Institut für Nutzpflanzenwissenschaften und Ressourcenschutz (INRES)
Fachbereich Pflanzenzüchtung

Genome Wide Association Study and Genomic Selection Models for Nitrogen Use Efficiency in Bread Wheat

Dissertation
zur Erlangung des Grades
Doktor der Agrarwissenschaften (Dr. agr.)

der Landwirtschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Mohammad Bahman Sadeqi

aus

Daikundi, Afghanistan

Bonn 2024

Angefertigt mit Genehmigung der Landwirtschaftlichen Fakultät der Universität Bonn

Referent:	Prof. Dr. Jens Léon
Koreferent:	Dr. Thomas Gaiser
Fachnahes Mitglied:	Prof. Dr. Gabriel Schaaf
Vorsitzender:	Prof. Dr. Mathias Becker
Tag der mündlichen Prüfung:	16.04.2024

Acknowledgement

First and foremost, my utmost gratitude to my supervisor, Prof. Dr. Jens Léon for his encouragement, patience, motivation, enthusiasm, immense knowledge and continuous support throughout my PhD study. I consider it an honor to work with him and he has been my inspiration in the completion of this research work. I owe my deepest gratitude to Prof. Dr. Jens Léon for his excellent supervision. I would also like to special thanks Dr. Thomas Gaiser, Prof. Dr. Gabriel Schaaf and Prof. Dr. Mathias Becker for accepting my request to be part of the examination committee.

I also express my deepest gratitude to Dr. Agim Ballvora for his guidance, kind support, valuable suggestions and discussions that laid the foundations for my PhD Work.

Also I would like to express my sincere gratitude to Dr. Boby Mathew for his guidance and valuable suggestions throughout my research study.

Meanwhile I am grateful to Dr. Said Dadshani, Dr. Ahossi Patrice Koua and Dr. Md Nurealam Siddiqui for their technical supports. Also I would like to thank to Prof. Dr. Annaliese Mason, Dr. Fei He and Prof. Dr. Ali Naz for the scientific discussions. I offer thanks to Karin Woitol and Anne Reinders who helped me to have a nice working atmosphere.

I am grateful to my colleagues Dr. Mohammad Kamruzzaman, Dr. Carolyn Mukiri Kambona and Gullar Gadimaliyeva for their mutual support and friendly working atmosphere. Many thanks to my Indian friends for their kind support. For the official support I am thankful to the Theodor-Brinkman-Graduate School of the Faculty of Agriculture at the University of Bonn.

Mohammad Bahman Sadeqi

Dedication

I dedicate this thesis to my belated mother, who was my best friend. I lost her during my study time. May Almighty Allah (SWT) rewards her departed soul in Jannah!

Table of Contents

List of Tables

List of Figures

Summary

Nitrogen (N) as an essential element in the structure of proteins, nucleic acids and chlorophyll plays an important role in grain yield. Nitrogen fertilizer, which is most commonly used in cereal production, is necessary to increase the shoot biomass and dry matter of bread wheat. Increasing the amount of nitrogen fertilizer contributes greatly to yield stability in bread wheat, but soil and environmental pollution due to significant greenhouse gas emissions from nitrogen fertilizer production, the cost of nitrogen fertilizer and the energy required in agricultural practice can be considered as major negative consequences of nitrogen fertilizer. For all these reasons, it is important to identify nitrogen use efficiency (NUE) as a complex target trait in breeding programs. Based on this definition, investigating GY under different N applications could be a practical approach to model NUE. The simultaneous optimization of NUE and GY under different N applications could be the main goal of breeding for use efficiency. Characterization of agronomic traits to model NUE provides useful information and guidance for genomic selection programs. Allelic variation for GY at low and high N could be high due to the large target size of mutations in candidate genes. Therefore, the major challenge in genome-wide association study (GWAS) models is to find a significant and reliable association for the complex trait. To address this dilemma, two precise and efficient computational approaches, local FDR correction and Bayesian survival analysis, were developed as different filters to determine the best GWAS model and obtain a reliable association in the output of the best selected model. GWAS models for GY under low and high N levels have shown that the local FDR correction based only on maximum likelihood estimation is not more accurate than the local Bayesian FDR correction to determine the effect size and power of a large-scale genomic file. Currently, phenotyping with the aim of identifying high yielding genotypes is still expensive compared to genomic selection (GS) approaches. GS models consist of a whole genome genotyping file (SNPs) and a phenotyping file (individuals) in the reference population (training population). Statistical machine learning algorithms such as classical methods, kernel regression and ensemble learning algorithms are used to predict the phenotypes or breeding values (BVs) of the candidates for selection in the test population (validation). In modern GS models, there are two types of traits, including genetic parameters with random effects and hyper-parameters with fixed effects, which determine the results. Linear GS models such as *rrBLUP* and *gBLUP* are specified without having to worry too much about the assumptions. Bayesian inference, however, is more flexible when it comes to assumptions, and it has a distribution of responses that can change with each run. The main challenge with *BGLR* and *LASSO* is to ensure that the distribution of statistical estimators follows the genetic parameters of the population. In *SVM*, hyper-parameter optimization can be done by various methods available for hyper-parameter optimization in *SVM*, but the most commonly used and convenient method is grid search. Our study has shown that focusing on the definition and optimization of the regularization parameters is crucial for the performance and accuracy of the GS model, which has not been sufficiently addressed in previous GS studies. However, in the *BOOST* model, the regularization parameter is only used to control the bias of the model. By adjusting the regularization parameter, the model can be made less complex and less prone to overfitting. In the *BAGG* model, the regularization parameter is used to control the variance of the model. By reducing the variance, the model can be made more stable and less susceptible to noise. In *STACK*, both bias and variance are taken into account by adjusting the regularization parameter to find a balance between model complexity and stability. Thus, our study confirmed the results of the bias-variance trade-off and the adaptive error of prediction for the *STACK* model was in the mid-range compared to other models. This remarkable result for the *STACK* model is consistent with previous results. For all ensemble models, especially the *STACK* model, the number of epochs and the stack must be specified as hyper-parameters together with the activation process. Ultimately, a smaller learning rate in the training dataset with a desired batch size leads to maximum SNP heritability and genomic estimated breeding values (GEBVs) at the mean of the given GS model.

