# Genetic Analysis of Potato Scab Disease.

Quan Yuan

B00923505

Supervised by Dr. Hong Gu

Dalhousie University

April 24th, 2024

# Abstract

This project analyzed gene expression and Single Nucleotide Polymorphisms (SNP) from 300 potato samples and employed a random forest model to predict the relationship between sample gene expression, SNP, and the severity of common scab disease. The sample data were categorized into two groups based on the severity of potato common scab disease and the coverage area of scabs, and then predictions were made accordingly. The final predictive accuracies of the random forest model were 60.46% and 67.3%, respectively, both lower than the default accuracy. Therefore, we believe that potato common scab disease seems to be unrelated to gene expression and SNP, and the infection status of potato common scab disease may be associated with environmental factors in potato cultivation.

## Introduction

Potatoes, with their high yield, rich nutritional value, adaptability, diverse culinary applications, and ease of cultivation and management, rank as the world's third-largest staple food crop, following rice and wheat. Data reveals that over a billion people worldwide consume potatoes, with more than a hundred million directly or indirectly reliant on potatoes as their primary food source (International Potato Center, 2023). Given the escalating population growth and hunger rates, potatoes play a crucial role in ensuring food security. Potato Common Scab (PCS), caused by Streptomyces spp. bacteria, poses a significant threat to both the yield and quality of potatoes. Afflicted potatoes develop scabs on their skin, appearing as raised or sunken lesions, thereby compromising their appearance and commercial value. Severe infections may even result in partial crop loss, leading to decreased potato yields. Moreover, scab lesions not only impact appearance but can also deepen, affecting the taste and edible quality of potatoes. Infected potatoes undergo shortened storage periods, becoming prone to rot and deterioration, further reducing their market value. Statistics from the International Potato Center indicate that approximately 10% to 20% of global potato yields are affected by common scab disease, with some regions experiencing even more severe losses. According to a survey, economic losses due to potato common scab disease in Canada range between CAD 15.3 million and CAD 17.3 million annually, with an average loss per grower of CAD 7,500 to CAD 8,500 and an average loss per hectare of CAD 90 to

CAD 102 (Canadian Growers Association, 2005). The detrimental effects of potato common scab disease present significant challenges to both the potato industry's development and global food security.

## Data

The dataset utilized in this study comprises infection statuses of common scab disease in 100 potato varieties planted in three different regions after 70 days, along with the quantities of 6248 mRNA transcripts and 1603 single nucleotide DNA sequences.

Variables relevant to the infection status of potato common scab disease include:

- Types of scab lesions, categorized into six classes based on severity, where 1 indicates no lesions and 6 represents extremely severe lesions.
- The coverage area of potato common scab, divided into nine classes according to coverage area size, where 1 signifies a scab coverage area of 0% and 9 indicates a scab coverage area of 100%.

Counts of 6248 mRNA transcripts was obtained using the Serial Analysis of Gene Expression (SAGE) method. It's a molecular biology technique used to quantify the expression levels of many genes simultaneously in a biological sample. In SAGE, short tags are generated from mRNA molecules, and these tags are then sequenced and mapped to specific genes in the genome. By counting the number of times each tag is observed, the number of corresponding mRNA transcripts can be recorded in the sample. SAGE has been widely used in gene expression profiling studies to understand gene regulation, identify biomarkers, and characterize disease mechanisms.

Single nucleotide polymorphism (SNP) primarily refers to the DNA sequence polymorphism caused by variation in a single nucleotide at the genomic level. Some SNP loci can affect gene function, leading to changes in biological traits or even disease susceptibility. Therefore, single nucleotide polymorphisms serve as crucial indicators for studying genetic variations in both animal and plant strains.

## Method

### Random Forest

The Random Forest algorithm (RF), proposed by Leo Breiman and Adele Cutler in 2001, is an ensemble learning method. Random Forest constructs multiple decision trees, each generated by random sampling of the training data and random selection of features. This approach reduces model variance and the risk of overfitting. During prediction, each decision tree in the Random Forest produces a prediction, and these predictions are combined through voting or averaging to obtain a single prediction. Random Forest can be applied to both classification and regression tasks, handling high-dimensional data and large sample sizes with high accuracy. Due to its robust performance and ease of implementation, Random Forest has found wide applications in fields such as data mining, pattern recognition, and bioinformatics, becoming a powerful machine learning tool.

