UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PREDICTING SARCOIDOSIS DISEASE INCIDENCE USING SINGLE NUCLEOTIDE POLYMORPHISMS AND SUPERVISED MACHINE LEARNING

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

NICHOLAS I. CEJDA

Norman, Oklahoma

2021

PREDICTING SARCOIDOSIS DISEASE INCIDENCE USING SINGLE NUCLEOTIDE POLYMORPHISMS AND SUPERVISED MACHINE LEARNING

A THESIS APPROVED FOR THE

GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF:

Dr. Talayeh Razzaghi, Chair

Dr. Courtney Montgomery

Dr. Charles Nicholson

Dr. Chongle Pan

Table of Contents

Abstract

Predicting disease incidence based on Single Nucleotide Polymorphisms (SNPs) for a complex multi-factorial disease like sarcoidosis remains a difficult prediction problem. If disease prediction could be improved, genetic screening could be implemented to assist identifying disease early, potentially improving patient outcomes.

In this thesis, we examine the predictive performance of several supervised machine learning models to assess if genetic variability can be used to accurately predict disease incidence in an African American patient population (n = 2,915). Further, we consider the use of SNP “functional scores” such as Combined Annotation Dependent Deletion (CADD) scores and FATHMM-XF scores to see if they can improve predictive ability.

Here we show that support vector machine (SVM), and random forest (RF) models can significantly outperform the naïve baseline model ($p < 0.05$) in terms of accuracy and achieve area under the ROC curve (AUC) values of 0.6016 and 0.6019, respectively. A neural network (NN) model had the optimal AUC value of 0.6103 but was slightly non-significant ($p = 0.05$) when compared to the naïve model in terms of accuracy. The overall impact of adding functional scores was minimal to negative on predictive performance.

This work reveals that supervised machine learning based on SNPs can significantly outperform random chance when predicting sarcoidosis incidence and supports the idea that genetic screening and disease modeling prior to disease incidence could improve preventative care.

Keywords: Supervised Machine Learning, Disease Prediction, Single Nucleotide Polymorphisms (SNPs), Sarcoidosis, Disease Modeling, Random Forest, Support Vector Machine (SVM), Neural Network

Chapter 1 – Introduction

1.1 - Predicting disease is a key challenge that can be aided by machine learning

Predicting disease incidence prior to symptom presentation is a key challenge in modern medicine, as achieving highly accurate disease predictions could lead to earlier detection of initial disease, which has potential to improve patient outcomes. For example, clinicians routinely screen patients deemed at high risk for colorectal cancer due to family history or prior disease, which has been shown to save countless lives due to early detection (Kahi et al., 2018). However, over-screening patients for suspected disease when none is present could burden patients with unnecessary cost, invasive screening procedures, psychological stress, false positives, and could potentially initiate the “nocebo” effect (Colloca & Miller, 2011), triggering negative symptoms despite absence of disease. It is imperative therefore that any disease prediction or risk assessment method be highly accurate prior to implementation in the clinic. Due to the benefits of early disease detection and the risks of inaccurate prediction, novel methods of identifying patients at high risk for life-threatening diseases should be explored and optimized to maximize accuracy.

One such method to identify people at high risk for disease is calculating a weighted polygenic risk score (Ho et al., 2019; *Polygenic Risk Scores*, 2020). This approach sums the number of individual risk alleles the person carries, weighted by how strongly each risk allele associates with disease in a genome-wide association study (GWAS), to generate a score which describes the patient’s relative risk of developing a specific disease. This approach is attractive because it does not require a person to be sick or have a family history to assess disease risk. Additionally, it only requires a simple blood draw, DNA extraction, and a genotyping assay, which is becoming increasingly inexpensive (Li et al., 2008). However, polygenic risk scores have received criticism for being unable to model complex interactions which occur in many diseases (Ho et al., 2019), overly training on European ancestry patients which limits their utility in other populations (De La Vega & Bustamante, 2018), and achieving only moderate accuracy in predicting disease outcome for several diseases (Belsky et al., 2013; Lewis & Vassos, 2020).

An alternative approach is to utilize supervised machine learning algorithms trained on a large cohort of case and control patients, using the most disease-associated patient single nucleotide polymorphisms (SNPs) identified by GWAS to generate a predictive model (Ho et al., 2019). This approach shares the advantages of polygenic risk scores yet produces more complex models capable of identifying otherwise unseen structure in the data, which has been shown to improve disease prediction accuracy compared to polygenic risk scores alone (Joseph et al., 2018; Kruppa et al., 2012; Paré et al., 2017). This approach still suffers from relying on data from predominately individuals with European ancestry, which limits utility in non-European populations, as well as demands a large amount of high-quality genotyped case and control data to effectively train the model.

1.2 - Hypothesis 1 – Machine learning can classify sarcoidosis cases with better than random chance

We hypothesized that machine learning could be useful in improving our ability to predict sarcoidosis disease incidence. Sarcoidosis is a complex disease with both genetic and environmental components contributing to pathology (Moller et al., 2017) and can be life-threatening in severe cases (Baughman & Lower, 2011). Additionally, accurately diagnosing sarcoidosis remains challenging due to minimal signs and symptoms in early stages of disease as well as similarity of symptoms to other common diseases (*Sarcoidosis - Diagnosis and Treatment - Mayo Clinic*, 2019). The defining feature of sarcoidosis is formation of clusters of inflammatory cells called granulomas, which typically form in the lungs but can also be found in other organs (*Learn About Sarcoidosis - American Lung Association*, 2020). However, several other diseases also feature granulomas, such as Crohn’s disease (Molnár et al., 2005), rheumatoid arthritis (Imadojemu et al., 2016), and tuberculosis (Silva Miranda et al., 2012), among others. A battery of tests is typically required to confirm sarcoidosis and rule out other disorders, such as chest X-ray, computerized tomography (CT), pulmonary function tests, and invasive lung or skin biopsy to collect granuloma samples (*Sarcoidosis - Diagnosis and Treatment - Mayo Clinic*, 2019). The difficulty of diagnosis underscores the need for robust genetic prediction tools to aid in identifying high-risk patients.

The etiology of sarcoidosis is currently unknown, but it is widely believed that environmental pollutants and/or bacterial or fungal infection triggers the onset of granulomatous formations, then the autoimmune system fails to resolve the granulomas once the pollutant or bacteria is cleared (Starshinova et al., 2020). *Mycobacterium* species may play a role, as well as working in environments with high mold/mildew, or exposure to some inorganic aerosols such as insecticides, certain metals, or wood smoke and ash (Judson, 2020; Moller et al., 2017).

Due to the combination of genetic and environmental contributions to sarcoidosis pathology, there is a theoretical limit to the performance of any genetics-only approach to predicting disease incidence. According to one study of 210 Danish and Finnish twin pairs, the heritable component of sarcoidosis was calculated to be 66% (95% C.I. 0.45 to 0.8) (Sverrild et al., 2008). A Swedish familial aggregation study looking at 23,888 cases and 171,891 general-population controls estimated the heritability to be 39% (95% C.I. 0.12 to 0.65) (Rossides et al., 2018). Both studies reach the conclusion that genetics plays a significant but non-exclusive role in sarcoidosis disease incidence. An optimal model therefore would require using “blended” data types, including both genetics, and known environmental exposures to disease-associated pollutants and/or pathogens.

Construction of this ideal “blended” dataset remains challenging, due to lack of knowledge about which specific pollutants contribute most to disease combined with the impracticality or impossibility of quantitatively measuring how much exposure a person has had to a particular pollutant such as wood smoke or aerosolized metal. Qualitative assessment of known exposure to a given pollutant could be obtained via questionnaire (e.g., asking a person how often they are around wood burning stoves or fires), which may still contain predictive value and improve modeling performance in combination with genotyping data. Incorporating gene-environment interaction in predictive models has been shown to improve phenotype predictive performance in yeast and plants (Grinberg et al., 2020) as well as Parkinson's disease prediction (Jacobs et al., 2020); however, a robust dataset of this type has not been created for sarcoidosis to our knowledge.

Despite the limitations of implementing a genetics-only approach to model disease incidence of sarcoidosis, we attempted to obtain model sensitivity near the estimated heritability of ~40% using a dataset of 2,915 genotyped African American sarcoidosis patients and controls with supervised machine learning. We first utilized genome-wide association (GWA) to identify the most correlated SNPs with disease, then used these top SNPs as features to build three classifiers based on random forest (RF), support vector machine (SVM), and neural network (NN). We then evaluated the performance of the models on unseen test data [See Chapter 2.2– Data Preparation]. The best model's most important features were identified and could be used to inform future studies or experiments. However, simply because a SNP is considered useful for classification does not necessarily mean that SNP plays an important biological role in disease incidence. The individual SNP in question may be in Linkage Disequilibrium (LD) with a nearby SNPs that plays a more causal role, it may be useful for classification by random chance, or it may be interacting with another important SNP which has a more causal role.

1.3 - Determining if single nucleotide polymorphisms have functional consequences or if they are benign is an active area of research

Assessing if a SNP has a functional role in forming a person's phenotype is an active area of research, both experimentally and computationally. Over 335 million SNPs have been identified from humans across the globe and a single individual contains around 4 to 5 million differences compared to the reference genome (Auton et al., 2015). Understanding which of these SNPs play functional roles in determining human traits or disease and which are inconsequential is an ongoing scientific effort.

Projects such as the Encyclopedia of DNA Elements (ENCODE), the Roadmap Epigenomics Project, RegulomeDB, and HaploReg aim to help address this large challenge (Abascal et al., 2020; Boyle et al., 2012; Roadmap Epigenomics Consortium et al., 2015; Ward & Kellis, 2016).

Determining functional consequences of SNPs in protein-coding regions, while still challenging, is more straightforward than non-coding SNPs, because changes to protein-coding regions of a genome can result in predictable changes to amino acid sequence based on the standard genetic codon table. If an early stop codon is introduced for example, entire proteins or chunks of proteins can be lost, leading to diseases such as Duchenne muscular dystrophy and cystic fibrosis (Keeling et al., 2013). A compendium of known inherited genetic disorders based on changes to protein coding sequences, known as the Online Mendelian Inheritance in Man (OMIM) was created and is maintained to assist in identifying which disorders (phenotypes) are generated from which specific changes to protein coding sequences (Amberger et al., 2019).

In general, phenotypic characterization of amino acid changes due to SNPs still requires experimental and/or observational validation, because predicting function involves predicting protein folding, then predicting protein-protein interactions, then predicting how those altered interactions will affect the rest of the organism’s biology, which remains an immensely challenging task. Great strides have been made recently to improve protein folding predictions using deep learning (Senior et al., 2020). Still, the challenge of *in silico* predicting phenotypic changes based on changes to protein coding regions remains standing.

The problem of predicting functional effects is compounded when non-protein coding SNPs are considered. Most SNPs occur outside the protein-coding regions of the genome, primarily due to increased selection pressure on the coding regions compared to the non-coding regions (Barreiro et al., 2008). Areas of the non-coding regions can still play vital regulatory roles by enabling transcription factor or promoter or silencer binding, acting as a cis- or trans-regulatory elements, being involved in epigenetic regulation, or regulating telomeres (Carroll, 2008; Cusanelli & Chartrand, 2014; Kasowski et al., 2010). Other areas of the non-coding region may have lost their function entirely and thus carry no evolutionary consequences when mutations occur in these regions, leading to accumulation of SNPs without functional effects (Zheng et al., 2007). Due to this variability across the non-coding regions of the genome, it is unclear if a given SNP will have a functional consequence or not simply based on nucleotide change alone. Additional context, such as adjacency to a coding region, location in known regulatory regions, or experimental modification in cell or tissue culture, is required to estimate the likelihood of a SNP's potential to have a functional effect.

Several machine learning efforts have been conducted to classify each known SNP as either functional or non-functional. One such effort has been the generation of Combined Annotation Dependent Depletion (CADD) scores (Kircher et al., 2014). In this study, researchers trained a support vector machine (SVM) to differentiate between 14.7 million high-frequency human alleles versus 14.7 million simulated variants using a suite of over 60 different features as predictors. Some features used include measurements of evolutionary conservation, open chromatin, acetylation or methylation, distance from the nearest transcription factor binding site, human genetic frequency, along with many more. The logic behind comparing actual observed high-frequency alleles with simulated alleles is that deleterious mutations that reduce an organism's fitness tend to decrease over time due to natural selection, but these deleterious mutations will not be reduced in the simulation. The CADD-score, or "C-score," therefore measures how likely a given SNP is to be deleterious to an organism's fitness. This measurement correlates with changes in both molecular functionality and pathogenicity of a particular SNP. The researchers applied their trained model to generate C-scores for every human SNPs and have continued to update their scores for the latest version of the GRCh38 genome assembly (Rentzsch et al., 2019). One limitation of this approach is that disease-causing SNPs can survive natural selection if they cause disease in middle or old age, past the point of reproductive pressure.

Another approach that has been attempted to assess the deleteriousness of a specific SNPs is called FATHMM-MKL (Shihab et al., 2015). This approach leverages a manually curated database of known, heritable, disease-causing variants, the Human

Gene Mutation Database (Stenson et al., 2017), to generate a robust list of deleterious SNPs that are known to cause disease. They then generated a list of SNPs unlikely to cause disease by pulling SNPs from the 1000 genomes project (Altshuler et al., 2012) which were presumed to be benign due to their absence in the Gene Mutation Database. The features they used to train this model included: vertebrate sequence conservation, histone modification, transcription factor binding sites, open chromatin, local GC content 5bp around the SNP, and more. They trained their model using SVM with multiple kernel learning, and demonstrated superior performance compared to CADD on unbiased test samples. In early 2018, an updated model called FATHMM-XF was published, improving the results from FATHMM-MKL by utilizing additional features during model training (Rogers et al., 2018). Prediction scores are available for all GRCh37 and GRCh38 SNPs .

A meta-analysis conducted in 2018, prior to the release of FATHMM-XF, compared 15 different genome-wide deleterious prediction scores and 8 conservation scores to evaluate which models perform the best when predicting non-coding deleteriousness and found that the FATHMM-MKL model outperformed the competition (Liu et al., 2017). However, this result is perhaps biased toward FATHMM-MKL compared to CADD because the test samples used in the meta-analysis were pulled from the Human Gene Mutation Database. Nonetheless, *in silico* functional characterization of SNPs has demonstrated remarkable accuracy in predicting variants likely to cause deleterious effects in humans, and hopefully can be leveraged to potentially improve sarcoidosis modeling accuracy.

1.4 - Hypothesis 2 – Incorporating functional scores can improve sarcoidosis disease incidence predictive accuracy

Since sarcoidosis has a strong heritability component (~40-66%), some disease-causing variants must also be heritable and therefore might be identifiable *via* FATHMM-XF scores or C-scores. We hypothesized that incorporating functional prediction scores into our sarcoidosis prediction model would improve prediction accuracy, specifically by helping to identify and prioritize causal variants over variants that are associated with disease simply by chance or simply by being passengers in LD with the causal variants. This approach also has the advantage of generating novel lists of SNPs that can be used to investigate mechanisms of disease incidence in future studies. We generated a single “blended” association plus pathogenicity score, which rewards a SNP for being highly correlated with disease as well as having a high C-score or FATHMM-XF score [See Chapter 4.1 – Functional Score Assignment].

Chapter 2 – Data Preparation

2.1 - Code Availability

All R code, Bash shell commands, and trained model files used in this project are available at: https://github.com/cejdan/sarc-predictions

2.2 - Data Preparation and Quality Control

2,918 African American patient samples were obtained as part of the ACCESS sarcoidosis study (Freemer & King, 2001), the SAGA sarcoidosis study (Rybicki et al., 2005), or the Henry Ford Health System (HFHS) and were genotyped. More details on sample collection and genotyping can be found in previous work (Adrianto et al., 2012). Nucleotides not captured by the original sequencing arrays were imputed with the TOPmed imputation server (Taliun et al., 2021) using the following settings: r2 reference panel, GRCh38 reference build, no R-squared filter, and Eagle v2.4 phasing. TOPmed imputations were then quality controlled to remove any SNPs that were in high linkage disequilibrium with other SNPs (defined as r-squared < 0.5). Next, we replaced imputed nucleotides with observed nucleotides when available. This data set contained 2,918 individuals from 1,969 unique families, 819 males and 2,099 females, 1,273 cases and 1,645 controls, and 69,887,691 SNP variants.

The following quality controls were then used to filter and clean the data to make it more suitable for GWAS and subsequent modeling:

1) Remove any individuals with < 90% genotyping rate (10% missing rate).

2) Remove SNPs that have any missing values for any individual. This step helps simplify downstream modeling by ensuring all data is free of missing values.

3) Remove minor alleles with < 1% frequency.

4) Remove variants that reach 0.0001 significance on the Hardy-Weinberg equilibrium test.

5) Remove individuals that have more than 5% Mendelian errors.

After quality control, the remaining dataset contained 2,915 individuals from 1,969 unique families, 818 males and 2,097 females, 1,272 cases and 1,643 controls, and 7,723,467 SNP variants.

2.3 - Creation of Test and Training data subsets

Prior to any modeling, we separated test and training datasets to ensure independence of samples. Careful consideration had to be taken during this step to ensure that whole families were kept entirely within either the training set or the test set, as family members present in both training and test sets would reduce the independence of the test set.

