Plant Genetics, Genomics, & Biotechnology 2025, pp. 120–127
https://doi.org/10.5147/pggb.266

Published in the United States of America
https://www.atlas-publishing.org

Dissecting Root, Shoot, and Water-Use Traits in Soybean Using Genomic Prediction and Explainable Machine Learning

Dounya Knizia[1], Khalid Meksem[1], and My Abdelmajid Kassem[2*]

[1] School of Agricultural Sciences, Southern Illinois University, Carbondale, IL 62901, USA; [2] Plant Genomics and Bioinformatics Lab, Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC 28301, USA.

Received: May 8, 2025 / Accepted: September 28, 2025

Abstract

Improving drought resilience in soybean requires a deep understanding of the genetic basis underlying root architecture, shoot biomass, and water-use traits. Using the well-characterized Essex × Forrest recombinant inbred line (RIL) population (n = 94), we combined classical composite interval mapping (CIM) with interpretable machine learning (ML) models to dissect 15 traits related to early vigor and drought tolerance. Genomic prediction models—including Ridge Regression and XGBoost—were trained on 370 molecular markers. XGBoost achieved superior predictive accuracy (R^2 up to 0.72), especially for biomass-related traits. SHAP (SHapley Additive exPlanations) analysis provided interpretable insights into marker contributions, identifying both previously known QTL and novel loci with directional effects. Several high-importance markers aligned with QTL reported by Williams et al. (2012) and Salvador et al. (2012), supporting the biological validity of the ML-based approach. Traits such as relative water content (RWC), root fresh weight (RFW), and shoot dry weight (SDW) were effectively modeled, and markers on chromosomes 1, 8, 10, and 18 emerged as pleiotropic hotspots. This integrative framework showcases the power of explainable AI in plant genomics and offers a robust pipeline for future marker-assisted selection in soybean breeding.

Keywords: Soybean, Genomic Prediction, Root Traits, Shoot Biomass, Water-Use Efficiency, QTL Mapping, Machine Learning.

* Corresponding author: mkassem@uncfsu.edu

1. Introduction

Soybean (Glycine max (L.) Merr.) plays a central role in global agriculture, serving as a major source of plant-based protein and oil for food, feed, and industrial products. As climate variability increases, improving complex physiological traits—particularly those related to water uptake, biomass accumulation, and early vigor—has become a priority for sustaining soybean productivity under abiotic stress conditions such as drought. Traits like basal root thickness (BRT), lateral root number (LRN), maximum root length (MRL), plant height (PH), seed weight (SW), root fresh weight (RFW), root dry weight (RDW), shoot fresh weight (SFW), shoot dry weight (SDW), and (RFW/RDW) are vital for nutrient acquisition, water-use efficiency, and overall plant performance.

To better understand the genetic control of these traits, we focused on the well-characterized Essex × Forrest (ExF) recombinant inbred line (RIL) population, which includes 94 lines derived from a cross between two agronomically distinct cultivars. Essex contributes higher yield and oil content, while Forrest offers strong disease resistance and moderately elevated protein levels. Since its development, the ExF population has been extensively phenotyped across multiple environments (Meksem et al., 2001; Cho et al., 2002; Njiti et al., 2002; Yuan et al., 2002; Lightfoot et al., 2005; Kassem et al., 2004a,b, 2006, 2007a,b; Alcivar et al., 2007; Jacobson et al., 2007; Karangula et al., 2009; Ivey et al., 2011; Williams et al., 2012; Salvador et al., 2012), providing a rich dataset for trait dissection. Prior work by Williams et al. (2012) identified significant quantitative trait loci (QTL) controlling root and shoot traits using composite interval mapping (CIM), with key genomic regions clustered on chromosomes 3, 6, 8, 13, 14, and 18. Many of these QTL exhibited pleiotropic effects, influencing multiple traits simultaneously (Williams et al., 2012).

In our study, we extended this analysis by incorporating four additional physiological traits related to leaf tissue: leaf fresh weight (LFW), leaf dry weight (LDW), leaf turgid weight (LTW), and relative water content (RWC) from Salavador et al. (2012). These parameters provide complementary insights into drought tolerance mechanisms at the seedling stage, particularly in the context of water retention and turgor maintenance. Among them, RWC is widely recognized as a robust indicator of leaf hydration status, integrating aspects of water uptake, transpiration, and cellular elasticity.

To quantify RWC, we adopted the standard formula described by Salvador et al. (2012):

RWC=(Leaf Fresh Weigh-Leaf Dry Weigh)/(Leaf Turgid Weight-Leaf Dry Weight)

Previous QTL mapping in the ExF population identified genomic regions associated with RWC, particularly under water-limited conditions, highlighting the genetic basis of physiological drought tolerance mechanisms (Salvador et al., 2012; Grant et al., 2010).

While traditional QTL mapping approaches like CIM have successfully identified major loci for complex traits, they face limitations in detecting small-effect QTL, modeling nonlinear relationships, and handling genotype-by-environment (G×E) interactions. These limitations have led to increasing adoption of genomic prediction (GP) and machine learning (ML) approaches in plant genetics. GP models use genome-wide marker information to predict phenotypic performance, enabling selection before full phenotyping and thereby accelerating breeding cycles (Meuwissen et al., 2001; Crossa et al., 2017). Among machine learning methods, Ridge Regression Best Linear Unbiased Prediction (RR-BLUP) is widely used due to its simplicity and robustness. However, nonlinear models such as extreme gradient boosting (XGBoost) offer the added advantage of modeling epistatic and interaction effects (Chen and Guestrin, 2016).

The integration of explainable AI tools like SHAP (SHapley Additive exPlanations) further enhances ML-based genomic analysis by attributing trait variation to specific markers in an interpretable manner (Lundberg and Lee, 2017). This capability is particularly valuable for breeding programs, as it provides biological insight and allows for prioritization of markers with real-world predictive utility.

Given the depth of phenotypic and genotypic data available for the ExF RIL population, including both classical and newly evaluated traits, there is an opportunity to revisit the genetic basis of these traits using state-of-the-art machine learning and interpretation frameworks. By comparing traditional QTL regions with marker importance derived from SHAP and ML-based models, we can assess both predictive power and biological relevance.

In this study, we aimed to (i) perform machine learning-based genomic prediction for 15 root, shoot, and leaf traits in the ExF RIL population, (ii) optimize model performance using hyperparameter tuning, (iii) apply SHAP for marker-level interpretation, and (iv) compare important markers identified through ML to those previously reported via CIM. This integrative approach seeks to enhance our understanding of the genetic architecture underlying biomass partitioning and water-use efficiency in soybean and provide valuable tools for marker-assisted selection.

