Random Forest Outperforms Other Phenotype Prediction Algorithms

EPI 511: Advanced Population and Medical Genetics

Prof. Alkes Price, Spring 2019

Harvard T. H. Chan School of Public Health

JACOB ROSENTHAL*

Using simulated genotypes from four HapMap 3 populations and simulated genotypes of 1000 loci, I compare the performance of three methods of phenotype prediction (k-nearest neighbors, random forest, and polygenic risk scoring) under two effect size distributions (10 and 100 causal loci). Random forest performed nearly 3.5 times better than polygenic risk scoring on average, as measured by Pearson correlation between predicted and true phenotypes in the test set. Polygenic risk scores performed slightly better for 100 causal loci than for 10, while the predictive power of the random forest was worse with more causal loci. Although results of this simulation may not extend to the cohort sizes and number of markers used in modern studies, these results demonstrate how powerful machine learning models can be, and the random forest in particular. Further work is needed to understand how to best leverage them alongside more typical models from statistical genetics.

1 INTRODUCTION

The study of genetics is the study of the relationship between genotype and phenotype. In some cases this relationship is straightforward, such as flower color in Gregor Mendel's pea plants. For the vast majority of traits, however, the relationship is much more complex. As evidence has grown in recent years indicating that many clinically important phenotypes are simulataneously influenced by hundreds or thousands of different genes, the state of the art for phenotype prediction has moved away from simplistic Mendelian models towards polygenic risk score models, incorporating genotype information from many loci throughout the genome. Although much progress has been made, widespread clinical implementation of polygenic risk scoring is still not yet a reality [1]. In this analysis, we use simulated genotypes and phenotypes to compare the performance of three methods of phenotype prediction under two effect size distributions.

The models used in this analysis are k-nearest neighbors, random forest, and polygenic risk scoring. Broadly, these models can be separated into two groups: statistical learning models, and statistical genetics models. The two types of models are built with different goals and each have their own advantages and disadvantages. Statistical learning models are designed to maximize predictive power by defining a loss function and searching a parameter space in an attempt to find a global minimum. Machine learning models have achieved remarkable performance on many tasks, but they run the risk of overfitting and are often not interpretable. Methods from statistical genetics, on the other hand, are typically based on conducting a set of univariate analyses on all SNPs to identify the association of each locus to the trait of interest, and then building a score based on the estimated effect sizes from association analysis. These models are more interpretable and their results can therefore be used to gain insight on other biological processes. However, they are based on a stronger set of assumptions about effect size distribution and other parameters and are therefore less flexible than non-parametric machine learning models. Additionally, single locus tests lack power to detect genetic interactions compared to many machine learning models [2]. In sum, the two classes of models come from different paradigms, each with their own strengths and weaknesses, and provide complementary information. This analysis attempts to quantify the performance of the three models under a variety of conditions, to provide more information for researchers choosing which to use for their studies.

*rosenthal1@hsph.harvard.edu

Table 1. Populations used in this analysis

Population	Continental Population	N
CEU	EUR	112
JPT	EAS	86
CHB	EAS	84
CHD	EAS	85

2 METHODS

2.1 Simulation

Simulations are based on genotypes from the HapMap 3 project [3]. We restrict our analysis to common SNPs only (SNPs with minor allele frequency > 5% in all eleven HapMap 3 populations). For this analysis, we consider four of the HapMap 3 populations: CEU (Utah residents of Northern and Western European ancestry), JPT (Japanese residents of Tokyo), CHB (Han Chinese in Beijing, China), and CHD (Residents of Denver with Chinese ancestry) (Table 1). The JPT, CHB, and CHD populations can all be assigned to the East Asian continental population (EAS), while the CEU population is part of the European continental population (EUR). From these populations, we construct five different training sets to investigate the impacts on model performance of (a) the composition of the continental populations contained in the training set and (b) the number of individuals in the training set.

