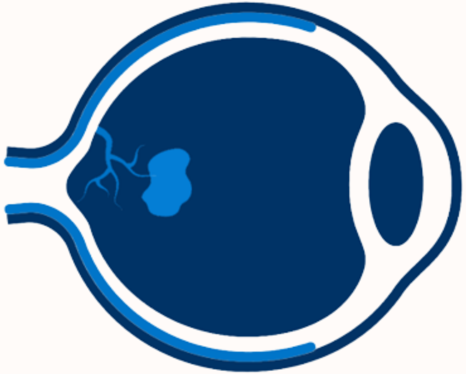


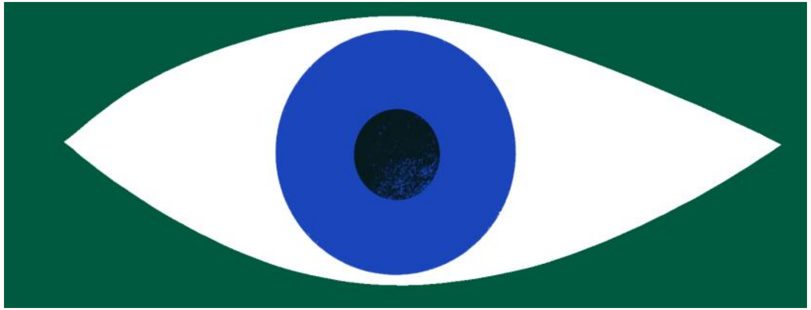
Leveraging ML Solutions to Identify Genomic Signatures in AMD

Sponsor	Dr. Rinki Ratnapriya
Mentor	Maryam Khalid
Team	Duy Ha, Patrick Yee, Sujitha Ravichandran, Qingxin Yuan, Wanying Xu, Tian Xia

RICE | D2K Data to Knowledge Lab

Baylor Medicine

Age- Related Macular Degeneration (AMD) is an eye disease that can blur central vision.



200 million people worldwide suffer from AMD

About **1 in 10 Americans aged 50** and older have the early form of AMD.

Prevalence of age-related macular degeneration (AMD) - Prevent blindness. Prevent Blindness. https://preventblindness.org/amd-prevalence-vehss/ , Nov 2022.

Causes of AMD

Lifestyle & Environmental Risk Factors

Genetics

- **Large # of genes (~18K)** → difficult to identify contributing genes
- **Gene Expression Regulation** → Uncertainty in how disease-associated DNA variants regulate gene expression affecting AMD risk.

Objectives

- **(1) Feature extraction, dimensionality reduction, and engineering.**

  Identify critical genomic signatures related to AMD.

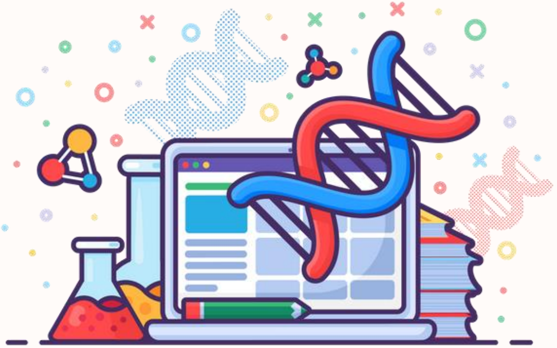
- **(2) Interpretable modeling of AMD labels from reduced data.**

  Use statistical and ML methods to predict AMD.

  Explain those models

- **(3) Deployable Python package.**

  Wrap our pipeline in a generalizable Python package to be used by medical professionals.