Kurzfassung

Stickstoff (N) als essentielles Element in der Struktur von Proteinen, Nukleinsäuren und Chlorophyll spielt eine wichtige Rolle im Getreideertrag (GY). Stickstoffdünger als die am häufigsten angewendete Methode in der Getreideproduktion ist notwendig, um die Schießmasse und Trockenmasse im Weizen zu erhöhen. Eine Erhöhung der Menge an Stickstoffdünger trägt maßgeblich zur Ertragssicherheit im Weizen bei, aber Boden- und Umweltverschmutzung aufgrund signifikanter Treibhausgasemissionen aus der Stickstoffdüngerproduktion, die Kosten für Stickstoffdünger und die erforderliche Energie in landwirtschaftlichen Praktiken können als signifikante negative Folgen von Stickstoffdünger betrachtet werden. Aus all diesen Gründen ist es wesentlich, die Stickstoffnutzungseffizienz (NUE) als komplexes Zielmerkmal in Züchtungsprogrammen zu identifizieren. Gemäß der Definition könnte die Untersuchung des GY unter verschiedenen N-Anwendungen ein praktischer Ansatz zur Modellierung der NUE sein. Die Optimierung von NUE und GY gleichzeitig unter verschiedenen N-Anwendungen könnte das Hauptziel der Züchtung zur Nutzungseffizienz sein. Die Charakterisierung von agronomischen Merkmalen in Bezug auf die Modellierung von NUE liefert nützliche Informationen und Richtungen für genomische Selektionsprogramme. Die allelische Variation für GY unter niedrigen und hohen N-Levels könnte hoch sein aufgrund der großen Mutationszielgröße innerhalb der Kandidatengene. Daher besteht die Hauptherausforderung bei Genomweiten Assoziationsstudien (GWAS)-Modellen darin, signifikante zuverlässige Assoziationen bei komplexen Merkmalen zu finden. Im Gegensatz zu diesem Dilemma sind die lokale FDR-Korrektur und die bayesianische Überlebensanalyse zwei präzise und effiziente rechnergestützte Ansätze als unterschiedliche Filter, um das beste GWAS-Modell zu bestimmen, um zuverlässige Assoziationen in den Ergebnissen des besten ausgewählten Modells zu erhalten. GWAS-Modelle für GY unter niedrigen und hohen N-Levels haben gezeigt, dass die lokale FDR-Korrektur nur auf Basis der Maximum-Likelihood-Schätzung nicht genauer ist als die bayesianische lokale FDR, um die Effektgröße und die Leistung von groß angelegten genomischen Dateien durchzuführen. Derzeit ist das Phänotypisieren, um Genotypen mit hohem Ertrag zu identifizieren, im Vergleich zu genomischen Selektions (GS)-Ansätzen immer noch teuer. GS-Modelle bestehen aus der gesamten Genom-Genotypisierungsdatei (SNPs) und der Phänotypisierungsdatei (Individuen) in der Referenz (Trainings-)Population und prognostizieren mithilfe von statistischen maschinellen Lernalgorithmen wie klassischen Methoden, Kernel-Regression und Ensemble-Lernalgorithmen die Phänotypen oder Zuchtwerte (BVs) der Kandidaten für die Auswahl in der Test (Validierungs-)Population. In den modernen GS-Modellen gibt es zwei Arten von Merkmalen, die genetische Parameter mit zufälligen Effekten und Hyperparameter mit festen Effekten umfassen, die die Ergebnisse bestimmen. GS-Linearmodelle wie *rrBLUP* und *gBLUP* werden spezifiziert, ohne sich allzu sehr um die Annahmen zu kümmern. Aber die Bayes'sche Inferenz ist flexibler, wenn es um Annahmen geht, und sie wird eine Verteilung von Antworten haben, die sich von jedem Durchlauf ändern kann. Die Hauptherausforderung bei *BGLR* und *LASSO* besteht darin, sicherzustellen, dass die Verteilung der statistischen Schätzer den genetischen Parametern der Population folgt. In *SVM* kann die Optimierung der Hyperparameter durch verschiedene Methoden zur Hyperparameter-Optimierung in *SVM* erreicht werden, aber die am häufigsten verwendete und praktischste Methode ist die Gittersuche. Unsere Studie hat gezeigt, dass die Definition und Optimierung des Regularisierungsparameters entscheidend ist, um die Leistung und Genauigkeit des GS-Modells zu demonstrieren, was in früheren GS-Studien nicht ausreichend berücksichtigt wurde. Im *BOOST*-Modell wird der Regularisierungsparameter jedoch nur verwendet, um die Verzerrung des Modells zu steuern. Durch Anpassung des Regularisierungsparameters kann das Modell weniger komplex und weniger anfällig für Overfitting gemacht werden. Im *BAGG*-Modell wird der Regularisierungsparameter verwendet, um die Varianz des Modells zu steuern. Durch Verringerung der Varianz kann das Modell stabiler gemacht werden und weniger empfindlich auf Rauschen reagieren. In *STACK* werden sowohl Verzerrung als auch Varianz berücksichtigt, indem der Regularisierungsparameter angepasst wird, um ein Gleichgewicht zwischen Modellkomplexität und Stabilität zu finden. Unsere Studie hat bestätigt, dass die Ergebnisse des Bias-Varianz-Tradeoffs und des adaptiven Fehler der Vorhersage für das *STACK*-Modell im Vergleich zu anderen Modellen intermediär waren. Dies bemerkenswerte Ergebnis beim *STACK*-Modell ist konsistent mit früheren Ergebnissen. In allen Ensemble-Modellen, insbesondere im *STACK*-Modell, müssen die Anzahl der Epochs und Batches als Hyperparameter zusammen mit dem Aktivierungsprozess spezifiziert werden. Letztendlich führt eine kleinere Lernrate im Trainingsdatensatz mit einer gewünschten Batch-Größe zu maximaler SNP-Erblichkeit und genomischen geschätzten Zuchtwerten (GEBVs) im gegebenen GS-Modell.

Abbreviations

Abbreviation	Explanation
AIC	Akaike Information Criterion
AM	Association Mapping
BF	Bayes Factor
BGLR	Bayesian Generalized Linear Regression
BIC	Bayesian Information Criterion
BLUPs	Best Linear Unbiased Predictors
BV	Breeding Value
CI	Confidence Interval
CV	Cross Validation
DAPC	Discriminate Analysis of Principle Components
EBVs	Estimated Breeding Values
EBayes	Empirical Bayesian algorithm
FCR	False Coverage Rate
FDR	False Discovery Rate
FWER	Family Wise Error Rate
gBLUP	Genomic Best Linear Unbiased Prediction
GEBVs	Genomic Estimated Breeding Values
GLM	Generalized Mixed Linear Models
GP	Genomic Prediction
GRM	Genomic Relationship Matrix
GS	Genomic Selection
GV	Genotypic Value
GWAS	Genome Wide Association Study
GY	Grain Yield
HN	High Nitrogen
HWE	Hardy-Weinberg Equilibrium
IE	Irreducible Error
KNN	K-Nearest Neighbor KNN
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium

Abbreviation	Explanation
LN	Low Nitrogen
LFDR	Local FDR
LR	Learning Rate
MAF	Minor Allele Frequency
MAS	Marker Assisted Selection
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
MTAs	Marker-Trait Associations
N	Nitrogen
NGS	Next Generation Sequencing
NUE	Nitrogen Use Efficiency
NUpE	Nitrogen Uptake Efficiency
NUtE	Nitrogen Utilization Efficiencies
PCA	Principle Component Analysis
QTL	Quantitative Trait Loci
REML	Restricted Maximum Likelihood REML
RF	Random Forest
RKHS	Regression Reproducing Kernel Hilbert Spaces
rrBLUP	Ridge Regression Best Linear Unbiased Prediction
SE	Standard Error
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
VI	Variable Importance

This PhD thesis consists of five (5) chapters. It starts with a general introduction (Chapter 1), three experimental studies thus each experimental study focuses on specific aims and objectives stated under sections of chapters (Chapters 2 to 4), and at the end general discussion (Chapter 5). These three studies form the most important parts of this thesis. They have been published as peer-reviewed journal articles: (Sadeqi et al., 2023a: International Journal of Molecular Science, doi: https://doi.org/10.3390/ijms241814011, Sadeqi et al., 2023b: International Journal of Molecular Science, doi: https://doi.org/10.3390/ijms241814275), or are under review as manuscript. My contributions to each paper are specified and listed under the publications section of this thesis.