Simulated genotypes are created by selecting every 125th SNP of the first 125,000 common SNPs, allowing the simplifying assumption that the SNPs are unlinked due to their distance. Note that there may still be some amount of LD despite the distance between markers; however, this is outside the scope of this project and from this point on we assume that the markers are independent.

Effect sizes are simulated under two scenarios, both assuming that $h_g^2 = 1$; that is, that the variance in the phenotype can entirely

be explained by the genotyped SNPs. h_g^2 represents the upper bound that can be achieved by any model that only considers genetic information [4], so the ideal model should be able to achieve perfect performance ($r^2(\hat{\varphi}, \varphi) = 1$) in this simulation study.

In the first scenario, we take the first 10 SNPs to be causal, each with a per-allele effect size $\beta \sim \mathcal{N}(0, 0.1)$. In the second, we take the first 100 SNPs to be causal, each with a per-allele effect size $\beta \sim \mathcal{N}(0, 0.01)$. The case of 100 causal loci is potentially of special interest because it is more similar to recent polygenic (even omnigenic [5]) frameworks where effect sizes are distributed across all or almost all loci. In both scenarios, we let all non-causal SNPs have effect size 0. Phenotypes are then simulated under an additive model:

$$\varphi_i = \vec{\beta} \cdot X_i$$

2.2 Association Testing

The Armitage trend test [6] is used to test for association between genotype and phenotype in the training data:

$$\text{ATT}\chi_1^2 = N\rho(X_i, \varphi_i)^2$$

If genotype and phenotype have both been normalized to mean 0 and variance 1, then we can get the estimated effect sizes directly by computing the Pearson correlation coefficient:

$$\hat{\beta} = \rho(X_i, \varphi_i)$$

2.3 k-Nearest Neighbors

For a test observation x_t, a k-nearest neighbors regression model identifies the K training points that are "closest" to x_t, and then estimates $\hat{f}(x_t) = \hat{y}_t$ by averaging across the y values of those K closest points. More formally, given some distance function $d(x_i, x_j)$ and a set of n labeled training points $(x_1, y_1), \ldots, (x_n, y_n)$ and some value of K where $1 \leq K \leq n$ and a test observation x_t, we make a set $\mathcal{N}_0$ of the K training points with the smallest distance from the test point, such that $d(x_t, x_i) \geq d(x_t, x_{(K)})$ for all $x_i \notin \mathcal{N}_0$. Then the k-nearest neighbors regression estimator is given by:

$$\hat{f}(x_t) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

[7]. There are two hyperparameters in this model: the choice of K and the choice of distance function. For this analysis, I use 10-fold cross validation to select the optimal value from among a set of choices of values to use as K (Figure 1). The choice of distance function is also critical: options include the naive norm of the difference in genotypes $d(x_i, x_j) = ||x_i - x_j||_2$ as well as distance functions based on information theory such as Jensen-Shannon distance [8, 9].

For this analysis, I devise a new distance metric:

$$d(x_i, x_j) = ||\vec{\chi} \odot (x_i - x_j)||_2$$

where $\odot$ is the Hadamard product and $\vec{\chi}$ is the vector of χ_1^2 association statistics from association analysis. In other words, differences in genotypes are weighted more for SNPs that are more strongly associated with the phenotype. The underlying assumption behind this model is that individuals with similar genotypes are predicted to have similar phenotypes, and that phenotype similarity is more important at more strongly associated loci.

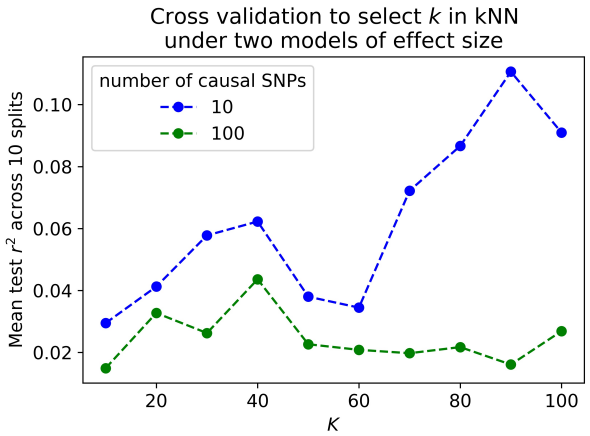

Fig. 1. Cross-validation to select K for k-nearest neighbors model. For 10 causal loci, the optimal performance occurs when $K = 90$. For 100 causal loci, peak prediction r^2 occurs when $K = 40$.