In R, Random Forest has two important parameters: ntree and mtry. ntree determines the number of decision trees in the Random Forest, while mtry determines the number of variables randomly selected at each decision. Typically, increasing the number of trees by adjusting ntree can improve model performance but increases computational costs. Adjusting mtry to select an appropriate number of variables helps prevent overfitting.

This project utilizes the Random Forest model to predict mRNA transcripts and SNPs associated with the severity of potato common scab disease. It aims to identify which gene expression and SNP have the greatest impact on the severity of potato common scab disease.

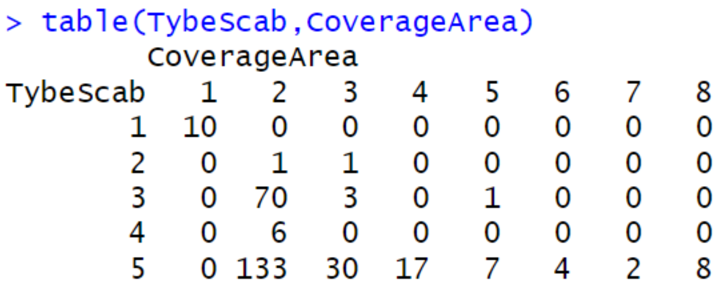
### Cross Validation

Cross-validation is a statistical method used to evaluate the performance of a model. It involves splitting the dataset into a training set and a test set. The model is trained on the training set, and then the error of the trained model is calculated based on a separate test set to assess the model's performance. Common forms of cross-validation include hold-out, k-fold cross-validation, and leave-one-out.

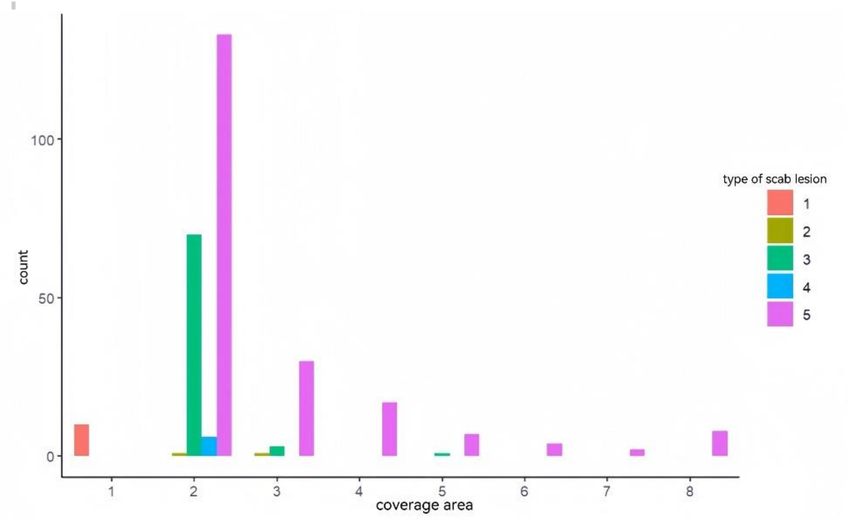
In this project, 10-fold cross-validation be used. This method involves dividing the data into 10 equally sized datasets. Nine of these datasets are used as the training set to train the model, while the remaining one dataset serves as the test set to calculate the model's error. This validation process is repeated 10 times, with each iteration using a different dataset as the test set. Subsequently, we obtain 10 prediction errors, and the final cross-validation error is calculated by averaging these 10 prediction errors. This approach ensures robust evaluation of the model's performance across multiple iterations and provides a reliable estimate of its generalization capability.

