# Genetic Mutations in Sassafras: Investigating Environmental Impacts through Single Nucleotide Polymorphisms

Capstone

Samantha Harper

December 9 2024

## Contents

# Abstract

# Background and Question

Every living thing is governed by genetic material, usually DNA, that provides a 'blueprint' for that organism. Mutations and diversity due to recombination are passed through DNA in individuals that survive and reproduce. Identifying mutations that persist allows scientists to look into the genetic history of an organism; how and why those specific changes persist in certain populations can lead to important new findings about species and their adaptability. One method for examining these mutations is Genotyping by Sequencing (GBS), which identifies single nucleotide polymorphisms (SNPs) within the genome (Guan et al., 2024). However, the analysis of SNPs has historically been difficult.

The focus of this research is the connection between SNPs and environmental factors in Sassafras. Sassafras is a species of tree found throughout China and known for its value as an beautiful ornimental species as well as a source of lumber (Guan et al, 2024). Since SNPs can help identify genes that are important and that differ between organisms, they can be used to identify important gene differences that may be linked to environmental conditions. The research question is: Can we use machine learning algorithms to mine SNPs in order to find genes or gene regions of interest between natural cultivars of Sassafras? This question addresses the need for protecting vulnerable populations of Sassafras that are highly sought after due to their ornamental, lumber, and medicinal value (Guan, et al., 2024). It has also been suggested that climate change will have an impact on the habitat availability for Sassafras in China (Zhang, et al., 2020). Future research could lead to a drought or cold resistant strain of Sassafras for industrial uses. Although numerous methods to study SNPs and identify gene regions exist, none of these methods are without their drawbacks. SNPs are notoriously difficult to study and machine learning techniques may provide new insight towards this problem. One challenge posed by this type of research is that genetic data is very high dimensional data, leading to the 'curse of dimensionality'; high dimensional data is often computationally costly and may not yield the best results as not all of the data is useful to our aims (Silva et al.,2022). Using methods such as k-fold cross-validation can help create more accurate models and prevent overfitting. These results could lead to actionable insights that could guide future research towards protecting this species and may even be able to improve forestry approaches towards minimizing the impacts of climate change and deforestation.

Hypothesis: Underlying genes, as identified by SNPs, in Sassafras are influenced by environmental factors because environmental pressure can cause mutations to persist in a population that is unique to each area.

Prediction: Populations of Sassafras that are under high environmental pressure are more likely to have many predictive SNPs due to evolutionary influences. This analysis will be conducted using the data collected from Guan, et al. (2024).

# Data

## Data Aquisition

These data were collected by Guan et al. (2024). DNA was extracted from dried floral leaves of 106 individual trees in across China. Genotyping-by-sequencing was performed to isolate SNPs into a database. Using variables from the environment from which these plants were collected, including altitude gleaned from their latitude and longitude, weather data such as temperature, precipitation, humidity, and average days of sunshine, and possible data about soil that we can find from the latitudes and longitudes, we will attempt to find SNPs, and therefore gene or gene regions, that predict key differences between Sassafras cultivars.