To achieve these objectives, we design a 5-step Pipeline

Data Description

Postmortem Retinas from 453 Patients

18056 genes

1. Initial filtering

18,056 Genes

Statistical Filtering
Variance threshold
+ Kolmogorov Smirnov test
+ Wald test
+ Rank-sum test

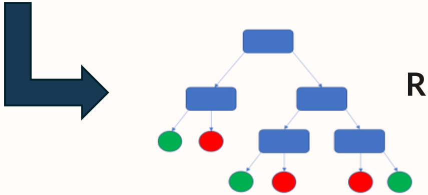
652 Genes

2. Ranking genes

18,056 Genes

Statistical Filtering
Variance threshold
+ Kolmogorov Smirnov test
+ Wald test
+ Rank-sum test

652 Genes



Ranked list of 652 Genes

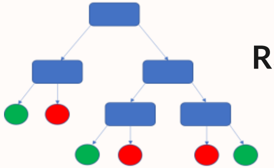
Gene ranking
PCA, MRMR, Random Forest

3. Top *k* genes

18,056 Genes

Statistical Filtering
Variance threshold
+ Kolmogorov Smirnov test
+ Wald test
+ Rank-sum test

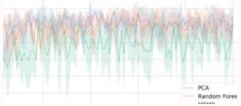
652 Genes



Gene ranking
PCA, MRMR, Random Forest

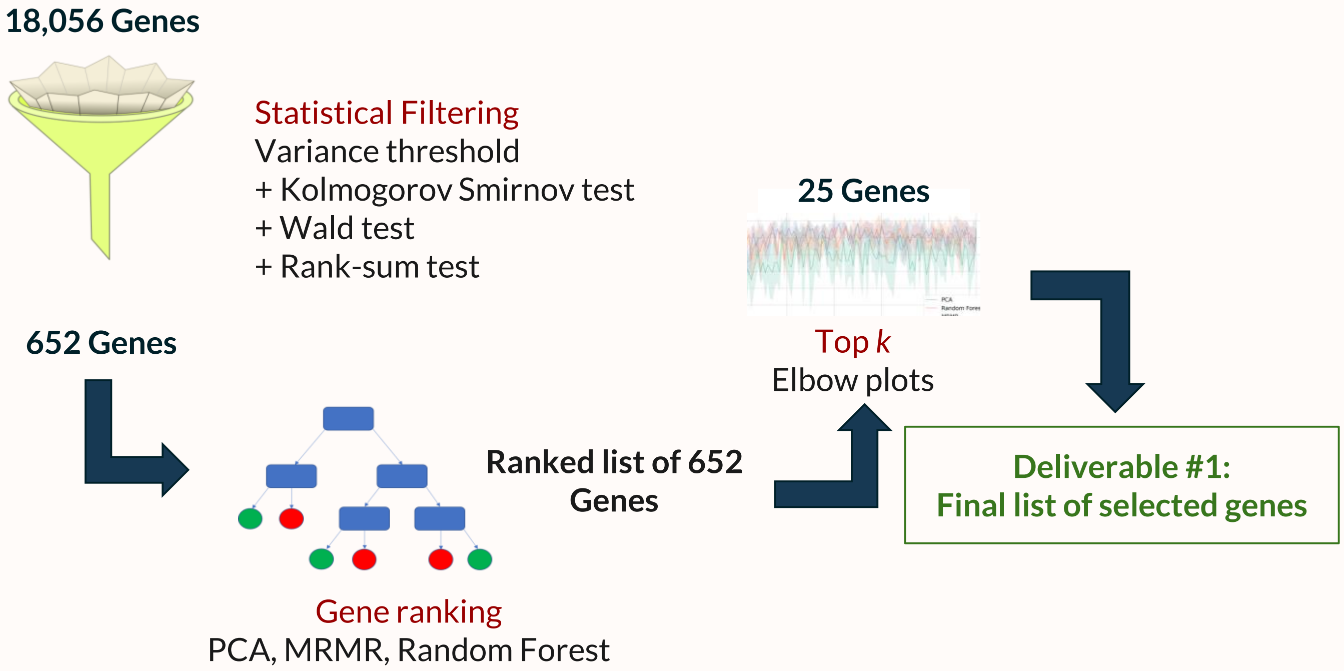
Ranked list of 652 Genes

25 Genes



Top *k*
Elbow plots

3. Top *k* genes



18,056 Genes

Statistical Filtering
Variance threshold
+ Kolmogorov Smirnov test
+ Wald test
+ Rank-sum test