2. Materials and Methods

2.1. Plant Materials

The study utilized 94 recombinant inbred lines (RILs) developed from a cross between the soybean cultivars Essex and Forrest (ExF, n = 94). The population was advanced to the F6:8 generation via single-seed descent (Lightfoot et al., 2005). Field evaluations were carried out in 2010 at two contrasting environments, Carbondale, Illinois, and Spring Lake, North Carolina, using a randomized complete block design with two replications at each site (Williams et al., 2012; Salvador et al., 2012).

2.2. Phenotyping Evaluation

Eleven (11) traits related to root and shoot architecture: basal root thickness (BRT), lateral root number (LRN), maximum root length (MRL), root fresh weight (RFW), root dry weight (RDW), shoot fresh weight (SFW), shoot dry weight (SDW), plant height (PH), and the ratios of RFW/SFW and RDW/SDW were evaluated as described earlier (Williams et al., 2012). Briefly, plants were harvested at maturity growth stage, root systems were gently washed, and fresh weights were recorded immediately. Dry weights were obtained after oven-drying samples at 60°C for 72 hours. BRT and MRL were measured manually using digital calipers and rulers. LRN was determined by visual counting.

We also evaluated an additional four (4) traits: leaf fresh weight (LFW), leaf dry weight (LDW), leaf turgid weight (LTW), and relative water content (RWC). RWC was calculated by the formula: RWC = (Leaf Fresh Weight – Leaf Dry Weight)/(Leaf Turgid Weight – Leaf Dry Weight) as described earlier (Salvador et al., 2012).

2.3. Genotyping and Data Processing

DNA was extracted from young trifoliate leaves using a modified cetyltrimethylammonium bromide (CTAB) method, following the protocol previously described by Kassem et al. (2006). Genotyping of the Essex × Forrest (ExF) recombinant inbred line (RIL) population (n = 94) was performed using a panel of 370 molecular markers, including 195 simple sequence repeat (SSR) markers and 175 additional markers comprising restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), and morphological loci. This marker set was selected to ensure comprehensive coverage across all 20 soybean chromosomes (Kassem et al., 2006).

Genotype calls were quality-checked, and markers with ambiguous scoring were re-evaluated. Individuals or markers with more than 10% missing data were excluded from the analysis. Remaining missing genotypes were imputed using the most frequent allele observed at each marker within the population. The final genotype matrix was organized with individuals as rows and markers as columns, and marker data were coded categorically based on allele origin—Essex, Forrest, or heterozygous where applicable (Kassem et al., 2006).

2.4. Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted using Python 3.11 (Pandas, Seaborn, Matplotlib libraries). Summary statistics (Mean, Standard Deviation, Minimum, and Maximum) were computed for each trait. Histograms were plotted to assess trait distributions, and Pearson correlation coefficients among traits were calculated. No significant outliers were detected.

2.5. QTL Mapping: Composite Interval Mapping

Quantitative trait loci (QTL) mapping was conducted using composite interval mapping (CIM) as previously described earlier (Williams et al., 2012; Salvador et al., 2012). Analyses were performed in QTL Cartographer version 2.5 (Wang et al., 2012), employing standard model 6 (forward regression) with a walking speed of 1 cM and a window size of 10 cM. Significance thresholds for QTL detection were determined using 1,000 permutations at a significance level of P = 0.05.

For each trait, the genomic positions of significant QTL were reported in centimorgans (cM), and corresponding statistics—including LOD score, additive effect, and the proportion of phenotypic variance explained (R^2)—were recorded. This mapping strategy follows the protocols established in earlier studies on the same population to ensure methodological consistency and comparability of results (Williams et al., 2012; Salvador et al., 2012).

2.6. Machine Learning-Based Genomic Prediction

Two ML models were implemented for genomic prediction:

2.6.1. Ridge Regression Best Linear Unbiased Prediction (RR-BLUP)

Ridge regression was applied using RidgeCV (scikit-learn), with alpha (regularization strength) optimized across a log-spaced range from 10−610−6 to 106106. Five-fold cross-validation (CV) was used for model selection.

2.6.2. Extreme Gradient Boosting (XGBoost)

XGBoost regression models were implemented using the xgboost Python package. Hyperparameters were optimized via grid search (GridSearchCV) over the following ranges:
n_estimators: [300, 500], max_depth: [4, 6, 8], learning_rate: [0.01, 0.05, 0.1], subsample: [0.8, 1.0], and colsample_bytree: [0.8, 1.0]. The optimal model was selected based on mean 5-fold CV R^2 score.

2.7. Model Evaluation

Model performance was evaluated on a 20% held-out test set. Performance metrics included coefficient of determination (R^2) and Root Mean Square Error (RMSE) (Willmott and Matsuura, 2005). Additionally, five-fold CV R^2 scores were reported to assess model robustness across the full dataset.

2.8. Feature Importance and Explainable AI

Feature importance from the best XGBoost model was extracted and ranked based on gain scores. SHAP (SHapley Additive exPlanations) analysis was conducted using the TreeExplainer from the shap package to assess the contribution of individual markers to trait predictions. Global SHAP summary plots and marker-specific dependence plots were generated.

2.9. Software and Computational Resources

All analyses were performed using Python 3.11 (Van Rossum and Drake, 2009). Packages used included: Pandas (McKinney, 2010), Numpy (Harris et al., 2020), Scikit-learn (Pedregosa et al., 2011), XGBoost (Chen and Guestrin, 2016), SHAP (Lundberg and Lee, 2017), Matplotlib (Hunter, 2007), Seaborn (Waskom, 2021), Statsmodels (Seabold and Perktold, 2010). Jupyter Notebbok (Kluyver et al., 2016) was used as the interactive environment. Computations were performed on a MacBook Pro equipped with an Apple M1 Pro processor and 16 GB of RAM, running macOS Sonoma version 14.4.1.