2.4 Random Forest

The random forest is a method of ensemble learning that works by creating a large number of decision trees on the training data and averaging across them to make predictions on test data. During the training process, at each node in each tree, a random subset of m predictors is chosen, and the tree selects its next branching criterion from among them according to some criteria (e.g. for regression trees, finding the binary split that minimizes mean squared error). The random sampling of predictors at each branch point means that the trees are decorrelated, which improves performance (relative to bagging, for example) [7]. Random forests have shown good performance in a variety of domains, including in genetics [10–13]. Important hyperparameters to consider for random forest methods are the number of trees to grow, the maximum depth of each tree, and the number of predictors to sample at each node. I use 5-fold cross validation in a dataset of combined CEU and CHD to select the best hyperparameters (Figure 2).

2.5 Polygenic Risk Score

In general, a polygenic risk score estimate is given by:

$$\hat{\varphi}_k = \sum_i \hat{\beta}_i x_{ik}$$

where i indexes all of the SNPs included in the predictor. The choice of what SNPs to include in the polygenic risk score is crucial: naive approaches are to include all SNPs, and to include only genome-wide significant SNPs. However, performance can be improved by instead choosing a P-value threshold P_T and including all SNPs with a P-value below the threshold in association analysis [14]. I use this thresholding method, and use 10-fold cross-validation to select P_T for both effect size distributions (Figure 3).

3 RESULTS

All three models were fit on each training population, for each of the two effect size distributions. Model performance is quantified by computing r^2 between predicted and true phenotypes (Table 2,

Table 2. Model performance under various simulation conditions. Reported values are r^2 between predicted and true phenotypes in test set. kNN = k-nearest neighbors; RF = random forest; PRS = polygenic risk score.

		10 causal loci			100 causal loci		
Train Pop.	Test Pop.	kNN	RF	PRS	kNN	RF	PRS
CEU	CHB	0.0483	0.4595	0.1237	0.0017	0.1249	0.1969
CHD	CHB	0.0009	0.4615	0.1053	0.0059	0.1897	0.0482
CEU + CHD	CHB	0.0491	0.5609	0.0623	0.0014	0.3264	0.0701
CEU + CHD + JPT	CHB	0.0781	0.6800	0.0751	0.0872	0.3395	0.2003
CHD + JPT	CHB	0.0857	0.6103	0.1372	0.0872	0.3218	0.0996

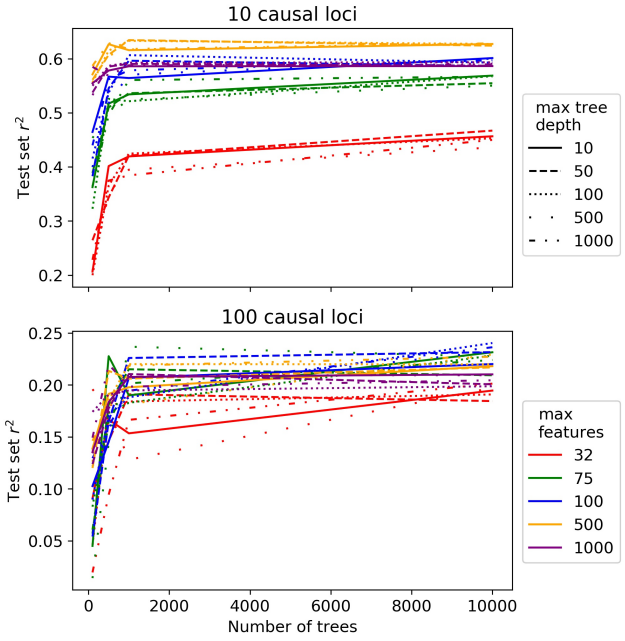

Fig. 2. Cross-validation to select hyperparameters for random forest models. Optimal parameters for 10 causal loci are: max_depth=50, max_features=500, n_estimators=1000. For 100 causal loci, optimal parameters are: max_depth=100, max_features=100, n_estimators=10000.

