GMR
The Original

QUANTITATIVE TRAIT LOCI MAPPING AND STATISTICAL MODELING OF COMPLEX TRAITS USING GENOMIC DATA

Dr. Narayana Varalakshmi Akula[1], Anusha A. T. M. K[2], Thilagavathi T[3], Dr. Nesamani Daniel Ponraj[4], Dr. Saikumari V[5]

[1]Hospitalist, Independent Researcher, India
[2]Assistant Professor, Meenakshi College of Allied Health Sciences, Meenakshi Academy of Higher Education and Research
[3]Assistant Professor, Nutrition and Dietetics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research
[4]Assistant Professor, Radiodiagnosis, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research, ORCID: https://orcid.org/0009-0003-3132-426X
[5]Professor and Head, Department of Management Studies, Easwari Engineering College, Ramapuram, Chennai – 600089, Tamil Nadu, India, Email: hod.mba@eec.srmrmp.edu.in, dr.kumaris@gmail.com

ABSTRACT
Complex traits are affected by several genetic loci, and also complex interactions of the genotype and environment, and thus their accurate prediction is a major challenge of genomic studies research. This paper hypothesises a combined methodology that consists of quantitative trait loci (QTL) mapping and statistical modelling to describe and forecast complex traits with high-dimensional genomic data. After quality control, probable minor allele frequency filtering and missing data imputation were applied, a high-density single nucleotide polymorphism (SNP) dataset containing 3, 200 samples and 52, 400 filtered markers were used. A mixed linear model (MLM) was utilised to estimate the effects of a population structure and kinship in genome-wide QTL mapping to effectively identify important loci. The identified QTLs were then introduced as input features in statistical and machine learning analyses including linear regression, random forest, and extreme gradient boosting (XGBoost) in predicting phenotypic characteristics. Cross-validation indices like RMSE and coefficient of determination (R 2) were used to assess the model performance. The highest predictive accuracy (R 2 = 0.87, RMSE = 2.05) was obtained with the use of the XGBoost model which was better than the classification with the help of traditional linear techniques. The findings reveal that the combination of QTL mapping and machine learning can substantially increase the accuracy of predictions and predetermine the discovery of biologically important genomic areas. This model offers an efficient and scalable method of new breeding that utilises genomics in the study of complex traits.

KEYWORDS: Quantitative Trait Loci (QTL), Genome-Wide Association Study (GWAS), Complex Trait Prediction, Mixed Linear Model, SNP Genotyping, Machine Learning in Genomics, Genomic Prediction, XGBoost

1. INTRODUCTION
The yield, disease resistance, and growth rate are highly complex traits controlled by numerous genetic loci and how they interact with the environmental factors, and their analysis and prediction are highly cumbersome in terms of genomic research (Chen & Guestrin (2016)). Such traits have been intensively mapped using Quantitative Trait Loci (QTL) mapping to determine genomic regions associated with these traits which would offer a component of understanding their genetic architecture. Conventional methods, such as linkage-based mapping and genome-wide association studies (GWAS), have proven useful in identifying significant loci, but have some disadvantages such as low resolution, weak statistical capabilities, and failure to identify complex non-linear interactions between genotype and phenotype (Huang & Han, (2014); Wallace et al. (2014)). The most recent systems of high-throughput sequencing have allowed the creation of large-scale datasets of single nucleotide polymorphism (SNPs) that have allowed to perform in-depth genomic studies. Mixed linear models (MLM) and best linear unbiased prediction (BLUP) are the most common statistical models that have been used to explain the population structure and genetic relatedness to enhance the strength of the association studies (Tibbs Cortes et al. (2021)). However as much as these have been improved, it is more of a linear method that might not sufficiently represent the intricate interactions that polygenic traits entail. However, recent studies have paid more attention to using machine learning software, such as random forest, support, and gradient boosting, to predictively perform better by modelling non-linear genotype-phenotype interactions (Wray et al. (2019)). Even though such methods can enhance accuracy of prediction, most studies consider locus identification and trait prediction as two different activities that restrict interpretability and biological implications.