To accomplish this, we randomly sampled the quality-controlled data based on family ID, not individual ID. We also wanted to maintain similar ratio of cases and controls in both training and test sets. We randomly sampled ~10% of the 935 families containing at least one member with sarcoidosis (94 families sampled of the 935) as well as ~10% of the 1,367 families containing at least one member as a control (137 families sampled out of 1367 families), which left us 231 families in the test set. However, 4 of these families were sampled twice by random chance and were dropped from the test set, which left 227 families in the test set. In total, 412 individuals (229 controls and 183 cases) from 227 families were used as the test set (55.6% controls, 44.4% cases), while 2,503 (1,414 controls and 1,089 cases) from 1742 families were used as the training set (56.5% controls, 43.5% cases). No families were separated. New training-specific and test-specific binary plink files were generated to ensure that the test and train samples remained separated for all downstream analysis.

2.4 - Principle Component Analysis, Logistic regression, and LASSO regression

We performed principal component analysis (PCA) on the training samples to help adjust for correlated ancestry during the logistic regression step and used the first 4 PCs as covariates in the model. It is common practice in GWA studies to account for correlated ancestry by using PCs as covariates in your model, as it improves the robustness of the results (C. Chen, 2019).

Next, to help narrow the list of SNPs down from 7.7 million, we performed logistic regression using Plink v1.9 (Chang et al., 2015) to generate odds ratios and p-values for each SNP and plotted the resulting GWAS results in a Manhattan plot (Figure 1).

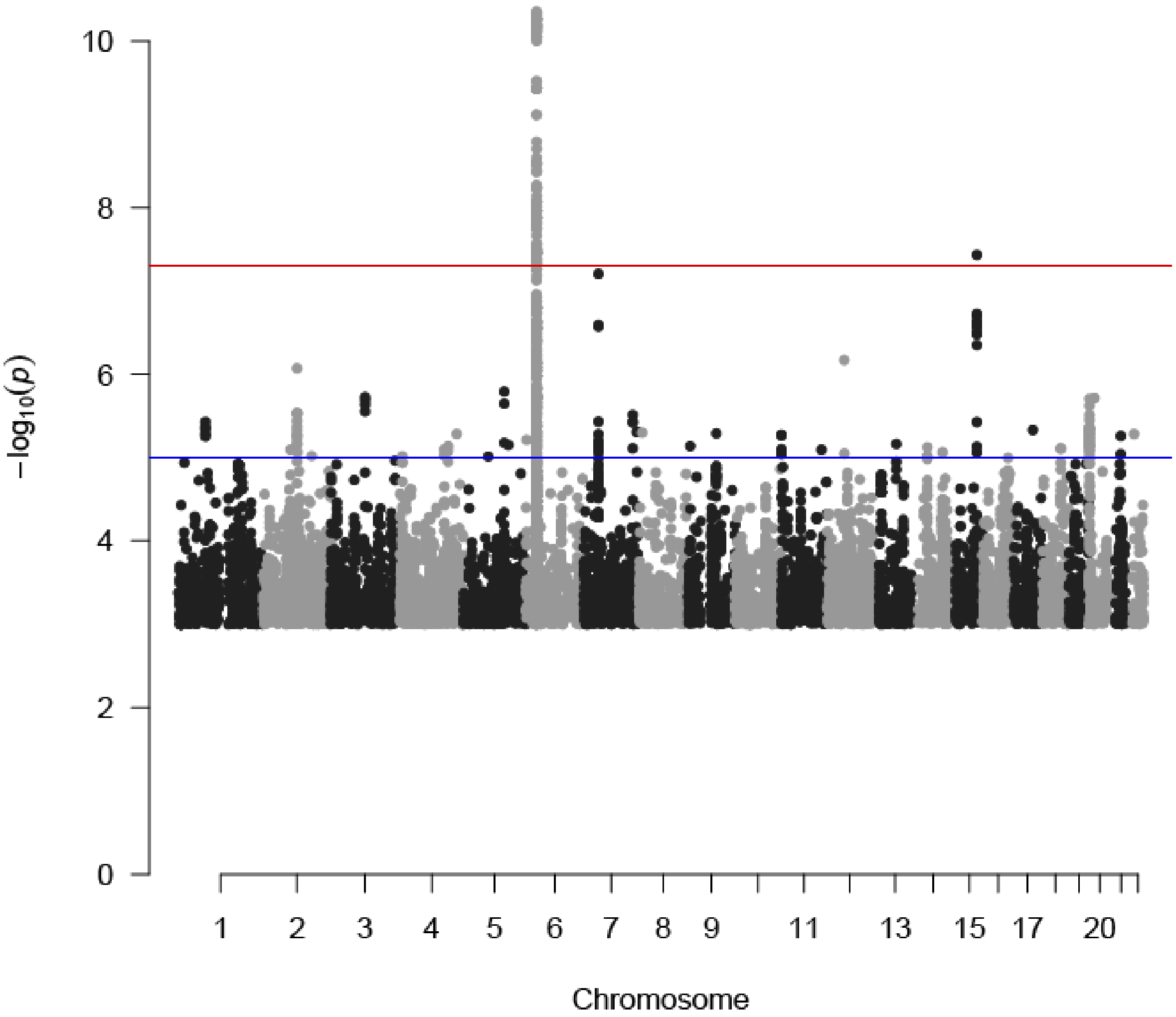

Figure 1. Manhattan plot of the training samples GWAS. Blue line at p = 1x10^{-5} indicates genome-wide suggestive SNPs. Red line at p = 5x10^{-8} indicates genome-wide significant SNPs. The large number genome-wide significant SNPs on chromosome 6 correspond to the major histocompatibility complex (MHC) region of the genome, which is highly variable and involved in antigen presentation. SNPs ranked below p > 1x10^{-3} were omitted from the plot for clarity.

The resulting Manhattan plot displayed similar significant SNPs to those previously published (Adrianto et al., 2012). The main advantage of performing the logistic regression was that we were able to obtain a list of the most highly linearly correlated SNPs with the disease. In essence, this analysis was feature selection, helping to narrow our initial list of 7.7 million features down to only the top features expected to contribute to predictive modeling. We did not limit ourselves to using only the genome-wide significant SNPs while modeling, because one major advantage of using machine learning to model disease incidence is that using SNPs with smaller effect sizes can still improve performance.

After performing logistic regression, we also performed a least absolute shrinkage and selection operator (LASSO) regression, which performs an L1 regularization, a linear modeling technique that reduces the effect sizes of unimportant features down to zero while keeping features with non-zero effect sizes. (Tibshirani, 1996). This approach offered an alternative way to prioritize SNPs. After performing LASSO, we had 484 SNPs with non-zero effect sizes remaining.

2.5 - Preparation for modeling and exploratory data analysis

After sorting the logistic regression results by p-value, we extracted the top 10 SNPs, top 100 SNPs, top 500 SNPs, top 1000 SNPs, and top 2000 SNPs. Using Plink v1.9, we then extracted the total number of minor alleles carried by each patient at each SNP, generating a matrix defined by:

$$M_{i,j} = \begin{cases} 0 & if\ person\ i\ has\ no\ minor\ alleles\ at\ SNP\ j \\ 1 & if\ person\ i\ has\ 1\ minor\ alleles\ at\ SNP\ j \\ 2 & if\ person\ has\ 2\ minor\ alleles\ at\ SNP\ j \end{cases}$$

In addition, the individual’s sex (based on XX or XY genotype) was added as a feature to the model, and the phenotype (class) column was added so that we could perform supervised machine learning. This matrix contained the raw data used for modeling. Some data cleaning was necessary prior to modeling because many of the features were very highly correlated due to linkage disequilibrium (LD) and thus contained redundant information. LD occurs because genetic recombination, the process of mixing an organism’s alleles during meiosis, occurs with decreasing probability the closer two sections of DNA are in the genome. Nearby SNPs on the same chromosome are unlikely to have recombination occur between them, so they will usually be inherited as a unit. In our list of top 10 SNPs, almost all were SNPs on Chromosome 6 in proximity, and therefore high LD, with each other (Figure 2). In the case of the top 100 SNP correlation matrix, we find that large blocks of SNPs have high correlation coefficients within the block (Figure 3). These are known as haplotype blocks, sections of DNA that are often inherited together due to LD (Zhu et al., 2004).

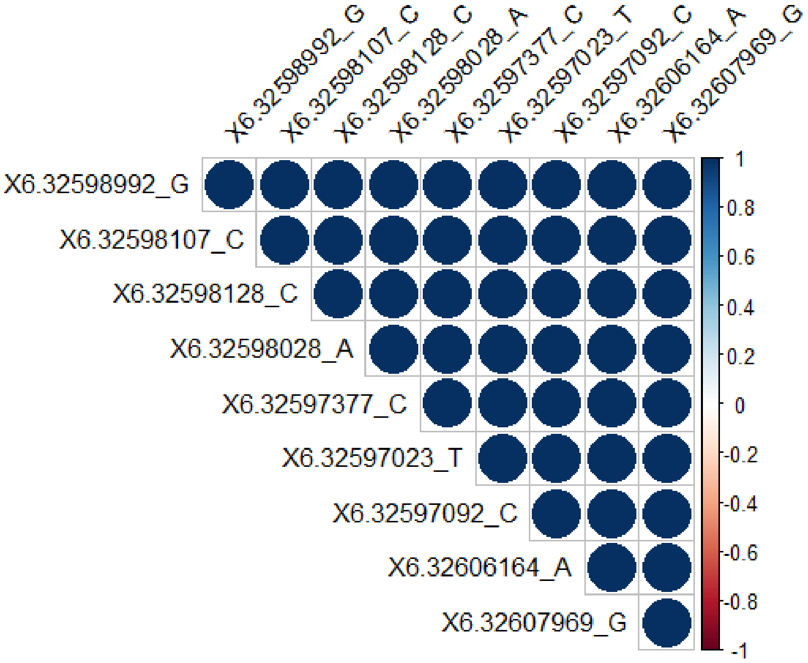

Figure 2. Pearson's r^2 correlation matrix of the top 10 SNPs extracted from the logistic regression model. Dark blue circles indicate that all SNPs have near-perfect correlations with each other ($r^2 \approx 1.0$), and thus contain redundant information for downstream modeling.

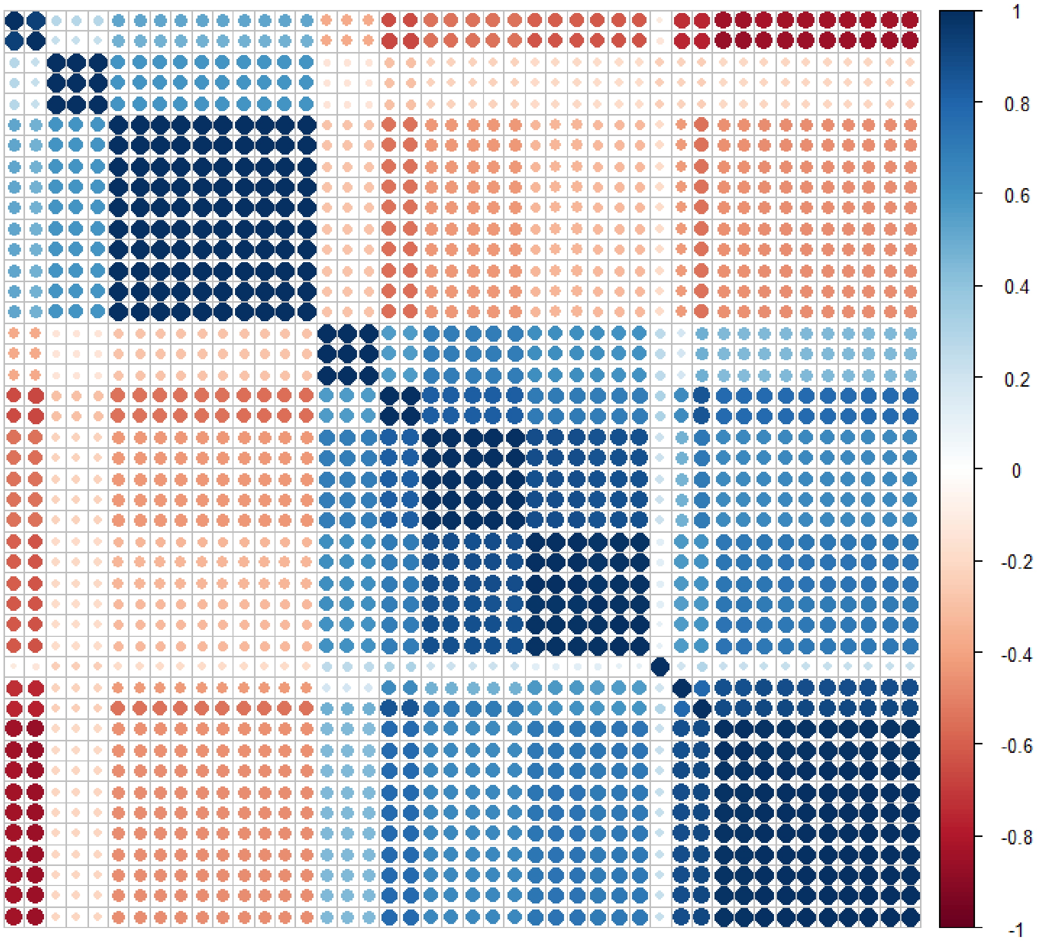

Figure 3. Pearson's r^2 correlation matrix of the top 100 SNPs (only top 50 SNPs are shown here for clarity). Large "blocks" of SNPs tend to be highly correlated as they are in close proximity to each other in the genome and thus are inherited together. These are known as haplotype blocks and contain large sections of redundant information that is not useful for downstream modeling.

2.6 - Removal of correlated features

To remove redundant information and improve model performance, the most highly correlated variables needed to be removed. To address this issue for the logistic regression models, we utilized the findCorrelation method in the caret package (Kuhn, 2008) to eliminate correlated variables. This function works by checking each pairwise correlation value, and if the value is above a specified r^2 threshold, the column with the higher mean correlation across all rows is selected for elimination. Several models were trained with differing correlation values (0.75, 0.8, 0.85, 0.9, 0.95) to find the optimal correlation removal threshold (see Chapter 2 – Random Forest modeling). An example using an r^2 cutoff of 0.95 on the top 100 SNPs is shown (Figure 4). The number of features retained for each subset of the initial data matrix across the different correlation cutoffs is described in Table 1.

For the LASSO model, SNPs were eliminated by comparing each pairwise Pearson r^2 value and eliminating the column with the lower absolute value effect size. This was done to preserve the maximum amount of useful information in the final model. The SNPs retained for each Pearson r^2 threshold cutoff for the LASSO model is also summarized in Table 1. In general, fewer SNPs were eliminated from the LASSO model because the L1 regularization process reduces effect sizes to zero or near-zero for highly correlated / redundant variables.

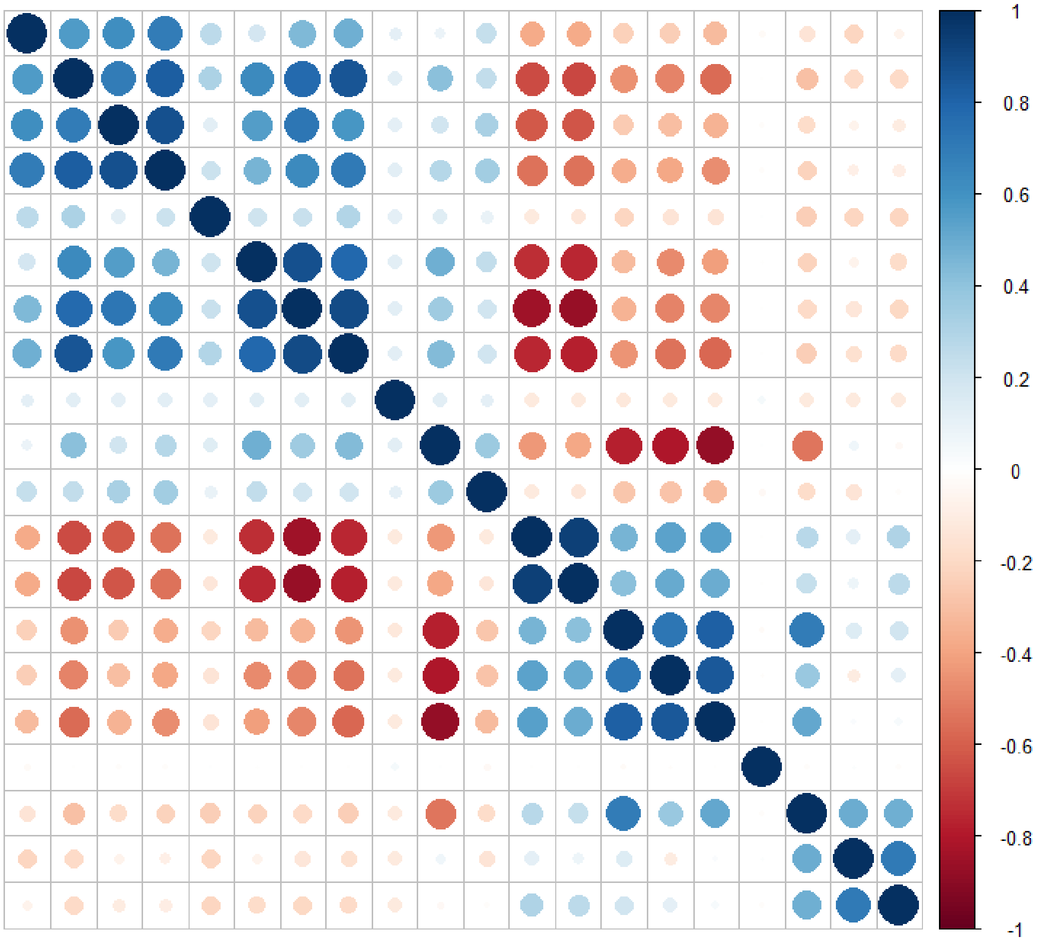

Figure 4. Top 100 SNP correlation matrix with Pearson r^2 correlations > 0.95 removed. 18 SNPs survive this cutoff and are shown here. Most of the highly correlated haplotype blocks are removed or reduced.

