1:V **do**
-Grow the zth decision tree in the random forest,T_z.
-Observe the $X_{optimal}$ splitting variable at the root of T_z
-construct $F_{j,z}$

$$\text{where} \quad F_{j,z} = \begin{cases} 1 & X_{optimal} = X_j \text{ for root}(T_z) \\ 0 & otherwise \end{cases}$$

end for

- Test Hypothesis $H_0 : Pr(X_j) \leq p_{root}$ $H_A : Pr(X_j) > p_{root}$
- $F_j = \sum_{z=1}^{V} F_{j,z} \sim \text{Binomial}(V, p_{root})$
- Rejection Region $R = \{X_j : F_j < Q_(, V, p_{root})\}$

To generalize to identify K-set interactions, denoted by $\otimes X_{i=1}^{K}$, replace $F_{j,z}$ with

$$F_{\otimes \boldsymbol{X}_{i=1}^{K}, z} = \begin{cases} 1 & \{X_i\}_{i \in K} \\ 0 & otherwise \end{cases},$$

where $\{X_i\}_{i \in K}$ denotes the interaction sequence. Then replace p_{root} with

$$p_{K-way} = \frac{1}{2^{K-1}} \prod_{k=1}^{K} \left(1 - \left(\prod_{g=1}^{m} \frac{(L-k)-g}{(L-k)-(g-1)} \left(\frac{1}{m} \right) \right) \right).$$

Then, $F_{\bigotimes \boldsymbol{X}_{i=1}^{K}} = \sum_{z=1}^{V} F_{\boldsymbol{X}_{\bigotimes}^{K},z} \sim Binomial\left(V, p_{K-way}\right)$, and the hypothesis test follows from this.

A2: climbToRoot (DTP) Algorithm

Note, in binary split trees, the nodes are ordered such that each the label of each left daughter node is twice the label of its parent node, and the label of each right daughter -1 is twice the label of its parent node. Under this such ordering, the K_{terminal} nodes are identified as $K_{\text{terminal}} = \{2^{K-1}: (2^{K} - 1)\}$, and lines 20-23 are required for cases when true terminal (leaf) nodes occur before the K_{terminal} nodes.

```
3.  climbToRoot <- function(final.node, Tree, K) {
4.
5.    Tree = as.data.frame(Tree)
6.    path <- c(final.node)
7.
8.    while(final.node != 1){
9.      if(final.node %% 2 ==0){
10.       new.parent <- which(Tree[,'left daughter'] == final.node)
11.       path <- c(path, new.parent)
12.       final.node <- new.parent
13.
14.     } else if(final.node %% 2 != 0){
15.       new.parent <- which(Tree[,'right daughter'] == final.node)
16.       path <- c(path, new.parent)
17.       final.node <- new.parent
18.     }
19.
20.   }
21.
22.   if(length(path)>K){
23.     new.path <- rev(path)[-1]
24.
25.     return(new.path)
26.
27.   } else {
28.     return(path)
29.   }
30. }
```

A3. Asthma Validation study validation: Predicted Pathways by binomialRF

The table below compares our Asthma Validation study to the original classifier obtained by Gardeux et al's random forest classifier. The pathways below are the ones determined significant by the binomialRF algorithm. The first two columns show the GO pathway identifier and description. The third one determines whether it was validated in the HRV by Gardeux, and

