ST592 - Statisitcal Genomics

Final Project Report

Group 3
Sebastian Mueller
Benson Cyril Nana Boakye
Oluwaseun Daniel Fowotade

Instructor: Dr. Xinzhou Shawn Ge

Department of Statistics
Oregon State University
March 19th, 2025

1 Introduction

Deoxyribonuclease Acid (DNA) is a fundamental building block of life. DNA can be thought of as the instructions to the "assembly line" of complex life. From the information encoded in DNA, different cellular activities can produce biological subunits like ribonucleic acid (RNA), amino acids, and proteins, which inurn can be used to build larger biological units. It is through this iterative processes of biological sub units instructing and combining to build larger biological units that complex life emerges [1].

All known biological life uses this cellular machinery to function. One process of the biological "assembly line" many types of life rely on is the termed the Central Dogma of Biology. This process is as follows: DNA is transcribed in RNA which is then translated into protein. Information about the function and form of an organism is encoded by the quantities and qualities of the biological subunits found at each step of the central dogma (DNA, RNA, protein) [3].

One organism of interest is Maize. This crop is found all across the world and has been cultivated for millennium. For this reason, understanding the form and function of maize is particularly important [5].

One area of research is exploring genotype to phenotype predictive models. One group member (Sebastian Mueller) has been exploring how to predict maize pollen phenotype from genotype. Through their work they have identified pollen specificity as a critically important parameter for predicting maize pollen fitness. Currently, Mueller uses RNA and proteomic profiling data to build a predicitve model. However; relying on RNA and proteomic profiling data can cause experimental bottlenecks. This is because RNA and proteomic profiling data require experimental rigor and in turn are more expensive and time consuming to measure. Comparatively, with the advent of next generation sequencing, DNA sequencing has become cheaper and more accessible [2]. For this reason, a predictive model that uses only genome sequence information would be a more accessible tool. Thus, the goal of this project is to build a model that can predict pollen specificity from genomic sequence information. Specifically, this model will take information about a gene and predict if the given gene will express specifically to pollen or not (binary outcome).

2 Data Description

The data for this project is sourced from the public repository MaizeGDB.org. There are 12,725 genes available with expression data present. Through previous work (not done for this project) Mueller derived a binary pollen specificity metric for all 12,725 genes using available expression data. This binary pollen specificity feature was the ground truth label for the dataset used in this project.

For all 12,725 genes there were 2626 different genomic sequence metrics available. All 2626 metrics were used in the modeling project. All 2626 genomic sequence metrics are encoding a type of information from one of the following categories: Sequence Features, Gene Structure, Chromatin Features, Count, Correlation, Varionomic, ”Other”. After binarization there were 1282 pollen specific instances and 11,443 Non-pollen specific instances.

3 Methods

3.1 Feature Selection and Dimensionality Reduction

The dataset consists of 12,725 genes, each with 2,626 genomic sequence descriptors. With such a large number of features, selecting the most relevant ones was crucial for improving model performance and making the results interpretable. We started by removing features with near-zero variance using the `caret` package in R, as these features contribute little useful information and can lead to overfitting. We also eliminated highly correlated features ($|r| > 0.9$) to avoid redundancy and ensure the model focuses on unique, meaningful information.

To explore how the data is structured, we applied Principal Component Analysis (PCA). However, the first two principal components explained only 17% of the total variance, with PC1 capturing just 11.7%. This means that much of the dataset's complexity was not captured by PCA alone. The PCA plot confirmed this, showing no clear separation between pollen-specific and non-pollen-specific genes. Given these results, we concluded that PCA was insufficient as a standalone method for classification, and more advanced feature selection techniques, like LASSO, or non-linear models were necessary.