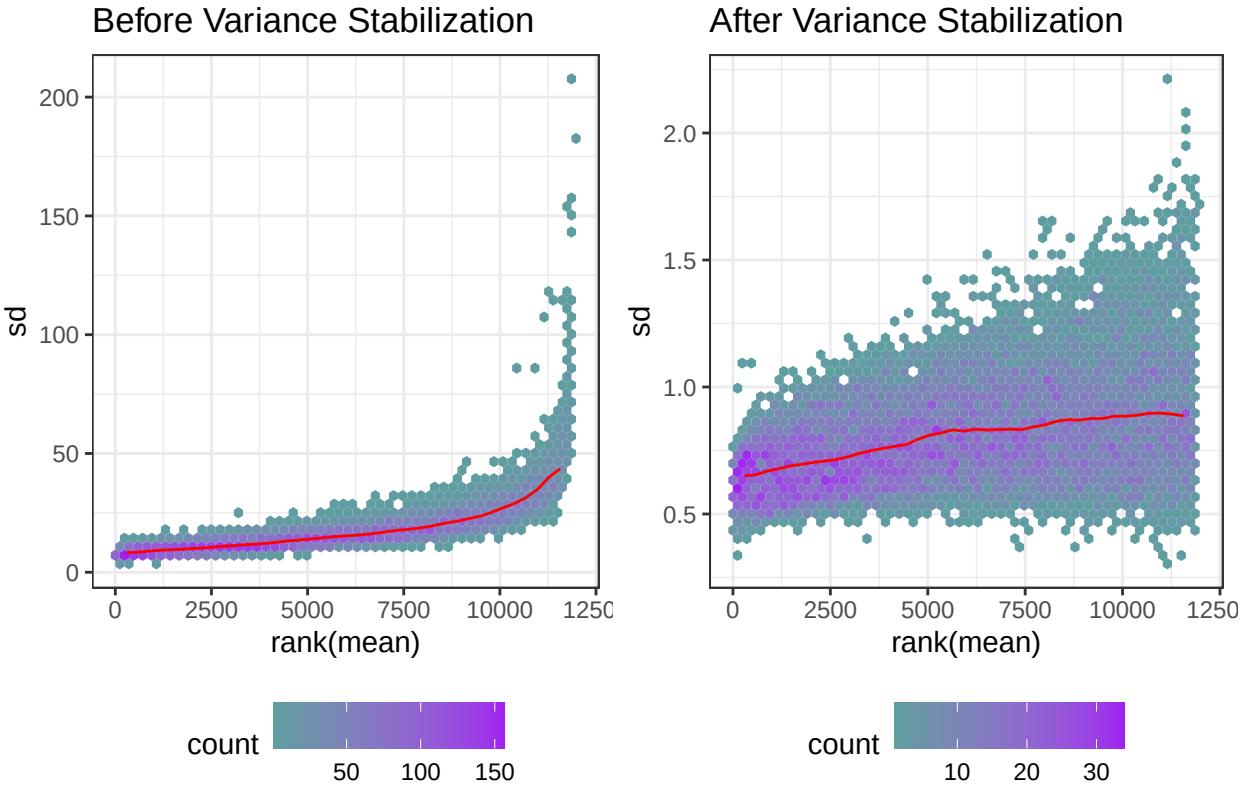
This data is in variant call format (VCF) and contains a 'meta' section that provides the necessary metadata including chromosome locations. This data contains the SNPs of 13 Sassafras tzumu (Lauraceae) populations with 106 individuals. There are 11,862 rows that contain Single Nucleotide Polymorphisms.

For this analysis, each of the SNPs will be considered features, while 'id' will be used as the label. In this matter, we can identify which SNPs are most linked with the each of the populations.

## Data Cleaning

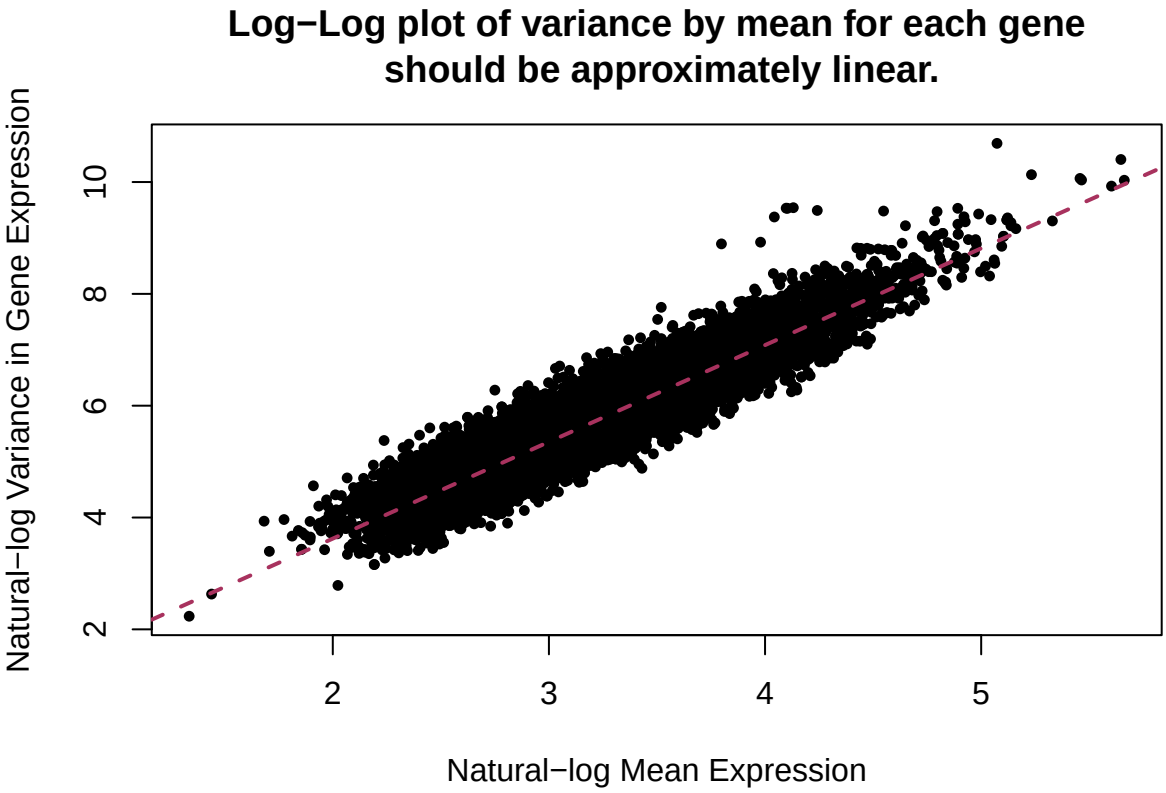
The aims for cleaning this data were mainly to organize the genetic data into a form compatible with various objects that can facilitate analysis specific to genetic data. The first step was to transform the VCF data into a tibble for easier manipulation in R. The 'gt' section of the data contains the counts from the Sassafras samples, so this section was isolated. Metadata was joined with the counts data and variables were then cleaned, including removing white space and leading numbers from variable names. Once the variables were compatible, the data were pivoted wider to create a matrix compatible with the design object from DESeq2. Following this cleaning, a variance stabilizing transformation was applied to the data. The results are shown below in figure 1.