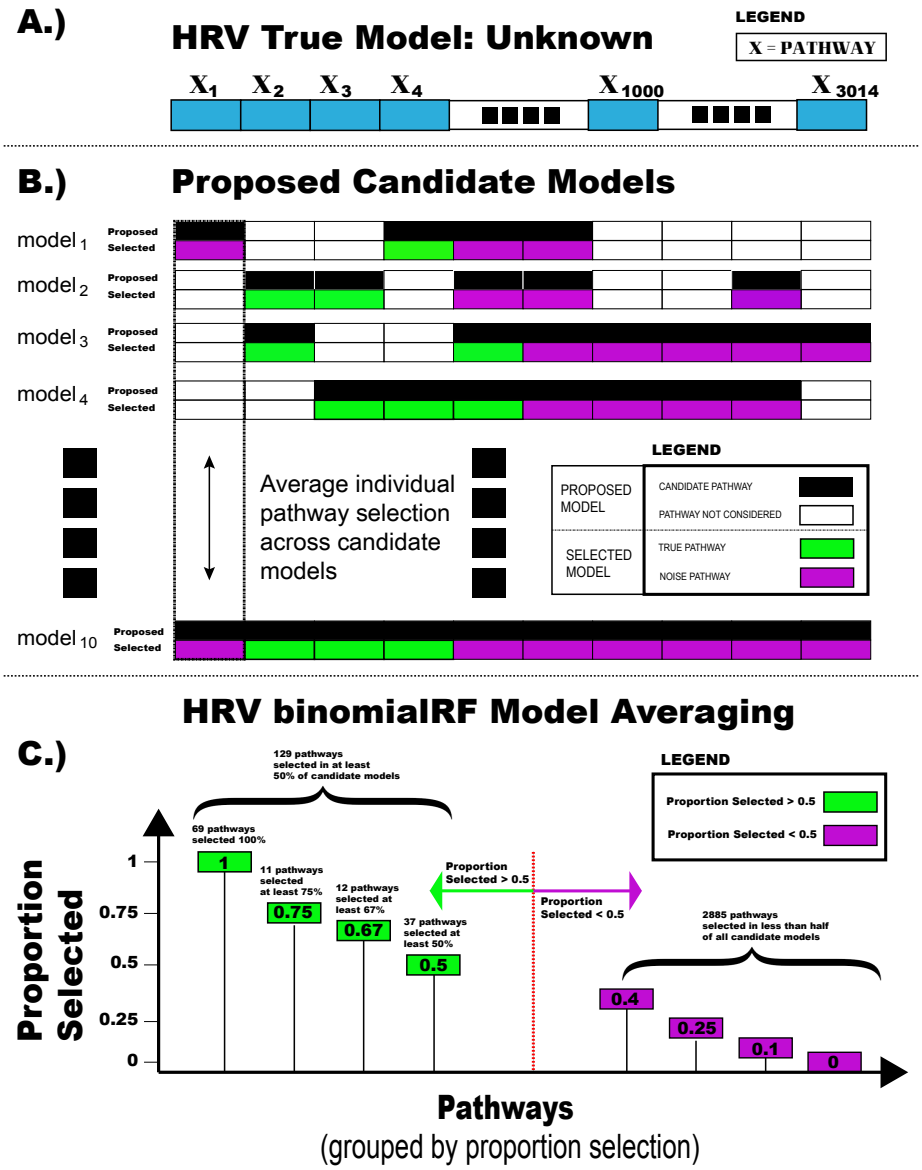
987 the last column corresponds to their 5 distinct pathway classes. As seen by some of the pathways
988 not validated in the HRV study, (e.g., "GO:0001763") even though they were not part of the
989 original discoveries, they correspond to the Class III "Morphogenesis" pathways, thus
990 identifying physiologically relevant and related candidate pathway discoveries.
991

Pathway	**Freq.**	**Pvalue**	**FDR**	**description**
GO:0016570	823	Pvalue < 10^{-6}	FDR< 10^{-6}	histone modification
GO:0016569	221	Pvalue < 10^{-6}	FDR< 10^{-6}	covalent chromatin modification
GO:0045087	197	Pvalue < 10^{-6}	FDR< 10^{-6}	innate immune response
GO:0034340	129	Pvalue < 10^{-6}	FDR< 10^{-6}	response to type I interferon
GO:0060337	124	Pvalue < 10^{-6}	FDR< 10^{-6}	type I interferon-mediated signaling pathway
GO:0071357	122	Pvalue < 10^{-6}	FDR< 10^{-6}	cellular response to type I interferon
GO:0016568	78	Pvalue < 10^{-6}	FDR< 10^{-6}	chromatin modification
GO:0006913	68	Pvalue < 10^{-6}	FDR< 10^{-6}	nucleocytoplasmic transport
GO:0007017	60	Pvalue < 10^{-6}	FDR< 10^{-6}	microtubule-based process
GO:0006325	57	Pvalue < 10^{-6}	FDR< 10^{-6}	chromatin organization
GO:0006954	48	Pvalue < 10^{-6}	FDR< 10^{-6}	inflammatory response
GO:0051028	46	Pvalue < 10^{-6}	FDR< 10^{-6}	mRNA transport
GO:0009617	41	Pvalue < 10^{-6}	FDR< 10^{-6}	response to bacterium
GO:0060688	32	Pvalue < 10^{-6}	FDR< 10^{-6}	regulation of morphogenesis of a branching structure
GO:0007565	31	Pvalue < 10^{-6}	FDR< 10^{-6}	female pregnancy
GO:0009615	29	Pvalue < 10^{-6}	FDR< 10^{-6}	response to virus
GO:0045814	28	Pvalue < 10^{-6}	FDR< 10^{-6}	negative regulation of gene expression, epigenetic
GO:0097028	28	Pvalue < 10^{-6}	FDR< 10^{-6}	dendritic cell differentiation
GO:0048568	27	Pvalue < 10^{-6}	FDR< 10^{-6}	embryonic organ development
GO:0009581	26	Pvalue < 10^{-6}	FDR< 10^{-6}	detection of external stimulus
GO:0042274	26	Pvalue < 10^{-6}	FDR< 10^{-6}	ribosomal small subunit biogenesis
GO:0050900	25	Pvalue < 10^{-6}	FDR< 10^{-6}	leukocyte migration
GO:0051272	23	Pvalue < 10^{-6}	FDR< 10^{-6}	positive regulation of cellular component movement
GO:0046888	21	Pvalue < 10^{-6}	FDR< 10^{-6}	negative regulation of hormone secretion
GO:0006342	17	Pvalue < 10^{-6}	FDR< 10^{-6}	chromatin silencing
GO:0007050	16	Pvalue < 10^{-6}	FDR< 10^{-6}	cell cycle arrest
GO:0000226	15	Pvalue < 10^{-6}	FDR< 10^{-6}	microtubule cytoskeleton organization
GO:0006260	15	Pvalue < 10^{-6}	FDR< 10^{-6}	DNA replication
GO:0048638	15	Pvalue < 10^{-6}	FDR< 10^{-6}	regulation of developmental growth