652 Genes

Gene ranking
PCA, MRMR, Random Forest

Ranked list of 652 Genes

25 Genes

Top *k*
Elbow plots

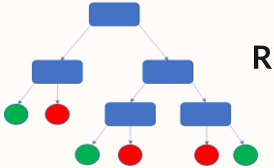
Deliverable #1:
Final list of selected genes

4. Modeling

18,056 Genes

Statistical Filtering
Variance threshold
+ Kolmogorov Smirnov test
+ Wald test
+ Rank-sum test

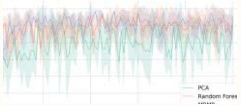
652 Genes



Ranked list of 652 Genes

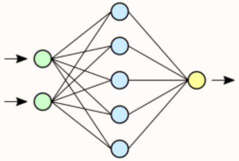
Gene ranking
PCA, MRMR, Random Forest

25 Genes



Top *k*
Elbow plots

Modeling
XGB, RF, Neural Networks



> 95% AUC on test set

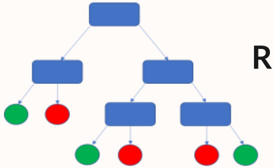
Deliverable #1:
Final list of selected genes

5. Interpretation

18,056 Genes

Statistical Filtering
Variance threshold
+ Kolmogorov Smirnov test
+ Wald test
+ Rank-sum test