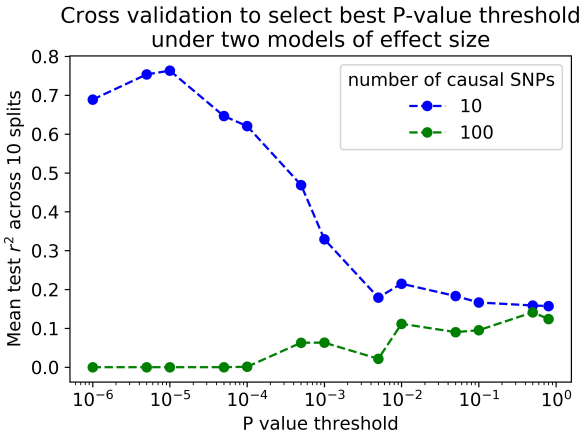

Fig. 3. Cross-validation to select P_T for polygenic risk scoring. For 10 causal loci, the optimal threshold value is 1×10^{-5}. For 100 causal loci, the optimal threshold is 0.5.

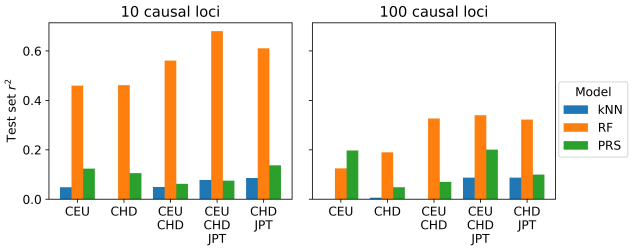

Fig. 4. Model performance under various simulation conditions. Reported values are r^2 between predicted and true phenotypes in test set. kNN = k-nearest neighbors; RF = random forest; PRS = polygenic risk score.

Figure 4) [15]. The random forest trained on CEU + CHD + JPT achieved peak performance out of all models under both effect size distributions ($r^2_{10} = 0.6800$ and $r^2_{100} = 0.3395$). Random forest was only outperformed by another model in one instance: polygenic risk scoring showed better predictive power under 100 causal loci, with CEU as the training population. Across all simulations, the random forest model achieved a test set r^2 that was 3.48-fold higher than that of the polygenic risk score. Overall, k-nearest neighbors performed worse than the other models, with average r^2 less than one half that of polygenic risk scoring and less than one fifth that of random forest. Predictive power is diminished across all three models under the scenario of 100 causal loci, relative to the scenario of 10 causal loci (Table 3).

4 DISCUSSION

The random forest models drastically outperforms that of standard polygenic risk prediction scores, with model performance up to 9-fold better than that of polygenic risk scoring in the most extreme case. On average, the performance boost of random forest is nearly 3.5-fold relative to polygenic risk scoring. Although the performance boost is somewhat attenuated in the scenario of 10 causal loci, it still represents a big improvement across the board.

Table 3. Average test set r^2 across simulations.

	10 causal loci	100 causal loci	overall
kNN	0.052434	0.036657	0.044546
RF	0.554452	0.260461	0.407457
PRS	0.100713	0.123023	0.111868
All models	0.235867	0.140047	0.187957

We also observe that testing populations affect model performance differently for different models. Random forest performs better when the training set contains more people, with large performance improvement when going from one population to two and then modest improvement when adding a third. k-nearest neighbors, on the other hand, also improved when trained on more individuals, but was more sensitive to the continental populations of training samples. Relative to performance when trained on CHD, performance is increased more by adding another EAS population (CHD + JPT) than by adding a EUR population (CEU + CHD). In addition to the worse overall performance, this lack of robustness for k-nearest neighbors is another reason to generally avoid using it for phenotype prediction.

