**Random Forest Analysis of Age-Related Macular Degeneration**

Hannah Lewis and Eric Mallmann
Faculty Mentor: Dr. Abra Brisbin
University of Wisconsin-Eau Claire

---

**Introduction**

**Definitions**

**Genotype:** the genetic constitution of an individual organism - what an organism's gene codes for.

**Phenotype:** the set of observable characteristics of an organism, a result of the interaction of genotype with the environment

**Single Nucleotide Polymorphisms (SNP):** changes to a single base pair in the DNA

---

**Background**

Nineteen genetic variants have been well-established as associated with risk for AMD; many of these are in genes related to atherosclerosis, angiogenesis, and the innate immune system. These known variants explain approximately 65% of the genetic basis of the disease [2]. There is a need to develop statistical methods to identify additional genetic variants that contribute to AMD, to improve understanding of the disease etiology and enable treatments to be targeted to specific versions of the disease. For example, some evidence suggests that a treatment of antioxidants with zinc slows the progression of AMD in patients with the low-risk variant of the CFH gene, but not in patients with the high-risk variant [3].

---

**Objectives**

Our research goal is to find undiscovered genetic variants which may be associated with AMD using a two-phase approach. In phase 1 we will do a regression on previously known associated SNPs, and in phase 2 we will do a random forest on a different set of SNPs using residuals from the regression to find potential new associations. We will perform our statistical analysis first on a simulated data set and then apply this analysis to a set of real genetic data.

---

**Random Forests**

A random forest is a machine learning technique to predict quantitative or categorical values. Random forests create decision trees for a single response variable [4]. The algorithm then chooses optimal splits on the decision tree to minimize the residual sum of squares.

Next, random forests use bagging (bootstrap aggregation). Bagging takes a bootstrap sample from the data and builds a tree based on a random subset of predictors to predict the phenotype values of our sample. Then, the function estimates the phenotypic values that are not included in the sample with the decision tree. This is done many times to create many trees, or a forest. The random forest then averages the estimates for each data value when it was not included in the bootstrap sample.

---

**Simulation Results**

We simulated a case-control phenotype based on real genetic data on 1043 people from the HapMap project. Next, we ran a logistic regression on the simulated data set to model binary disease status based on genetic variants that are already known to be associated with AMD. Then we used random forest analysis to predict the residuals. We excluded the known genetic variants from the set of possible predictor variables from our random forests in hopes of finding undiscovered associated variants.

---

**Options for Configuring Random Forest**

Variable importance is a measure of how much each SNP contributes to the prediction accuracy of the random forest. Because our primary goal is the selection of relevant genetic variants, an optimal method should assign high scores of variable importance to genetic variants that directly influence the simulated disease, and low scores to other variants. We used the Receiver Operator Characteristic (ROC) curve of the variable importance scores to assess our different methods.

---

**Optimal Random Forest**

We ran random forest analysis using four different residuals: deviance, working, response, and Pearson residuals. We determined that none of the ROC curves were significantly different from each other. Thus, we chose to use the Pearson residuals in our optimal forest because they generated the greatest area under the ROC curve.

Using the Pearson residuals we ran random forest analysis on all individuals and on individuals with absolute residuals in the top 50% and top 25%. There was no significant difference between using all individuals and using individuals with absolute residuals in the top 50%. However, using individuals in the top 25% was significantly worse. Thus, we chose to include all individuals in our random forest.

---

**Real Data**

Our next task was to implement our regression and random forest on a real data set extracted from dbGap [5]. The data set included 370,404 SNPs and 1,533 participants, of whom 930 were diagnosed with AMD and 603 who were not.

We identified 16 genes which were known to be associated with AMD from medical literature. We decided to use 6 genes in our set of predictor variables for the logistic regression and use the other 10 for validation purposes.

We ran a logistic regression of case-control status using 53 SNPs from the 6 genes in the predictor set. We computed the Pearson residuals to use as a response variable for our random forest.

To prepare for the random forest, we filtered the data set using PLINK with the following criteria:

1. Remove 53 SNPs that belong to the predictor set
2. No missing data for the SNP values
3. Minor-allele frequency > 0.05
4. Linkage disequilibrium < 0.1

This gave us a final data set of 35,472 SNPs and 1,533 participants. For computational feasibility, we split the data set into four groups by chromosomes, with 8,469-10,377 SNPs in each.

Our first group consisted of chromosomes 1-3. The random forest analysis was run for this subset of the data, and the variable importance scores were examined. SNPs rs2019724 and rs7707472 were among the 10 most important SNPs and are already known to be associated with AMD [6,7]. Two other SNPs that were found to be important, rs3792234 and rs7565565, were from the 9 SNPs in our data set which appear in the 10 genes we saved for validation purposes.

---

**Future Work**

- Examine variable importance scores of SNPs used in random forests for the other three groups
- Use variable importance scores to determine if SNPs from the 10 known associated genes which were not included in the forest are detected as important predictors
- Conduct a literature review to identify functions of other SNPs with high variable importance scores
- SNPs in genes associated with eye development are high priority candidates for potential associations with AMD. Other researchers could use our work to test whether these potential variants are causally related to AMD.

---

**References**


---

**Acknowledgments**

UWEC Office of Research and Sponsored Programs
R Statistical Programming Software and PLINK
Jack Lazowski and Molly Petersen