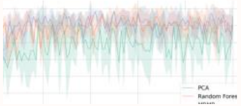
652 Genes



Gene ranking
PCA, MRMR, Random Forest

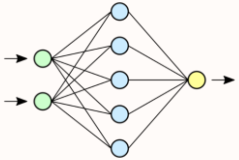
Ranked list of 652 Genes

25 Genes



Top *k*
Elbow plots

Modeling
XGB, RF, Neural Networks

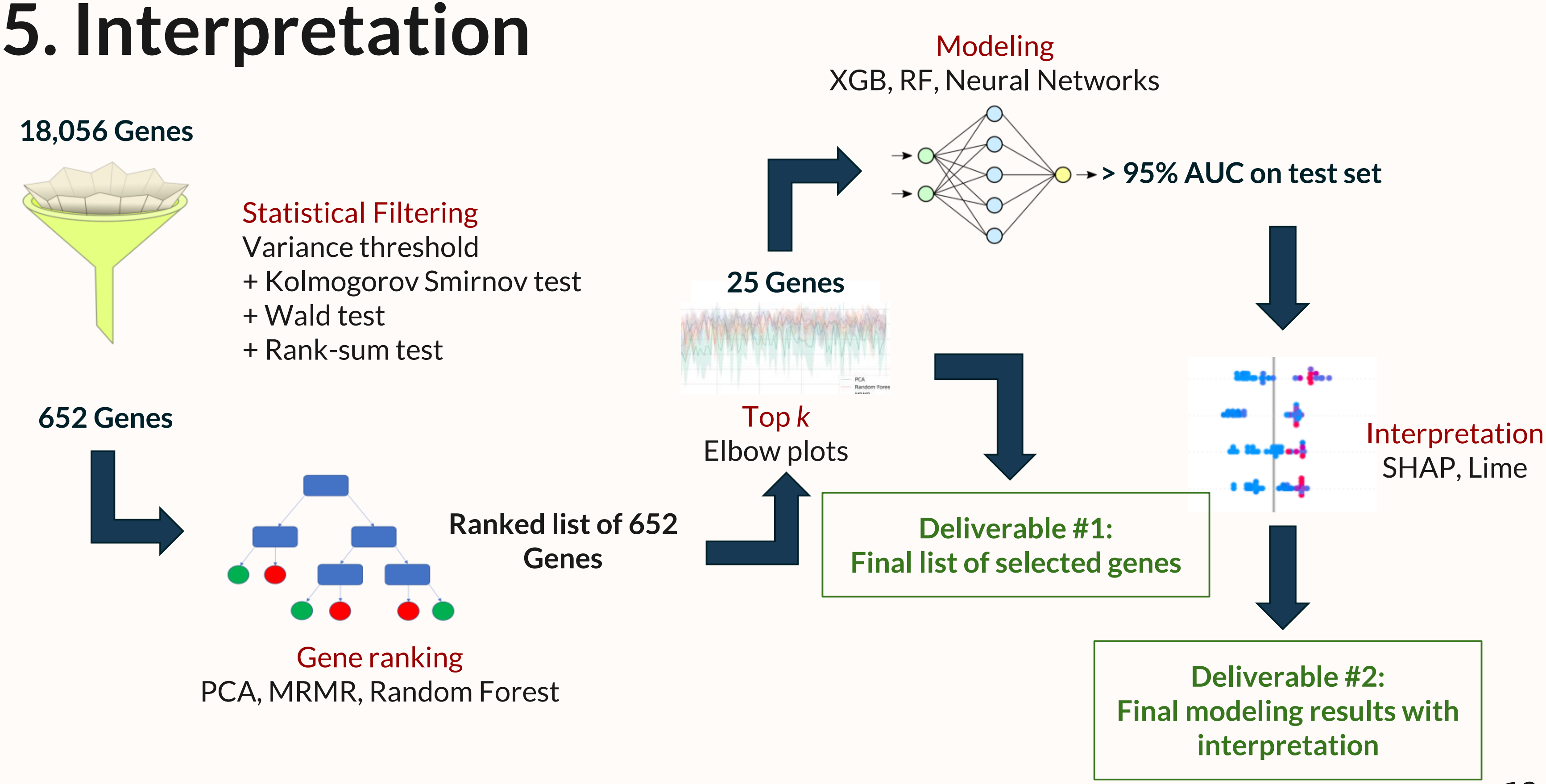


> 95% AUC on test set

Interpretation
SHAP, Lime

Deliverable #1:
Final list of selected genes

5. Interpretation



18,056 Genes

Statistical Filtering
Variance threshold
+ Kolmogorov Smirnov test
+ Wald test
+ Rank-sum test

652 Genes

Gene ranking
PCA, MRMR, Random Forest

Ranked list of 652 Genes

25 Genes

Top *k*
Elbow plots

Modeling
XGB, RF, Neural Networks

> 95% AUC on test set

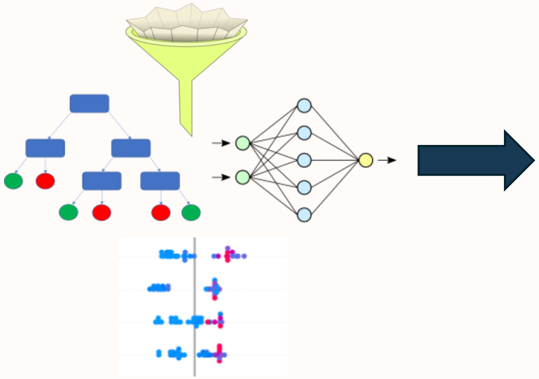
Interpretation
SHAP, Lime

Deliverable #1:
Final list of selected genes

Deliverable #2:
Final modeling results with interpretation

Introducing DREAM-R

Dimensionality Reduction, feature Extraction, and Modeling for RNA transcriptome data



Deliverable #3: Deployable and generalizable Python package

A deployable and generalizable genomic analytics package for medical professionals!