Number of SNPs retained	≤0.75	≤0.8	≤0.85	≤0.9	≤0.95
Logistic Top10 SNPs	1	1	1	1	1
Logistic Top100 SNPs	10	12	12	16	18
Logistic Top500 SNPs	69	78	90	104	114
Logistic Top1000 SNPs	171	180	197	221	248
Logistic Top2000 SNPs	451	469	507	558	625
LASSO 484 SNPs	430	431	433	437	441

Table 1. Number of SNPs retained for each “Top SNP” subset for various Pearson’s r^2 correlation cutoffs. Lower cutoffs are stricter and remove more SNPs.

Chapter 3 – Random Forest, SVM, and Neural Network modeling

3.1 - Random Forest Modeling – Background

After correlation removal, we could now begin addressing if machine learning could accurately predict sarcoidosis. We began by utilizing random forests (RFs) to construct the initial predictive models because RFs have been shown to be robust for learning structure in a variety of contexts, including genetics-based disease prediction for diseases like Type 2 Diabetes (López et al., 2018).

RFs are composed of numerous binary decision trees. The individual trees are generally constructed using the Classification and Regression Tree (CART) methodology, whereby splits are created based on maximizing "information gain", or maximum improvement to a given impurity index, usually the Gini Index. (Breiman et al., 1984). In a CART decision tree, the initial node begins with all data from both classes present. We generate a split using one of the data's columns based on the column that provides the maximum information gain, and the data is broken into two pieces, with the goal being to concentrate observations from class "A" on one side, and class "B" on the other. The more homogenous the class mixture in the new data, the "purer" the node is. The algorithm is repeated recursively, until all nodes reach a desired purity or until the tree reaches a desired depth or until there are no additional ways to separate the classes. If the tree grows to complete purity, all leaf nodes will contain data belonging to only a single class, but the tree may become very deep and complex. An example of a simple decision tree is shown in Figure 5.

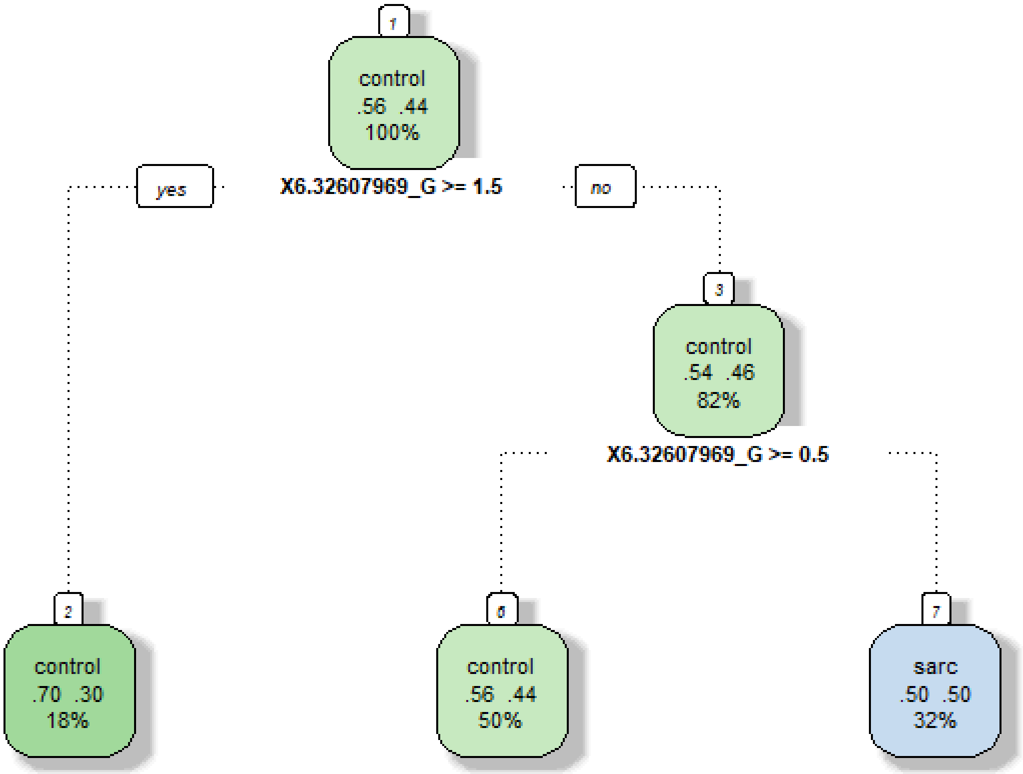

Figure 5. Simple binary decision tree using a single SNP, 6:32607969_G. The labels inside the nodes represent: majority class in that node, % of observations from the "control" class, % of observations from the "sarcoidosis" class, and % of total observations in that node. We follow the tree by evaluating the conditional on each branch, if TRUE go left, if FALSE go right. The class assignment in the leaf nodes can then be used for classification.

Decision trees benefit from being highly interpretable, however they suffer from an inherent instability, as decision trees are very sensitive to the sample chosen. Addition or deletion of even a single observation could change the resulting tree, and thus individual trees are considered unstable. Tree ensembles, or "Forest" methods have been developed to improve the robustness of decision tree classifiers, including methods such as Bagging (Breiman, 1996), Boosting (Freund & Schapire, 1996), and Random Forests (Breiman, 2001).

Bagging operates by generating bootstrapped samples (with replacement) from the input training data, then generating a tree for each sample. Class prediction is done by majority vote across all the trees. Boosting operates by generating weighted errors for each constructed tree, and generating a new tree based on the residual errors from the previous tree to minimize the final error. Random forests operate by also generating bootstrapped samples (with replacement), but instead of using all the columns to generate individual trees like Bagging, a randomly selected subset of columns is used. This enables more diversity across trees compared with Bagging and tends to improve the final prediction accuracy by “de-correlating” the individual trees. Construction of a RF model follows these steps (X. Chen & Ishwaran, 2012):

1) Sample *ntree* bootstrapped samples from the input data, with replacement.
2) Grow a decision tree using the CART methodology and the Gini Index for each *ntree* sample. At each node in the tree, consider a random set of *mtry* columns to decide how to split. Grow the tree until the sample size in each node reaches a specified *nodesize* value, or until the tree can no longer be split further.
3) Aggregate the information across all *ntree* nodes so that new data can be classified by majority vote.
4) Compute the out-of-bag (OOB) error rate for each of the *ntrees* by using the data left out of the bootstrapped sample as test data.

Another key advantage of forest-based approaches is that variable importance can be measured, and thus the algorithm can be used for feature selection. Variable importance is calculated by measuring how much a given variable improves the information gain across the trees. This importance metric is calculated by evaluating how much a given variable decreases the Gini Index, summed for every node where that variable is used in each tree and normalized by the number of trees. This procedure allows us to determine which variables have the most information to assist classification and rank their importance. As such, RF models have been used in “high-*p*, low-*n*” problems (high feature count, low sample count problems) such as genomic prediction with SNPs (X. Chen & Ishwaran, 2012).

3.2 - Random Forest Modeling – Methodology

All RF models were generated with 10-fold cross validation, repeated three times to reduce overfitting, using the *train* function in the Caret package in R. K-fold Cross validation is a technique which reduces overfitting by randomly separating the input data into *k* discrete partitions, then leaving one subset out of the training process to use as a validation set while the other *k-1* partitions are used as a training set. The process is repeated until all *k* subsets have been used as the validation set, and the results of all *k* models are averaged to generate a final model. One advantage of this approach is that every observation is guaranteed to be used exactly once in the validation set. We repeated the 10-fold cross validation three times and averaged the results from all three iterations to further reduce overfitting and to help create a more generalizable model on real test data.

Additionally, we optimized the hyperparameter *mtry* (the number of columns randomly selected for each node in each individual decision tree). We used a range of five evenly spaced numbers ranging from one to the total number of columns, to try and find a near-optimal value of *mtry* in reasonable computational time. Each value of *mtry* was tested on a full RF with 10-fold cross validation repeated three times, and the optimal Kappa value was measured for each of the fitted models, and the optimal *mtry* was selected for use in the final model. The Kappa statistic is a modified accuracy statistic that also considers the expected accuracy based on random chance. Kappa ranges from -1 to 1, with a Kappa of 1 indicating perfect classification, Kappa of 0 indicating accuracy exactly in line with the expected values (a perfectly random classifier), and a Kappa of -1 indicating a perfectly opposite classification. Optimal values of *mtry* are described in Table 2.

Optimal *mtry* values	≤0.75	≤0.8	≤0.85	≤0.9	≤0.95
Logistic Top10 SNPs	2	2	2	2	2
Logistic Top100 SNPs	1	3	3	1	1
Logistic Top500 SNPs	57	65	74	85	93
Logistic Top1000 SNPs	70	73	160	134	201
Logistic Top2000 SNPs	92	284	205	449	505
LASSO 484 SNPs	87	174	175	352	266

Table 2. Optimal values for the random forest hyperparameter *mtry* after 10-fold cross validation, repeated three times. These values were generated without the use of test data to ensure test data remained independent from model construction. Trained models were created with these optimal values and applied to the test data to evaluate final model performance.

3.3 - Random Forest Modeling – Results

Each trained model was applied to the test data to evaluate model performance. The test data was not involved in the training or validation or hyperparameter optimization process, and thus represents a real-world test of the model's generalizability to new data. We plotted the area under the receiving operating characteristic curve (AUC), which is a common metric used to evaluate model performance (Figure 6), as well as plotted the Kappa values (Figure 7). Each model was also evaluated in comparison to the "no-information model", which is the accuracy you would expect if you simply always predicted the majority class. In this case, our test data contained 55.58% controls, which means that a no-information model would simply always predict that a person is a control and reach exactly 55.58% accuracy automatically. Therefore, we are only interested in models that can outperform this baseline. In this experiment, only one model statistically outperformed the no-information model, the top500 SNP model with an r-squared cutoff of 0.9. This model achieved an absolute accuracy of 0.5995 (p = 0.041), a Kappa of 0.1599, a sensitivity of 0.377, and a specificity of 0.777. See Table 3 for a full breakdown of the results from each model.

Random Forest Results	Accuracy	Accuracy > Null Accuracy? (P-value)	Kappa	Sensitivity	Specificity	AUC
Logistic Top10 - 0.95	0.5364	0.8005	0.0275	0.306	0.7205	0.5367
Logistic Top100 – 0.95	0.5801	0.1732	0.1121	0.3169	0.7904	0.5785
Logistic Top100 – 0.90	0.5777	0.1998	0.1044	0.3005	0.7991	0.5759
Logistic Top100 – 0.85	0.5777	0.1998	0.1225	0.3989	0.7205	0.5827
Logistic Top100 – 0.80	0.5631	0.4028	0.0933	0.388	0.7031	0.5843
Logistic Top100 – 0.75	0.5607	0.4416	0.0678	0.2787	0.786	0.5827
Logistic Top500 – 0.95	0.5752	0.2288	0.1059	0.3333	0.7686	0.5844
Logistic Top500 – 0.90	**0.5995**	**0.041 (*)**	**0.1599**	**0.377**	**0.7773**	**0.6019**
Logistic Top500 – 0.85	0.5898	0.0901	0.1375	0.3552	0.7773	0.5914
Logistic Top500 – 0.80	0.5947	0.0618	0.1477	0.3607	0.7817	0.5859
Logistic Top500 – 0.75	0.5825	0.1489	0.1217	0.3443	0.7729	0.5792
Logistic Top1000 – 0.95	0.585	0.127	0.1152	0.2842	0.8253	0.5931
Logistic Top1000 – 0.90	0.5825	0.1489	0.1126	0.2951	0.8122	0.5916
Logistic Top1000 – 0.85	0.5898	0.0901	0.1246	0.2842	0.8341	0.5973
Logistic Top1000 – 0.80	0.5874	0.1074	0.1209	0.2896	0.8253	0.6073
Logistic Top1000 – 0.75	0.5704	0.2932	0.0767	0.2295	0.8428	0.5921
Logistic Top2000 – 0.95	0.5631	0.4028	0.0594	0.2131	0.8428	0.5769
Logistic Top2000 – 0.90	0.5825	0.1489	0.1012	0.235	0.8603	0.5886
Logistic Top2000 – 0.85	0.585	0.127	0.1017	0.2131	0.8821	0.5973
Logistic Top2000 – 0.80	0.5752	0.2288	0.0731	0.1639	0.9039	0.5983
Logistic Top2000 – 0.75	0.5752	0.2288	0.072	0.1585	0.9083	0.5993
LASSO – 0.95	0.5631	0.4028	0.0461	0.1475	0.8952	0.5868
LASSO – 0.90	0.5874	0.1074	0.1044	0.2022	0.8952	0.6014
LASSO – 0.85	0.5752	0.2288	0.0687	0.1421	0.9214	0.5952
LASSO – 0.80	0.568	0.3283	0.06	0.1694	0.8865	0.5846
LASSO – 0.75	0.568	0.3283	0.0567	0.153	0.8996	0.5947

Table 3. Full results from the Random Forest models. Logistic Top500 with 0.9 cutoff achieved significant improvement (p=0.041) over the no-information model.

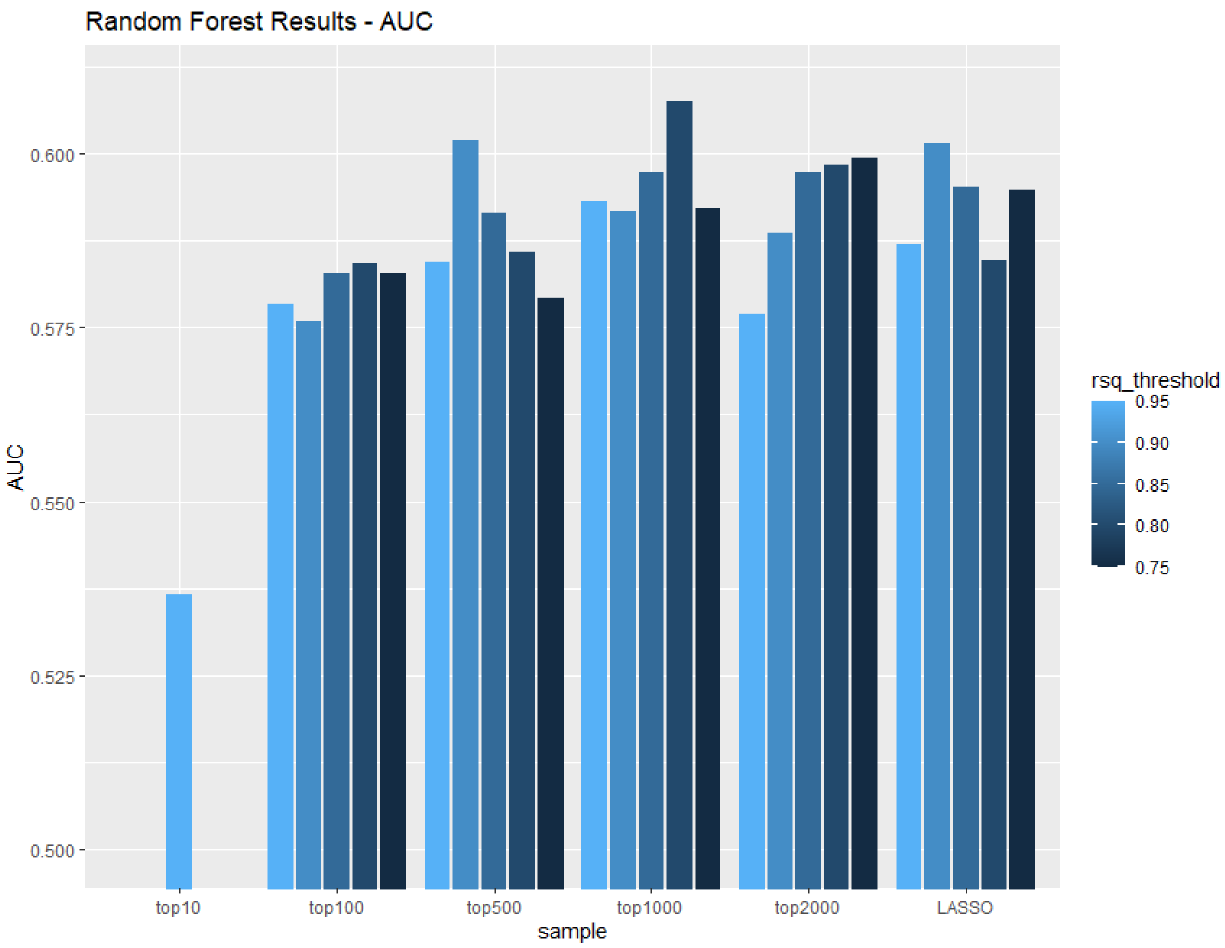

Figure 6. Area under the ROC curve for each SNP subset across five Pearson r^2 threshold cutoffs. Each bar represents a fully trained 10-fold cross validated 3x repeated random forest model.