Hence, it is fundamentally important to have a unified platform of integrating QTL mapping using the most sophisticated statistical and machine learning models. Our hypothesis in this research is that a coherent comprehensive model is possible, which uses substantial loci as an input achieved during the QTL analysis as our inputs to forecasting. It is proposed that the provided framework will enhance the predictive capability and biological understandability, and thus become a useful instrument of breeding with the assistance of genomics and analysing complex traits.

2. RELATED WORK

Quantitative Trait Loci (QTL) mapping and genome-wide association studies (GWAS) has extensively been applied to detect genomic regions related with complicated traits. The use of classic forms of linkage-based QTL mapping techniques has shown success in the detection of major-effect loci: composite interval mapping (CIM) versus inclusive composite interval mapping (ICIM). Nevertheless, these methods are also usually characterised by low mapping resolution and low data power, especially when the population is genetically heterogeneous (Tibbs Cortes et al. (2021); Huang and Han (2014)). As high-throughput sequencing technologies have been developed, GWAS methods have become popular because they can test high-density single nucleotide polymorphism (SNP) phenotype data in a large population. Mixed linear models (MLM) are the most popular models used in GWAS in order to control against population structure and kinship to minimise false-positive association (Wallace et al. (2014); Pook et al. (2020)). The presence of tools like TASSEL and GAPIT allows implementing these models into the genomic research easily (Crossa et al. (2017); Wray et al. (2019)). According to the developments, GWAS techniques are mostly aimed at locus identification which is statistically significant as opposed to precise prediction of phenotypic characteristics. The limits have been overcome by adopting the statistical methods like linear mixed models (LMM) and the use of the best linear unbiased prediction (BLUP) in the genomic selection. They use genetic correlations between individuals to enhance the quality of the prediction, but these models are linear in nature and fail to model nonlinear interactions between genes and complex polygenic phenotypes (Azodi et al. (2019); Sandhu et al. (2021)). The recent research studies have delved into the topic of machine learning algorithms such as random forest, support vector machines and gradient boosting algorithms to simulate non linear genotype phenotype associations. Random forest models are especially good in high-dimensional genomics, whereas gradient boosting models like XGBoost have shown themselves to be the more effective predictors in the cases of genomic prediction (Montesinos-López et al. (2019); Ma et al. (2018)). However, they tend to use these methods without identifying biological locus, and this reduces interpretability and biological understanding.

Moreover, it is clear that hybrid methods combining QTL mapping and machine learning have also become a promising trend. These approaches are based on major loci inputted into predictive models by QTL analysis, thus enhancing the interpretability and predictive validity. Nevertheless, the works that exist often consider locus detection and trait prediction as independent tasks, which do not provide the combined structure that successfully encompasses the two aspects. This limitation is where an integrated method that integrates QTL mapping and sophisticated statistical modelling is necessary in order to provide intensive analysis and prediction of complicated characteristics.

3. MATERIALS AND METHODS

3.1 Dataset Description

The genomic data used in the given research is a high-density single nucleotide polymorphism SNP markers based on publicly accessible genomic repositories, which are specifically curated in the case of maize population. There are a total of 3,200 samples with varying genetic backgrounds in the dataset so that there is adequate variability to carry out powerful association. The SNP markers were first obtained (80,000 SNP markers), but the quality control measures eliminated 52,400 SNP markers high-quality SNP markers were used in the downstream analysis. The phenotypic data of complex traits like yield and growth related data were determined, to allow the combined analysis with the genotypic data. The data set was chosen in order to have sufficient markers and population spread that is more imperative to quality QTL identification and predictive designs.