3. Results

3.1 Overview of Previously Identified QTL

Extensive prior work on the Essex × Forrest (ExF) recombinant inbred line (RIL) population has characterized the genetic architecture of root, shoot, and water-related traits using traditional QTL mapping approaches (Kassem, 2021). Williams et al. (2012) identified significant QTL controlling basal root thickness (BRT), lateral root number (LRN), maximum root length (MRL), root fresh weight (RFW), root dry weight (RDW), shoot fresh weight (SFW), and shoot dry weight (SDW) through composite interval mapping (CIM) implemented in QTL Cartographer. Major QTL clusters were localized on chromosomes 3, 6, 8, 13, 14, and 18, with some loci exhibiting pleiotropic effects across multiple traits (Williams et al., 2012). Additional work by Salvador et al. (2012) expanded this analysis to include leaf-related traits, identifying significant QTL for leaf fresh weight (LFW), leaf dry weight (LDW), and relative water content (RWC), particularly on chromosomes 2, 3, 4, 6, 8, 10, 11, 17, and 18. Collectively, these studies highlighted key genomic regions contributing to plant biomass partitioning, root system development, and water retention efficiency in soybean. Building upon these foundational results, the present study aims to re-assess the genetic basis of these traits using ML-based genomic prediction models and SHAP-based marker interpretation, providing a complementary and potentially higher-resolution understanding of trait architecture in the ExF RIL population.

3.2 Phenotypic Variation Among RILs

Phenotypic evaluation of the ExF RIL population revealed substantial variation across all 15 measured traits (Figure 1). Root biomass (RFW, RDW), shoot biomass (SFW, SDW), and water-related indices (RWC, LFW, LTW) demonstrated continuous, polygenic-like distributions (Figure 2). Summary statistics (Table 1) indicated wide ranges and moderate to high coefficients of variation (CV), particularly for shoot traits. Traits such as BRT and RDW exhibited narrower distributions, suggesting more constrained genetic control or measurement resolution.

To further quantify trait distribution characteristics, we calculated skewness and kurtosis values (Table 2). Most traits showed moderate right skew, apart from extreme values for biomass ratios. RFW/RDW exhibited the highest skewness (7.66) and kurtosis (67.01), indicating a highly asymmetric and leptokurtic distribution. SFW/SDW and RWC also displayed substantial deviations from normality (kurtosis > 12). In contrast, leaf traits such as LFW and LTW were more normally distributed, with near-zero skewness and mesokurtic kurtosis values. These deviations from normality support the use of nonlinear models, such as XGBoost, which can capture non-additive genetic effects and handle skewed phenotypic traits more effectively than linear models.

3.3 Trait Relationships and Correlations

Correlation analyses revealed biologically intuitive relationships within trait categories. Root traits were moderately to strongly correlated with one another, particularly RFW and RDW (r = 0.89), as shown in the root-specific pairplot (Supplementary Figure S1A). Shoot traits (SFW, SDW, SFW/SDW) formed a tight cluster (Supplementary Figure S1B), and leaf traits (LFW, LTW, LDW, RWC) displayed strong interdependence (Supplementary Figure S1C). The full correlation heatmap (Figure 4) confirmed these intra-group associations and revealed a few moderate cross-category relationships, such as between SFW and LFW or RFW and RWC.

Table 1. Summary statistics of phenotypic traits in the ExF RIL population. Includes mean, standard deviation, min, max, and coefficient of variation (CV) for each of the 15 measured traits.

	MRL	LRN	BRT	PH	RFW	RDW	SFW	SDW
Count	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00
Mean	31.87	31.78	0.82	105.66	6.02	3.80	37.93	25.20
Std	8.77	11.55	0.32	19.44	2.30	1.36	17.56	12.75
Min	3.86	10.00	0.04	61.40	1.09	0.90	1.86	1.29
25%	31.20	28.19	0.82	104.25	6.02	3.54	28.77	19.11
50%	31.87	31.78	0.82	105.66	6.02	3.80	37.93	25.20
75%	31.87	31.78	0.82	105.66	6.02	3.80	37.93	25.20
Max	62.00	77.00	2.10	174.70	17.75	9.60	128.38	76.71

	SW	RFW/RDW	SFW/SDW	LFW	LTW	LDW	RWC
Count	94.00	94.00	94.00	94.00	94.00	94.00	94.00
Mean	4.33	1.65	1.71	0.42	0.56	0.17	0.64
Std	2.04	0.61	0.51	0.10	0.12	0.04	0.08
Min	0.10	1.05	1.05	0.19	0.30	0.09	0.37
25%	3.63	1.53	1.50	0.42	0.56	0.17	0.64
50%	4.33	1.65	1.71	0.42	0.56	0.17	0.64
75%	4.33	1.65	1.71	0.42	0.56	0.17	0.64
Max	11.77	7.02	4.32	0.71	0.89	0.31	1.07

Plant Genetics, Genomics, & Biotechnology - ISSN 2332-2012. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