Model Evaluation

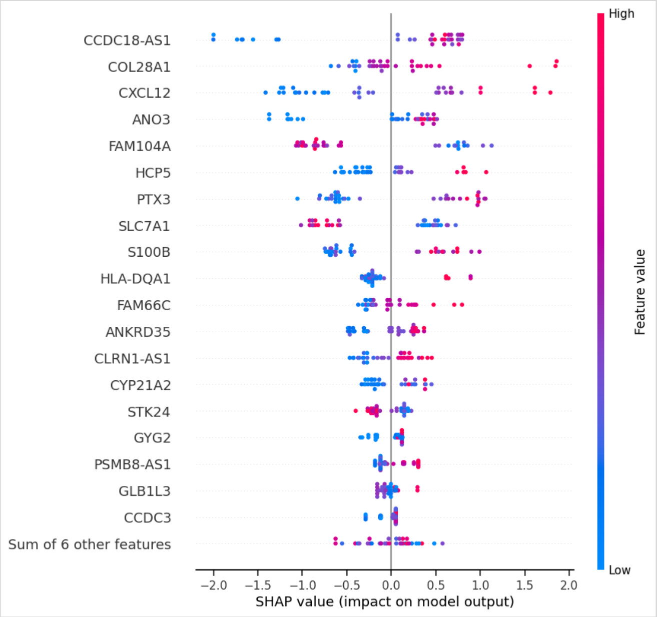
Bootstrapped over 100 iterations on selected top k = 25 genes
Random Split train:test = 80:20

Method + XGBoost	Precision	Sensitivity	Specificity	F1 Score	AUC
MRMR	0.900 ± 0.052	0.894 ± 0.054	0.922 ± 0.065	0.893 ± 0.054	0.946 ± 0.056

- MRMR + XgBoost gives the highest performance ~ 95% AUC

SHapley Additive exPlanations (SHAP): Interpreting the model

- Game theoretic approach to explain the output of any ML model.
- SHAP breaks down a prediction to show the impact of each feature



SHAP scores on test set for XGBOOST model with MRMR feature selection

Conclusion: take the top 10 candidates for an example

Highly expressed genes in AMD subjects	Highly expressed genes in Control subjects
CCDC18-AS1, COL28A1, CXCL12, HCP5, PTX3, S100B, HLA-DQA1, FAM66C	FAM104A, SLC7A1

Balancing Act: ML systems and Gene Discovery Challenges

- **Ensemble models: high prediction performance, but sensitive to number of input genes**

- **MRMR mitigates this issue but is not suitable for *Complete* Gene Discovery:**
  - Tendency to mask important genes which are redundant with others
  - E.g: Biologically significant MOXD1 removed due to previously selected CCDC18-AS1.

- **Models sensitive to data normalization are not recommended:**
  - Different genes are expressed on a different scale, thus data can not be normalized.

- **Statistical filtering is critical for stability of all ML methods.**

- **High prediction performance is a double-edged sword:**
  - The ability to accurately predict AMD from small set of genes detracts from the discovery of all crucial genes.

- **Future Directions:**
  - Incorporating domain knowledge in ML models
  - Extending the problem to multiple stages of AMD
  - Integrating other risk factors

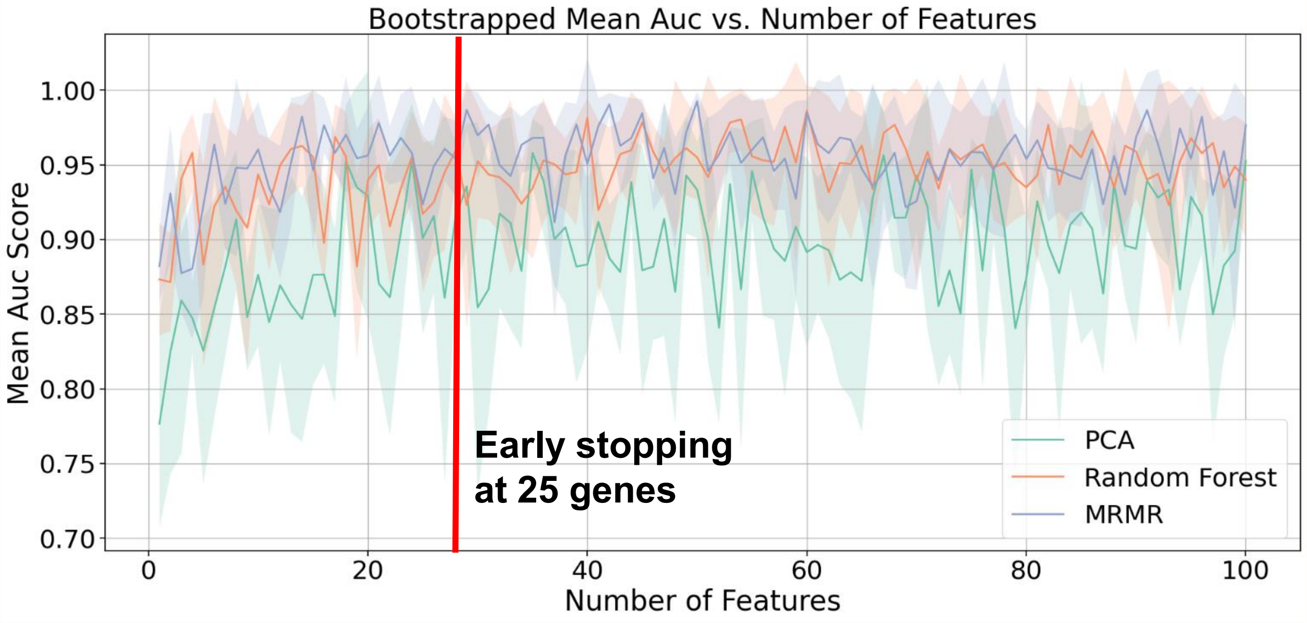
Questions?

