Imports and Setup

This block imports all required Python libraries for data processing, visualization, statistical analysis, machine learning, and the Streamlit app framework.

Streamlit UI Title

Displays the app title using Streamlit.

File Upload

Allows users to upload a VCF file (genomic data) and a CSV file (phenotype data) through the Streamlit interface.

VCF and Phenotype Data Preprocessing

Reads VCF using scikit-allel, extracts genotypes, and merges them with phenotypic data based on sample IDs.

GWAS Results Calculation

For each trait and SNP, runs a linear regression using statsmodels to compute p-values. Stores results in a DataFrame.

GWAS Visualization - Manhattan Plot

Plots -log10(p-values) for each SNP per trait as a Manhattan plot to visualize significant associations.

Genomic Selection

Selects top 50 significant SNPs and trains two ML models (RandomForest and Ridge Regression) to predict pest resistance using cross-validation.

QTL Mapping

Performs simple linear regression between each SNP and the pest resistance trait to identify QTLs and plots -log10(p-values) by genomic position.

PCA Clustering

Applies PCA to the selected SNP data and visualizes trait distribution by coloring points in 2D PCA

space.

Candidate SNPs

Extracts SNPs with p-values below a significance threshold (1e-5) as top candidates.

Functional Annotation Placeholder

Placeholder text indicating where functional annotation logic using Biopython or NCBI Entrez API could be added.