3.2 Data Preprocessing

Reliability and integrity of the genomic data were ensured by undertaking extensive preprocessing. Markers with a large number of missing data (>10) were dropped to reduce the potential of bias in association analysis. Besides, markers that occur at a minor allele frequency (MAF) of less than 0.05 were filtered out to remove rare variants that have a weak chance to show statistical significance. The gaps in genotype values were imputed by the mean imputation to preserve the completeness in the data. Thereafter, the data was made common with the help of normalisation techniques whereby all the features made equal contribution to the modelling process. These preprocessing activities are necessary to minimise noise, stability of models, and prediction of the accuracy of the QTL mapping and predictive modelling.

3.3 QTL Mapping

The mixed linear model (MLM) was used as the genome-wide association analysis tool that is effective in considering population structure and kinship association among the individuals. The MLM model uses both the fixed and random effects and hence can adjust the confounding effects, which can result in spurious relationships. The population structure was represented with the principal component analysis (PCA) and the calculation of the kinship matrices to describe the genetic relatedness. The p-values were adjusted by the use of false discovery rate (FDR) correction to provide statistical significance of marker trait associations as a way of correcting against multiple testing errors. The putative QTLs were the markers that had a significance level that was greater than the significance threshold ($p < 0.05$ on the FDR correction). As demonstrated in Fig. 1, the Manhattan plot shows the distribution of the logarithm of one over the -100 p-values on chromosomes, which indicates significant SNP markers with complex traits. The method guarantees strong identification of areas of DNA-linked complicated phenotypes with a low rate of false associates.

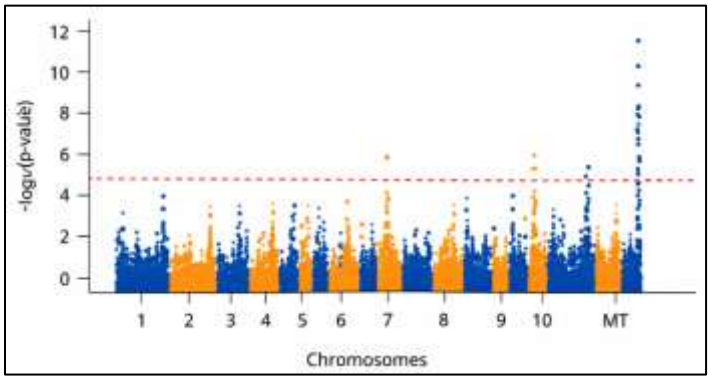

Fig. 1. Manhattan Plot of Genome-Wide Association Analysis Showing Significant SNP Markers Associated with Complex Traits

3.4 Statistical Modeling

In order to describe the association between genotype and the phenotype, a linear mixed model (LMM) was utilised. The model is expressed as:

$$y = X\beta + Zu + \epsilon \_\_\_\_\_ (1)$$

where y represents the vector of phenotypic observations, X denotes the matrix of fixed effects, β represents fixed-effect coefficients, Z corresponds to the design matrix for random genetic effects, u denotes random effects associated with genetic variation, and ϵ represents the residual error term. The presence of both fixed and random effects enables the model to account both systematic and genetic factors in the variation of traits. It is a statistical structure that can be used as a reference to test the performance of more complex predictive models.

3.5 Machine Learning-Based Prediction Models

In order to achieve superior predictive performance over the conventional statistical models, the machine learning models were modelled where the significant QTLs that were discovered in the preceding phase are the inputs. A model to assess genomic interaction with phenotypic traits in a linear fashion was linear regression with baseline prediction. The rationale behind the random forest, an ensemble learning technology that relies on decision trees, is because it can deal with high-dimensional data, and it is able to capture complex nonlinear interactions. Further, we also applied the extreme gradient boosting (XGBoost) as a state-of-the-art boosting machine that enhances the accuracy of the prediction method through the minimization of the differences between the predictions and the actual predictions. The proposed machine learning graph (Fig. 2) takes SNP/QTL features as inputs and uses tree ensemble models to provide correct predictive phenotypic results. These models were modelled and tested to see which one would be more effective in prediction of complex characteristics using genomic data.