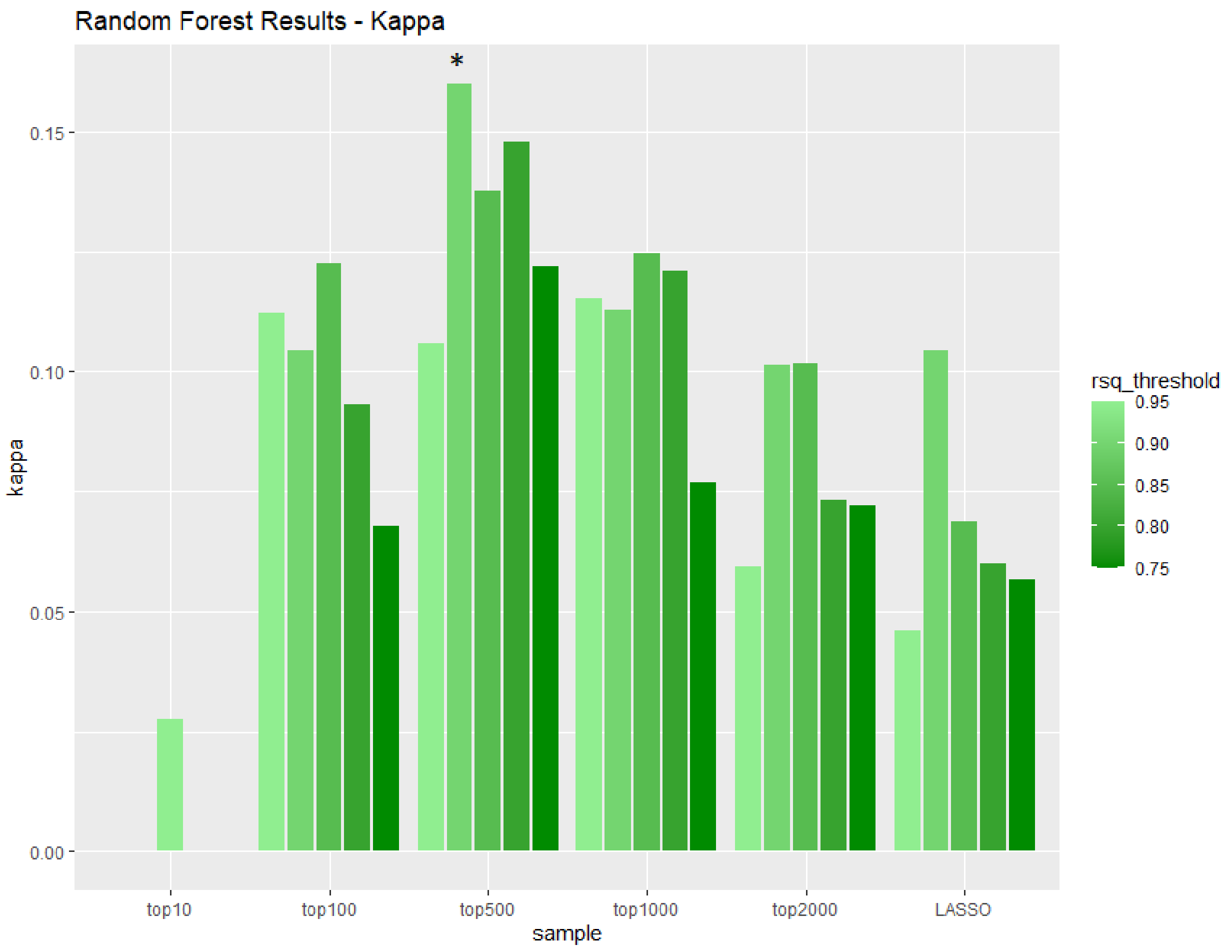

Figure 7. Random forest models - Kappa values for each SNP subset across five Pearson r^2 threshold cutoffs. Star indicates that the model was significantly more accurate ($p < 0.05$) compared to the no-information model. Each bar represents a fully trained 10-fold cross validated 3x repeated random forest model.

3.4 - Support Vector Machine with Radial Kernel – background

After running the RF models, we wondered if we could improve the results by utilizing support vector machines (SVM). SVMs are a powerful class of machine learning algorithms that have been used successfully in genomic prediction of traits in plant and animal breeding, analyzing genetic subtypes of cancer, discovery of active epigenic regions of the genome, and prediction of coronary heart disease, chronic kidney disease, and diabetes, among others diseases (Harimoorthy & Thangavelu, 2020; Huang et al., 2018; Zhang et al., 2017; Zhao et al., 2020). SVMs operate by drawing a hyperplane through your data to maximally separate the classes. To select an optimal hyperplane separator, a SVM works to maximize the distance between the hyperplane and the nearest datapoints. The nearest datapoints to the hyperplane are called the *support vectors*, and SVMs use the support vectors when calculating the class of a new observation. This technique relies on the ability to draw a linear hyperplane capable of separating your observations. However, for complex non-linear problems, a hyperplane drawn in the current problem dimension will generally not result in a good separator. To enable non-linear classification, SVMs have adopted the "Kernel Trick", which applies a function to each observation which projects the data to a higher dimensional space. Good linear separations can usually be found in higher-dimensional space, enabling the SVM to operate normally but still learn non-linear relationships.

Several kernel functions exist which can project data to a higher dimension. Polynomial functions, the radial-basis function, and the sigmoid function are three examples widely used by SVMs. In general, the radial-basis function works well on a variety of data types and was selected as the kernel function for use in this project. Each kernel has unique parameters that can be tuned to further optimize the model's performance. In the case of the radial-basis kernel SVM, the hyperparameters needed to tune are the *sigma* and the *soft margin cost parameter*. Sigma controls how much influence a support vector has on determining the final class of the predicted data point. Lower sigma means less influence of individual support vectors, which increases the variance but decreases the bias. The soft margin cost parameter, or *C*, controls how soft or hard the decision boundary is, larger values of C allow less mis-classified datapoints, but too hard of a decision boundary can hurt the overall generalizability of the model on new data. As with all hyperparameters, a range of values should be tested to determine the optimal values for a given problem. To optimize *C* and *sigma* for the SVM, we utilized the *train* function in the Caret package in R. We used a grid search with 15 total combinations of *C* and *Sigma* and ran 10-fold cross validation with 3x repeats to find the optimal combination of parameters (Table 4).

Optimal *sigma* and *C* values	≤0.75	≤0.8	≤0.85	≤0.9	≤0.95
Logistic Top100 SNPs	(0.069, 0.5)	(0.056, 8)	(0.056, 8)	(0.045, 8)	(0.040, 2)
Logistic Top500 SNPs	(0.008, 0.5)	(0.007, 0.25)	(0.006, 0.25)	(0.005, 0.25)	(0.005, 0.5)
Logistic Top1000 SNPs	(0.003, 0.25)	(0.003, 1)	(0.003, 1)	(0.002, 1)	(0.002, 1)
Logistic Top2000 SNPs	(0.001, 0.5)	(0.001, 0.5)	(0.001, 0.5)	(0.001, 0.5)	(0.001, 1)
LASSO 484 SNPs	(0.001, 0.25)	(0.001, 0.25)	(0.001, 0.25)	(0.001, 0.25)	(0.001, 0.25)

Table 4. Optimal values for the SVM radial kernel hyperparameter sigma and C after 10-fold cross validation, repeated 3x. These values were generated without the use of test data to ensure test data remained independent from model construction. Trained models were created with these optimal values and applied to the test data to evaluate final model performance.

3.5 - Support Vector Machine – Results

As with RF, we plotted both the AUC (Figure 8) and the Kappa value (Figure 9) for each model to visually inspect relative quality. In general, we saw an increase in performance up to the Top1000 models which performed the best. Performance then decreased at the Top2000 model and the LASSO model, both of which contain many more features. We found that SVM utilizing the top1000 SNP 0.75 r^2 model outperformed the best RF model in terms of absolute accuracy, achieving accuracy of

0.6068 (p = 0.0207), kappa of 0.1858, sensitivity of 0.4484, specificity of 0.7336, and an AUC of 0.6016. Full results, including accuracy, Accuracy p-value, Kappa, sensitivity, specificity, and AUC, are available for each model (Table 5).

SVM Radial Kernel Results	Accuracy	Accuracy > Null Accuracy? (P-value)	Kappa	Sensitivity	Specificity	AUC
Logistic Top100 – 0.95	0.5874	0.1074	0.125	0.3115	0.8079	0.5885
Logistic Top100 – 0.90	0.551	0.5985	0.0526	0.2951	0.7555	0.58
Logistic Top100 – 0.85	0.5558	0.5205	0.0597	0.2842	0.7729	0.5775
Logistic Top100 – 0.80	0.5607	0.4416	0.0646	0.2623	0.7991	0.5799
Logistic Top100 – 0.75	0.5704	0.2932	0.0905	0.3005	0.786	0.564
Logistic Top500 – 0.95	0.5801	0.1732	0.1261	0.3934	0.7293	0.5823
Logistic Top500 – 0.90	0.5752	0.2288	0.118	0.3989	0.7162	0.5928
Logistic Top500 – 0.85	0.5777	0.1998	0.1235	0.4044	0.7162	0.59
Logistic Top500 – 0.80	0.5801	0.1732	0.129	0.4098	0.7162	0.5904
Logistic Top500 – 0.75	0.5558	0.5205	0.0777	0.377	0.6987	0.5724
Logistic Top1000 – 0.95	**0.6044**	**0.0262 (*)**	**0.1766**	**0.4208**	**0.7511**	**0.5877**
Logistic Top1000 – 0.90	**0.6044**	**0.0262 (*)**	**0.1785**	**0.4317**	**0.7424**	**0.5885**
Logistic Top1000 – 0.85	0.5922	0.0749	0.1518	0.4098	0.738	0.5884
Logistic Top1000 – 0.80	0.5801	0.1732	0.129	0.4098	0.7162	0.5917
Logistic Top1000 – 0.75	**0.6068**	**0.0207 (*)**	**0.1858**	**0.4481**	**0.7336**	**0.6016**
Logistic Top2000 – 0.95	0.551	0.5985	0.0676	0.3716	0.6943	0.5629
Logistic Top2000 – 0.90	0.5461	0.6728	0.0628	0.3934	0.6681	0.5643
Logistic Top2000 – 0.85	0.5437	0.7077	0.0562	0.3825	0.6725	0.5645
Logistic Top2000 – 0.80	0.5485	0.6363	0.0673	0.3934	0.6725	0.5722
Logistic Top2000 – 0.75	0.5388	0.7717	0.0461	0.377	0.6681	0.5701
LASSO – 0.95	0.534	0.8269	0.0489	0.4372	0.6114	0.5625
LASSO – 0.90	0.5267	0.8922	0.0335	0.4262	0.607	0.5613
LASSO – 0.85	0.5267	0.8922	0.0335	0.4262	0.607	0.5605
LASSO – 0.80	0.5364	0.8005	0.0544	0.4426	0.6114	0.5619
LASSO – 0.75	0.5316	0.8511	0.0434	0.4317	0.6114	0.5615

Table 5. Full results from the SVM models. Logistic Top1000 with 0.95 and 0.90 cutoff (p=0.0262), as well as Top1000 with 0.75 cutoff (p=0.0207) achieved significant improvement over the no-information model.

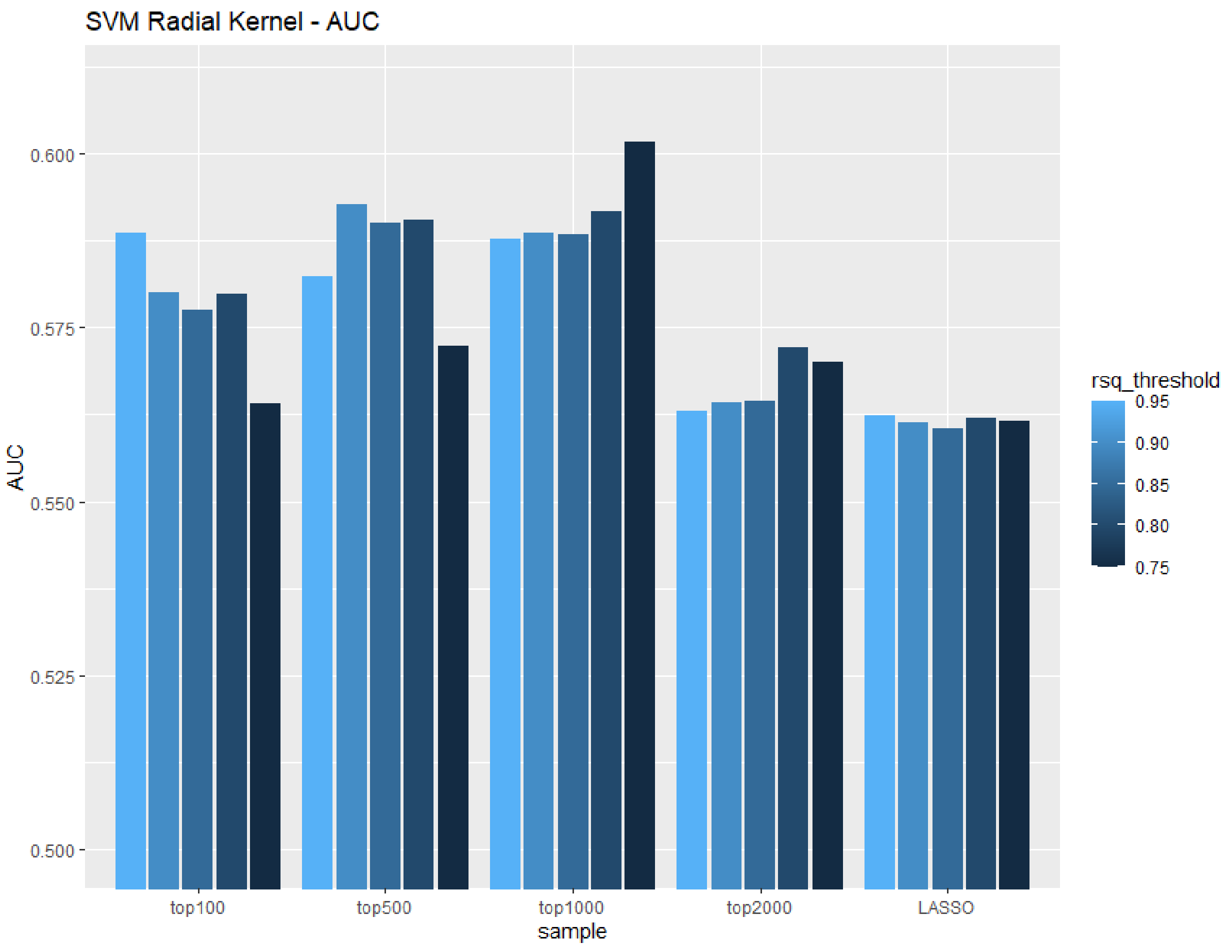

Figure 8. Support vector machine models - Area under the ROC curve for each SNP subset across five Pearson r^2 threshold cutoffs. Each bar represents a fully trained 10-fold cross validated repeated three times repeated SVM model.