Appendix

A1: Comparative Analysis

A comparative analysis between models to identify optimal number of genes.
Our best model was XGBoost when trained on genes selected by MRMR.
This model achieves >95% AUC after 10 bootstrap resamples.

Conclusion: choosing 25 features is enough to reach 95% AUC for MRMR



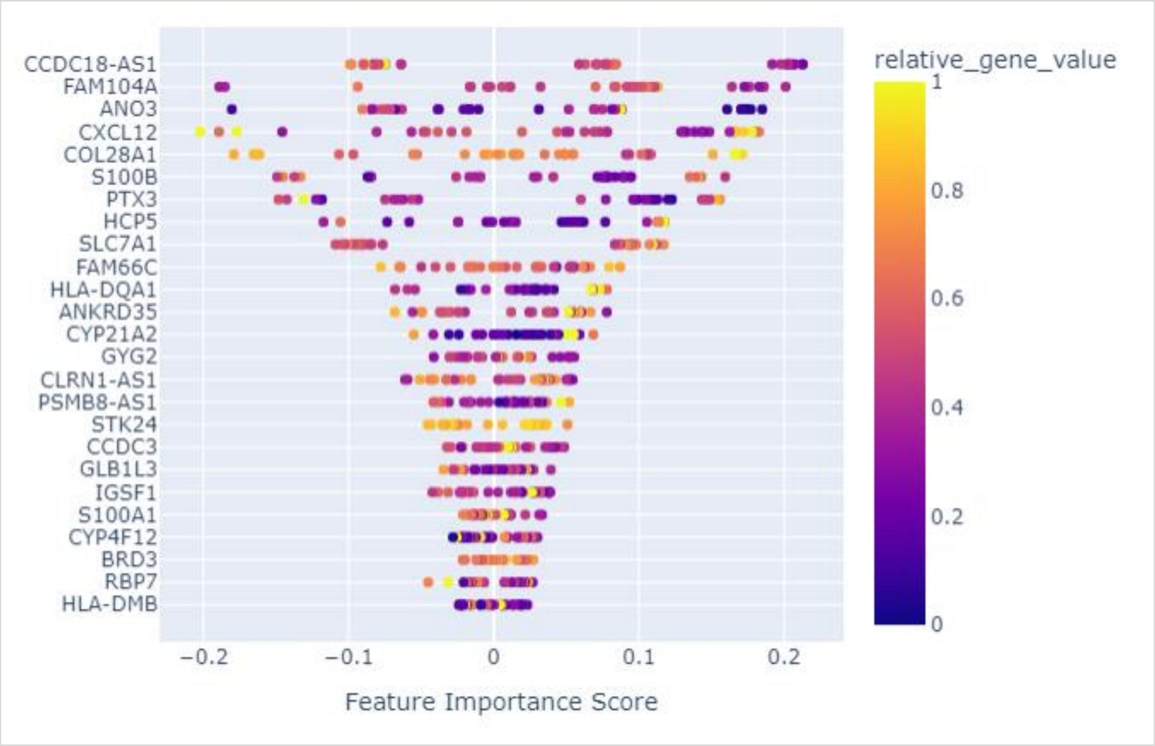
* Zoomed in version from upto 100 features along x-axis. Early stopping at 25 genes.