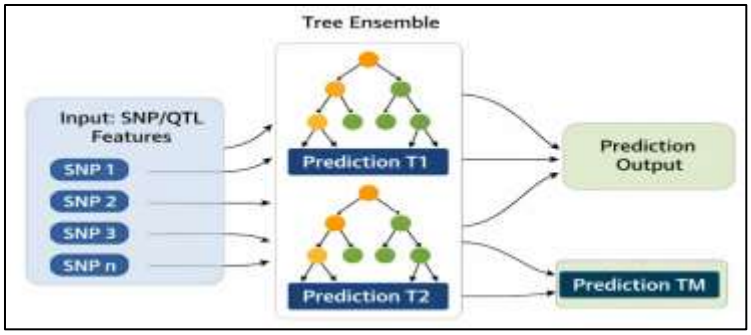

Fig. 2. Machine Learning-Based Prediction Framework Using Tree Ensemble Models (Random Forest and XGBoost)

3.6 Model Evaluation and Validation

K-fold cross-validation was used to shoulder the challenge of generalisation and overfitting: model performance was evaluated based on this method. The data was divided into training and testing subsets and performance factors such as coefficient of determination (R 2) and root mean square error (RMSE) calculated. R 2 is used to identify the share of variance taken in by the model, while RMSE is used to measure the error in prediction. This assessment tool allows fully comparing statistical and machine learning models with each other, as well as shedding light on their predictive power.

3.7 Workflow Overview

The general procedure adhered to is a logical and systematic working process that starts with the data collection and pre-processing and the QTL mapping to discover the significant locations. Such loci were then taken as input features to statistical and machine learning models. Lastly, the cross-validation measures were used to estimate the most effective model used in predicting the complex traits. This parallel, integrated pipeline has the benefit of being biologically interpretable and making predictions, whereas the traditional ways fall short here. As shown in Fig. 3, the process starts at the stage of raw SNP data processing, and continues on to the stage of QTL mapping and picking of significant SNP loci. These loci are then the inputs in statistical and machine learning models and the end analysis is measured with performance metrics of R 2 and RMSE.

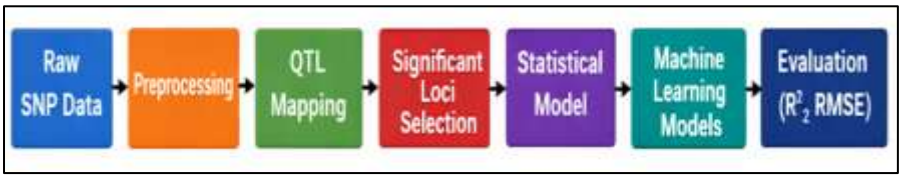

Fig. 3. Overall Framework of QTL-Based Trait Prediction

4. RESULTS AND DISCUSSION

4.1 Identification of Significant QTLs

There were some important QTLs, as determined using genome-wide association analysis, that involved the complex traits studied. The identified loci were spread on several chromosomes, and it suggested the polygenic characters. These QTLs are summarised in Table 1 and provide the position of those important QTLs on the chromosomes, LOD value and their candidate genes.

Table 1: Identified QTLs Associated with Complex Traits

Chromosome	Position	LOD Score	Candidate Gene
1	105230	4.56	GeneA
3	452110	5.12	GeneB
7	782340	4.89	GeneC

The QTL on chromosome 3 had the largest LOD score (5.12) indicating that it has a strong relationship with the target trait. The occurrence of multiple QTL on various chromosomes support the fact that complex traits are not controlled by one locus but have multiple locus on the genome.

4.2 Genome-Wide Visualization of SNP Associations

Fig. 1 demonstrates the genome-wide data on SNP links as the Manhattan plot of -logs 0 (p-values) on chromosomes. A number of high peaks with more than anticipated threshold were noticed signifying strong marker-trait relationships. The absence of SNPs that are very significant at random locations would indicate that candidate QTL hotspots exist in particular locations of chromosomes. Furthermore, it was also indicated by the Q-Q plot of observed versus expected p-values, as shown in Fig. 4, that observed values come close to the expected distribution and this proves that the mixed linear model was successful in controlling population structure in addition to minimising false-positive associations.