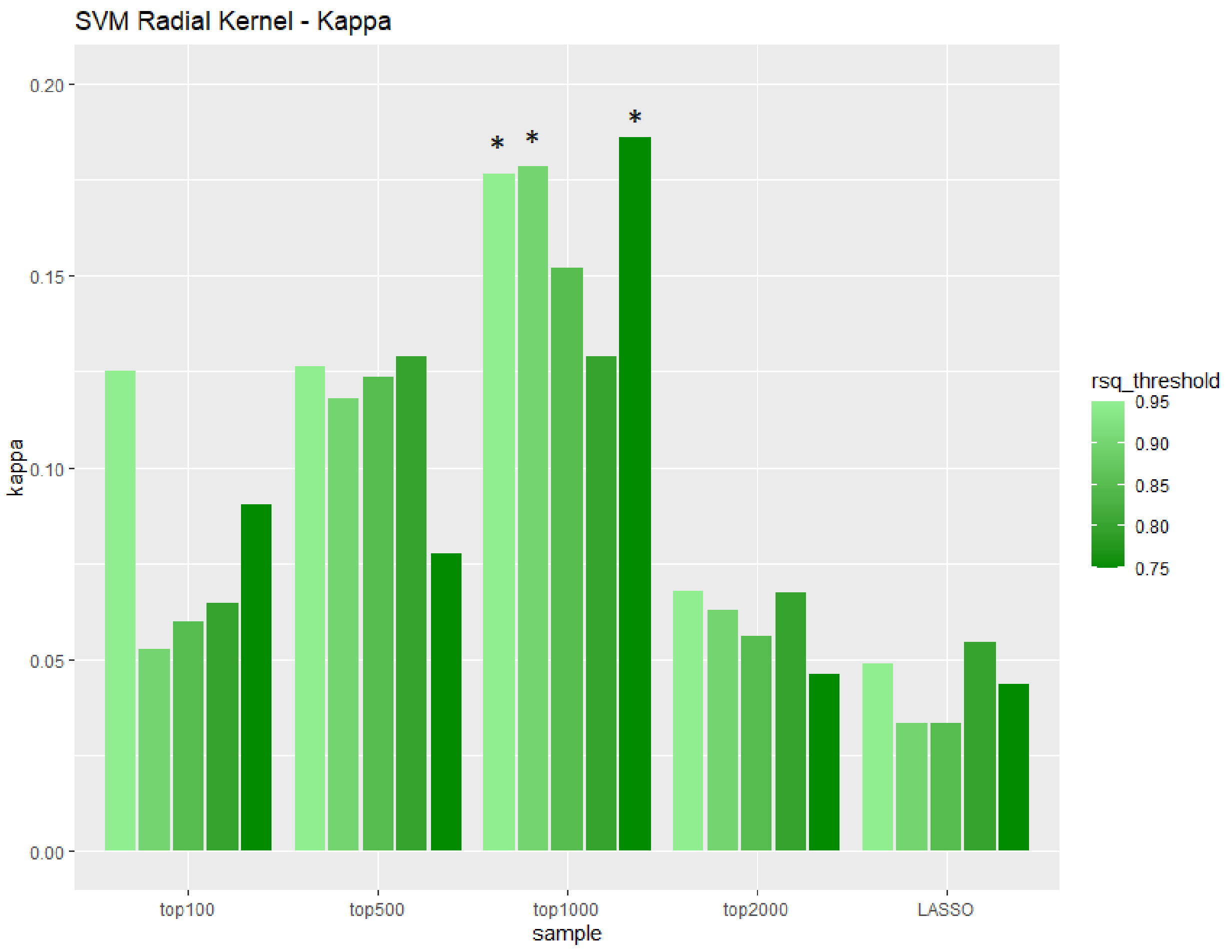

Figure 9. Support vector machine models - Kappa values for each SNP subset across five Pearson r^2 threshold cutoffs. Star indicates that the model was significantly more accurate ($p < 0.05$) compared to the no-information model. Each bar represents a fully trained 10-fold cross validated repeated three times random forest model.

3.6 - Neural Network modeling – background

After SVM modeling showed some improvement compared to RF, we wondered if we could gain even more performance by utilizing neural networks (NNs). NNs have become increasingly popular supervised machine learning tools, in part due to their ability to flexibly learn non-linear functions, and have widespread use and applications in many domains, including disease prediction. At the most basic level, a NN is a directed graph with weighted edges which convert raw input with *n*-dimensional features into a single-valued outcome, usually a probability of that observation belonging to a particular class. The basic until of many NNs is the sigmoid neuron, which converts a vector of input values (multiplied by their respective weights) into a (0,1) interval. If the sigmoid neuron is the final neuron in the network, the output value represents the probability of that observation belonging to a given class. For the model to learn, we must supply class labels for each observation in the training set and evaluate the accuracy of the final output based on an evaluation function. If the output does not match the class label, we must update the weights to minimize the evaluation function's error through a process known as *back-propagation*. Usually, back-propagation is achieved with a process known as gradient descent. Gradient descent is a non-linear optimization technique which calculates the derivative of the evaluation function with respect to the weights and updates the weights to lower the overall error (the function takes a step "downhill" to reach a lower error value). This process of backpropagation is repeated until weights are stabilized (they have reached convergence at the local minima), or until a set number of backpropagation cycles have occurred (known as early-stopping).

A single-layer NN is one in which input features are directly connected to the output neuron. A network of this type is only capable of learning direct linear relationships. To increase the flexibility of NNs, multi-layer networks have been developed which include the addition of one or more layers of "hidden neurons", which are intermediate sigmoid neurons. The addition of a hidden layer(s) enables the model to flexibly learn non-linear functions. In general, the more hidden layers you add, the more abstract the function is you can learn. However special consideration must be made to preserve the gradient across deep networks, which is known to rapidly deplete (known as the *vanishing gradient problem*). Recent advances in the field have helped to alleviate this issue, enabling many-layer networks (known as *deep learning*) to become useful and powerful techniques for learning solutions to very complex problems, like how to select winning moves in the board games Go and Chess (Silver et al., 2018), or how to drive a car in full self-driving vehicles (*Autopilot AI - Tesla*, 2021). However, deep learning NNs generally require and greatly benefit from large volumes of training examples to accurately learn their very abstract functions.

In this work, we applied a multi-layer NN with sigmoid neurons and a single hidden layer to our training data, using the *nnet* package in R. An example of the network architecture on the Top100 0.9 r^2 model is shown (Figure 10). The hyperparameters we were able to optimize were *size* and *decay*. *Size* is the number of neurons in the hidden layer. *Decay* is an additional parameter used to control how large the weights can become, which can help avoid overfitting your model by preventing a few very large weights dominate the model. Specifically, *decay* multiplied by the L2 norm of all the weights is added to the evaluation function at each iteration of backpropagation, which effectively keeps the weights at a small size. A grid of hyperparameters was tested, using *size* = 1, 3, or 5, and *decay* = 0, 0.01, 0.1, 0.2, and 0.5. The size of the hidden layer was kept small to reduce computation time. A table showing the optimized hyperparameters is shown in Table 6.

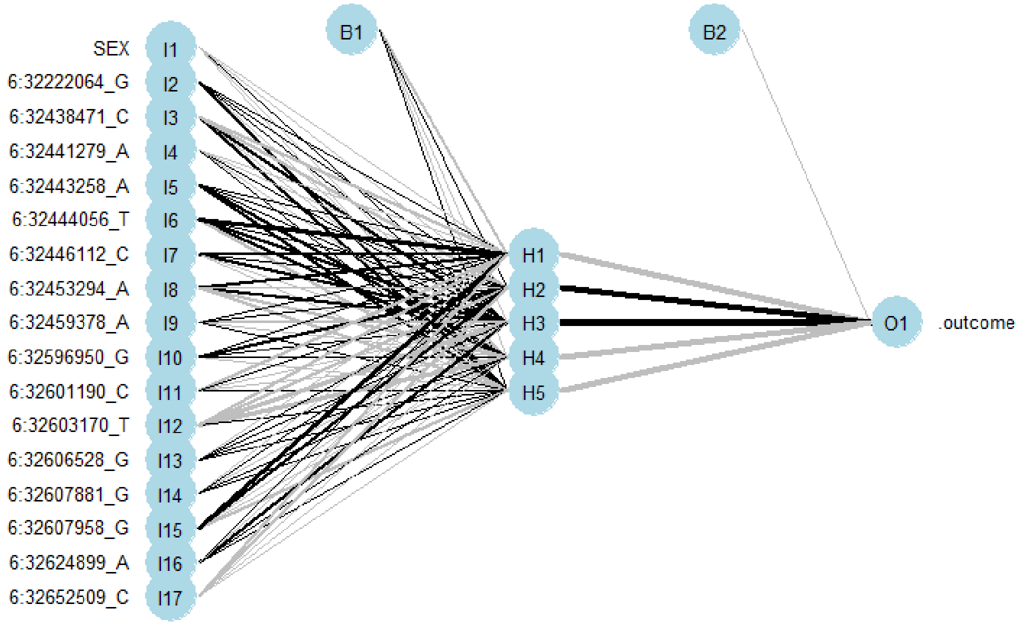

Figure 10. Single hidden layer neural network architecture for the Top100 0.9 r^2 cutoff model (16 SNPs + Sex used as features).

Optimal *size* and *decay* values	≤0.75	≤0.8	≤0.85	≤0.9	≤0.95
Logistic Top10 SNPs	(1, 0.5)	(1, 0.5)	(1, 0.5)	(1, 0.5)	(1, 0.5)
Logistic Top100 SNPs	(1, 0.00)	(1, 0.01)	(3, 0.1)	(5, 0.5)	(5, 0.5)
Logistic Top500 SNPs	(1, 0.5)	(1, 0.1)	(1, 0.2)	(1, 0.1)	(1, 0.1)
Logistic Top1000 SNPs	(1, 0.5)	(1, 0.5)	(1, 0.5)	(1, 0.5)	(1, 0.5)
Logistic Top2000 SNPs	(5, 0.5)	(5, 0.5)	(5, 0.00)	(5, 0.5)	(5, 0.5)
LASSO 484 SNPs	(5, 0.5)	(5, 0.5)	(5, 0.5)	(5, 0.5)	(5, 0.5)

Table 6. Optimal values for the neural network hyperparameters *size* and *decay* after 10-fold cross validation, repeated three times. These values were generated without the use of test data to ensure test data remained independent from model construction. Trained models were created with these optimal values and applied to the test data to evaluate final model performance.

3.6 - Neural Network modeling – results

As with RF and SVM, we recorded the full results for each model (Table 7) as well as plotted the AUC (Figure 11) and the Kappa value (Figure 12) to visually inspect the quality of each model. No models achieved significantly better accuracy compared with the naïve model; however, the Top100 0.85 and Top100 0.80 models came very close, achieving accuracies of 0.5971 (p = 0.051). The top100 0.85 model had a Kappa of 0.1657, AUC of 0.6103, sensitivity of 0.4372, and specificity of 0.7249, while the top100 0.8 model had a Kappa of 0.1759, AUC of 0.6042, sensitivity of 0.4973, and specificity of 0.6769. Overall performance of the NNs tended to decrease as the feature space grew. One possible reason for this decrease could be that the hidden layer was too small relative to the number of features, which limited the total number of interactions that could be modeled which in turn hurt performance. We would like to investigate this question in more detail in the future.

Neural Network Results	Accuracy	Accuracy > Null Accuracy? (P-value)	Kappa	Sensitivity	Specificity	AUC
Logistic Top10 – 0.95	0.5485	0.6363	0.0348	0.2295	0.8035	0.5375
Logistic Top100 – 0.95	0.5825	0.1489	0.1277	0.377	0.7467	0.5957
Logistic Top100 – 0.90	0.5922	0.0749	0.1441	0.3661	0.7729	0.5938
Logistic Top100 – 0.85	**0.5971**	**0.0505**	**0.1657**	**0.4372**	**0.7249**	**0.6103**
Logistic Top100 – 0.80	**0.5971**	**0.0505**	**0.1759**	**0.4973**	**0.6769**	**0.6042**
Logistic Top100 – 0.75	0.5874	0.1074	0.1504	0.4536	0.6943	0.5986
Logistic Top500 – 0.95	0.5316	0.8511	0.0528	0.4809	0.5721	0.5545
Logistic Top500 – 0.90	0.5267	0.8922	0.0388	0.4536	0.5852	0.5564
Logistic Top500 – 0.85	0.5583	0.481	0.0822	0.377	0.7031	0.5627
Logistic Top500 – 0.80	0.5485	0.6363	0.0735	0.4262	0.6463	0.5674
Logistic Top500 – 0.75	0.5485	0.6363	0.0662	0.388	0.6769	0.5609
Logistic Top1000 – 0.95	0.5655	0.3649	0.1039	0.4208	0.6812	0.574
Logistic Top1000 – 0.90	0.5752	0.2288	0.1249	0.4372	0.6856	0.5792
Logistic Top1000 – 0.85	0.5947	0.0618	0.1611	0.4372	0.7205	0.5819
Logistic Top1000 – 0.80	0.5655	0.3649	0.1029	0.4153	0.6856	0.5892
Logistic Top1000 – 0.75	0.5752	0.2288	0.1219	0.4208	0.6987	0.5855
Logistic Top2000 – 0.95	0.5485	0.6363	0.0735	0.4262	0.6463	0.5635
Logistic Top2000 – 0.90	0.5413	0.7407	0.057	0.4098	0.6463	0.5423
Logistic Top2000 – 0.85	0.5558	0.5205	0.0777	0.377	0.6987	0.5576
Logistic Top2000 – 0.80	0.5631	0.4028	0.0943	0.3934	0.6987	0.5774
Logistic Top2000 – 0.75	0.5461	0.6728	0.0607	0.3825	0.6769	0.5539
LASSO – 0.95	0.5461	0.6728	0.0731	0.4481	0.6245	0.5513
LASSO – 0.90	0.5437	0.7077	0.0635	0.4208	0.6419	0.544
LASSO – 0.85	0.5558	0.5205	0.095	0.4699	0.6245	0.5465
LASSO – 0.80	0.5607	0.4416	0.1029	0.4645	0.6376	0.5628
LASSO – 0.75	0.5461	0.6728	0.0721	0.4426	0.6288	0.5517

Table 7. Full results from the neural network models. No models achieved significant improvement $p< 0.05$ over the no information model, however Logistic Top100 with 0.85 and 0.80 cutoff were very close with $p=0.0505$. Despite this, the AUC was quite high for these models, with 0.6103 and 0.6042, respectively.

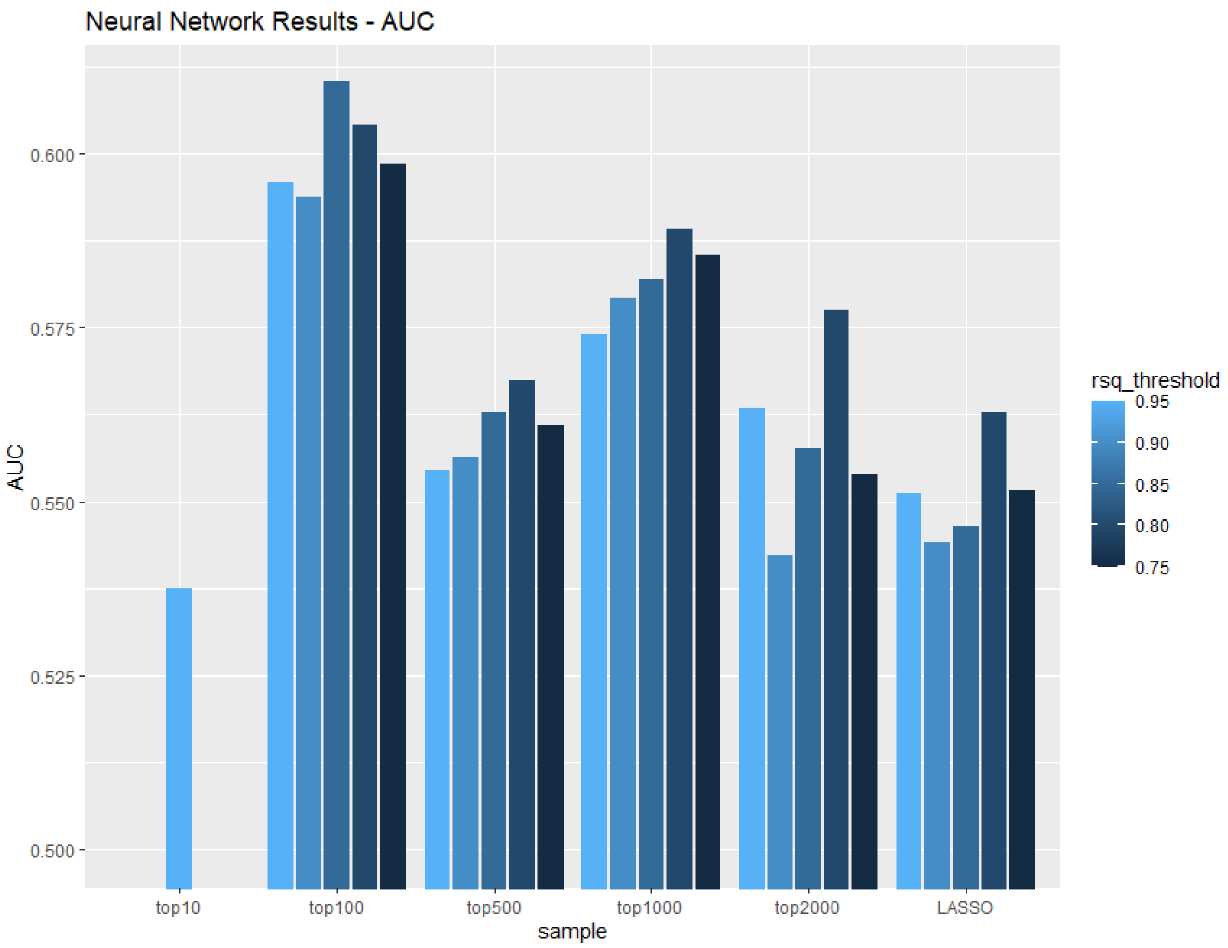

Figure 11. Neural network models - Area under the ROC curve for each SNP subset across five Pearson r^2 threshold cutoffs. Each bar represents a fully trained 10-fold cross validated repeated three times NN model.