A1.2

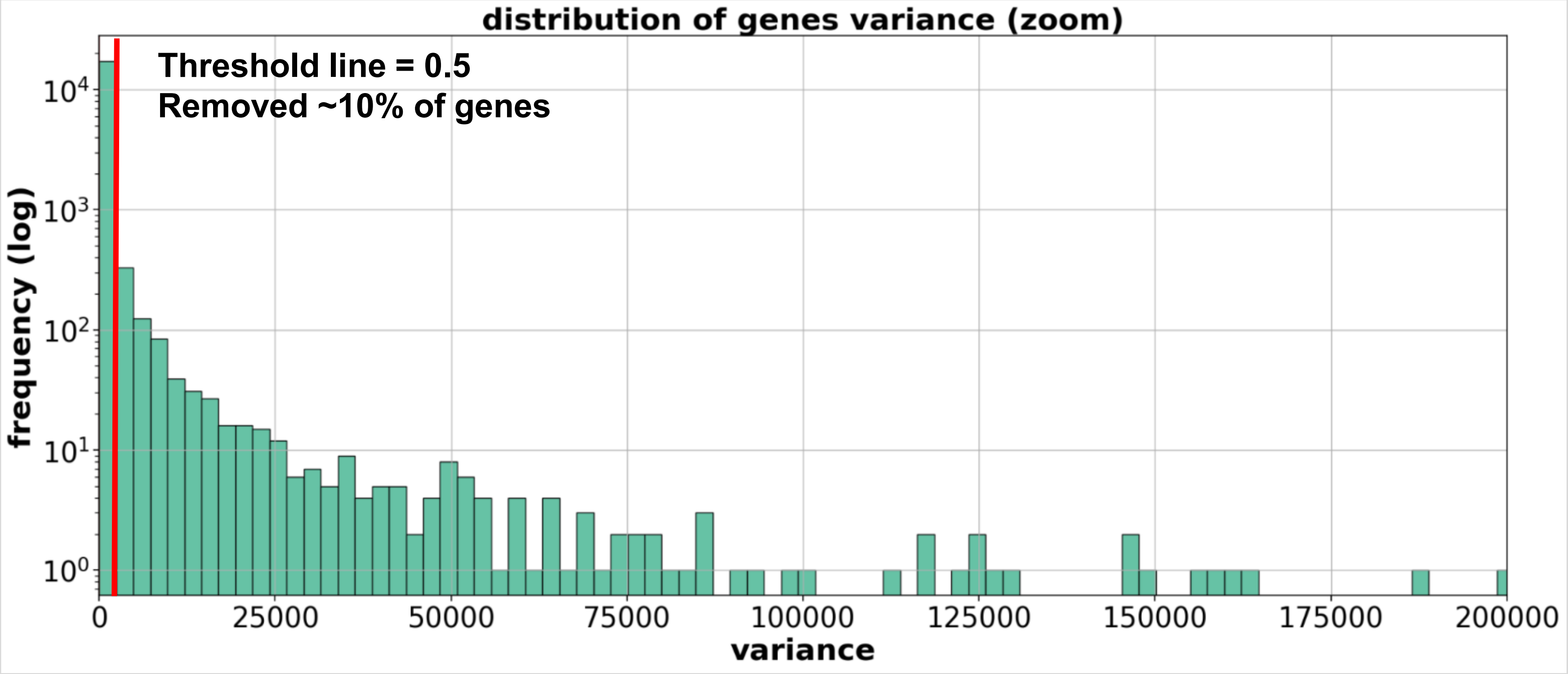
Final list of 25 genes after ranking + early stopping.