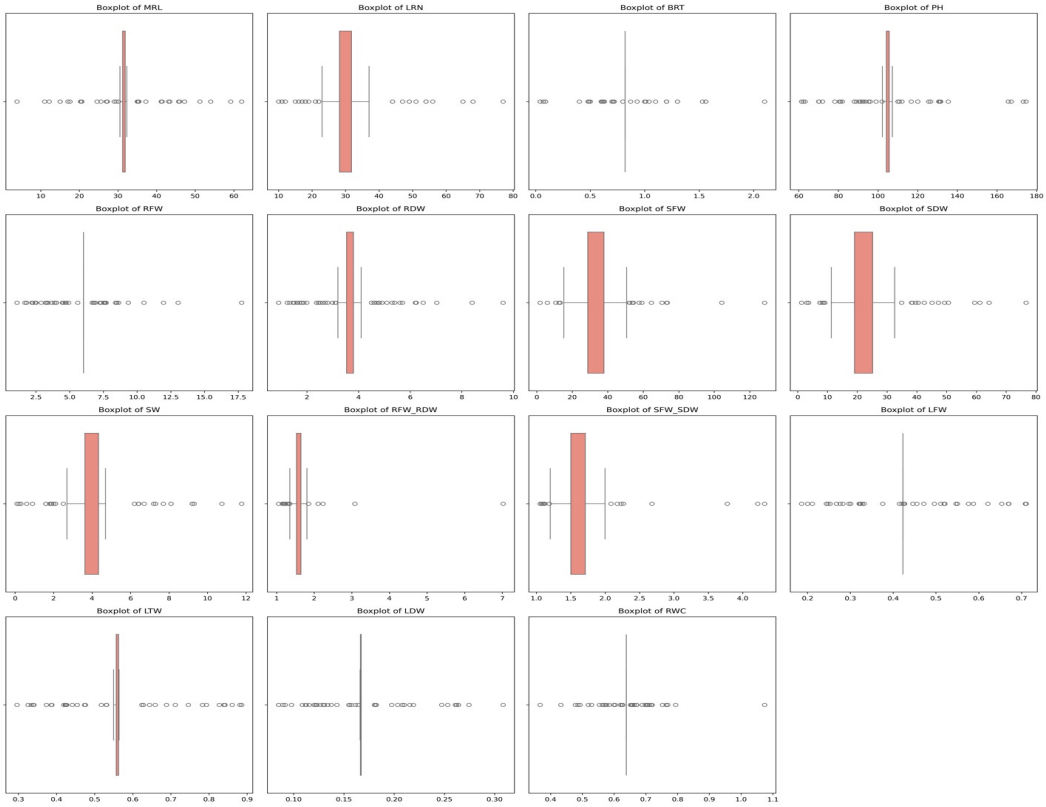

Figure 1. Boxplots of eleven root, shoot, and leaf traits measured in the Essex × Forrest (ExF) recombinant inbred line (RIL) population. Traits include maximum root length (MRL), lateral root number (LRN), basal root thickness (BRT), plant height (PH), root fresh weight (RFW), root dry weight (RDW), shoot fresh weight (SFW), shoot dry weight (SDW), shoot weight (SW), the ratios of root fresh weight to shoot fresh weight (RFW/SFW) and root dry weight to shoot dry weight (RDW/SDW), leaf fresh weight (LFW), leaf dry weight (LDW), and relative water content (RWC). Outliers are indicated as individual points beyond the whiskers.

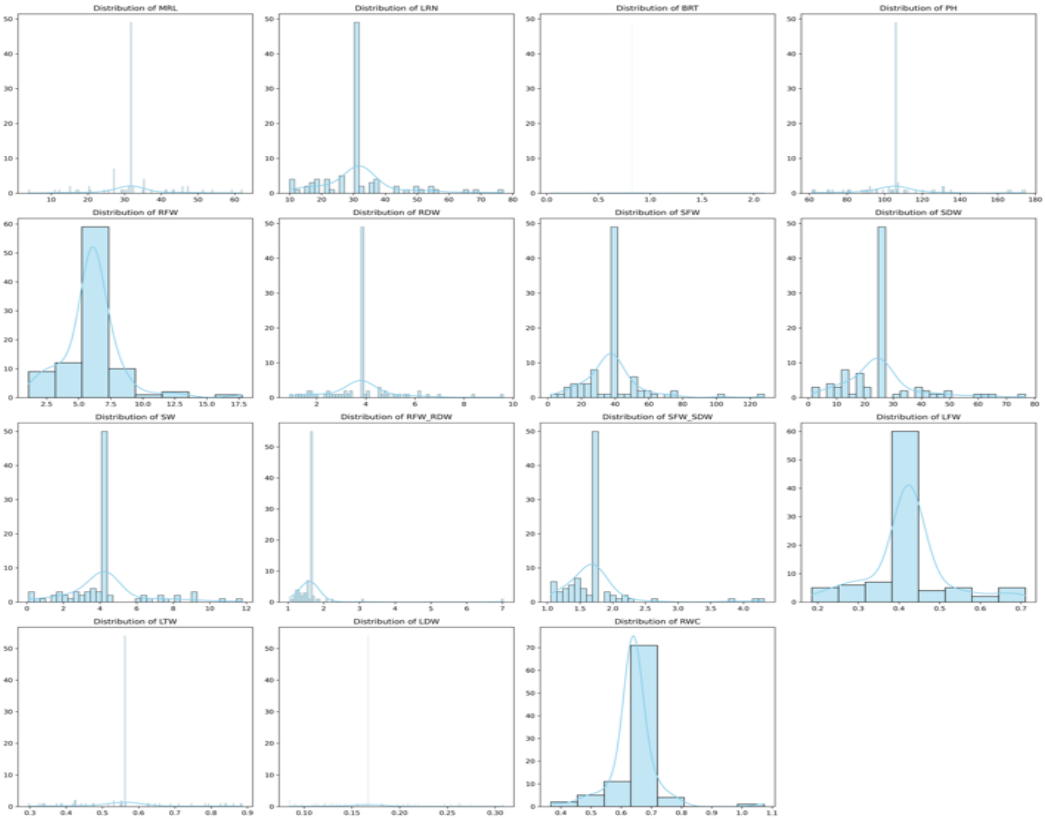

Figure 2. Distribution histograms and density plots of eleven root, shoot, and leaf traits in the Essex × Forrest (ExF) RIL population. Histograms and kernel density estimates (KDE) illustrate the distribution shape and normality for each trait. Traits exhibit varying degrees of symmetry and skewness.

Table 2. Skewness and Kurtosis of Phenotypic Traits in the Essex × Forrest RIL Population.

	Skewness	Kurtosis
MRL	0.38	3.22
LRN	1.13	3.20
BRT	0.21	3.27
PH	1.10	4.20
RFW	1.51	7.46
RDW	1.12	4.21
SFW	2.02	8.68
SDW	1.41	3.54
SW	1.07	2.76
RFW/RDW	7.66	67.01
SFW/SDW	3.34	14.44
LFW	0.50	1.89
LTW	0.63	1.74
LDW	0.88	2.67
RWC	1.21	12.41

Table 3. Summary of common QTL identified across Williams et al. (2012), Salvador et al. (2012), and Knizia et al. (2025, this study). The table lists markers previously associated with root, shoot, and leaf traits, along with their corresponding chromosomes and overlapping trait associations identified in the current study using machine learning (ML) and SHAP-based interpretation.

Williams et al. (2012) Trait QTL	Marker	Chr.	Common QTL with this study (Knizia et al., 2025)
SFW	Satt368	1	SFW/SDW
SFW	Satt267	1	SFW/SDW, LTW
SFW	Satt485	3	SFW
SFW	Satt239	6	SFW/SDW
SFW	Satt177	8	RFW, SW
SFW	Satt424	8	SW
Salvador et al. (2012) Trait QTL	**Marker**	**Chr.**	**Common QTL with this study (Knizia et al., 2025)**
LDW	Satt399	4	BRT, MRL, LDW, SFW, SDW
LFW	Satt358	10	LFW, SFW, RWC, SDW, RFW
LFW	Satt132	10	LRN, LFW, RWC, RFW/RDW
LDW	Satt324	18	LTW, MRL, RFW/RDW
LDW	Satt356	18	LDW
LDW	Satt122	18	SFW/SDW