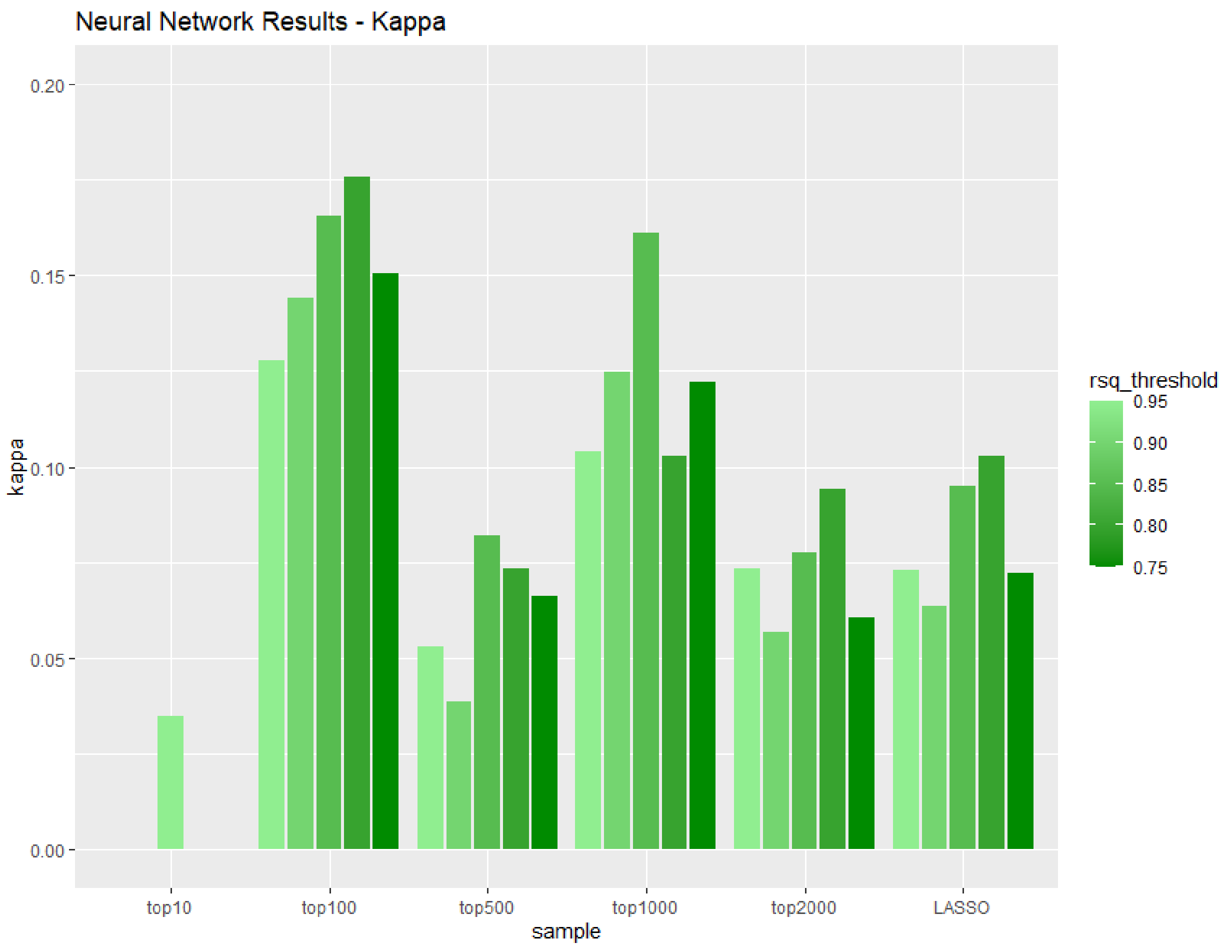

Figure 12. Neural network models - Kappa values for each SNP subset across five Pearson r^2 threshold cutoffs. Each bar represents a fully trained 10-fold cross validated repeated three times NN model.

Chapter 4 – Functional Score Assignment and Modeling

4.1 Functional Score Assignment – CADD and FATHMM-XF

We then determined if blending the logistic regression feature selection with a measurement of the likelihood of a SNP to contribute to disease would improve modeling performance. We chose CADD scores and FATHMM-XF scores for this task, in part because they are both available for most of the SNPs in the GRCh38 genome assembly, allowing annotation of most SNPs. Both resources also have online tools (https://cadd-staging.kircherlab.bihealth.org/score and http://fathmm.biocompute.org.uk/fathmm-xf/) which enable a user to simply upload a list of SNPs and download the relevant scores.

For CADD scores, we first filtered out poorly correlate SNPs by sorting the logistic regression results by p-value and extracting all SNPs with $p < 0.05$. With this subset of 429,870 SNPs, we used the SNP ID to extract CADD scores and obtained 80.5% of SNPs (346,384 / 429,870) with labeled CADD scores. All FATHMM scores were downloaded from the website and merged with all the logistic regression SNPs. A total of 92.57% of SNPs (7,149,617 / 7,723,468) had labeled FATHMM-XF scores. 91.3% of these SNPs (316,276 / 346,384) had both CADD and FATHMM-XF scores. Figure 13 shows the density of FATHMM-XF and CADD scores for all 316,276 of these SNPs. Most have very low functional scores, which is to be expected since most SNPs in the genome have low functional scores.

We then merged the CADD scores or the FATHMM-XF scores with the logistic regression p-values in R and applied the following formulas to generate a “blended” score.

$$cadd\_logistic\_score \ = log_c\left(\frac{1}{p.val}\right) + cadd\_score$$

$$fathmm\_logistic\_score \ = log_c\left(\frac{1}{p.val}\right) + fathmm\_score$$

$$c \in (1, \infty)$$

This formulation allows us to adjust the proportion of p-value contributing to the overall score versus the proportion of functional score. As *c* increases, the contribution of the p-value to the overall score decreases. As *c* approaches 1, the contribution of the p-value to the overall score becomes increasingly large. For all initial experiments, the value of c was set to 2.

It should be noted that the CADD scores range from 0 to 99 based on their pathogenicity rank relative to all other possible 8.6 billion substitutions in the human genome, while the FATHMM-XF scores range from 0 to 1. In practice, this difference in potential values meant that in all initial experiments, the relative contribution from the logistic p-value was higher for the *fathmm_logistic* scores compared to the *cadd_logistic* scores (Figure 14). Future experiments could correct this issue by increasing the log base *c* of the *fathmm_logistic* blended score.

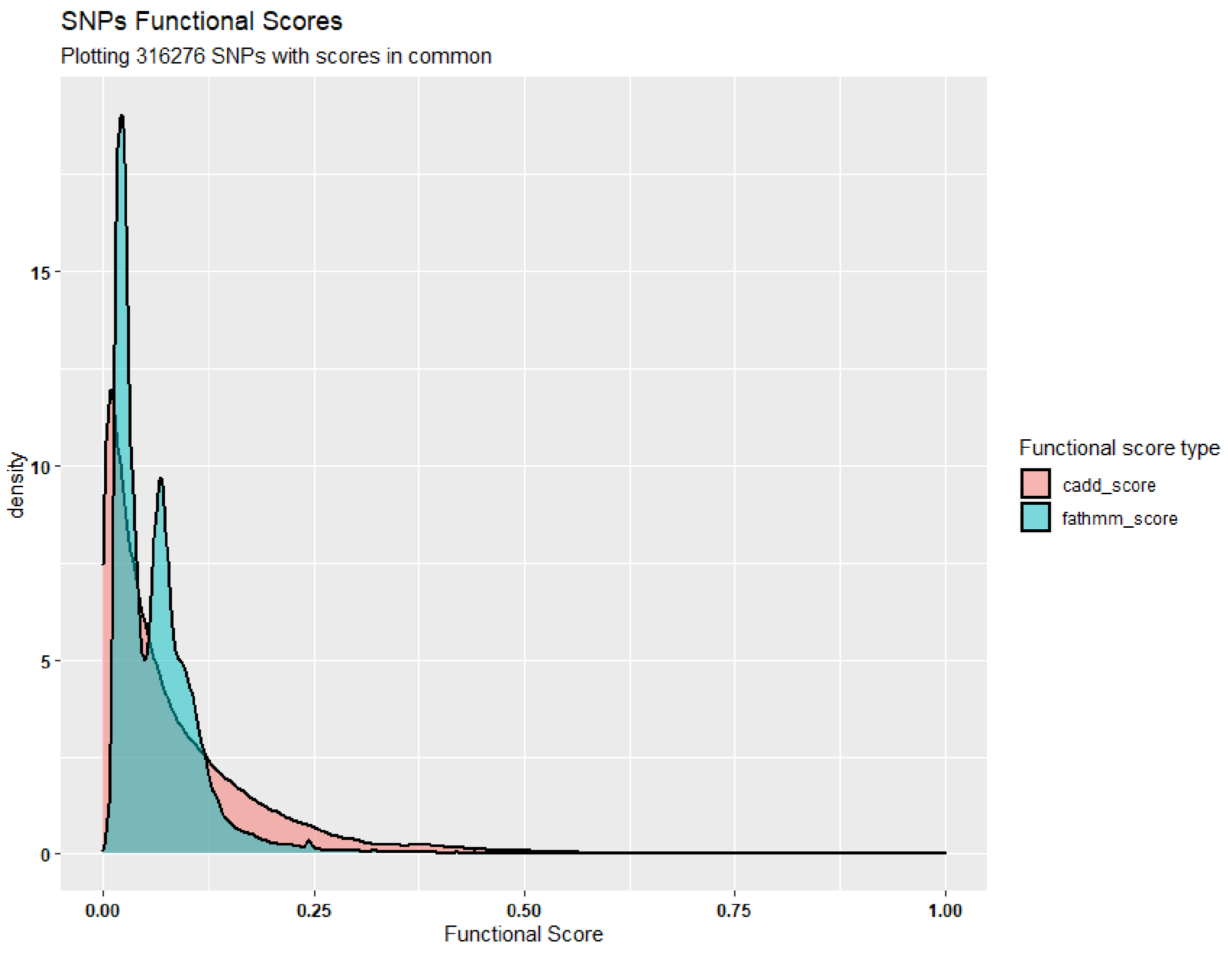

Figure 13. Density plot of the distribution of CADD and FATHMM scores across all 316,276 SNPs with both scores in common. CADD scores were normalized to fit within a 0 to 1 range so that the two scores could be compared.

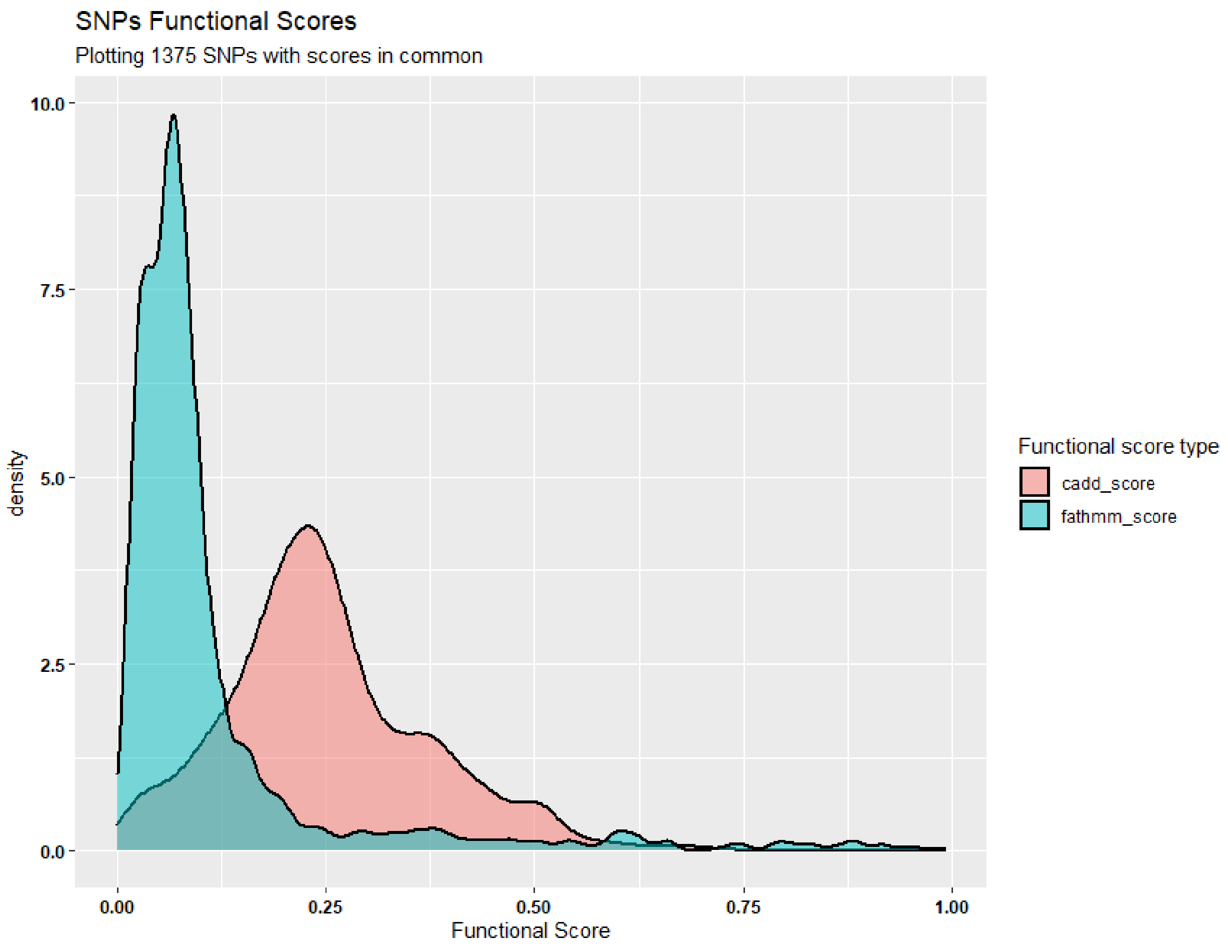

Figure 14. Density plot of the distribution of CADD and FATHMM scores across the top 10,000 SNPs based on the top blended functional scores. 1,375 SNPs were found in both subsets. CADD scores were normalized to fit within a 0 to 1 range. It is clear we are enriching for higher CADD score SNPs, but not enriching as much for higher FATHMM-XF score SNPs.

4.2 CADD + Logistic blended results – Random Forest, SVM, and Neural Network

The top 10, 100, 500, 1000, and 1000 SNPs based on the *cadd_logistic* blended score was extracted and modeled with RF, SVM, and NN as before. To reduce the total computation time, a single Pearson’s r^2 correlation cutoff of 0.9 was used for these experiments, with a slight difference in the method used for SNP removal. Previously, if a pair of SNPs had a correlation coefficient above the threshold, the SNP with the higher mean correlation across all columns was selected for removal. This is an appropriate strategy when the two SNPs are otherwise equal. However, SNPs can have variable functional scores even when they are otherwise correlated, and we are interested in testing the hypothesis that adding functional scores improves predictive performance. Therefore, when a pair of SNPs has a correlation cutoff above the threshold, we would like to keep the SNP with the higher functional score. The total number of SNPs retained in the *cadd_logistic* models are summarized in Table 8. The number of retained SNPs is significantly higher than in the no-functional score models used in Chapter 3. This is an indication that there are fewer correlated variables in the top SNPs because the functional score component of the blended score is boosting the rank of uncorrelated lower p-value SNPs.

Number of SNPs retained CADD + Logistic	≤ 0.9 r^2
Top10 SNPs	9
Top100 SNPs	76
Top500 SNPs	405
Top1000 SNPs	832
Top2000 SNPs	1677

Table 8. Number of SNPs retained after filtering correlated variables for the CADD + logistic blended models. There are significantly more SNPs retained after correlation filtering compared to the baseline no-functional score models.

The models were run as before, optimizing hyperparameters using 10-fold 3x repeated cross validation on the training data before applying the optimal model / hyperparameters to the test data. Hyperparameter optimization is summarized in Table 9. Full results for all models are reported in Table 10. AUC measurements are plotted in Figure 15, and Kappa measurements are plotted in Figure 16. No models achieved significantly better performance compared to the naïve model. In general, the performance on average tended to worsen with this strategy over simply using the most linearly correlated SNPs by logistic regression.

Optimal hyperparameters	NNET – *size* and *decay*	RF – *mtry*	SVM – *sigma* and *C*
cadd_logistic Top10	(3, 0.5)	3	(0.075, 0.25)
cadd_logistic Top100	(5, 0)	47	(0.007, 0.25)
cadd_logistic Top500	(5, 0.5)	163	(0.001, 0.5)
cadd_logistic Top1000	(5, 0.1)	501	(0.001, 0.5)
cadd_logistic Top2000	(5, 0.2)	1008	(0.0003, 0.5)

Table 9. Optimized hyperparameters for all *cadd_logistic* models.

CADD + Logistic Results	Accuracy	Accuracy > Null Accuracy? (P-value)	Kappa	Sensitivity	Specificity	AUC
RF – Top10 0.9	0.5825	0.1489	0.1197	0.3333	0.7817	0.5781
RF – Top100 0.9	0.5583	0.481	0.0707	0.3169	0.7511	0.591
RF – Top500 0.9	0.5825	0.1489	0.1147	0.306	0.8035	0.5859
RF – Top1000 0.9	0.5898	0.0901	0.1185	0.2514	0.8603	0.6005
RF – Top2000 0.9	0.5898	0.0901	0.1112	0.2131	0.8908	0.6027
NNET – Top10 0.9	0.5583	0.481	0.0512	0.2186	0.8297	0.4504
NNET – Top100 0.9	0.5825	0.1489	0.1461	0.4809	0.6638	0.5807
NNET – Top500 0.9	0.5607	0.4416	0.1233	0.5792	0.5459	0.5683
NNET – Top1000 0.9	0.5437	0.7077	0.0718	0.4645	0.607	0.5409
NNET – Top2000 0.9	0.551	0.5985	0.081	0.4426	0.6376	0.5404
SVM – Top10 0.9	0.5777	0.1998	0.0982	0.2678	0.8253	0.566
SVM – Top100 0.9	0.5631	0.4028	0.1034	0.4426	0.6594	0.5921
SVM – Top500 0.9	0.5631	0.4028	0.1103	0.4809	0.6288	0.57
SVM – Top1000 0.9	0.5437	0.7077	0.0759	0.4863	0.5895	0.5874
SVM – Top2000 0.9	0.5728	0.26	0.1272	0.4754	0.6507	0.5996

Table 10. Full results table for the *cadd_logistic* blended models. No models were significantly more accurate compared to the naïve model.