## Analysis

After conducting a correlation analysis on the severity of common scab disease lesions and the extent of scab coverage, it can be inferred that there is an association between the lesion severity and scab coverage area in potato common scab disease. The two variables exhibit a positive correlation, as determined by a chi-square test with a significance level of P < 0.01.



```
> table(TybeScab,CoverageArea)
        CoverageArea
TybeScab   1   2   3   4   5   6   7   8
       1  10   0   0   0   0   0   0   0
       2   0   1   1   0   0   0   0   0
       3   0  70   3   0   1   0   0   0
       4   0   6   0   0   0   0   0   0
       5   0 133  30  17   7   4   2   8
```



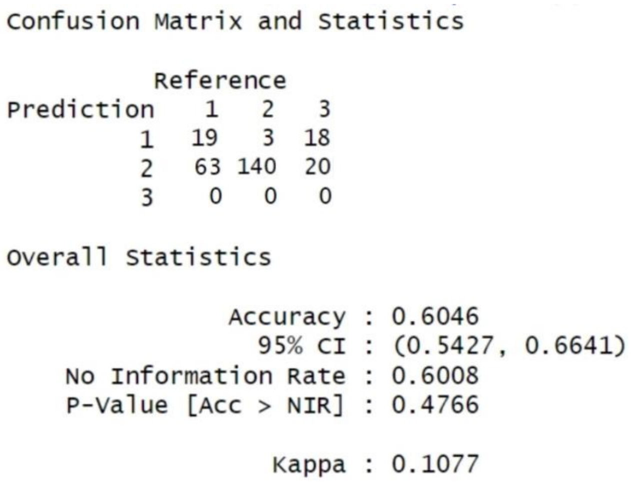
Therefore, aiming to enhance the predictive accuracy and interpretability of the random forest model, the following categorization criteria are applied:

- •Potato samples with scab lesion severity less than or equal to 4 and the scab coverage area less than or equal to 3 are considered less susceptible to common scab disease.

- Potato samples with scab lesion severity greater than 4 and the scab coverage area less than or equal to 3 are classified as moderately susceptible to common scab disease.

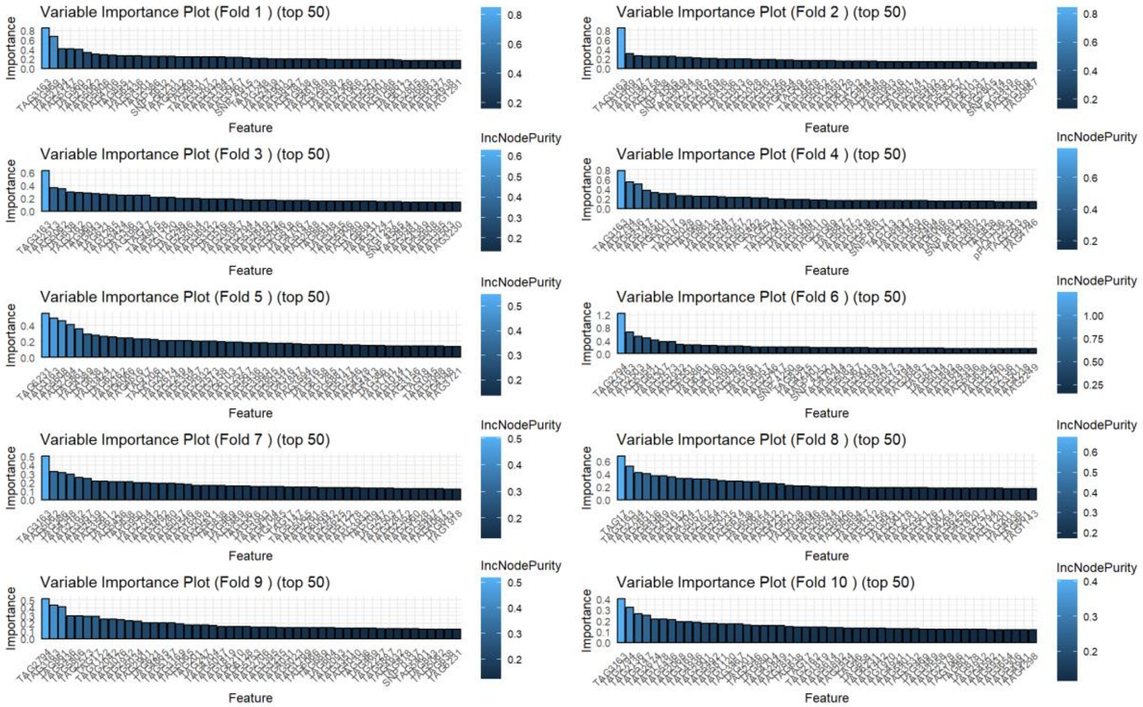
- Potato samples with scab lesion severity greater than 4 and the scab coverage area greater than 3 are categorized as highly susceptible to common scab disease.