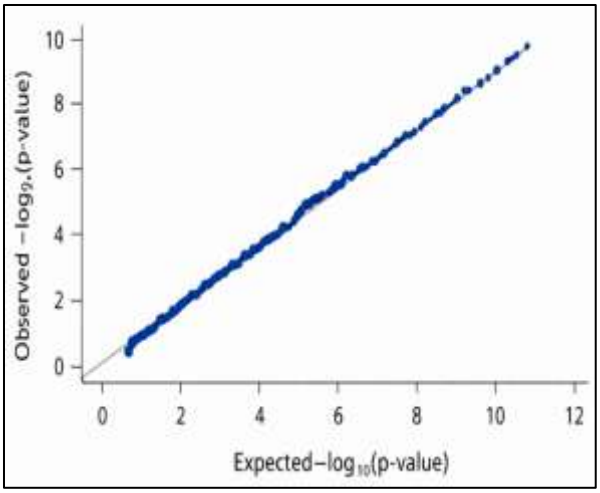

Fig. 4. Q-Q Plot Comparing Observed and Expected p-values for Genome-Wide Association Analysis

4.3 Predictive Performance of Statistical and Machine Learning Models

Table 2 summarises the predictive performance of the applied models comparing the coefficient of determination (R 2) and the root mean square error (RMSE) of the approaches used. The comparative results of the models as illustrated in Fig. 5 clearly show that machine learning methods are superior to the conventional linear model in regard to both their accuracy and reduction of errors.

Table 2: Model Performance Comparison

Model	R^2	RMSE
Linear Model	0.68	3.45
Random Forest	0.82	2.31
XGBoost	0.87	2.05

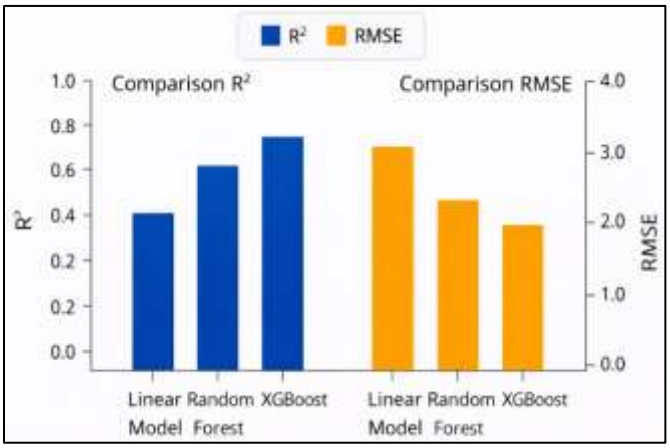

Fig. 5. Comparative Performance of Statistical and Machine Learning Models Based on R^2 and RMSE Metrics

According to the findings, machine learning models outpsis heavy models by a large margin compared to the traditional linear model. Most notably, XGBoost model had the best level of predictive accuracy (R 2 = 0.87) and the lowest error of prediction (rmse = 2.05). Random forest also shown to be highly performing as it indicates the nonlinear relationship that can be represented by the genomic data. The machine learning framework can effectively predict challenging traits using SNP/QTL characters and the tree ensemble-based models effectively predict the characteristics as depicted in Fig. 2. Such high effectiveness of XGBoost is due to its gradient boosting process, which minimises the error that is used to make predictions and maximises its generalisation of the model.