Models also performed differently under the two effect size simulations. The random forest saw performance drop by more than half when the number of causal loci was increased from 10 to 100, yet the polygenic risk score actually improved for a greater number of causal loci. This performance bump was an unexpected result, given the results of the hyperparamter tuning 3. Future work should address methods for choosing hyperparameter tuning, as the cross-validation scheme used in this analysis is imperfect and the choice of hyperparameters can have a big impact on how the model performs. A more exhaustive search of hyperparameter space may also lead to improvements for random forest.

This analysis makes a number of simplifying assumptions, due to restraints on time and computation. One way to improve this simulation would be to incorporate linkage disequilibrium into the predictions, instead of assuming that all loci are unlinked. Another would be to increase sample size, both by adding more individuals and by considering more loci. In addition, it may be more realistic to account for the fact that per-allele effect size may differ between populations. Finally, this analysis would benefit from the inclusion of more models, such as LDPred [16] and neural networks [17].

In sum, I show that random forest performs much better than the typical polygenic risk score for phenotype prediction. I also show that k-nearest neighbors performs much worse than the other two methods, and should probably not be used. Although random forest shows much promise in this study, it must be noted that this is a very small simulation and that the results may not hold for the cohort sizes and number of markers used in modern studies. Nevertheless, machine learning clearly has a place in phenotype prediction and genetics in general. The challenge going forward will be how to incorporate statistical learning and statistical genetics, so as to leverage the complementary advantages of both.

REFERENCES

[1] Torkamani, Wineinger, and Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19:581–590, 2018.

[2] Okser, Pahikkala, Airola, Salakoski, Ripatti, and Aittokallio. Regularized machine learning in the genetic prediction of complex traits. *PLOS Genetics*, 10:e1004754, 2014.

[3] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–59, 2010.

[4] Wray, Yang, Goddard, and Visscher. The genetic interpretation of area under the roc curve in genomic profiling. *PLoS Genetics*, 6:e1000864, 2010.

[5] Boyle, Li, and Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169:1177–1186, 2017.

[6] Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386, 1955.

[7] James, Witten, Hastie, and Tibshirani. *An Introduction to Statistical Learning*. Springer, New York, NY, USA, 1st edition, 2013.

[8] Kim and Kim. Empirical prediction of genomic susceptibilities for multiple cancer classes. *PNAS*, 111:1921–1926, 2014.

[9] Kim and Kim. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. *PNAS*, 115:1322–1327, 2018.

[10] Paré, Mao, and Deng. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Scientific Reports*, 7:12665, 2017.

[11] Chuang and Kuo. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. *Scientific Reports*, 7:39943, 2007.

[12] Goldstein, Hubbard, Cutler, and Barcellos. An application of random forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics*, 11, 2010.

[13] Bureau, Dupuis, Falls, Lunetta, Hayward, Keith, and Van Eerdewegh. Identifying snps predictive of phenotype using random forests. *Genet Epidemio)*, 28:171–182, 2005.

[14] The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–752, 2009.

[15] Daetwyler, Villanueva, and Woolliams. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE*, 3:e3395, 2008.

[16] Vilhjálmsson, Yang, Finucane, Gusev, Lindström, Ripke, Genovese, Loh, Bhatia, Do, Hayeck, Won, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology,, Risk of Inherited Variants in Breast Cancer (DRIVE) study, Kathiresan, Pato, Pato , Tamimi , Stahl, Zaitlen, Pananiuc, Belbin, Kenny, Schierup, De Jager, Patsopoulos, McCarroll, Daly , Purcell, Chasman, Neale, Goddard, Visscher, Kraft, Patterson,, and Price. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics*, 97:576–592, 2015.

[17] Ho, Schierding, Wake, Saffery, and O'Sullivan. Machine learning snp based prediction for precision medicine. *Frontiers in Genetics*, 10:267, 2014.