Classification of Disease-Associated Genomic Variants in Rice Crops using Machine Learning.

Team Members

Aditya R Auradkar - BL.SC.U4AIE24005

Bhuvaneswar Reddy - BL.SC.U4AIE24019

P Varun Sathwik - BL.SC.U4AIE24035

1. Introduction

This report describes the curation, validation, and integrity analysis of a genomic dataset constructed for studying disease-associated variation in rice (*Oryza sativa*). The dataset includes healthy reference gene sequences and disease-associated variant-bearing sequences derived from public genomic resources. The emphasis of this section is on dataset preparation, redundancy analysis, and sequence validation.

2. Dataset Sources

2.1 Healthy Gene Sequences

A total of 3000 full-length rice gene sequences were collected in FASTA format. These sequences represent complete gene-level nucleotide data and serve as the healthy reference dataset.

2.2 Disease-Associated Sequences

Disease-associated sequences were curated using two complementary sources:

1. Public genomic variant information from **Ensembl Plants**, which provides plant genomic annotations and variation data.
2. Biologically realistic nucleotide substitutions applied to simulate disease-associated mutation burden while maintaining sequence integrity.

Each diseased sequence corresponds to a healthy reference gene and contains controlled nucleotide-level variations.

3. Dataset Summary

The final curated dataset contains:

Total Samples: 6000

Healthy Samples: 3000

Diseased Samples: 3000

Each sequence is full-length, with representative gene lengths averaging approximately 7833 base pairs.

Feature extraction produced 65 sequence-derived features per sample (k-mer frequencies and GC content), prepared for downstream analysis.

Train-test stratified split:

Train Samples: 4800

Test Samples: 1200

4. Redundancy Analysis

To ensure dataset integrity, a redundancy check was performed by comparing all sequence strings.

Results:

- Unique Sequences: 5996
- Duplicate Sequences: 4

The duplicate rate corresponds to approximately 0.067% of the total dataset. This minimal duplication level is statistically negligible and does not affect overall dataset balance or validity.

The redundancy analysis confirms that:

- The dataset is overwhelmingly unique.
- No large-scale sequence duplication is present.
- Healthy and diseased sequences remain distinct in almost all cases.

5. Alignment Verification

To validate that disease-associated sequences contain genuine nucleotide-level differences relative to healthy counterparts, pairwise sequence alignment was performed.

Alignment summary:

Sequence Length: 7833

Matching Positions: 7819

Differing Positions: 14

The alignment output (see alignment_example.txt) confirms:

- Controlled mutation incorporation.
- No structural distortion of sequence length.
- Presence of nucleotide substitutions consistent with expected mutation burden.

The number of differing positions aligns with the curated mutation range introduced during diseased dataset construction.

6. Dataset Integrity Validation

The dataset curation process ensured:

- Balanced class distribution (3000 healthy / 3000 diseased)
- Full-length gene sequences retained
- Controlled mutation burden applied
- Minimal redundancy
- Verified nucleotide-level differences via alignment

The curated dataset is therefore structurally sound, biologically consistent, and suitable for downstream genomic analysis.

Feature Extraction and Biological Relevance

1. k-mer Frequency (k = 3)

Trinucleotide (3-mer) frequency was used to capture local nucleotide composition patterns within each gene sequence. A single nucleotide substitution affects multiple overlapping 3-mers, making this representation highly sensitive to mutation-induced compositional shifts. Since trinucleotides correspond to codons in coding regions, this feature also indirectly reflects potential codon-level alterations. Thus, k-mer frequency provides a biologically meaningful measure of sequence perturbation at a fine-grained structural level.

2. GC Content

GC content represents the proportion of guanine (G) and cytosine (C) bases within a gene sequence. GC composition influences DNA stability, melting temperature, and transcriptional behaviour. Even modest nucleotide substitutions can shift overall GC proportion, reflecting broader compositional imbalance. Unlike k-mers, which capture localized patterns, GC content provides a global compositional indicator, allowing detection of mutation-driven shifts in nucleotide bias across the entire gene.

3. Mutation Density

Mutation density quantifies the number of nucleotide substitutions present within each diseased gene sequence. This feature directly represents variant burden and serves as a measurable indicator of genomic perturbation. Higher mutation density implies greater deviation from the reference sequence and increased disruption of nucleotide composition. Including this feature allows explicit modelling of mutation magnitude, complementing compositional features such as k-mer frequency and GC content.

4. Sequence Length

Sequence length reflects the total number of nucleotides within each gene. Although no insertions or deletions were introduced in this dataset, length remains a biologically relevant structural attribute. Gene length influences mutation probability, regulatory complexity, and structural organization.

Including length ensures methodological completeness and maintains compatibility with future datasets that may incorporate structural variations such as indels or truncations.

7. Discussion

The curated dataset demonstrates a structured approach to modelling disease-associated genomic variation in rice gene sequences. By integrating full-length healthy genes with variant-bearing counterparts, the dataset captures nucleotide-level differences while preserving sequence structure and biological realism.

The use of Ensembl Plants as a genomic reference ensures that disease-associated variation is grounded in publicly available plant genomic resources. Furthermore, the introduction of biologically realistic mutation biases favouring transition mutations over transversions aligns with known mutation patterns observed in plant genomes. This design choice strengthens the biological plausibility of the dataset.

Redundancy analysis confirmed a negligible duplication rate (0.067%), indicating strong dataset integrity. Such validation is critical, as duplicate sequences can introduce bias and artificially inflate downstream analytical performance.

Pairwise alignment verification further demonstrated that nucleotide substitutions were successfully incorporated without affecting overall sequence length or structure. The observed mutation burden (approximately 8–20 substitutions per gene) represents a small proportion of total gene length, maintaining biological plausibility while enabling measurable sequence variation.

8. Conclusion

This project presents a systematically curated genomic dataset for analysing disease-associated nucleotide variation in rice genes. The dataset consists of 6000 balanced full-length gene sequences, with controlled variant incorporation and validated sequence integrity.

Redundancy analysis confirmed minimal duplication, and alignment verification demonstrated accurate mutation integration. The dataset design preserves biological plausibility while enabling structured computational analysis.

Overall, the curated dataset provides a reliable foundation for genomic sequence classification and mutation burden analysis in plant genomics. Future extensions may incorporate functional variant annotation, gene expression integration, and phenotype-linked genomic data to enhance biological interpretability.