GO:0006813	14	Pvalue < 10^{-6}	FDR< 10^{-6}	potassium ion transport
GO:0051169	13	Pvalue < 10^{-6}	FDR< 10^{-6}	nuclear transport
GO:0051606	13	Pvalue < 10^{-6}	FDR< 10^{-6}	detection of stimulus
GO:0006363	12	Pvalue < 10^{-6}	FDR< 10^{-6}	termination of RNA polymerase I transcription
GO:0034341	12	Pvalue < 10^{-6}	FDR< 10^{-6}	response to interferon-gamma
GO:0050851	12	Pvalue < 10^{-6}	FDR< 10^{-6}	antigen receptor-mediated signaling pathway
GO:0008015	11	Pvalue < 10^{-6}	FDR< 10^{-6}	blood circulation
GO:0001763	10	Pvalue < 10^{-6}	FDR< 10^{-6}	morphogenesis of a branching structure
GO:0043966	10	Pvalue < 10^{-6}	FDR< 10^{-6}	histone H3 acetylation
GO:0050853	10	Pvalue < 10^{-6}	FDR< 10^{-6}	B cell receptor signaling pathway
GO:0051216	10	Pvalue < 10^{-6}	FDR< 10^{-6}	cartilage development
GO:0060038	10	Pvalue < 10^{-6}	FDR< 10^{-6}	cardiac muscle cell proliferation
GO:0007423	9	Pvalue < 10^{-6}	FDR< 10^{-6}	sensory organ development
GO:0019221	9	Pvalue < 10^{-6}	FDR< 10^{-6}	cytokine-mediated signaling pathway
GO:0050807	9	Pvalue < 10^{-6}	FDR< 10^{-6}	regulation of synapse organization
GO:0001658	8	Pvalue < 10^{-6}	0.000053	branching involved in ureteric bud morphogenesis
GO:0006361	8	Pvalue < 10^{-6}	.000053	transcription initiation from RNA polymerase I promoter
GO:0006473	8	Pvalue < 10^{-6}	.000053	protein acetylation
GO:0007586	8	Pvalue < 10^{-6}	.000053	digestion
GO:0043484	8	Pvalue < 10^{-6}	.000053	regulation of RNA splicing
GO:0001890	7	0.000081	0.00041	placenta development
GO:0002768	7	0.000081	0.00041	immune response-regulating cell surface receptor signaling pathway
GO:2000027	7	0.000081	0.00041	regulation of organ morphogenesis
GO:0000722	6	0.00058	0.0026	telomere maintenance via recombination
GO:0002429	6	0.00058	0.0026	immune response-activating cell surface receptor signaling pathway
GO:0043279	6	0.00058	0.0026	response to alkaloid
GO:0060326	6	0.00058	0.0026	cell chemotaxis
GO:0060351	6	0.00058	0.0026	cartilage development involved in endochondral bone morphogenesis
GO:0070925	6	0.00058	0.0026	organelle assembly
GO:0000910	5	0.0036	0.014	cytokinesis