The random forest model is employed to fit 6248 mRNA transcripts, 1603 SNPs and the infection status of potato common scab disease. Initially, the data is randomly partitioned into ten equally sized datasets, and ten random forest models are utilized to fit the mRNA transcripts to the infection status of potato common scab disease. 10-fold cross-validation was performed when fitting each random forest model to the data. According to the results of 10-fold cross-validation, the final prediction accuracy is 60.46%.



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3
         1  19   3  18
         2  63 140  20
         3   0   0   0

Overall Statistics

               Accuracy : 0.6046
                 95% CI : (0.5427, 0.6641)
    No Information Rate : 0.6008
    P-Value [Acc > NIR] : 0.4766

                  Kappa : 0.1077
```

Following this, the mRNA transcripts are ranked by importance based on "incNodePurity" (increase in node purity), where a higher value indicates greater importance of the variable. Subsequently, the top 50 most important mRNA transcripts from each random forest model are selected。

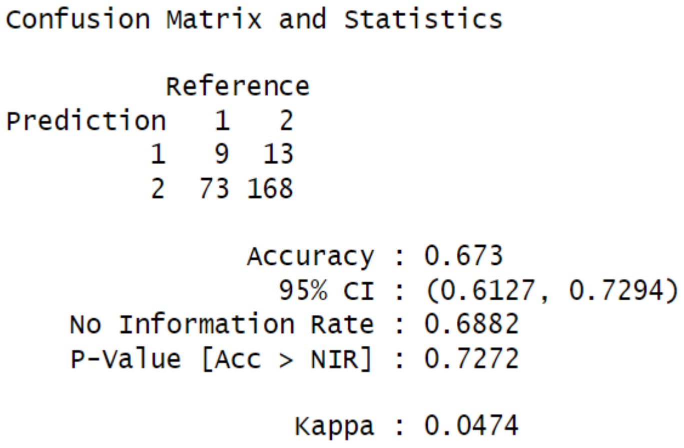


As the predictive accuracy of the random forest model was not satisfactory after dividing the data into three categories, the data was subsequently divided into two categories.

•Potato samples with scab lesion severity less than or equal to 4 are divided into one class.

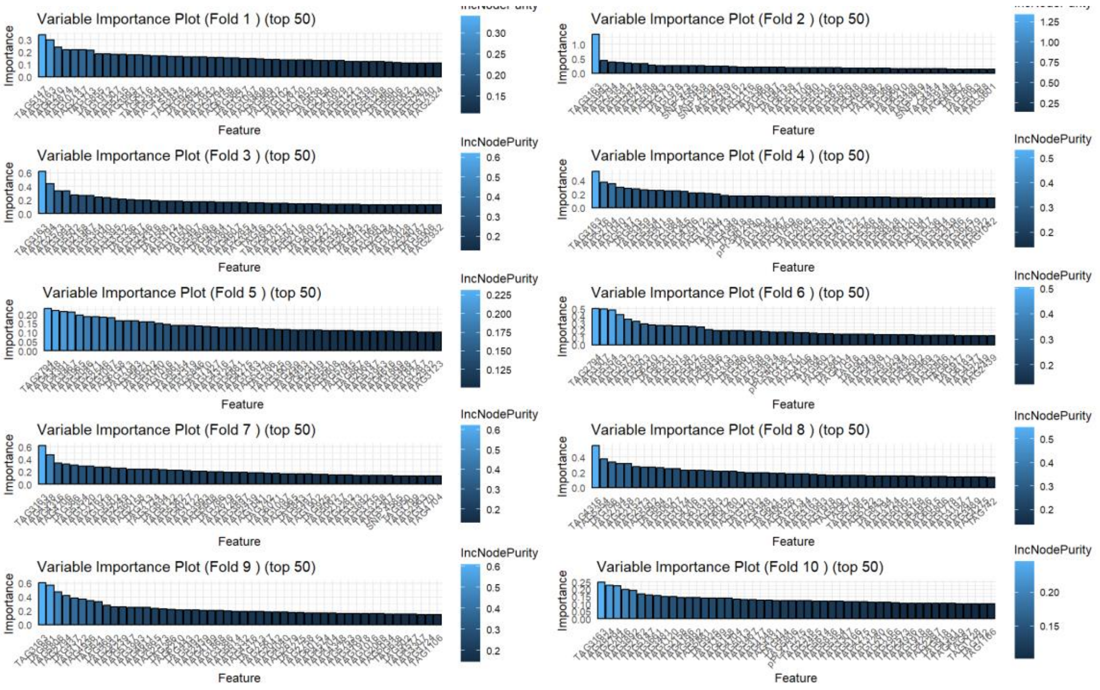
•Potato samples with scab lesion severity greater than 4 are divided into one class

The random forest model was then used to fit 6248 mRNA transcripts, 1603 SNPs, and the infection status of potato common scab disease. After conducting 10-fold cross-validation, the optimal mtry parameter was selected to maximize the predictive accuracy of the random forest model. The final predictive accuracy achieved was 67.3%.



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1   9  13
         2  73 168

               Accuracy : 0.673
                 95% CI : (0.6127, 0.7294)
    No Information Rate : 0.6882
    P-Value [Acc > NIR] : 0.7272

                  Kappa : 0.0474
```