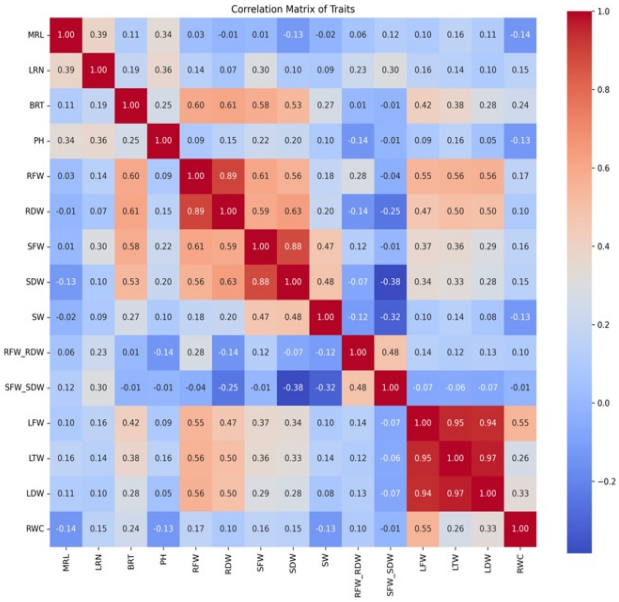

Figure 3. Trait correlation matrix. Color-scaled heatmap of Pearson correlation coefficients among all 15 traits. Positive and negative associations are indicated in red and blue, respectively.

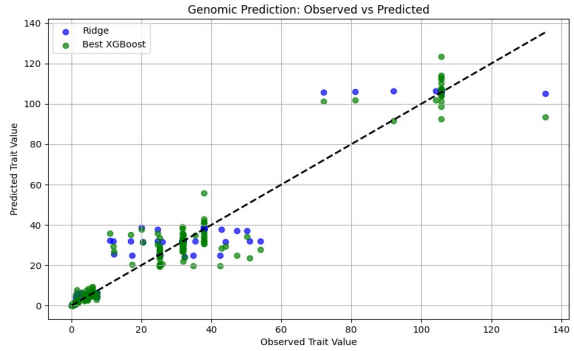

Figure 4. Genomic prediction: Observed vs. predicted trait values for Ridge Regression and XGBoost. XGBoost models outperformed Ridge in nearly all traits, as shown by tighter clustering around the 1:1 dashed line.

3.4 Genomic Prediction Performance

Genomic prediction models were trained using Ridge Regression and XGBoost. Ridge models achieved moderate accuracy across traits (mean $R^2 \approx 0.45$), while XGBoost consistently outperformed Ridge, with R^2 values exceeding 0.70 for traits like SFW and RFW. Figure 5 illustrates observed vs. predicted values, with XGBoost showing tighter clustering along the 1:1 diagonal, indicating superior fit and capacity to capture nonlinear and interaction effects.

3.5 Trait-Specific Marker Importance

Using feature importance scores from XGBoost, we identified the top 15 contributing markers per trait (Figure 6). A full list of the top 50 most important markers for each trait is provided in Supplementary Tables S1–S15. These tables offer a detailed view of the ranked marker contributions, allowing downstream users to prioritize candidate loci for validation or marker-assisted selection. Markers such as Satt153, Sat_262, Satt107, and Sat_299 emerged as highly informative across multiple traits, consistent with previously reported pleiotropic QTL on chromosomes 8 and 18 LGs G (Williams et al., 2012; Salvador et al., 2012). The presence of shared markers between shoot and root biomass traits supports the existence of genomic hotspots influencing general biomass allocation.

3.6 Interpretable SHAP Marker Contributions

To complement feature importance scores, we applied SHAP to quantify each marker's directional effect on trait prediction (Figure 7). SHAP beeswarm plots revealed both positive and negative effects of markers on trait values. For example, markers such as Sat_320 and DABF-6a-450 contributed positively to root biomass, while others like Sat_239 negatively influenced shoot ratios. SHAP interpretation enables identification of markers that not only contribute strongly to prediction but also affect trait expression directionally crucial for downstream functional validation.

3.7 Comparison to Previous QTL Mapping Results

To assess the robustness and biological relevance of the markers identified in this study, we compared our machine learning-derived results—based on SHAP and XGBoost feature importance—with previously published QTL identified through composite interval mapping (CIM) in the same ExF population (Williams et al., 2012; Salvador et al., 2012). Notably, several top-ranked markers from our analysis co-localized with QTL reported by Williams et al. (2012) and Salvador et al. (2012), highlighting strong concordance between traditional and ML-based approaches. For example, markers on chromosome 1 such as Satt267 and Sat368, which were highly predictive of leaf turgid weight (LTW) and SFW/SDW in this study, had also been associated with biomass traits (SFW) in earlier CIM study (Williams et al., 2012). Similarly, Satt399 on chr. 4 showed strong associations with BRT, MRL, LDW, SFW, and SDW, echoing results reported by Salvador et al. (2012). These overlaps reinforce the validity of the ML approach in detecting biologically meaningful loci, including some that may have moderate effects or pleiotropic influence across traits. Additional informative markers, listed in Supplementary Tables S1–S15, represent novel associations not previously detected using CIM, further expanding the landscape of trait-associated genomic regions.

To investigate this overlap more systematically, we compiled a summary of common QTL across studies (Table 3), focusing on shared markers and trait associations. Several markers identified by Williams et al. (2012) for shoot fresh weight (SFW)—including Satt177 and Satt424 on chromosome 8—also emerged in our study as important for RFW and SN, suggesting potential pleiotropic effects or tight linkage. Markers such as Satt239 on chromosome 6, initially associated with SFW (Williams et al., 2012), is among the most influ-