GO:0007600	5	0.0036	0.014	sensory perception
GO:0018105	5	0.0036	0.014	peptidyl-serine phosphorylation
GO:0018205	5	0.0036	0.014	peptidyl-lysine modification
GO:0035337	5	0.0036	0.014	fatty-acyl-CoA metabolic process
GO:0044057	5	0.0036	0.014	regulation of system process
GO:0046949	5	0.0036	0.014	fatty-acyl-CoA biosynthetic process
GO:0050795	5	0.0036	0.014	regulation of behavior
GO:0050953	5	0.0036	0.014	sensory perception of light stimulus

A4. binomialRF Model Averaging in Asthma Validation Study



Supplemental Figure A4: Model Averaging in the Asthma Validation Study. In contrast to the simulation example in Section 2.7, in the Asthma Validationstudy the "true" model is not known therefore, (A) is unknown, and we use the model averaging framework (B) to make the most educated guess as to which pathways are significant. Out of the 3,014 pathways expressed from the GO-BP ontology, 129 pathways were selected via model averaging and these were used to estimate the validation accuracy in the pediatric asthma validation cohort (C).

A5. Pathway-Pathway Interactions in Asthma Validation Study

Pathway1	Pathway2	Adj.Pval	Description Pathway 1	Description Pathway 2

GO:0016570	GO:0009581	FDR< 10^{-6}	histone modification	detection of external stimulus
GO:0016570	GO:0007423	FDR< 10^{-6}	histone modification	sensory organ development
GO:0016570	GO:0001763	FDR< 10^{-6}	histone modification	morphogenesis of a branching structure
GO:0016570	GO:0048568	FDR< 10^{-6}	histone modification	embryonic organ development
GO:0016570	GO:0006184	FDR< 10^{-6}	histone modification	GTP catabolic process
GO:0016570	GO:0009164	FDR< 10^{-6}	histone modification	nucleoside catabolic process
GO:0016570	GO:0009615	FDR< 10^{-6}	histone modification	response to virus
GO:0071357	GO:0097028	FDR< 10^{-6}	cellular response to type I interferon	dendritic cell differentiation
GO:0045087	GO:0097028	FDR< 10^{-6}	innate immune response	dendritic cell differentiation
GO:0045087	GO:0046949	FDR< 10^{-6}	innate immune response	fatty-acyl-CoA biosynthetic process
GO:0016570	GO:0051272	FDR< 10^{-6}	histone modification	positive regulation of cellular component movement
GO:0034340	GO:0046949	FDR< 10^{-6}	response to type I interferon	fatty-acyl-CoA biosynthetic process
GO:0016570	GO:0050795	FDR< 10^{-6}	histone modification	regulation of behavior
GO:0060337	GO:0046949	FDR< 10^{-6}	type I interferon-mediated signaling pathway	fatty-acyl-CoA biosynthetic process
GO:0016570	GO:2000027	FDR< 10^{-6}	histone modification	regulation of organ morphogenesis
GO:0016569	GO:0060688	FDR< 10^{-6}	covalent chromatin modification	regulation of morphogenesis of a branching structure
GO:0016570	GO:0060038	FDR< 10^{-6}	histone modification	cardiac muscle cell proliferation
GO:0016570	GO:0043279	FDR< 10^{-6}	histone modification	response to alkaloid
GO:0016570	GO:0060688	FDR< 10^{-6}	histone modification	regulation of morphogenesis of a branching structure
GO:0060337	GO:0097028	FDR< 10^{-6}	type I interferon-mediated signaling pathway	dendritic cell differentiation
GO:0016569	GO:0051216	FDR< 10^{-6}	covalent chromatin modification	cartilage development
GO:0016569	GO:0007423	FDR< 10^{-6}	covalent chromatin modification	sensory organ development
GO:0034340	GO:0097028	FDR< 10^{-6}	response to type I interferon	dendritic cell differentiation
GO:0016569	GO:0051272	FDR< 10^{-6}	covalent chromatin modification	positive regulation of cellular component movement
GO:0016569	GO:0006184	FDR< 10^{-6}	covalent chromatin modification	GTP catabolic process
GO:0016570	GO:0046888	FDR< 10^{-6}	histone modification	negative regulation of hormone secretion
GO:0016570	GO:0051216	FDR< 10^{-6}	histone modification	cartilage development
GO:0016570	GO:0001658	FDR< 10^{-6}	histone modification	branching involved in ureteric bud morphogenesis