Chapter 1:

General Introduction

1.1. Research Context

Nitrogen (N) plays an important role in plant production as an essential element in the structure of proteins, nucleic acids, and chlorophyll. Moreover, N is the main nutrient for canopy growth and photosynthesis, which determines grain yield and quality (Beres et al, 2018; Walsh et al, 2018 and Ondoua et al, 2019). N fertilizer as the most common application in grain production is necessary to increase shoot biomass and dry matter of bread wheat (Saleem et al., 2021). Considering nearly 8 billion people, the consumption of N fertilizer is about 123 million tons, with 45% applied in developed countries and the remaining in developing countries, which clearly shows that GY has increased simultaneously with fertilizer use (FAOSTAT, 2019). This upward trend in the world population forecast is mainly due to the demand, especially from developing countries, to maintain food security with population growth. According to Malthus' theory, population grows in geometric or nonlinear progression and food production grows in arithmetic or linear progression, resulting in an imbalance between population and food supply (Tisdell and Svizzero, 2017). Increasing production of wheat (*Triticum aestivum* L.) is needed to provide staple foods for about half of the world's population to achieve global food security (Curtis and Halford, 2014). Wheat production in Germany increased from 7.99 million tons in 1960 to 23.1 million tons in 2019 with an average annual growth rate of 2.33% (Knoema, 2019). Globally, about 83 million tons of N are used, which is about a 100-fold increase in the last 100 years. About 60% of global N fertilizer is used for the production of the world's three major cereals: rice, wheat, and maize (Chang et al. 2021). To meet the needs of the growing population, wheat varieties require sustainable management and facilities that can cope with all environmental factors such as biotic and abiotic stresses and maintain expected yields. In addition, a reduction in N fertilization could reduce yield and quality while the crop faces N deficiency (Figure 1.1).

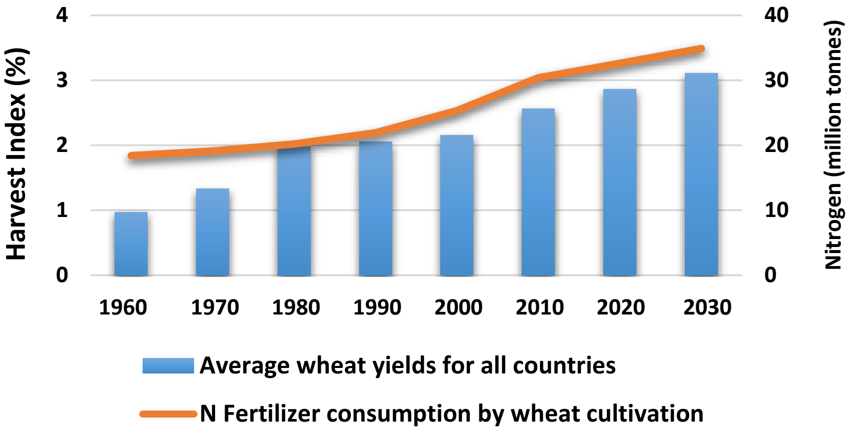


Figure 1.1- Total average of wheat grain yield and N fertilizer consumption (past and projected) in the world from 1960 to 2030. Yields are projected to grow by averagely 41 kg per year, which shows the GY gap based on Malthus' theory.
Note: All countries include major developed country experts for wheat production.

Increased N fertilizer rates contribute greatly to yield stability in bread wheat (Zemichael et al. 2017), but soil and environmental pollution from significant greenhouse gas emissions from N fertilizer production (Garnett et al., 2015), nitrate leaching (Pathak et al. 2011), volatilization, surface runoff, and denitrification from the soil-plant system (Yadav et al. 2017) can be considered major negative consequences of high N fertilizer use in wheat production. Recovery of N fertilizer in cereals is generally poor, and only 33% of the applied N is actually harvested in the grain, with the remaining portion (67%) remaining in the soil (Sharma and Bali, 2017; Doe, 2015). In addition, the cost of N fertilizer and the energy required in agricultural practices (Mahjourimajd et al. 2016) make it essential to determine nitrogen use efficiency (NUE) in wheat breeding programs. Recent breeding programs in Europe to improve N uptake as a major aspect (Grinsven et al., 2013; Allard et al., 2013; Bingham et al., 2012) have resulted in Europe becoming one of the largest wheat producers in the world with an average yield of more than 5.01 t ha-1 (FAOSTAT, 2019). Over the past three decades, higher grain yields have been achieved in commercial bread wheat varieties, and breeding progress has led to improvements in both yield performance and baking quality (Laidig et al., 2016). Nevertheless, there is still a large demand for bread wheat in developing and developed countries, and it is important to improve NUE to minimize additional N input. Therefore, there is great interest in wheat varieties with high NUE, as these high yield lines are expected to minimize environmental damage and production costs, which is economically attractive to breeders.

1.2. Breeding for NUE in wheat

When GY is a function, NUE is defined as GY divided by total available N from fertilizer and soil (Barraclough et al, 2010; Hawkesford 2017). NUE as a quantitative trait is divided into N uptake efficiency and N utilization efficiency (NUtE). N uptake efficiency (NUpE) indicates the maximum capacity of uptake of N by the plant from the soil. NUtE is GY per unit of N in the plant. By definition, both NUpE and NUtE result in GY.

Therefore, modeling GY based on field experiment is an integral part of breeding for NUE in wheat. Simultaneous optimization of NUE and GY under different N applications could be the main objective of breeding programs. However, this optimization is not easy and requires a trade-off between the actual bias and variance for each component of NUE. This trade-off requires powerful and accurate statistical models that can effectively identify the principal components of NUE. Applying frequentist inference, such as maximum likelihood estimation (MLE), to each agronomic trait provides a general judgment about the input of related traits as components in the NUE model. However, when the number of environmental features such as soil and weather, increases, the accuracy of the MLE is low. Since some of the related traits have an exponential probability distribution, the empirical Bayes algorithm may be an alternative inference in this case to model NUE based on agronomic data recorded from multi years and locations trials. Characterization of agronomic traits to improve NUE and genetic variation provides useful information and directions to wheat breeders (Gaju et al., 2011; Asplund et al., 2015) and selection programs. Many studies on agronomic traits related to NUE have found a significant interaction between genotypes and N levels (G × N), indicating that this is an important source for efficient genotype selection (Lei et al, 2018; Guttieri et al, 2017; Barraclough et al, 2010). The significance of G × N interactions directly affects the correlations of genetic value for each genotype, implying that the best varieties for NUE at high N (HN) may not be the best at low N (LN). To estimate the genetic progress of NUE, a historical study comparing 193 old and new varieties at their response to optimal N levels was conducted (Cormier et al., 2013), and a genetic gain of 0.30-0.37% per year was observed between 1960 and 2010 for this elite European panel, indicating very low genetic progress. Due to the low genetic variance in the panel, genetic improvement was considered only marginally, as there is no targeted selection trait for NUE.