- 'S100B',
- 'STK24',
- 'GLB1L3',
- 'BRD3',
- 'CXCL12',
- 'FAM104A',
- 'PSMB8-AS1',
- 'IGSF1',
- 'COL28A1',
- 'PTX3',
- 'CYP21A2',
- 'FAM66C',
- 'HCP5',
- 'GYG2',
- 'CCDC18-AS1',
- 'CCDC3',
- 'HLA-DMB',
- 'ANO3',
- 'CYP4F12',
- 'CLRN1-AS1',
- 'HLA-DQA1',
- 'ANKRD35',
- 'SLC7A1',
- 'S100A1',
- 'RBP7'

A2: Local Interpretable Model-Agnostic Explanations (LIME) on MRMR & XGBoost

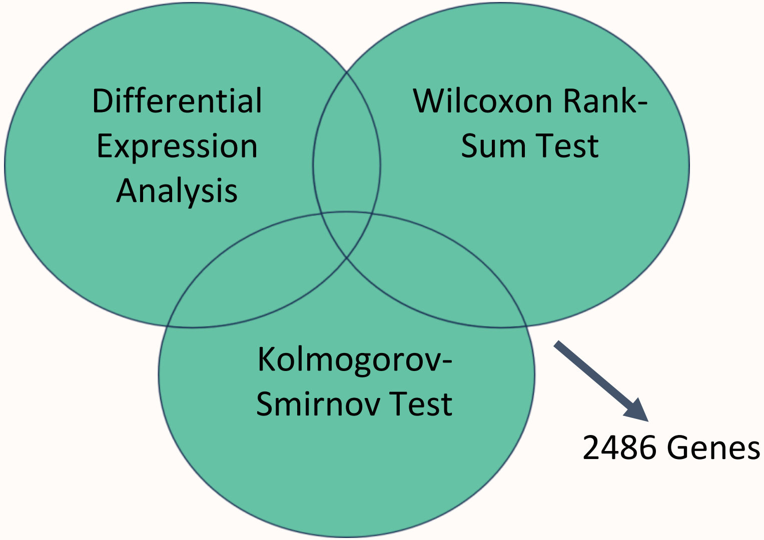


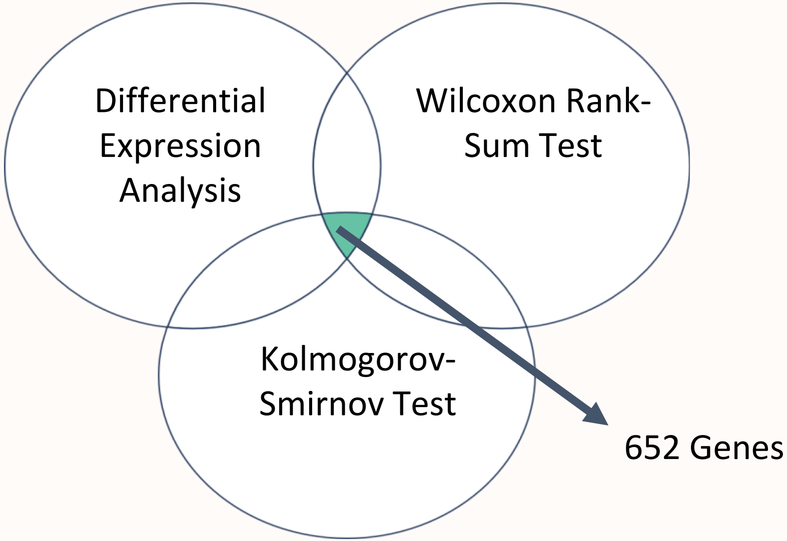
A3: Filtering through variance



Threshold line = 0.5
Removed ~10% of genes

A4: Filtering Through Methods


Differential Expression Analysis
Wilcoxon Rank-Sum Test
Kolmogorov-Smirnov Test
2486 Genes


Differential Expression Analysis
Wilcoxon Rank-Sum Test
Kolmogorov-Smirnov Test
652 Genes