GO:0016569	GO:0006813	FDR< 10^{-6}	covalent chromatin modification	potassium ion transport
GO:0016570	GO:0006813	FDR< 10^{-6}	histone modification	potassium ion transport
GO:0016569	GO:0060038	FDR< 10^{-6}	covalent chromatin modification	cardiac muscle cell proliferation
GO:0045087	GO:0050853	FDR< 10^{-6}	innate immune response	B cell receptor signaling pathway
GO:0016569	GO:0001763	FDR< 10^{-6}	covalent chromatin modification	morphogenesis of a branching structure
GO:0016568	GO:0046949	FDR< 10^{-6}	chromatin modification	fatty-acyl-CoA biosynthetic process
GO:0016569	GO:0009164	FDR< 10^{-6}	covalent chromatin modification	nucleoside catabolic process
GO:0016570	GO:0006325	FDR< 10^{-6}	histone modification	chromatin organization
GO:0016568	GO:0097028	FDR< 10^{-6}	chromatin modification	dendritic cell differentiation
GO:0016569	GO:0007565	FDR< 10^{-6}	covalent chromatin modification	female pregnancy
GO:0016570	GO:0007565	FDR< 10^{-6}	histone modification	female pregnancy
GO:0097028	GO:0045814	1.62E-05	dendritic cell differentiation	negative regulation of gene expression, epigenetic
GO:0016569	GO:2000027	1.62E-05	covalent chromatin modification	regulation of organ morphogenesis
GO:0009617	GO:0006363	1.62E-05	response to bacterium	termination of RNA polymerase I transcription
GO:0071357	GO:0046949	1.62E-05	cellular response to type I interferon	fatty-acyl-CoA biosynthetic process
GO:0016569	GO:0048568	1.62E-05	covalent chromatin modification	embryonic organ development
GO:0045087	GO:0006361	1.62E-05	innate immune response	transcription initiation from RNA polymerase I promoter
GO:0016570	GO:0071357	1.55E-04	histone modification	cellular response to type I interferon
GO:0016569	GO:0009581	1.55E-04	covalent chromatin modification	detection of external stimulus
GO:0071357	GO:0050853	1.55E-04	cellular response to type I interferon	B cell receptor signaling pathway
GO:0045087	GO:0006363	1.55E-04	innate immune response	termination of RNA polymerase I transcription
GO:0006913	GO:0046949	1.55E-04	nucleocytoplasmic transport	fatty-acyl-CoA biosynthetic process
GO:0007017	GO:0016568	1.55E-04	microtubule-based process	chromatin modification
GO:0097028	GO:0009615	1.55E-04	dendritic cell differentiation	response to virus
GO:0000722	GO:0046949	1.35E-03	telomere maintenance via recombination	fatty-acyl-CoA biosynthetic process
GO:0060688	GO:0046949	1.35E-03	regulation of morphogenesis of a branching structure	fatty-acyl-CoA biosynthetic process

GO:0007017	GO:0009615	1.35E-03	microtubule-based process	response to virus
GO:0009617	GO:0046949	1.35E-03	response to bacterium	fatty-acyl-CoA biosynthetic process
GO:0007017	GO:0006913	1.35E-03	microtubule-based process	nucleocytoplasmic transport
GO:0034340	GO:0048568	1.35E-03	response to type I interferon	embryonic organ development
GO:0007017	GO:0060688	8.37E-03	microtubule-based process	regulation of morphogenesis of a branching structure
GO:0045087	GO:0060688	8.37E-03	innate immune response	regulation of morphogenesis of a branching structure
GO:0006954	GO:0097028	8.37E-03	inflammatory response	dendritic cell differentiation
GO:0007017	GO:0043279	8.37E-03	microtubule-based process	response to alkaloid
GO:0006325	GO:0050853	8.37E-03	chromatin organization	B cell receptor signaling pathway
GO:0007017	GO:0009164	8.37E-03	microtubule-based process	nucleoside catabolic process
GO:0009617	GO:0097028	8.37E-03	response to bacterium	dendritic cell differentiation
GO:0006325	GO:0006363	8.37E-03	chromatin organization	termination of RNA polymerase I transcription
GO:0006913	GO:0097028	8.37E-03	nucleocytoplasmic transport	dendritic cell differentiation
GO:0060337	GO:0006361	8.37E-03	type I interferon-mediated signaling pathway	transcription initiation from RNA polymerase I promoter
GO:0006342	GO:0046949	8.37E-03	chromatin silencing	fatty-acyl-CoA biosynthetic process
GO:0016569	GO:0050807	8.37E-03	covalent chromatin modification	regulation of synapse organization
GO:0051028	GO:0046949	8.37E-03	mRNA transport	fatty-acyl-CoA biosynthetic process
GO:0007017	GO:0051169	8.37E-03	microtubule-based process	nuclear transport
GO:0007017	GO:0051272	8.37E-03	microtubule-based process	positive regulation of cellular component movement
GO:0002768	GO:0046949	8.37E-03	immune response-regulating cell surface receptor signaling pathway	fatty-acyl-CoA biosynthetic process
GO:0060337	GO:0034340	8.37E-03	type I interferon-mediated signaling pathway	response to type I interferon
GO:0097028	GO:0006342	8.37E-03	dendritic cell differentiation	chromatin silencing
GO:0097028	GO:2000027	8.37E-03	dendritic cell differentiation	regulation of organ morphogenesis
GO:0016568	GO:0046888	8.37E-03	chromatin modification	negative regulation of hormone secretion
GO:0016568	GO:0009581	4.28E-02	chromatin modification	detection of external stimulus