Quantitative trait loci (QTL) analysis using molecular markers is a well-known statistical genetic analysis for identifying genetic loci associated with quantitative traits such as NUE and GY. Two of the most commonly used QTL mapping approaches are: i) linkage analysis (LA): also known as family-based linkage mapping approach for experimental population or QTL mapping and ii) association mapping (AM): with linkage disequilibrium (LD) mapping for an associated panel, the genetic map is viewed at higher resolution. In AM studies, especially in the form of genome wide association studies (GWAS), LD decay of chromosomal segments and genomic relationship matrix (kinship) are important parameters to identify regions associated with interested trait (Dadshani et al, 2021; Mathew et al, 2018). Several studies reported desirable QTLs for shoot biomass and GY (Zhang et al, 2021; Mahjourimajd et al, 2016; Xu et al, 2013) and root weight (Ren et al, 2017; Horn et al, 2016; Shorinola et al, 2018) in wheat under HN applications. However, most of them, especially the identified QTLs for root morphology in soil, are not precise and stable. Therefore, further studies on epistatic interactions between root and shoot biomass cofactors under different environmental factors are needed (Zhang et al, 2018; Li et al, 2014). Meanwhile, there are not many GWAS studies for NUE and GY in European wheat population under different N applications (Saini et al, 2021; Monostori et al, 2017; Gouis, 2011). Basically, NUE is a complex trait determined by many other related agronomic traits, each of which is controlled by many genes in cooperation with low effects and many environmental factors. As a result, genetic progress and narrow sense heritability (h^2) are very low for this type of quantitative trait per year. In GWAS and genomic prediction (GP), one side of the model is assigned to the complex trait. Outliers in the NUE vector could be the first cause of pseudo-marker trait associations (MTAs) in the GWAS results. The second reason is due to the type of GWAS and GP models used for the trait in question. Most GWAS models for single loci and multiple loci exhibit collinearity and over- or under-fitting of results due to pairwise comparisons between single nucleotide polymorphisms (SNP). This problem will be intensified, especially, when the value of epistatic variance (V_I) in genetic variance (V_g) is high for the trait of interest (Wang et al, 2016, Kärkkäinen et al, 2015; Li and Sillanpää 2012). Allelic variation for NUE could be high due to the large mutation target size within candidate genes. In addition, it is of considerable importance to identify all involved variants between traits (Robinson et al., 2014). Therefore, the main challenge in GWAS is to find a significant and reliable association in a complex trait. To address this dilemma, false discovery rate (FDR) correction and Bayesian survival analysis, two precise and efficient computational approaches, serve as different filters to determine the best GWAS and GP models and consequently obtain a reliable association in the result of the best selected model. In addition, the application of genotyping by sequencing (GBS) as a new technique, with high SNP density and aligned to the wheat reference genome, has the potential to improve the accuracy of the associations determined for NUE and related traits (Brasier et al., 2020).

1.3. Background of the study

To investigate genetic progress based on NUE, usually two approaches are used: i) historical trial analyses and ii) direct comparisons of old and modern varieties in the same location. First approach is commonly with bias as elimination of year effects and it's leaded to inadequate consideration of G × Y interactions (Lopes et al, 2012; Graybosch and Peterson, 2010). Direct comparisons between old and modern genotypes is limited by few available genotypes assessed in few environments and then, size of experiment will be big and genetic progress is earning gradually (Migliorini et al, 2016; Ceccarelli, 2014). The calculation of genetic progress based on NUE across variety release year was done using the following equation: $Y = \beta_0 + \beta X + \varepsilon$, where Y is the mean NUE for two N applied include High N (HN) equal by 220 kg ha^{-1} N fertilizer and Low N (LN) equal by only natural minimum N available in the soil at each year of variety ith, X is the year in which variety was released, β is a fixed regression coefficient for NUE trend, β_0 the intercept of equation which a random normal deviation of Y from NUE trend and ε is residual error.

The genetic progress in both N applications were significant (ρ_{HN} = *2.2e-14* and ρ_{LN} = *1.1e-7*) and did result from a change in variety ranking among years (Figure 1.2), and this difference may be explained partially the G × N and significant allelic variation on NUE related traits, which delicates the high potential of genetic gain in the population. So NUE in wheat requires to be more genetically improved.

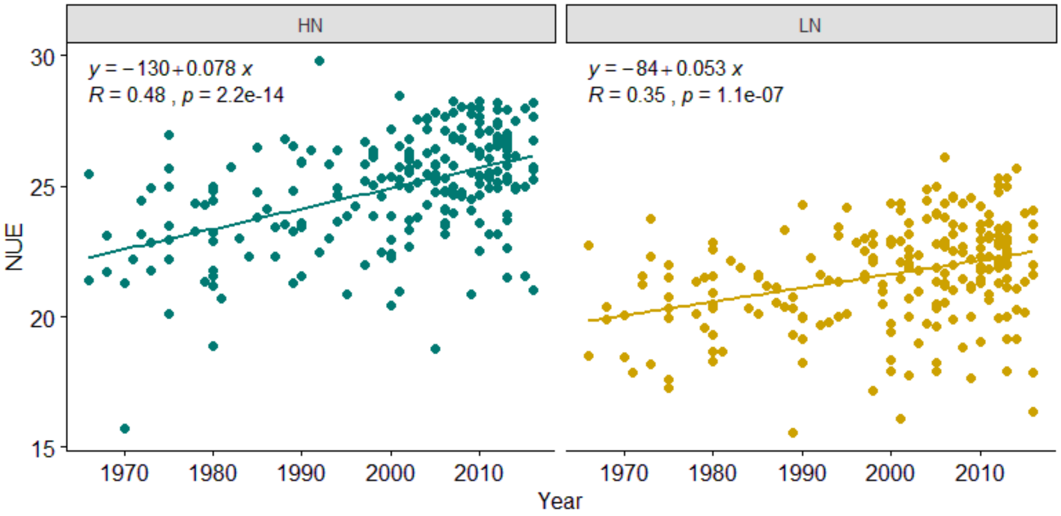


Figure 1.2- Adjusted mean of genetic progress based on NUE (%) in 221 wheat varieties across two levels of applied N across variety release year (1960-2020), Regression lines for both HN and LN are plotted to display NUE trend in the equations.