3.2 t-SNE for Nonlinear Structure Analysis

Since PCA didn't reveal clear patterns, we turned to **t-distributed Stochastic Neighbor Embedding (t-SNE)** to capture non-linear structures in the data. In the t-SNE visualization, each gene is represented as a dot, colored according to whether it is pollen-specific (cyan) or non-pollen-specific (red). While some clusters formed, there was still considerable overlap, making classification challenging.

t-SNE is particularly useful for visualizing complex relationships that PCA might miss. However, the lack of a clear distinction between classes reinforced the need for more advanced modeling approaches. Techniques such as LASSO, Random Forest, or XGBoost may be better suited for identifying meaningful patterns. Additionally, incorporating biological insights through feature engineering could further refine classification accuracy.

3.3 Model Training: Logistic Regression with LASSO Regularization

To refine our predictors, we applied **LASSO (Least Absolute Shrinkage and Selection Operator)** regularization. LASSO helps by both selecting important features and preventing overfitting, ensuring that only the most relevant genomic features are used for classification.

The logistic regression model estimates the probability that a gene is pollen-specific using the equation:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}} \tag{1}$$

where β represents the model coefficients optimized during training. LASSO introduces an L1 penalty to enforce sparsity:

$$\min_{\beta} - \sum_{i=1}^{n} \big(y_i \log p_i + (1 - y_i) \log(1 - p_i) \big) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

This penalty forces some coefficients to be exactly zero, effectively removing unimportant features while maintaining predictive accuracy. By leveraging these techniques, we aimed to build a robust model capable of accurately predicting maize pollen specificity using genomic sequence data.

4 Results

4.1 Confusion Matrix

The confusion matrix provides insight into the model's performance in distinguishing between pollen-specific and non-pollen-specific genes. Here's a breakdown of the results:

- **True Positives (TP = 301)**: Genes that were correctly identified as pollen-specific.
- **False Positives (FP = 1339)**: Non-pollen-specific genes that were mistakenly classified as pollen-specific.
- **False Negatives (FN = 84)**: Pollen-specific genes that the model failed to recognize.
- **True Negatives (TN = 2094)**: Genes that were correctly identified as non-pollen-specific.

Although the model performs well in distinguishing non pollen-specificic genes, the relatively high number of false positives indicates that it sometimes misclassifies nonpollen-specific genes as pollen-specific. Reducing this misclassification could involve refining feature selection methods or adjusting the classification threshold to achieve a better balance between sensitivity and specificity.

4.2 Sensitivity vs. Specificity

The model's sensitivity and specificity were evaluated using three cross-validation (CV) folds to assess its robustness across different data splits.

- **Sensitivity (Red bar)**: Measures the model's ability to correctly identify pollen-specific genes. A lower sensitivity means some pollen-specific genes are missed.
- **Specificity (Green bar)**: Indicates how well the model correctly classifies non-pollen-specific genes. High specificity suggests the model is effective in avoiding false positives.

The results show that specificity remains consistently high, meaning the model is very reliable in identifying non-pollen-specific genes. However, the lower sensitivity suggests that some pollen-specific genes are being overlooked. Addressing this imbalance could involve fine-tuning classification thresholds or improving feature selection techniques to enhance the detection of pollen-specific genes without increasing false positives.