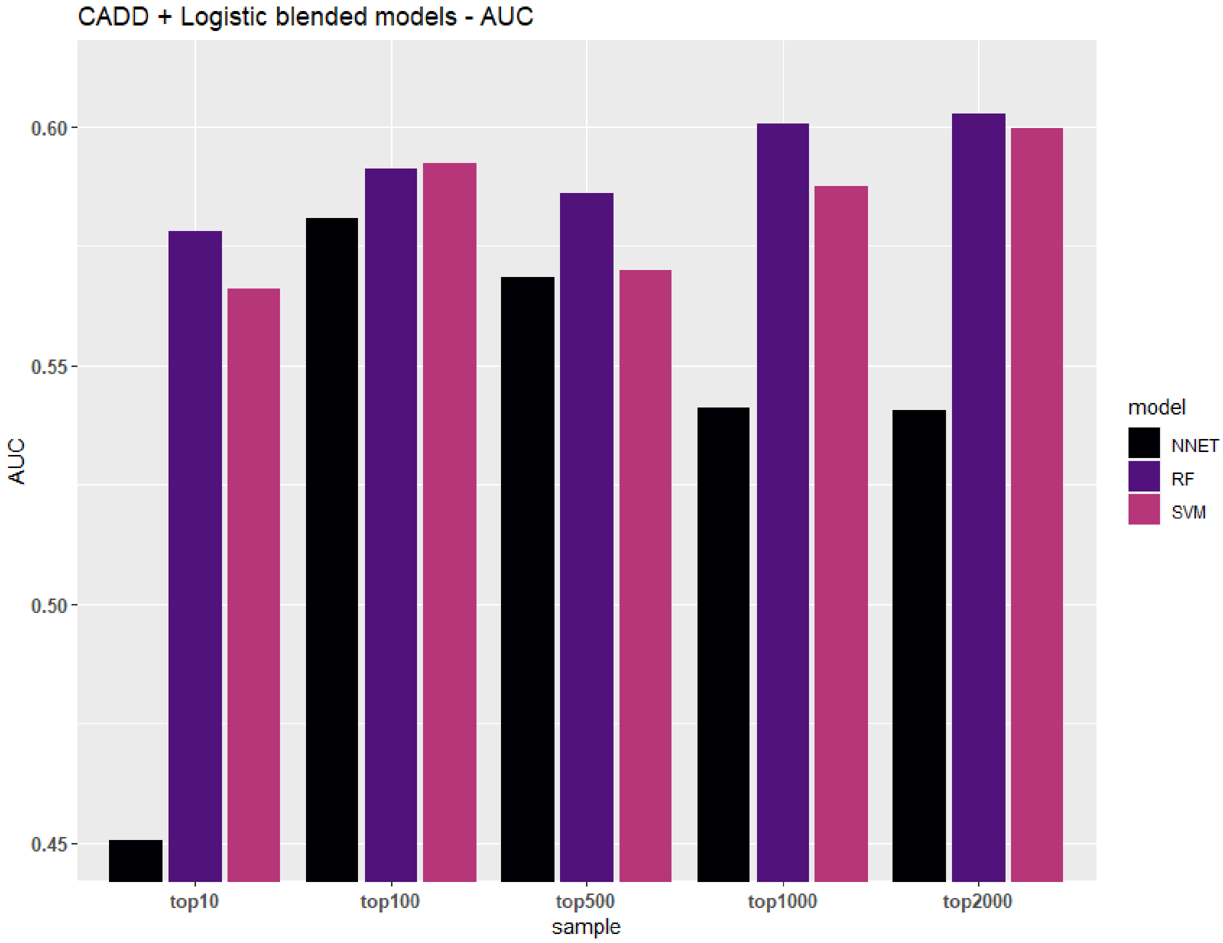

Figure 15. AUC values for all *cadd_logistic* models. In general, we see lower AUC values compared with the no-functional score models.

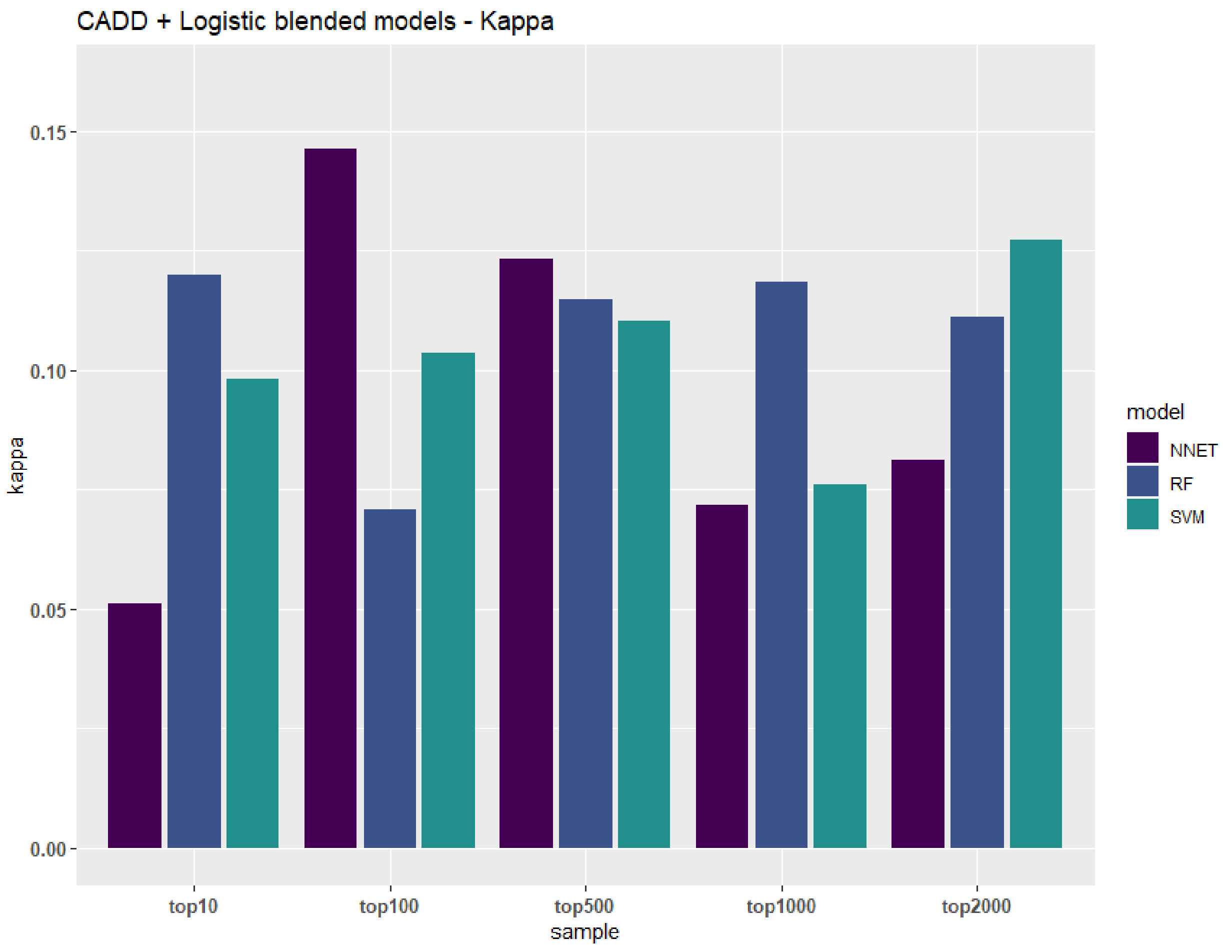

Figure 16. Kappa values for all *cadd_logistic* models. No models were significantly more accurate compared to the naïve model.

4.3 FATHMM-XF results – Random Forest, SVM, and Neural Network

Modeling with RF, NNET, and SVM was repeated for the *fathmm_logistic* blended score top10, top 100, top500, top1000, and top2000 SNPs, again using a single 0.9 r^2 correlation cutoff threshold and keeping the SNP with the higher functional score when choosing which SNP to keep. The number of SNPs retained was similar to those seen in the no-functional score models (Table 11), indicating that the balance of the *fathmm_logistic* score is too heavily in favor of the logistic p-value.

Number of SNPs retained FATHMM-XF + Logistic	≤ 0.9 r^2
Top10 SNPs	1
Top100 SNPs	15
Top500 SNPs	103
Top1000 SNPs	213
Top2000 SNPs	528

Table 11. Number of SNPs retained after filtering correlated variables for the *fathmm_logistic* blended models.

Hyperparameter optimization was conducted as before, with 10-fold cross validation repeated three times across a grid of hyperparameters and is summarized (Table 12), full *fathmm_logistic* results are reported (Table 13), and AUC values are plotted (Figure 17), as well as Kappa values (Figure 18). One model, top500 RF, achieved significance (p=0.026) compared with the naïve model, and achieved 0.6044 absolute accuracy, 0.1663 Kappa, 0.3607 sensitivity, 0.7991 specificity, and 0.5966 AUC.

Optimal hyperparameters	NNET – *size* and *decay*	RF – *mtry*	SVM – *sigma* and *C*
fathmm_logistic Top10	(1, 0.5)	2	(0.301, 256)
fathmm_logistic Top100	(5, 0)	1	(0.046, 0.5)
fathmm_logistic Top500	(5, 0.2)	64	(0.005, 0.5)
fathmm_logistic Top1000	(5, 0.2)	87	(0.002, 1.0)
fathmm_logistic Top2000	(5, 0.5)	425	(0.0001, 0.5)

Table 12. Optimized hyperparameters for the *fathmm_logistic* blended models.

FATHMM + Logistic Results	Accuracy	Accuracy > Null Accuracy? (P-value)	Kappa	Sensitivity	Specificity	AUC
RF – Top10 0.9	0.5534	0.5598	0.0452	0.235	0.8079	0.5364
RF – Top100 0.9	0.5801	0.1732	0.1171	0.3443	0.7686	0.5783
RF – Top500 0.9	**0.6044**	**0.0262 (*)**	**0.1663**	**0.3607**	**0.7991**	**0.5966**
RF – Top1000 0.9	0.5631	0.4028	0.0692	0.2623	0.8035	0.5866
RF – Top2000 0.9	0.5898	0.0901	0.1154	0.235	0.8734	0.5931
NNET – Top10 0.9	0.5534	0.5598	0.0452	0.235	0.8079	0.5378
NNET – Top100 0.9	0.551	0.5985	0.0655	0.3607	0.7031	0.5529
NNET – Top500 0.9	0.5291	0.8728	0.0411	0.4426	0.5983	0.5649
NNET – Top1000 0.9	0.5461	0.6728	0.0543	0.3497	0.7031	0.5563
NNET – Top2000 0.9	0.5583	0.481	0.0811	0.3716	0.7074	0.5619
SVM – Top10 0.9	0.5558	0.5205	0	0	1	0.5209
SVM – Top100 0.9	0.5655	0.3649	0.076	0.2732	0.7991	0.5777
SVM – Top500 0.9	0.585	0.127	0.1372	0.4044	0.7293	0.5931
SVM – Top1000 0.9	0.5801	0.1732	0.132	0.4262	0.7031	0.5972
SVM – Top2000 0.9	0.5898	0.0901	0.1482	0.4153	0.7293	0.5773

Table 13. Full results table for the *fathmm_logistic* blended models. The Top500 RF significantly outperformed the naïve model (p=0.026)

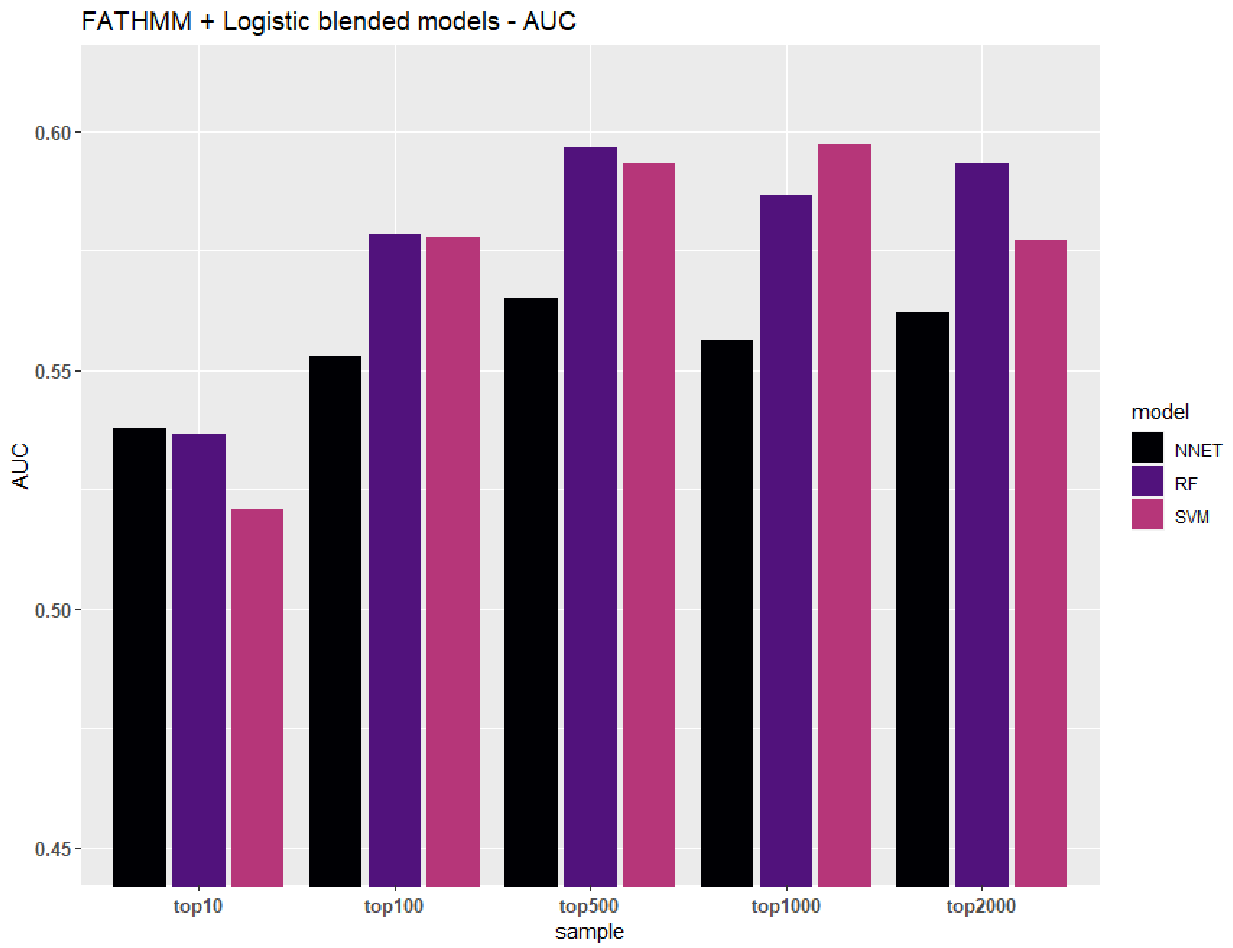

Figure 17. AUC values for all *fathmm_logistic* models. No models achieved > 0.6 AUC. In general, we see lower AUC values compared with the no-functional score models.