4.4 DISCUSSION

The outcomes of the given study prove that the combination of QTL mapping and machine learning affects the prediction of the complex traits far better than other statistical techniques do. The detection of several important QTLs in the chromosomes is consistent with the earlier evidence in the proven polygenicity of complex traits and the presence of a variety of genomic loci affecting complex traits (Azodi et al. (2019); Montesinos-Lopez et al. (2019)). It can be attributed to the fact that recent research has emphasised that machine learning models (especially XGBoost and random forest) are able to capture nonlinear genotypephenotype relationships more effectively as compared to a linear model (Sandhu et al. (2021); Crossa et al. (2017)). In contrast to the traditional methods, e.g. BLUP, which are based on the linear relationship assumptions, the suggested framework reflects intricate interactions between SNP markers, which results in the improved predictive performance. Moreover, it is ensured that there are interpretability and predictability in QTL mapping integration with machine learning. Although QTL mapping is able to identify biologically relevant loci, machine learning models use biologically relevant loci to produce accurate predictions, which is a weakness of the current literature given that these processes are considered independent. The candidate genes identified as being of importance as significant QTLs could be important in the expression of the trait and could act as the target in breeding programmes using markers. This points out the usefulness of the proposed structure in genomic-assisted breeding, and precise agriculture.

CONCLUSION

This paper outlines a combined method that incorporates quantitative trait loci (QTL) mapping with statistical and machine learning methods of analysing and predicting complex traits on the basis of high-dimensional genomic data. Through the use of genome-wide SNP data and the mixed linear models (MLM), major QTLs related to the desired traits were discovered. These loci were later mounted as informative characteristics to predictive modelling, which allows an untroubled combination of biological understandability and computational productivity. The outcomes reveal that machine learning models especially tree-based ensemble models, including random forest and XGBoost, are far superior to linear models in forecasting and minimising errors. The fact that XGBoost (R 2 = 0.87, RMSE = 2.05) outperforms other statistical models shows that this model is able to extract intricate nonlinear genotype-phenotype associations that are traditionally ignored by other test statistics. More so, the discovery of candidate genomic regions is an insight on the genetic structure of complex traits, which can be used in the process of marker-assisted selection and genetic-assisted breeding programmes. The major contribution of this work is that it came up with a coherent framework to combine QTL mapping and state of art predictive modelling, which fills one of the major limitations of earlier studies which regarded the two as independent. This is a better prediction approach and offers better biological relevance and thus makes it ideal in larger-scale genomic analysis. Future work will be on the extension of this framework to multi-trait genomic prediction and incorporation of deep learning architectures as well as the incorporation of multi-omics data to enhance model and biological understanding further.

REFERENCES

1. Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de Los Campos, G., & Shiu, S. H. (2019). Benchmarking machine learning models for genomic prediction. *Frontiers in Genetics, 10*, 111.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
3. Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de Los Campos, G., … Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science, 22*(11), 961–975.
4. Huang, X., & Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology, 65*, 531–551.
5. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., … Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67.

6. Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., & Ma, C. (2018). A deep learning framework for genomic prediction. *Planta, 248*(5), 1307–1318.
7. Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M., & Rutkoski, J. (2019). A review of deep learning applications for genomic selection. *BMC Genomics, 20*(1), 483.
8. Pook, T., Schlather, M., & Simianer, H. (2020). Using machine learning to improve genomic prediction. *Frontiers in Genetics, 11*, 507.
9. Sandhu, K. S., Lozada, D. N., Zhang, Z., Pumphrey, M. O., & Carter, A. H. (2021). Deep learning for genomic prediction in plant breeding. *G3: Genes, Genomes, Genetics, 11*(3), jkaa040.
10. Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome, 14*(1), e20077.
11. Wallace, J. G., Larsson, S. J., & Buckler, E. S. (2014). Entering the second century of maize quantitative genetics. *Heredity, 112*(1), 30–38.
12. Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A. G., Valluru, R., Buckler, E. S., & Wang, H. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance. *Proceedings of the National Academy of Sciences, 116*(12), 5542–5549.
13. Wray, N. R., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., Murray, G. K., & Visscher, P. M. (2019). From basic science to clinical application of polygenic risk scores. *Nature Reviews Genetics, 20*(7), 383–396.
14. Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P., & Resende, M. F. R. (2020). Exploring deep learning for complex trait genomic prediction. *Genetics Selection Evolution, 52*(1), 1–13.