1.4. Research strategy

NUE in the wheat, is a broad topic with different aspects that can be addressed to study on high throughput phenotyping of root and shoot related traits or agricultural practices on N fertilizer and its impacts on the environment or finding out the genetic basis of this complex trait to improve it with high efficiency. To define the breeding strategy in both phenotyping and genotyping phase of study, we have three chapters include: i) Field screen to model NUE and related component traits in wheat using frequentist, Bayesian and modern statistical inferences, ii) Allelic variation for NUE and related traits in wheat using different distinguish GWAS and GP models, and iii) FDR correction and Bayesian survival analysis to determine reliable associations based on result of GWAS and GP models. We had 221 genotypes of bread wheat with three N level and had 10 agronomic traits related to NUE. The general objectives of this study to provide a better understanding of wheat varieties (genotypes) and their responses to different levels of N treatment and to analyze the variance in agronomic traits related to NUE and then allelic variation and genetic gain of the panel. NUE is considered as complex trait to breed and it is controlled by many genes with minor effects or small *p-value*(s). Currently in wheat breeding, GWAS have successfully revealed the genetic basis of complex traits such as NUE (Rathan et al, 2022; Uffelmann et al, 2021). In GWAS, thresholding is common strategy to indicate deviation of false discovery rate (FDR) of high density markers under or over of test statistics. We will start the study by significant threshold selection for NUE as complex target based on wide genomic data of bread wheat. The limitation of common FDR thresholding approaches will be showed. After determination of best FDR threshold approach, in the second chapter, it will be utilized in the different GWAS and GP models. Therefore, in the third chapter, we will find the best GWAS model to receive reliable association for GY in bread wheat. Estimation of breeding values (BVs) based on genomic selection (GS) algorithms for complex traits is the main objective in wheat breeding program. Therefore, in the fourth chapter, we evaluate the modern GS models to find the best genotypes under low and high N levels based on genetic parameter and hyper-parameter estimation. Regarding the complementarity of study, we found it appropriate to present to the Jury Committee under chapter style. Finally this manuscript is presented as a compilation of dissertation linked by general discussion and further conclusion.

1.5. References

Allard, V., Martre, P., & Gouis, J. L. (2013). Genetic variability in biomass allocation to roots in wheat is mainly related to crop tillering dynamics and nitrogen status. *European Journal of Agronomy, 46*, 68-76. doi:10.1016/j.eja.2012.12.004

Asplund, L., Bergkvist, G., & Weih, M. (2015). Functional traits associated with nitrogen use efficiency in wheat *Acta Agriculturae Scandinavica, Section B, Soil & Plant Science, 66*(2), 153-169 doi:10.1080/09064710.2015.1087586

Beres, B., Graf, R., Irvine, R., O'Donovan, J., Harker, K., Johnson, E. Stevenson, F. (2018). Enhanced nitrogen management strategies for winter wheat production in the Canadian prairies. Canadian Journal of Plant Science, 98(3), 683-702. doi:10.1139/cjps-2017-0319

Bingham, I., Karley, A., White, P., Thomas, W., & Russell, J. (2012). Analysis of improvements in nitrogen use efficiency associated with 75 years of spring barley breeding. *European Journal of Agronomy, 42*, 49-58. doi:10.1016/j.eja.2011.10.003

Barraclough, P. B., Howarth, J. R., Jones, J., Lopez-Bellido, R., Parmar, S., Shepherd, C. E., & Hawkesford, M. J. (2010). Nitrogen efficiency of wheat: Genotypic and environmental variation and prospects for improvement. *European Journal of Agronomy, 33*(1), 1-11. doi:10.1016/j.eja.2010.01.005

Brasier, K., Ward, B., Smith, J., Seago, J., Oakes, J., Balota, M., Davis, P., Fountain, M., Brown-Guedira, G., Sneller, C., Thomason, W., & Griffey, C. (2020). Identification of quantitative trait loci associated with nitrogen use efficiency in winter wheat. *PLOS ONE*, *15*(2). https://doi.org/10.1371/journal.pone.0228775

Ceccarelli, S. (n.d.). Drought. *Plant Genetic Resources and Climate Change,* 221-235. doi:10.1079/9781780641973.0221

Chang, J., Havlík, P., Leclère, D., de Vries, W., Valin, H., Deppermann, A., Hasegawa, T., & Obersteiner, M. (2021). Reconciling regional nitrogen boundaries with Global Food Security. *Nature Food*, *2*(9), 700–711. https://doi.org/10.1038/s43016-021-00366-x

Cormier, F., Faure, S., Dubreuil, P., Heumez, E., Beauchêne, K., Lafarge, S., Gouis, J. L. (2013). A multi-environmental study of recent breeding progress on nitrogen use efficiency in wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics, 126*(12), 3035-3048. doi:10.1007/s00122-013-2191-9

Curtis, T., & Halford, N. G. (2014). Food security: The challenge of increasing wheat yield and the importance of not compromising food safety. Annals of Applied Biology, 164(3), 354 372. https://doi.org/10.1111/aab.12108

Dadshani, S., Mathew, B., Ballvora, A., Mason, A. S., & Léon, J. (2021). Detection of breeding signatures in wheat using a linkage disequilibrium-corrected mapping approach. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-85226-1

Doe, J. (2015). Ammonia and Nitrate Losses from Agriculture and Their Effect on Nitrogen Recovery in the EU and U.S. *CSA News*, 60(4), p.16

FAOSTAT (2019) Current world fertilizer trends and outlook to 2025. https://www.fao.org/faostat/en/#data/RFN

Gaju, O., Allard, V., Martre, P., Snape, J., Heumez, E., Legouis, J., Foulkes, M. (2011). Identification of traits to improve the nitrogen-use efficiency of wheat genotypes. *Field Crops Research, 123*(2), 139-152. doi:10.1016/j.fcr.2011.05.010

Garnett, T., Plett, D., Heuer, S., & Okamoto, M. (2015). Genetic approaches to enhancing nitrogen-use efficiency (NUE) in cereals: Challenges and future directions. *Functional Plant Biology, 42*(10), 921. doi:10.1071/fp15025

Graybosch, R. A., & Peterson, C. J. (2010). Genetic Improvement in Winter Wheat Yields in the Great Plains of North America, 1959–2008. *Crop Science, 50*(5), 1882. doi:10.2135/cropsci2009.11.0685

Grinsven, H. J., Spiertz, J. H., Westhoek, H. J., Bouwman, A. F., & Erisman, J. W. (2013). Nitrogen use and food production in European regions from a global perspective. *The Journal of Agricultural Science, 152*(S1), 9-19. doi:10.1017/s0021859613000853

Gouis, J. L. (2011). Genetic Improvement of Nutrient Use Efficiency in Wheat. *The Molecular and Physiological Basis of Nutrient Use Efficiency in Crops,* 121-138. doi:10.1002/9780470960707.ch7

Guttieri, M. J., Frels, K., Regassa, T., Waters, B. M., & Baenziger, P. S. (2017). Variation for nitrogen use efficiency traits in current and historical great plains hard winter wheat. *Euphytica, 213*(4). doi:10.1007/s10681-017-1869-5