The top 50 most important mRNA transcripts or SNPs from each random forest model are selected.



## Conclusion

After dividing the potato samples into three categories, the random forest model achieved a predictive accuracy of 60.46%, which was lower than the default accuracy. Upon categorizing the potato samples into two classes, the predictive accuracy of the random forest model slightly improved to 67.3%, yet still remained below the corresponding default accuracy. Consequently, we speculate that the infection status of potato common scab disease may be unrelated to gene expression and SNP. Instead, external environmental factors may play a significant role in causing potato common scab disease. Factors such as soil pH, humidity, the type and quantity of pathogens in the local soil, as well as pest infestation, could potentially lead to severe infection of potato common scab disease.

## Discussion

The random forest method, as a powerful machine learning technique, has been widely used in bioinformatics and genomics for its ability to handle high-dimensional data, nonlinear relationships, and interaction effects. Using the random forest method to screen genes related to specific traits can accurately identify key genes associated with resistance from large-scale genomic data. Additionally, the non-parametric nature of the random forest method and its ability to handle missing data confer robustness and accuracy in genotype-phenotype association analysis. However, potato, as a complex polyploid plant, still poses significant challenges in gene function elucidation and genotype-phenotype association analysis.

In the dataset used in this project, only a small number of potato samples exhibited scab coverage area exceeding fifty percent, while more than two-thirds of the observed samples showed severe scab lesion types. This uneven data distribution may lead to lower prediction accuracy of the random forest model. In future predictions, it is advisable to use more samples and explore different classification methods and models.

Furthermore, according to the conclusions of this project, the infection status of potato common scab disease may be significantly influenced by external environmental factors. Altering planting conditions such as soil pH, humidity, and the use of pesticides to eliminate pathogens and harmful insects can help suppress the occurrence of potato common scab disease.

## References

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7, 1-13.

Harris, P. M. (Ed.). (2012). The potato crop: the scientific basis for improvement. Springer Science & Business Media.

Hill, J., & Lazarovits, G. (2005). A mail survey of growers to estimate potato common scab prevalence and economic loss in Canada. Canadian Journal of Plant Pathology, 27(1), 46-52.

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. BMC bioinformatics, 15, 1-8.

Kreuze, J. F., Suomalainen, S., Paulin, L., & Valkonen, J. P. (1999). Phylogenetic analysis of 16S rRNA genes and PCR analysis of the nec1 gene from Streptomyces spp. causing common scab, pitted scab, and netted scab in Finland. Phytopathology, 89(6), 462-469.

Kwok, P. Y. (2001). Methods for genotyping single nucleotide polymorphisms. Annual review of genomics and human genetics, 2(1), 235-258.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Current opinion in plant biology, 5(2), 94-100.

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. Science, 270(5235), 484-487.