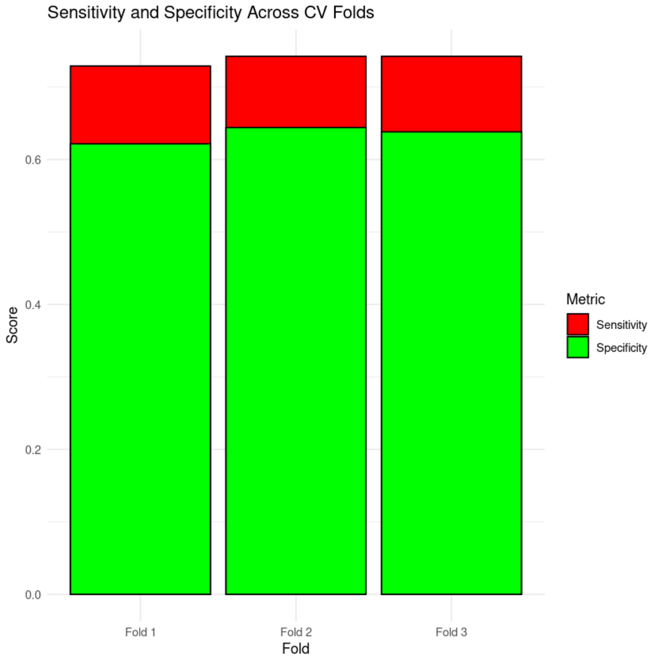

Figure 1

4.3 AUROC Performance for Cross-Validation Folds

The figure above presents the Train vs Validation AUROC scores for each of the three cross-validation (CV) folds in our model evaluation. The Area Under the Receiver Operating Characteristic Curve (AUROC) is a critical metric that evaluates the model's ability to discriminate between positive (pollen-specific) and negative (non-pollen-specific) classes.

In the chart, the blue bars represent the Train AUROC scores, while the orange bars correspond to the Validation AUROC scores. Both sets of AUROC scores are measured across three distinct cross-validation folds (Fold 1, Fold 2, and Fold 3), with each fold reflecting a different split of the training and validation data.

4.3.1 Observations

1. Consistent AUROC Scores: The Train AUROC scores are consistently high, ranging above 0.75 for all three folds. This suggests that the model is well-optimized and performs strongly on the training data, capturing meaningful patterns in the features used for classification.

2. Moderate Gap Between Train and Validation AUROC: A noticeable observation is the slight but consistent difference between the Train AUROC and Validation AUROC scores. The Train AUROC (blue) is higher than the Validation AUROC (orange) across all folds, indicating that the model is slightly overfitting to the training data. The model performs well in distinguishing pollen-specific genes during training, but this level of performance is not fully replicated on unseen data (validation set).

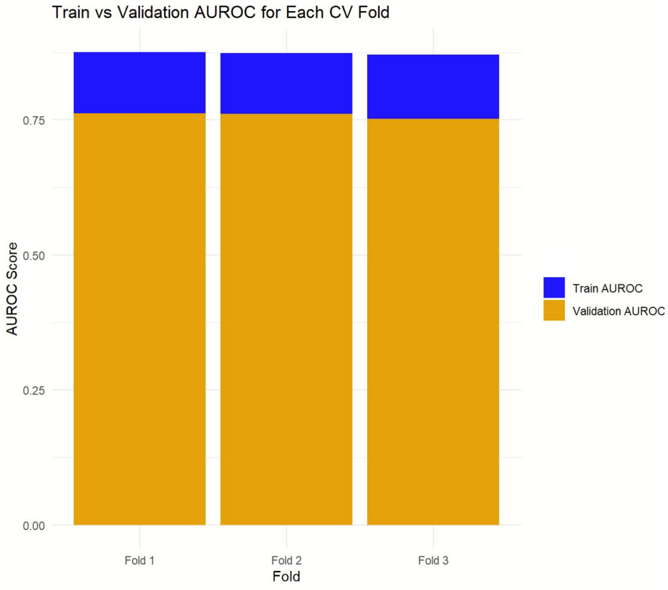

Figure 2

3. Generalization Ability: The Validation AUROC scores hover around 0.74, which indicates that the model has good predictive power but is not perfect. The relatively small gap between the training and validation AUROC scores suggests that the model generalizes well, as it is not heavily overfitting. If the model had a much higher Train AUROC compared to the Validation AUROC, it would indicate that it had memorized the training data and was struggling to generalize to new data.

4. Model Robustness: The results from all three folds show similar AUROC scores, implying that the model's performance is stable and reliable across different splits of the data. This consistency is a positive sign of the model's robustness and its ability to perform well across various data configurations.