Hawkesford, M. J. (2017). Genetic variation in traits for nitrogen use efficiency in wheat. *Journal of Experimental Botany, 68*(10), 2627-2632. doi:10.1093/jxb/erx079

Horn, R., Wingen, L. U., Snape, J. W., & Dolan, L. (2016). Mapping of quantitative trait loci for root hair length in wheat identifies loci that co-locate with loci for yield components. *Journal of Experimental Botany*, *67*(15), 4535–4543. https://doi.org/10.1093/jxb/erw228

Kärkkäinen, H. P., Li, Z., & Sillanpää, M. J. (2015). An Efficient Genome-Wide Multilocus Epistasis Search. Genetics, 201(3), 865–870. https://doi.org/10.1534/genetics.115.182444

Knoema (2019) Germany wheat production, 1961-2020. Retrieved October 10, 2021, from https://knoema.com/atlas/Germany/topics/Agriculture/Crops-Production-Quantity-tonnes/Wheat production.

Laidig, F., Piepho, H., Rentel, D., Drobek, T., Meyer, U., & Huesken, A. (2016). Breeding progress, environmental variation and correlation of winter wheat yield and quality traits in German official variety trials and on-farm during 1983–2014. *Theoretical and Applied Genetics, 130*(1), 223-245. doi:10.1007/s00122-016-2810-3

Lei, L., Li, G., Zhang, H., Powers, C., Fang, T., Chen, Y., . Yan, L. (2018). Nitrogen use efficiency is regulated by interacting proteins relevant to development in wheat. *Plant Biotechnology Journal, 16*(6), 1214-1226. doi:10.1111/pbi.12864

Li, Z., & Sillanpää, M. J. (2012). Estimation of Quantitative Trait Locus Effects with Epistasis by Variational Bayes Algorithms. Genetics, 190(1), 231–249. https://doi.org/10.1534/genetics.111.134866

Li, Z. K., Jiang, X. L., Peng, T., Shi, C. L., Han, S. X., Tian, B., Zhu, Z. L., & Tian, J. C. (2014). Mapping quantitative trait loci with additive effects and additive X additive epistatic interactions for biomass yield, grain yield, and straw yield using a doubled haploid population of wheat (triticum aestivum L.). *Genetics and Molecular Research*, *13*(1), 1412–1424. https://doi.org/10.4238/2014.february.28.14

Lopes, M. S., Reynolds, M. P., Manes, Y., Singh, R. P., Crossa, J., & Braun, H. J. (2012). Genetic Yield Gains and Changes in Associated Traits of CIMMYT Spring Bread Wheat in a "Historic" Set Representing 30 Years of Breeding. *Crop Science, 52*(3), 1123. doi:10.2135/cropsci2011.09.0467

Mahjourimajd, S., Taylor, J., Sznajder, B., Timmins, A., Shahinnia, F., Rengel, Z., Langridge, P. (2016). Genetic Basis for Variation in Wheat Grain Yield in Response to Varying Nitrogen Application. *Plos One, 11*(7). doi:10.1371/journal.pone.0159374

Mathew, B., Léon, J., & Sillanpää, M. J. (2018). Impact of residual covariance structures on genomic prediction ability in multi-environment trials. *PLOS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0201181

Migliorini, P., Spagnolo, S., Torri, L., Arnoulet, M., Lazzerini, G., & Ceccarelli, S. (2016). Agronomic and quality characteristics of old, modern and mixture wheat varieties and landraces for organic bread chain in diverse environments of northern Italy. *European Journal of Agronomy, 79*, 131-141. doi:10.1016/j.eja.2016.05.011

Monostori, I., Szira, F., Tondelli, A., Árendás, T., Gierczik, K., Cattivelli, L., Galiba, G., & Vágújfalvi, A. (2017). Genome-wide association study and genetic diversity analysis on nitrogen use efficiency in a Central European winter wheat (*Triticum aestivum* L.) collection. *PLOS ONE*, *12*(12). https://doi.org/10.1371/journal.pone.0189265

Ondoua, R. N., & Walsh, O. (2017). Varietal differences in nitrogen use efficiency among spring wheat varieties in Montana. *Crops & Soils*, *50*(5), 40–42. https://doi.org/10.2134/cs2017.50.0505

Pathak, R., Lochab, S., & Raghuram, N. (2011). Improving Plant Nitrogen-Use Efficiency. *Comprehensive Biotechnology*, 209-218. doi:10.1016/b978-0-08-088504-9.00472-4

Rathan, N. D., Krishna, H., Ellur, R. K., Sehgal, D., Govindan, V., Ahlawat, A. K., Krishnappa, G., Jaiswal, J. P., Singh, J. B., SV, S., Ambati, D., Singh, S. K., Bajpai, K., and Mahendru-Singh, A. (2022). Genome-wide association study identifies loci and candidate genes for grain micronutrients and quality traits in wheat (*Triticum aestivum* L.). *Scientific Reports*, 12(1). https://doi.org/10.1038/s41598-022-10618-w

Ren, Y., Qian, Y., Xu, Y., Zou, C. Q., Liu, D., Zhao, X., Zhang, A., & Tong, Y. (2017). Characterization of qtls for root traits of wheat grown under different nitrogen and phosphorus supply levels. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.02096

Robinson, M. R., Wray, N. R., & Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics*, *30*(4), 124–132. https://doi.org/10.1016/j.tig.2014.02.003

Saini, D. K., Chopra, Y., Pal, N., Chahal, A., Srivastava, P., & Gupta, P. K. (2021). Meta-qtls, ortho-mqtls and candidate genes for nitrogen use efficiency and root system architecture in bread wheat (*Triticum aestivum* L.). *Physiology and Molecular Biology of Plants*, *27*(10), 2245–2267. https://doi.org/10.1007/s12298-021-01085-0

Saleem Kubar, M., Feng, M., Sayed, S., Hussain Shar, A., Ali Rind, N., Ullah, H., Ali Kalhoro, S., Xie, Y., Yang, C., Yang, W., Ali Kalhoro, F., Gasparovic, K., Barboricova, M., Brestic, M., El Askary, A., & El-Sharnouby, M. (2021). Agronomical traits associated with yield and yield components of winter wheat as affected by nitrogen managements. *Saudi Journal of Biological Sciences*, *28*(9), 4852–4858. https://doi.org/10.1016/j.sjbs.2021.07.027

Sharma, L., & Bali, S. (2017). A Review of Methods to Improve Nitrogen Use Efficiency in Agriculture. *Sustainability*, 10(2), 51. doi:10.3390/su10010051

Shorinola, O., Kaye, R., Golan, G., Peleg, Z., Kepinski, S., & Uauy, C. (2018). Genetic screening for mutants with altered seminal root numbers in hexaploid wheat using a high-throughput root phenotyping platform. https://doi.org/10.1101/364018