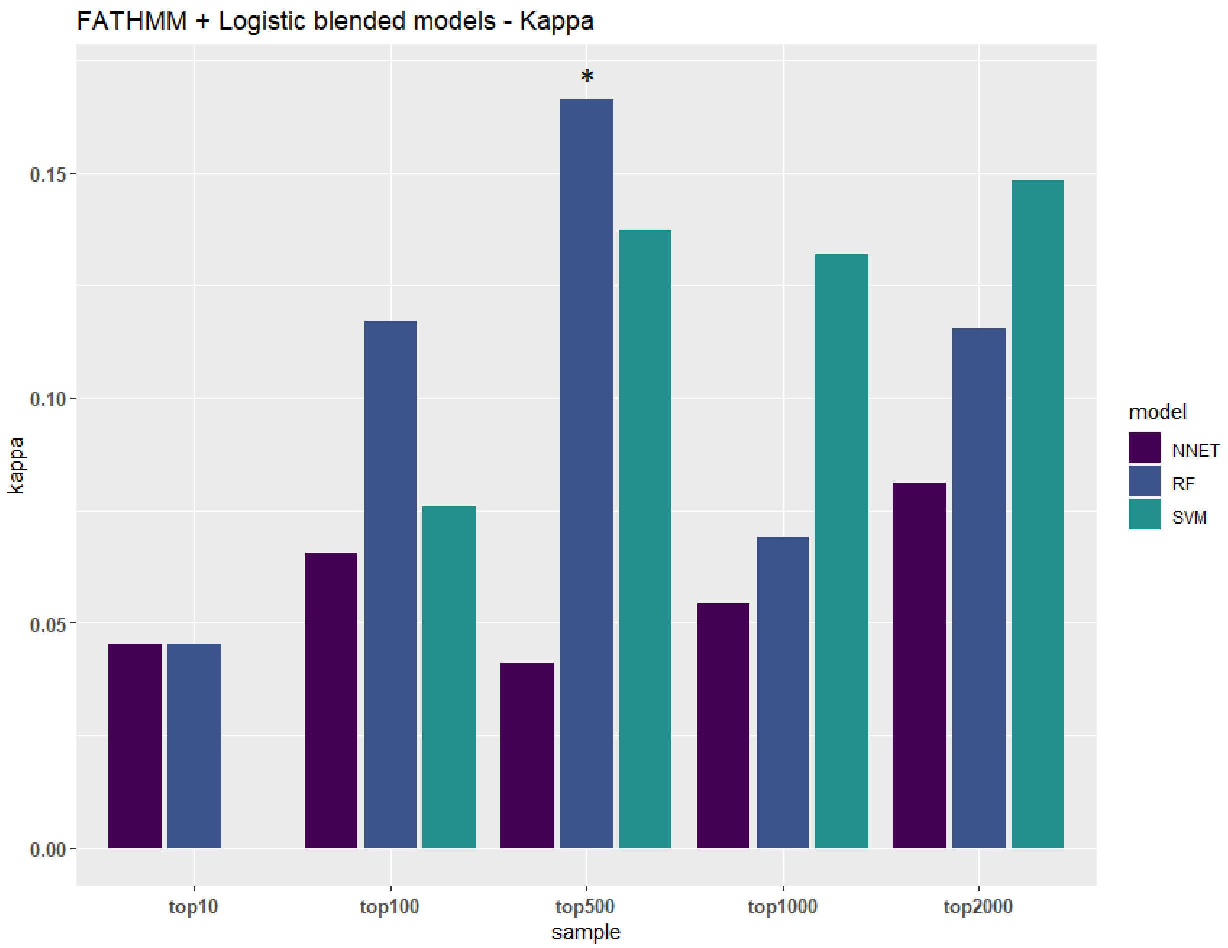

Figure 18. Kappa values for all *fathmm_logistic* models. The top500 RF model was significantly better than the naïve model (p=0.026).

Chapter 5 – Conclusions and Future Directions

5.1 – Conclusions

Here we have shown that sarcoidosis disease incidence in African Americans can be successfully modeled with various supervised machine learning techniques such as RF, SVM, and NN. We have generated multiple models that significantly outperform a naïve model in terms of absolute accuracy on test samples. The best model in terms of absolute accuracy and Kappa was the SVM Top1000 0.75 no-functional score model, with an accuracy of 0.6068 and a Kappa of 0.1858. The best model in terms of AUC was the NN Top100 0.75 with a 0.6103 AUC.

The top performing models were able to classify sarcoidosis cases with a sensitivity of 0.437 to 0.448, which are in-line with the expected theoretical heritability of 40-66% published in the literature for Scandinavian populations. While heritability for diseases can differ greatly between populations, these results support the idea that this estimate is plausible for African Americans with sarcoidosis.

The question then becomes, are we near the theoretical limit of predictive power using a genetics-only approach for a disease with an estimated 40-66% heritability or would more powerful machine learning techniques such as deep learning improve predictive accuracy further? Would increasing the sample size of the training or test sets result in a subsequent improvement to accuracy? Answers to these questions could help further establish the heritability estimate for sarcoidosis in African Americans as well as be of clinical utility.

Functional score incorporation with CADD and FATHMM-XF does not seem to improve predictive power. However, some SNPs with high functional scores and reasonable logistic regression p-values may still contain important clues about the mechanism of sarcoidosis granuloma formation. Further experimentation with these SNPs may prove fruitful in future work.

5.2 – Future Directions

We can envision several next steps for this project, both short-term and longer term. Most pressingly in the short term is the need to adjust log base *c* in the *fathmm_logistic* calculation so that functional score will take a higher proportion of the blended score. Additionally, we could consider running the functional score models with all the Pearson r^2 cutoffs used in the non-functional score section to further improve robustness of the results. Next, we need to increase the maximum *size* in the NN hyperparameter search grid, which could result in significant improvements to those models.

We would additionally like to employ deep learning to help identify more complex structure in the data and potentially improve prediction accuracy. Further, we are interested in consolidating the individual SNPs into their respective haplotype blocks, so that more information is contained within a single feature. We are also interested in comparing our predictive modeling results with previously established statistical methods which use polygenic risk scores to predict disease incidence, such as best linear unbiased prediction (BLUP), BayesA, and LDPred (Clark & Van Der Werf, 2013; Meuwissen et al., 2001; Vilhjálmsson et al., 2015). Finally, we plan to apply alternative feature selection algorithms beyond logistic regression and LASSO to the entire list of ~7.7 million SNPs, to see if alternate methods can yield SNPs which are even more useful for predictive modeling.

References

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. Al, Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A. A., Barnes, I. H. A., Barozzi, I., … Zimmerman, J. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*(7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4

Adrianto, I., Lin, C. P., Hale, J. J., Levin, A. M., Datta, I., Parker, R., Adler, A., Kelly, J. A., Kaufman, K. M., Lessard, C. J., Moser, K. L., Kimberly, R. P., Harley, J. B., Iannuzzi, M. C., Rybicki, B. A., & Montgomery, C. G. (2012). Genome-wide association study of African and European Americans implicates multiple shared and ethnic specific loci in sarcoidosis susceptibility. *PLoS ONE*, *7*(8), 43907. https://doi.org/10.1371/journal.pone.0043907

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, *47*(D1), D1038–D1043. https://doi.org/10.1093/nar/gky1151

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. In *Nature* (Vol. 526, Issue 7571, pp. 68–74). Nature Publishing Group. https://doi.org/10.1038/nature15393

Autopilot AI - Tesla. (2021). https://www.tesla.com/autopilotAI

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, *40*(3), 340–345. https://doi.org/10.1038/ng.78

Baughman, R. P., & Lower, E. E. (2011). Who dies from sarcoidosis and why? In *American Journal of Respiratory and Critical Care Medicine* (Vol. 183, Issue 11, pp. 1446–1447). Am J Respir Crit Care Med. https://doi.org/10.1164/rccm.201103-0409ED

Belsky, D. W., Moffitt, T. E., Sugden, K., Williams, B., Houts, R., McCarthy, J., & Caspi, A. (2013). Development and evaluation of a genetic risk score for obesity.

Biodemography and Social Biology, *59*(1), 85–100. https://doi.org/10.1080/19485565.2013.774628

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., & Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman & Hall.

Carroll, S. B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. In *Cell* (Vol. 134, Issue 1, pp. 25–36). Elsevier B.V. https://doi.org/10.1016/j.cell.2008.06.030

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, C. (2019, March 13). *Controlling for stratification in (meta-)GWAS with PCA: Theory, applications, and implications | Broad Institute*. https://www.broadinstitute.org/talks/controlling-stratification-meta-gwas-pca-theory-applications-and-implications

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. In *Genomics* (Vol. 99, Issue 6, pp. 323–329). Academic Press. https://doi.org/10.1016/j.ygeno.2012.04.003

Clark, S. A., & Van Der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. *Methods in Molecular Biology*, *1019*, 321–330. https://doi.org/10.1007/978-1-62703-447-0_13

Colloca, L., & Miller, F. G. (2011). The nocebo effect and its relevance for clinical practice. In *Psychosomatic Medicine* (Vol. 73, Issue 7, pp. 598–603). Lippincott Williams and Wilkins. https://doi.org/10.1097/PSY.0b013e3182294a50

Cusanelli, E., & Chartrand, P. (2014). Telomeric noncoding RNA: Telomeric repeat-containing RNA in telomere biology. In *Wiley Interdisciplinary Reviews: RNA* (Vol. 5, Issue 3, pp. 407–419). Blackwell Publishing Ltd. https://doi.org/10.1002/wrna.1220

De La Vega, F. M., & Bustamante, C. D. (2018). Polygenic risk scores: A biased prediction? *Genome Medicine*, *10*(1), 100. https://doi.org/10.1186/s13073-018-0610-x

Freemer, M., & King, J. (2001). The ACCESS study: Characterization of sarcoidosis in the United States. In *American Journal of Respiratory and Critical Care Medicine* (Vol. 164, Issue 10 I, pp. 1754–1755). American Lung Association. https://doi.org/10.1164/ajrccm.164.10.2109111b

Freund, Y., & Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm*. http://www.research.att.com/

Grinberg, N. F., Orhobor, O. I., & King, R. D. (2020). An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning*, *109*(2), 251–277. https://doi.org/10.1007/s10994-019-05848-5

Harimoorthy, K., & Thangavelu, M. (2020). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing*, *12*(3), 3715–3723. https://doi.org/10.1007/s12652-019-01652-0

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., & O’Sullivan, J. (2019). Machine learning SNP based prediction for precision medicine. In *Frontiers in Genetics* (Vol. 10, Issue MAR, p. 267). Frontiers Media S.A. https://doi.org/10.3389/fgene.2019.00267

Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics. In *Cancer Genomics and Proteomics* (Vol. 15, Issue 1, pp. 41–51). International Institute of Anticancer Research. https://doi.org/10.21873/cgp.20063

Imadojemu, S., Elenitsas, R., Chan, E., Wanat, K., & Rosenbach, M. (2016). Multiple granulomatous dermatitides in a patient with rheumatoid arthritis. *JAAD Case Reports*, *2*(1), 67–69. https://doi.org/10.1016/j.jdcr.2015.11.016

Jacobs, B. M., Belete, D., Bestwick, J., Blauwendraat, C., Bandres-Ciga, S., Heilbron, K., Dobson, R., Nalls, M. A., Singleton, A., Hardy, J., Giovannoni, G., Lees, A. J., Schrag, A. E., & Noyce, A. J. (2020). Parkinson’s disease determinants, prediction and gene-environment interactions in the UK Biobank. *Journal of Neurology, Neurosurgery and Psychiatry*, *91*(10), 1046–1054. https://doi.org/10.1136/jnnp-2020-323646

Joseph, P. V., Wang, Y., Fourie, N. H., & Henderson, W. A. (2018). A computational framework for predicting obesity risk based on optimizing and integrating genetic risk score and gene expression profiles. *PLoS ONE*, *13*(5). https://doi.org/10.1371/journal.pone.0197843

Judson, M. A. (2020). Environmental Risk Factors for Sarcoidosis. In *Frontiers in Immunology* (Vol. 11, p. 1340). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2020.01340

Kahi, C. J., Pohl, H., Myers, L. J., Mobarek, D., Robertson, D. J., & Imperiale, T. F. (2018). Colonoscopy and colorectal cancer mortality in the veterans affairs health care system: A case-control study. *Annals of Internal Medicine*, *168*(7), 481–488. https://doi.org/10.7326/M17-0723

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M. Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., & Snyder, M. (2010). Variation in transcription factor binding among humans. *Science*, *328*(5975), 232–235. https://doi.org/10.1126/science.1183621

Keeling, K. M., Du, M., & Bedwell, D. M. (2013). *Therapies of Nonsense-Associated Diseases*. https://www.ncbi.nlm.nih.gov/books/NBK6183/

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kruppa, J., Ziegler, A., & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. In *Human Genetics* (Vol. 131, Issue 10, pp. 1639–1654). Springer. https://doi.org/10.1007/s00439-012-1194-y

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Learn About Sarcoidosis | American Lung Association. (2020, October 24). https://www.lung.org/lung-health-diseases/lung-disease-lookup/sarcoidosis/learn-about-sarcoidosis

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. In *Genome Medicine* (Vol. 12, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-020-00742-5

Li, C., Li, M., Long, J. R., Cai, Q., & Zheng, W. (2008). Evaluating cost efficiency of SNP chips in genome-wide association studies. *Genetic Epidemiology*, *32*(5), 387–395. https://doi.org/10.1002/gepi.20312

Liu, X., Li, C., & Boerwinkle, E. (2017). The performance of deleteriousness prediction scores for rare non-proteinchanging single nucleotide variants in human genes. In *Journal of Medical Genetics* (Vol. 54, Issue 2, pp. 111–113). BMJ Publishing Group. https://doi.org/10.1136/jmedgenet-2016-104369

López, B., Torrent-Fontbona, F., Viñas, R., & Fernández-Real, J. M. (2018). Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. *Artificial Intelligence in Medicine*, *85*, 43–49. https://doi.org/10.1016/j.artmed.2017.09.005

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Moller, D. R., Rybicki, B. A., Hamzeh, N. Y., Montgomery, C. G., Chen, E. S., Drake, W., & Fontenot, A. P. (2017). Genetic, immunologic, and environmental basis of sarcoidosis. *Annals of the American Thoracic Society*, *14*(Suppl 6), S429–S436. https://doi.org/10.1513/AnnalsATS.201707-565OT

Molnár, T., Tiszlavicz, L., Gyulai, C., Nagy, F., & Lonovics, J. (2005). Clinical significance of granuloma in Crohn's disease. *World Journal of Gastroenterology*, *11*(20), 3118–3121. https://doi.org/10.3748/wjg.v11.i20.3118

Paré, G., Mao, S., & Deng, W. Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-13056-1

Polygenic Risk Scores. (2020, August 11). https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–329. https://doi.org/10.1038/nature14248

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., & Campbell, C. (2018). FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, *34*(3), 511–513. https://doi.org/10.1093/bioinformatics/btx536

Rossides, M., Grunewald, J., Eklund, A., Kullberg, S., Giuseppe, D. Di, Askling, J., & Arkema, E. V. (2018). Familial aggregation and heritability of sarcoidosis: A Swedish nested case-control study. *European Respiratory Journal*, *52*(2). https://doi.org/10.1183/13993003.00385-2018

Rybicki, B. A., Hirst, K., Iyengar, S. K., Barnard, J. G., Judson, M. A., Rose, C. S., Donohue, J. F., Kavuru, M. S., Rabin, D. L., Rossman, M. D., Baughman, R. P., Elston, R. C., Maliarik, M. J., Moller, D. R., Newman, L. S., Teirstein, A. S., & Iannuzzi, M. C. (2005). A sarcoidosis genetic linkage consortium: The sarcoidosis genetic analysis (SAGA) study. *Sarcoidosis Vasculitis and Diffuse Lung Diseases*, *22*(2), 115–122.

Sarcoidosis - Diagnosis and treatment - Mayo Clinic. (2019, January 30). https://www.mayoclinic.org/diseases-conditions/sarcoidosis/diagnosis-treatment/drc-20350363

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R., & Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, *31*(10), 1536–1543. https://doi.org/10.1093/bioinformatics/btv009

Silva Miranda, M., Breiman, A., Allain, S., Deknuydt, F., & Altare, F. (2012). The tuberculous granuloma: An unsuccessful host defence mechanism providing a safety shelter for the bacteria? In *Clinical and Developmental Immunology* (Vol. 2012). Clin Dev Immunol. https://doi.org/10.1155/2012/139127

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Starshinova, A. A., Malkova, A. M., Basantsova, N. Y., Zinchenko, Y. S., Kudryavtsev, I. V., Ershov, G. A., Soprun, L. A., Mayevskaya, V. A., Churilov, L. P., & Yablonskiy, P. K. (2020). Sarcoidosis as an Autoimmune Disease. *Frontiers in Immunology*, *10*, 2933. https://doi.org/10.3389/fimmu.2019.02933

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., & Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. In *Human Genetics* (Vol. 136, Issue 6, pp. 665–677). Springer Verlag. https://doi.org/10.1007/s00439-017-1779-6

Sverrild, A., Backer, V., Kyvik, K. O., Kaprio, J., Milman, N., Svendsen, C. B., & Thomsen, S. F. (2008). Heredity in sarcoidosis: A registry-based twin study. *Thorax*, *63*(10), 894–896. https://doi.org/10.1136/thx.2007.094060

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S. been, Tian, X., Browning, B. L., Das, S., Emde, A. K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P. R., Bhatia, G., Do, R., Hayeck, T., Won, H. H., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Ward, L. D., & Kellis, M. (2016). HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Research*, *44*(D1), D877–D881. https://doi.org/10.1093/nar/gkv1340

Zhang, S., Zhou, Z., Chen, X., Hu, Y., & Yang, L. (2017). pDHS-SVM: A prediction method for plant DNase I hypersensitive sites based on support vector machine. *Journal of Theoretical Biology*, *426*, 126–133. https://doi.org/10.1016/j.jtbi.2017.05.030

Zhao, W., Lai, X., Liu, D., Zhang, Z., Ma, P., Wang, Q., Zhang, Z., & Pan, Y. (2020). Applications of Support Vector Machine in Genomic Prediction in Pig and Maize Populations. *Frontiers in Genetics*, *11*, 1537. https://doi.org/10.3389/fgene.2020.598318

Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Siew, W. C., Lu, Y., Denoeud, F., Antonarakis, S. E., Snyder, M., Ruan, Y., Wei, C. L., Gingeras, T. R., Guigó, R., Harrow, J., & Gerstein, M. B. (2007). Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Research*, *17*(6), 839–851. https://doi.org/10.1101/gr.5586307

Zhu, X., Zhang, S., Kan, D., & Cooper, R. (2004). Haplotype block definition and its application. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 152–163. https://doi.org/10.1142/9789812704856_0015