GO:0097028	GO:0046949	4.28E-02	dendritic cell differentiation	fatty-acyl-CoA biosynthetic process
GO:0016569	GO:0007050	4.28E-02	covalent chromatin modification	cell cycle arrest
GO:0009581	GO:0007050	4.28E-02	detection of external stimulus	cell cycle arrest
GO:0007050	GO:0001763	4.28E-02	cell cycle arrest	morphogenesis of a branching structure
GO:0050852	GO:0046949	4.28E-02	T cell receptor signaling pathway	fatty-acyl-CoA biosynthetic process
GO:0016569	GO:0043279	4.28E-02	covalent chromatin modification	response to alkaloid
GO:0016569	GO:0009615	4.28E-02	covalent chromatin modification	response to virus
GO:0071357	GO:0006361	4.28E-02	cellular response to type I interferon	transcription initiation from RNA polymerase I promoter
GO:0060337	GO:0006363	4.28E-02	type I interferon-mediated signaling pathway	termination of RNA polymerase I transcription
GO:0006325	GO:0007565	4.28E-02	chromatin organization	female pregnancy
GO:0006325	GO:0009617	4.28E-02	chromatin organization	response to bacterium
GO:0045814	GO:0006363	4.28E-02	negative regulation of gene expression, epigenetic	termination of RNA polymerase I transcription
GO:0006913	GO:0050853	4.28E-02	nucleocytoplasmic transport	B cell receptor signaling pathway
GO:0034340	GO:0009145	4.28E-02	response to type I interferon	purine nucleoside triphosphate biosynthetic process
GO:0034340	GO:0050853	4.28E-02	response to type I interferon	B cell receptor signaling pathway
GO:0006342	GO:0006361	4.28E-02	chromatin silencing	transcription initiation from RNA polymerase I promoter
GO:0006325	GO:0046888	4.28E-02	chromatin organization	negative regulation of hormone secretion
GO:0046949	GO:0050795	4.28E-02	fatty-acyl-CoA biosynthetic process	regulation of behavior
GO:0009615	GO:0050853	4.28E-02	response to virus	B cell receptor signaling pathway
GO:0071357	GO:0008015	4.28E-02	cellular response to type I interferon	blood circulation
GO:0006325	GO:0009164	4.28E-02	chromatin organization	nucleoside catabolic process
GO:0060337	GO:0009145	4.28E-02	type I interferon-mediated signaling pathway	purine nucleoside triphosphate biosynthetic process
GO:0016569	GO:0050795	4.28E-02	covalent chromatin modification	regulation of behavior
GO:0007050	GO:0051216	4.28E-02	cell cycle arrest	cartilage development
GO:0006260	GO:0001763	4.28E-02	DNA replication	morphogenesis of a branching structure
GO:0050807	GO:0046949	4.28E-02	regulation of synapse organization	fatty-acyl-CoA biosynthetic process

GO:0034340	GO:0006361	4.28E-02	response to type I interferon	transcription initiation from RNA polymerase I promoter
GO:0007565	GO:0046888	4.28E-02	female pregnancy	negative regulation of hormone secretion

1001