Tisdell, C., & Svizzero, S. (2017). The ability in antiquity of some agrarian societies to avoid the Malthusian trap and develop. *Forum for Social Economics*, *49*(2), 202–227. https://doi.org/10.1080/07360932.2017.1356344

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1). https://doi.org/10.1038/s43586-021-00056-9

Walsh, O., Shafian, S., & Christiaens, R. (2018). Nitrogen Fertilizer Management in Dryland Wheat Cropping Systems. *Plants, 7*(1), 9. doi:10.3390/plants7010009

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J. M., Xu, S., & Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Scientific Reports, 6(1). https://doi.org/10.1038/srep19444

Xu, Y., Wang, R., Tong, Y., Zhao, H., Xie, Q., Liu, D., Zhang, A., Li, B., Xu, H., & An, D. (2013). Mapping qtls for yield and nitrogen-related traits in wheat: Influence of nitrogen and phosphorus fertilization on QTL expression. *Theoretical and Applied Genetics*, *127*(1), 59–72. https://doi.org/10.1007/s00122-013-2201-y

Yadav, M. R., Kumar, R., Parihar, C. M., Yadav, R. K., Jat, S., Ram, H. Jat, M. L. (2017). Strategies for improving nitrogen use efficiency: A review. *Agricultural Reviews,* (OF). doi:10.18805/ag.v0iof.7306

Zemichael, B., Dechassa, N., & Abay, F. (2017). Yield and Nutrient Use Efficiency of Bread Wheat (*Triticum Aestivum* L.) as Influenced by Time and Rate of Nitrogen Application in Enderta, Tigray, Northern Ethiopia. *Open Agriculture*, 2(1). doi:10.1515/opag-2017-0065.

Zhang, J., She, M., Yang, R., Jiang, Y., Qin, Y., Zhai, S., Balotf, S., Zhao, Y., Anwar, M., Alhabbar, Z., Juhász, A., Chen, J., Liu, H., Liu, Q., Zheng, T., Yang, F., Rong, J., Chen, K., Lu, M., … Ma, W. (2021). Yield-related QTL clusters and the potential candidate genes in two wheat DH populations. *International Journal of Molecular Sciences*, *22*(21), 11934. https://doi.org/10.3390/ijms222111934

Zhang, X., Sun, C., Zhang, Z., Dai, Z., Chen, Y., Yuan, X., Yuan, Z., Tang, W., Li, L., & Hu, Z. (2018). Correction: Genetic dissection of main and epistatic effects of QTL based on Augmented Triple Test Cross Design. *PLOS ONE*, *13*(5). https://doi.org/10.1371/journal.pone.0198562

Chapter 2:

Significant threshold selection for nitrogen use efficiency in whole genome of bread wheat

2.1. Introduction

Bread wheat (*Triticum aestivum* L.) is one of the most important cereals and covers the food requirements of a large part of the world's population. In general, the nitrogen use efficiency (NUE) of bread wheat is estimated to be only 30–40% of the total amount of nitrogen fertilizer applied that is actually harvested. Although genotype-environment (G×E) interaction and cultivation practices have a significant impact on NUE in bread wheat and a significant proportion of phenotypic variance is genetically based (Salim and Reza, 2019; Cormier et al, 2016). NUE is considered a complex breeding trait and is controlled by many genes with small effects or small p-values. In wheat breeding, genome-wide association studies (GWAS) have successfully uncovered the genetic basis of complex traits and biological processes (Rathan et al, 2022; Uffelmann et al, 2021). Genetic parameters such as population structure, genomic relationship matrix, marker density, sample size and minor allele frequency (MAF) have a major impact on the power and accuracy of the GWAS model (Wang et al., 2016). However, they play an important role in estimating the threshold for the significant false discovery rate (FDR). In GWAS, one side of the model is used for a complex trait such as NUE. However, to reduce multicollinearity and heteroscedasticity in the NUE observations, model evaluation and error measurement is required. Moreover, in GWAS models, thresholding is a common strategy to indicate the deviation from the expected range of p-values associated with single nucleotide polymorphisms (SNP) below or above the test statistic. Thus, the threshold plays a critical role in identifying significant features in as large a genome as possible while maintaining a relatively low proportion of false positives (Storey, 2002). In traditional GWAS models, linkage disequilibrium (LD) was tested locus by locus with a traditional p-value cutoff of 0.01 or 0.05 as a threshold to minimize false positives. LD analysis at each locus with a p = 0.05 probability of false positives is a stringent and conservative criterion because very few loci in the results show significant linkage (Zabaneh and Mackay, 2003) and this seems unlikely in a large association panel. The advent of high-throughput sequencing technologies and the availability of extensive genomic data have allowed researchers to explore GWAS models with different complex quantitative traits. As a result, mixed GWAS models with multiple loci have been proposed, focusing on a large number of hypotheses with small effects (Segura et al., 2012; Chen et al., 2016). The multiple comparisons problem arises when a statistical analysis involves multiple simultaneous statistical tests, each of which has the potential to be a discovery. There are several ways to tackle this problem, including family-wise error rate (FWER) and FDR, crude and adjusted p-values, consideration of threshold coherence and consonance, properties of proportional free and restricted hypothesis tests in threshold definition.

FDR thresholding based on linear approaches:

Bonferroni correction based on frequentist statistical inference is common method to deal with FWER which is appropriate for experimental studies such as quantitative trait locus (QTL) analysis. *Bonferroni* correction has intensive significance level (α) of null hypothesis to control the likelihood probability of false positive for multiple hypothesis test (Haynes 2013). Bonferroni correction is very simple and conservative technique with low power. It is depend on sample size and this adjustment does not impose a severe penalty on the range of effect sizes. The maximum likelihood (ML) of both type I and type II errors is increased, thereby with this threshold most of results are significant but does not exist truly (Perneger 1998; VanderWeele and Mathur 2018). *Holm* adjustment such as *Bonferroni* correction tries to estimate ML for threshold line in multiple comparisons hypotheses. It covers maximum tow way analysis of variance (ANOVA) in the same time (Giacalone et al, 2018; Lee and Lee 2020), which is appropriate for genome wide case and control studies, not association panel. *Hochberg* adjustment is an alternative approach to determine smallest FWER significant level and it is less sensitive than *Bonferroni* correction to decision about ML of FWER (Tan and Xu 2014; Chen et al, 2017). The idea thresholding with adjusted *p-value*(s) instead of raw *p-value*(s) was proposed for the first time in *Benjamini-Hochberg* approach (Reiner et al, 2003). Generalization of adjusted *p-value*(s) from raw *p-value*(s) in this approach has complexity in computation (Krzywinski and Altman 2014). Threshold based on adjusted *p-value*(s) has this potential to minimize error of FDR estimation, which is leaded to reveal marginal null hypotheses from peaks. Depending to the number of marginal null hypotheses that are true, threshold line in the *Benjamini-Hochberg* approach can be less conservative than Bonferroni correction (Benjamini and Bogomolov, 2013; Brinster et al, 2018). *Benjamini-Yekutieli* approach can be first tries to involve concept posterior distribution of FDR in threshold line definition. In practice there is many extreme raw p-value(s) even smaller than 2.7e-09 in the GWAS results that is a strict to estimate FDR with high accuracy. Therefore to deal with many simultaneous confidence intervals for FDR, it is less powerful than *Benjamini-Hochberg* approach (Furmańczyk 2013; Brinster et al, 2018). Such as *Bonferroni* correction, *Sidak* adjustment also is appropriate to deal with FWER threshold when there are low numbers of hypotheses are correlated (Blakesley et al, 2009). The origin of this correlation is coming back to genetic parameters such as minor MAF and level of LD in the genomic file. *Sidak* adjustment classifies null hypotheses using stepwise FWER controlling procedure, while threshold estimation is based on fixed bounds of raw *p-value*(s) (Midway et al, 2020).