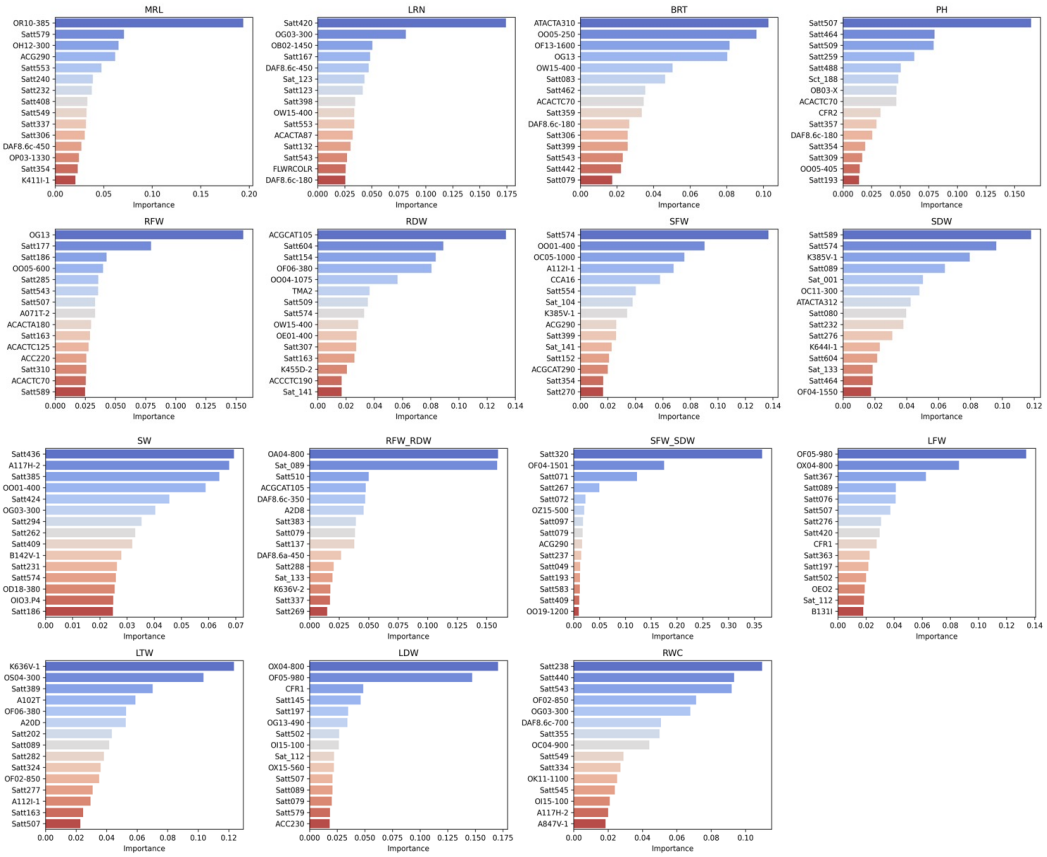

Figure 5. Top 15 most important markers for each trait based on XGBoost feature importance scores. Each subplot displays marker ranking and gain values. Markers like Satt153 and Sat_299 show high importance across multiple traits, indicating potential pleiotropic effects.

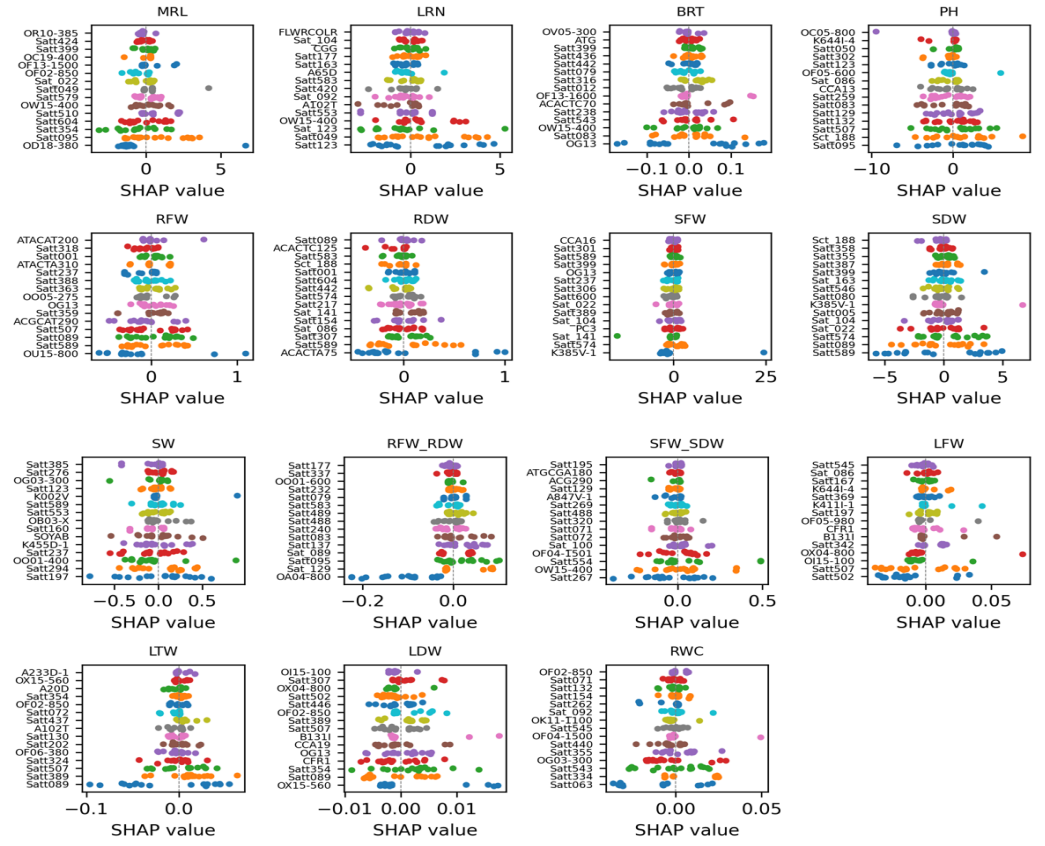

Figure 6. SHAP beeswarm plots showing the contribution of markers to model predictions. Each subplot visualizes SHAP values for top features, where color indicates the marker value (high = red, low = blue) and position along the x-axis indicates SHAP impact. Helps disentangle positive vs. negative marker effects.

ential feature for SFW/SDW in the current analysis. These findings highlight genomic regions of consistent importance across phenotyping methods and analytical frameworks.

Further agreement was observed with the findings of Salvador et al. (2012), particularly for leaf-related and water-use efficiency traits. On chromosome 10, Satt358 and Satt132 were strongly associated with LFW (Salvador et al., 2012) and were additionally linked to root-related and shoot-related traits (LFW, SFW, SDW, RWC, RFW, LRN) and biomass ratio (RFW/RDW) in this study. On chromosome 18, markers including Satt324, Satt356, and Satt122 were consistently implicated in multiple traits such as MRL, LTW, and biomass partitioning indices (RFW/RDW, SFW/SDW). This region appears to be a genomic hotspot for both structural and physiological traits related to drought response. The convergence of these findings across independent studies and analytical approaches underscores the stability of these QTL and their value for marker-assisted selection aimed at improving early vigor and drought resilience in soybean.

4. Discussion

This study applied ML-based genomic prediction and interpretable model analysis to dissect the genetic architecture of biomass and water-use traits in the well-characterized Essex × Forrest (ExF) soybean RIL population (Kassem, 2021). Building upon previous work using composite interval mapping (CIM) (Williams et al., 2012; Salvador et al., 2012), we integrated XGBoost modeling, SHAP analysis, and trait correlations to provide a multilayered view of trait architecture.