The Train vs Validation AUROC comparison demonstrates that the model achieves a good balance between fitting the training data and generalizing to unseen validation data. The consistent performance in both training and validation sets reflects that the model is capturing the underlying patterns in the data effectively. However, there is still room for improvement, particularly in enhancing the model's ability to identify pollen-specific genes with higher sensitivity.

4.4 Held out test set AUROC

- The x-axis represents the False Positive Rate (FPR), which indicates the proportion of non-pollen-specific genes that were incorrectly classified as pollen-specific. It ranges from 0 to 1.

- The y-axis represents the True Positive Rate (TPR), also known as sensitivity, which reflects the proportion of pollen-specific genes correctly identified by the model. This also ranges from 0 to 1.

4.4.1 Observation

1. AUROC (Area Under the Curve): The AUROC value is reported as 0.7437, which suggests the model performs better than random guessing (AUROC = 0.5). An AUROC of 0.74 is considered a decent result, showing that the model is capable of distinguishing between pollen-specific and non-pollen-specific genes in most cases.

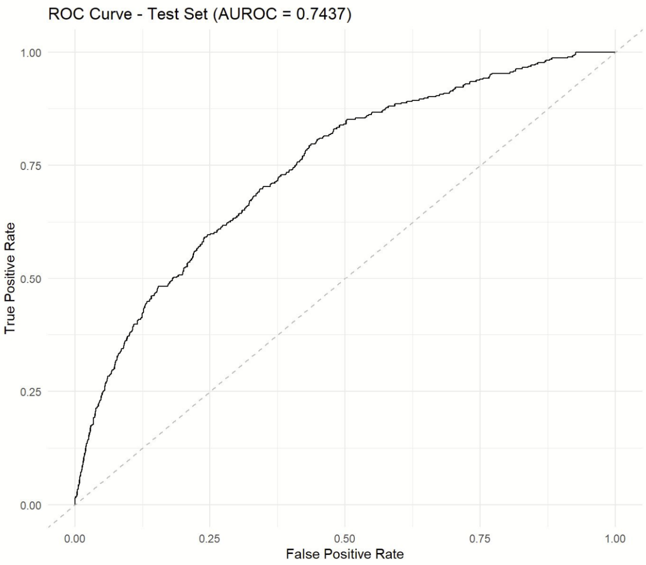

Figure 3

2. The Diagonal Dashed Line: This represents the baseline or random classifier. Any model whose ROC curve lies above this line is better than random, with the goal being for the curve to be as far from this diagonal line as possible.

3. Shape of the ROC Curve: The curve rises steeply at the beginning, indicating that the model is able to correctly identify a substantial number of pollen-specific genes with a relatively low number of false positives. As the curve continues upwards, it flattens out, which means that correctly classifying more pollen-specific genes starts to incur more false positives, reflecting a trade-off between sensitivity and specificity.

4. Performance Insight: The ROC curve indicates that while the model has good sensitivity in detecting pollen-specific genes (higher TPR), it also incurs a number of false positives, as reflected by the rise in FPR. The curve's steep increase at the beginning and its leveling off as it approaches higher FPR values highlight this trade-off.

The ROC curve shows that the model has a solid predictive ability for distinguishing pollen-specific from non-pollen-specific genes, with an AUROC of 0.7437, which is significantly better than random guessing.

4.5 Feature Weights Provide Biological Insight

The mathematical form of logistic regression allows for the inspection of feature coefficients after model training/optimization. These coefficients can be interpreted as a weight given to a features value when a prediction is made. Positive values can be understood as pulling a prediction instance to a higher probability of being a positive class call.

After model training and regularization there were 336 features left with weight. The top 10 highest magnitude positive and negative weights are showcased below in figure 4.