Figure 1




```
## class: DESeqTransform
## dim: 11862 106
## metadata(1): version
## assays(1): ''
## rownames(11862): S3155_111 S3816_65 ... S3725_108 S3900_10
## rowData names(6): baseMean baseVar ... dispGeneIter dispFit
## colnames(106): FD02 FD04 ... WYS03 WYS04
## colData names(9): Longitude Latitude ... Altitude sizeFactor
```

It is clear that the data benefited greatly from the variance stabilizing transformation, as the raw counts had a standard deviation of over 200, while the standard deviation of the VST transformed data was just over two. The transformed data also had a much more stable spread. This step prevents variables with naturally higher variance from over-influencing the future model.




The log-log plot shows an approximately linear relationship between the natural log of the variance and the natural log of the mean.

```
## [1] "After filtering, the number of genes remaining in the dataset are: 2470"
```

Finally, the data was split into training and test sets. The ratio was 80% training and 20% test set with each population split evenly among the two sets.

### Data Exploration

## Models

### Data Pre-Processing

### Algorithm Selection

My proposed analysis will be to use at least two different types of Machine Learning Analysis, including one random forest model and one Support Vector Machine. Depending on the outcome of those models, deep neural networks have also been demonstrated to be effective when working with genetic data (Elgart et al., 2022). The Random Forest algorithm was chosen for its robustness and ability to handle high-dimensional data, making it suitable for gene expression analysis. I also plan to use gini importance to attempt to identify SNPs that may be the most important for differences between cultivars in this dataset. SVMs are also capable of working with non-linear data. Appropriate feature selection and dimensionality reduction measures will probably be necessary. The nature of the available environmental and geographical data may influence the efficacy of the analysis. The analysis will predict which SNPs indicate cultivar differences due to environmental factors. The hypothesis can be supported if the predictive model trained on the data can accurately identify the same SNPs on unseen data.

### Final Model

## Conclusions

In order to access and manipulate the VCF file, several additional R packages were used. The vcfR package allows for loading and manipulation of the data, while the adegenet package was helpful in creating several of the visualizations.

## Discussion and Next Steps

## Code Availability

The code used for this research is available at: https://github.com/samanthaharper/sassafras_capstone

The code is available within the rmd version of this report or in the final code R file.

## References

Elgart, M., Lyons, G., Romero-Brufau, S. et al. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. Commun Biol 5, 856 (2022). https://doi.org/10.1038/s42003-022-03812-z Guan, B., Liu, Q., Liu, X. et al. Environment influences the genetic structure and genetic differentiation of Sassafras tzumu (Lauraceae). BMC Ecol Evo 24, 80 (2024). https://doi.org/10.1186/s12862-024-02264-9 Silva, P.P., Gaudillo, J.D., Vilela, J.A. et al. A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci. Sci Rep 12, 15817 (2022). https://doi.org/10.1038/s41598-022-19708-1 Zhang K, Zhang Y, Jia D, Tao J. Species Distribution Modeling of Sassafras Tzumu and Implications for Forest Management. Sustainability. 2020; 12(10):4132. https://doi.org/10.3390/su12104132