Our findings reaffirm the quantitative nature of traits such as root fresh weight (RFW), shoot dry weight (SDW), and relative water content (RWC), as evidenced by their continuous phenotypic distributions and modest to high levels of heritable variation. Consistent with earlier reports, traits within the same biological category–such as shoot and leaf biomass–were tightly correlated ($r > 0.80$), indicating potential pleiotropy or closely linked QTL.

Genomic prediction results demonstrated that nonlinear models like XGBoost significantly outperformed linear approaches such as Ridge regression. This is consistent with prior reports in soybean and other crops that highlight the advantages of ensemble-based tree models for complex trait prediction (Crossa et al., 2017; Montesinos-López et al., 2018; Kassem, 2025a,b). Traits with more pronounced heritability and less skew–such as SFW and LFW–were predicted with higher accuracy, while traits with skewed distributions or complex physiological determinants (e.g., BRT, RWC) were predicted with more variability.

XGBoost-derived feature importance metrics revealed markers of high predictive utility across multiple traits, with several overlapping previously mapped QTL regions. For instance, markers including Satt324, Satt356, and Satt122 were consistently implicated in multiple traits such as MRL, LTW, and biomass partitioning indices (RFW/RDW, SFW/SDW) (Williams et al., 2012). The high importance of these markers in ML models reinforces their central role in biomass accumulation and partitioning.

Importantly, SHAP analysis extended beyond feature importance to offer insight into directional marker effects. This interpretability is crucial for breeders aiming to select for markers that not only predict performance but also modulate traits in a favorable direction. SHAP plots showed that many markers contribute positively to trait values, such as those associated with higher RFW or SDW, while others had consistently negative impacts–information that is often lost in traditional QTL mapping outputs.

Our integrative approach also revealed strong agreement between markers identified through ML pipelines and those previously linked to QTL using CIM. For example, RWC and LDW shared high-importance markers in LGs E (chr. 15) and K (chr. 9), as previously identified by Salvador et al. (2012). This reinforces the biological validity of the ML-based methods and underscores their value in QTL re-discovery and refinement.

Together, these results highlight the potential of combining modern ML methods with explainable AI tools like SHAP to not only enhance prediction accuracy but also provide trait-level and marker-level biological insight. This framework is applicable beyond the current population and traits, offering a template for genomic dissection in other complex plant systems.

5. Conclusion

Through the integration of ML-based genomic prediction and SHAP interpretation, we refined the genetic architecture of complex biomass and water-use traits in soybean. XGBoost significantly outperformed linear models, capturing nonlinear interactions and epistatic effects. SHAP values enabled trait-specific, directional ranking of marker contributions, enhancing biological interpretability and marker prioritization. Concordance with previous CIM-based QTL mapping affirms the robustness of our approach, while newly identified loci expand the landscape of candidate regions. This pipeline presents a scalable strategy for genomic dissection across diverse plant populations and trait categories.

Data Availability Statement

Data and scripts are available upon reasonable request from the corresponding author.

Author Contributions

MAK conceptualized the study, led data integration, and supervised manuscript development. DK and KM contributed to writing and critical revision of the manuscript. All authors participated in data interpretation and approved the final version.

Funding Statement

This research was conducted using data originally generated through projects supported by Fayetteville State University (FSU). The foundational work, including phenotyping and QTL mapping reported in Williams et al. (2012) and Salvador et al. (2012), was made possible by institutional support from FSU. The current study, which builds on and re-analyzes this dataset using advanced ML approaches, received no specific grant from any funding agency.

Acknowledgments

The authors would like to thank Fayetteville State University (FSU) and Southern Illinois University Carbondale (SIUC) for supporting this project through laboratory resources and research infrastructure. We also acknowledge the long-term collaboration and data generation efforts that have made this study possible, including prior QTL mapping and phenotyping work in the Essex × Forrest RIL population.

Conflict of Interest Statement

The authors declare no conflict of interest.

Supplementary Material

Supplementary Figures (S1A-C) and Tables (S1–S15) are available upon reasonable request from the corresponding author.

References

Alcivar A, J Jacobson, J Rainho, K Meksem, DA Lightfoot, and MA Kassem (2007). Genetic Analysis of Soybean Plant Height, Hypocotyl and Internode Lengths. Journal of Agricultural, Food, and Environmental Sciences 1 (1): 1-20.

Chen T and C Guestrin (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; pp. 785-794. https://doi.org/10.1145/2939672.2939785.

Cho Y, Njiti VN, Chen X, Triwatayakorn K, Kassem MA, Meksem K, Lightfoot DA, Wood AJ (2002). Quantitative Trait Loci Associated with Foliar Trigonelline Accumulation in Glycine Max L. J. Biomed & Biotech 2(3): 151-157.

Crossa J, P Pérez-Rodríguez, J Cuevas, O Montesinos-López, D Jarquín, G de Los Campos, J Burgueño, JM González-Camacho, S Pérez-Elizalde, Y Beyene, S Dreisigacker, R Singh, X Zhang, M Gowda, M Roorkiwal, J Rutkoski, and RK Varshney (2017). Genomic selection in plant breeding: Methods, models, and perspectives. Trends in Plant Science, 22(11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011.

Grant D, RT Nelson, SB Cannon, and RC Shoemaker. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Research 2010, 38(Database issue): D843-6. https://doi.org/10.1093/nar/gkp798.

Harris CR, Millman KJ, van der Walt SJ, et al. (2020). Array programming with NumPy. Nature. 2020;585(7825):357-362. https://doi.org/10.1038/s41586-020-2649-2.

Hunter JD (2007). Matplotlib: A 2D graphics environment. Computing in Sciences and Engineering, 2007; 9(3):90-95. https://doi.org/10.1109/MCSE.2007.55.

Iqbal MJ, K Meksem, VN Njiti, MA Kassem, and DA Lightfoot (2001). Microsatellite markers identify three additional quantitative trait loci for resistance to soybean sudden death syndrome (SDS) in Essex x Forrest RILs. Theor Appl Genet 102(2/3): 187-192.

Ivey S, K Ouertani, E Washington, P Lage, SK Kantartzi, K Meksem, DA Lightfoot, and MA Kassem. Evaluation of Agronomic Traits in 'Essex' By 'Forrest' Recombinant Inbred Line Population of Soybean [Glycine max [L.] Merr.]. Atlas Journal of Plant Biology 1 (1): 13-17, 2011.