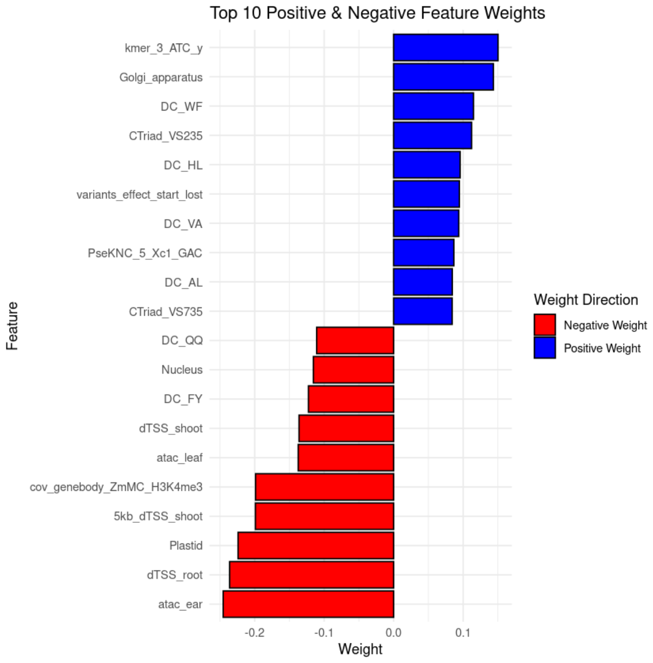

Figure 4: This figure showcases the 10 most highly positive and negative weighted features for our final model

Scrutinizing these features provides biological clues as to what might be important when predicting pollen specificity. For example, the top weighted positive feature is the normalized frequency of the 3-mer ATC. It has been found in past studies that maize pollen has a collection of specifically expressed genes [4], thus the high weight of the ATC kmer might indicate that it works as a cis-regulatory element in pollen specific gene promoters. Thus it may be important to explore kmers of length 4 or 5 to try and capture information about larger motifs. The weights provide clues as to what might be important and what to explore in future model refinement.

5 Conclusion

Predicting pollen specificity from genomic sequence data is a challenging task, as gene expression is influenced by complex biological mechanisms that simple models may not fully capture. Our analysis showed that while PCA helped reduce dimensionality, it did not provide a clear distinction between pollen-specific and non-pollen-specific genes, suggesting that non-linear methods could be more effective.

By applying LASSO regression, we identified 336 key genomic features that contribute to pollen specificity. This approach improved model interpretability while preventing overfitting. The final logistic regression model achieved an AUROC of 74.37%, indicating that it performs significantly better than random guessing but still has room for improvement. The model's high specificity suggests it is reliable in identifying non-pollen-specific genes, though its lower sensitivity indicates that some pollen-specific genes may be missed.

To improve performance, future work should explore alternative feature selection methods, ensemble models like Random Forest and XGBoost, and deep learning approaches that can better capture complex genomic relationships. Addressing class imbalance and integrating additional biological insights could further refine the model's predictive power. While this study represents a meaningful step toward using sequence data for maize pollen specificity classification, ongoing refinements will be essential to enhance its broader applicability in genomic research.

References

[1] L.V. &. *Fundamentals of Genetics*. EDTECH, 2019.

[2] Sam Behjati and Patrick S Tarpey. What is next generation sequencing? *Archives of Disease in Childhood - Education and Practice*, 98(6):236–238, 2013.

[3] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, August 1970.

[4] Yannan Shi, Yao Li, Yongchao Guo, Eli James Borrego, Zhengyi Wei, Hong Ren, Zhengqiang Ma, and Yuanxin Yan. A rapid pipeline for pollen- and anther-specific gene discovery based on transcriptome profiling analysis of maize tissues. *International Journal of Molecular Sciences*, 22(13), 2021.

[5] Kamlesh Prasad Tajamul Rouf Shah and Pradyuman Kumar. Maize—a potential source of human nutrition and health: A review. *Cogent Food & Agriculture*, 2(1):1166995, 2016.