Jacobson J, A Alcivar, J Rainho, and MA Kassem (2007). Genomic Regions Containing QTL for Plant Height, Internodes Length, and Flower Color in Soybean [Glycine max (L.) Merr.]. BIOS A Quarterly Journal of Biology 78 (4): 119-126.

Karangula UB, MA Kassem, L Gupta, and DA Lightfoot (2009). Locus Interactions Underlie Seed Yield in Soybeans Resistant to Heterodera glycines. Current Issue in Mol. Biol. 11 (Suppl. 1): i73–84.

Kassem MA (2021). Soybean Seed Composition: Protein, Oil, Fatty Acids, Amino Acids, Sugars, Mineral Nutrients, Tocopherols, and Isoflavones. Springer Nature. https://doi.org/10.1007/978-3-030-82906-3.

Kassem MA (2025a). Harnessing Artificial Intelligence and Machine Learning for Identifying Quantitative Trait Loci (QTL) Associated with Seed Quality Traits in Crops. Plants 2025, 14, 1727. https://doi.org/10.3390/plants14111727.

Kassem MA (2025b). QTL Mapping of Seed Quality Traits in Crops. Plants 2025; 14(3): 482. https://doi.org/10.3390/plants14030482.

Kassem MA, K Meksem, AJ Wood, and DA Lightfoot (2007a). Loci underlying SDS and SCN resistances mapped in the 'Essex' by 'Forrest' soybean recombinant inbred lines. Reviews in Biology and Biotechnology 6 (1): 2–10.

Kassem MA, K Meksem, AJ Wood, and DA Lightfoot (2007b). A Microsatellite Map Developed from Late Maturity Germplasm 'Essex' by 'Forrest' Detects Four QTL for Soybean Seed Yield Expected from Early Maturing Germplasm. Reviews in Biology and Biotechnology 6 (1): 11-19.

Kassem MA, K Meksem, CH Kang, VN Njiti, VY Kilo, AJ Wood, and DA Lightfoot (2004a). Loci Underlying Resistance to Manganese Toxicity Mapped in a Soybean Recombinant Inbred Line population of 'Essex' x 'Forrest'. Plant and Soil. 260 (1-2): 197-204.

Kassem MA, Meksem K, Iqbal MJ, Njiti VN, Banz WJ, Winters TA, Wood A, and Lightfoot DA (2004b). Definition of Soybean Genomic Regions That Control Seed Phytoestrogen Amounts. J. Biomed and Biotech 1(1): 52-60.

Kassem MA, Shultz J, Meksem K, Cho Y, Wood AJ, Iqbal MJ, and Lightfoot DA (2006). An updated 'Essex' by 'Forrest' linkage map and first composite map of QTL underlying six soybean traits. Theoretical and Applied Genetics 113: 1015–1026.

Kluyver T, B Ragan-Kelley, F Pérez, B Granger, M Bussonier, J Frederic, K Kelly, J Hamrick, J Grout, S Corlay, P Ivanov, D Avila, S Abdalla, and C Willing (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, 2016:87-90. https://doi.org/10.3233/978-1-61499-649-1-87.

Lightfoot DA, VN Njiti, PT Gibson, MA Kassem, JM Iqbal, and K Meksem. Registration of the Essex by Forrest Recombinant Inbred Line (RIL) Mapping Population. Crop Science 45 (4): 1678-1781, 2005. https://doi.org/10.2135/cropsci2004.0279.

Lundberg SM and SI Lee (2017). A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

McKinney W (2010). Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. Austin, TX; 2010:51-56. https://doi.org/10.25080/Majora-92bf1922-00a.

Meksem K, VN Njiti, WJ Banz, MJ Iqbal, MA Kassem, DL Hyten, J Yuang, TA Winters, and DA Lightfoot. Genomic regions that underlie soybean seed isoflavone content. J. Biomed. & Biotech. 1(1): 38-44, 2001.

Meuwissen THE, BJ Hayes, and ME Goddard (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Montesinos-López OA, A Montesinos-López, J Crossa, G de los Campos, G Alvarado, M Suchismita, and S Mondal (2018). Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. The Plant Genome, 11(1), 1–15. https://doi.org/10.3835/plantgenome2017.08.0075.

Njiti VN, Meksem K, Iqbal MJ, JE Johnson, Kassem MA, KF Zobrist, VY Kilo, D A Lightfoot. Common Loci Underlie Field Resistance to Soybean Sudden Death Syndrome in Forrest, Pyramid, Essex, and Douglas. Theor Appl Genet 104(2/3): 294-300, 2002.

Pedregosa F, G Varoquaux, A Gramfort, et al. (2011). Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.

Salvador V, Pagan P, Cooper M, Kantartzi SK, Lightfoot DA, Meksem K, and Kassem MA (2012). Genetic Analysis of Relative Water Content (RWC) in Two Recombinant Inbred Line Populations of Soybean [Glycine max (L.) Merr.]. Journal of Plant Genome Sciences 1 (2): 46–53. Https://doi.org/10.5147/jpgs.2012.0058.

Seabold S and J Perktold (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference.

Van Rossum G and FL Drake (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Wang S, CJ Basten, and ZB Zeng (2012) Windows QTL Cartographer 2.5, Department of Statistics, North Carolina State University, Raleigh, NC, USA. https://www.scirp.org/reference/referencespapers?referenceid=1995571].

Waskom ML (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 2021, 6(60):3021. https://doi.org/10.21105/joss.03021.

Williams B, Kantartzi SK, Meksem K, Grier RL, Barakat A, Lightfoot DA, and Kassem MA (2012). Genetic Analysis of Root and Shoot Traits in the 'Essex' By 'Forrest' Recombinant Inbred Line (RIL) Population of Soybean [Glycine max (L.) Merr.]. Journal of Plant Genome Sciences 1 (1): 1–9, 2012. https://doi.org/10.5147/jpgs.2012.0051.

Willmott CJ and Matsuura K (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30: 79-82. https://doi.org/10.3354/cr030079.

Yuan J, VN Njiti, K Meksem, MJ Iqbal, K Triwitayakorn, MA Kassem, GT Davis, ME Schmidt, DA Lightfoot. Quantitative trait loci in Two Soybean Recombinant Inbred Line Populations Segregating for Yield and Disease Resistance. Crop Science 42: 271-277, 2002.

Plant Genetics, Genomics, & Biotechnology - ISSN 2332-2012. Published By Atlas Publishing, LLC (www.atlas-publishing.org)