FDR thresholding based on non-linear parameters:

Currently, in the distinguished GWAS models such as single locus or multi locus associations, threshold is widely considered as linear parameter (Cook et al, 2016; Wen et al, 2017; Lozano et al, 2023). Thresholding based on linear estimator has only mean squared error as scale upon high dimensional genomic dataset. Usually, in the GWAS results scaled version of adjusted *p-value* (s) doesn't follow easily distribution of raw *p-value*(s) and it is risky to soft thresholding. The first attempts to avoid this risk was consideration of threshold as non-linear concept (Wilson 2019; McCaw et al, 2020; Asif et al, 2021) and *q-value*(s) was introduced as an alternative approach to the adjusted *p-value* (s) method for thresholding. The *q-value*(s) is still widely used in genome wide studies and it is expected proportion of false positives rather than the ML estimation of false positive rate (Storey and Tibshirani, 2003). While estimator has an important role in thresholding FDR, the main drawback of this non-linear estimator is that *q-value*(s) is random variable and it may underestimate FDR to unexpected false positives (Lai 2017; Menyhart et al, 2022). Also according to raw *p-value*(s), the *q-value*(s) increases using the function $\pi_0(\lambda) = 1$. This function controls the proportion of raw *p-value*(s) used for the null distribution (*z-value*(s)). However, λ is a non-linear parameter with 0 to 1 value, even by bootstrap or permutation techniques, the λ closer to 1 implies increasing variance in π_0, which is leaded to autocorrelation and heteroscedasticity in the FDR thresholding. The second attempts to avoid the risk of soft thresholding upon large scale simultaneous hypothesis testing, was local FDR (LFDR) threshold (Efron 2013), *LFDR* features tail areas of null distribution and provides both scale and power of computation for large scale genome wide studies. With given the adjusted *p-value*(s), *LFDR* measures posterior probability of the local false positives using empirical Bayesian (EBayes) techniques (Efron 2013; Korthauer et al, 2019). All three adjustments *q-value*(s), $\pi_0(\lambda)$ and *LFDR* are computed based on proportion of false positives received from adjusted *p-value* (s). It assumes that this comparisons are independent. However, the mean square error of comparisons is affected by the outliers in the adjusted *p-value* (s), especially when there are strong negative correlations between comparisons simultaneously.

FDR threshold optimization:

In FDR thresholding, poor generalization performance is the main problem in genome wide dataset with large scale null hypotheses, simultaneously and generalization algorithms required more attention (Sørensen et al, 2022; Montesinos-López et al, 2023). While a few studies confirmed that threshold is a non-linear parameter concept (Emmert-Streib and Dehmer, 2019; Cao et al, 2023), we believe that sparsity and scale of false positives in the large simultaneous hypothesis tests are two important features in the FDR approach. Sparsity assumption in FDR threshold is having only small number of true positives reliable associations in GWAS results with nonzero SNP effects (Hastie et al, 2020). Scaling assumption is helpful technique, when minimizing of error is goal, because it separates the hierarchical false positives from true positives. Both sparsity and scaling assumptions lead to soft thresholding with smoother null distribution and hard thresholding with preserve peaks and likelihood median of posterior distribution, using *Gibbs* sampling based on

Markov chain Monte Carlo (MCMC) algorithm. FDR optimization is particularly useful for hard and soft thresholding, when sparsity and scale of large simultaneous hypothesis testing is involved in the threshold definition and selection. However, dependency in wide genome dataset, threshold location and posterior mean or median thresholding for large simultaneous hypothesis tests are still problems in front of threshold selection. Therefore, in GWAS determining the significance threshold that separates true and reliable associations from random noises, required more revision in the approaches. Although, genetic parameters and hyper-parameters have significant impacts over definition and optimization of FDR threshold (Chen et al, 2019). In this study, we evaluate different thresholding approaches and their performance ability to optimize the false positive rates in the context of a phenotypic and genotypic NUE dataset of 221 bread wheat genotypes. We emphasize that the focus on threshold definition and its statistical inference could be able to determine significant threshold with reliable results. Therefore, to compare distinguished FDR thresholding approaches, the objectives of this study include: (i) Identifying an appropriate regularization parameter for a given FDR threshold approach and (ii) Optimizing genetic parameters and hyper-parameters in the given FDR threshold approach, and (iii) Demonstrating the performance of the best threshold approach through empirical Bayes coherent behavior and compliance estimation.

2.2. Method

Phenotypic dataset:

In this study, a set of 221 bread wheat genotypes received from breeding innovations in wheat for resilient cropping systems (BRIWECS) project were widely grown in agricultural research station Klein-Altendorf, University of Bonn, 50°37'8.5"N, 6°59'25.4"E, for three cropping seasons 2017-18, 2018-19 and 2019-20. In order to limit competition effects, all genotypes were sorted by maturity date and were planted in layout of split-plot design, in two replication, six blocks and 1326 plots and the related traits to NUE were recorded. NUE is defined wheat as GY per unit of N supplied from soil or fertilizer (Moll et al, 1982):

$$NUE = \frac{GY}{N_s} = (\frac{N_t}{N_s})(\frac{GY}{N_t})$$

Where, GY is grain yield (gr/m^2), N_s nutrient supplied and N_t total above-ground plant nutrient at maturity.To analyze N and genotypes as main effects and N × G, during three agronomic years, a combine analysis of variance (ANOVA) on grain yield (GY), grain nitrogen yield (GNY), straw yield (SY), straw nitrogen yield (SNY) and NUE were performed using additive main effects and multiplicative interaction models (AMMI) with *R/agricolae*. To deal with outliers in the NUE vector, they were kept, because they were reflecting the actual field values across all years. The residuals distribution was checked using Shapiro normality test. To focus more on the quality of the vector, 2000 times the repeated random samples with replacement from original NUE vector was simulated using R/*bootstrap*, then Bayesian bootstrap *p-value*(s), was calculated.

Genomic dataset:

In order to characterize NUE vector among 221 bread wheat genotypes, a platform of 150K affymetrix SNP Chip at TraitGenetics GmbH (SGS GmbH Gatersleben, Germany), was used. After checking SNPs deviated from the Hardy-Weinberg equilibrium (HWE) only 22489 polymorphic SNP markers, were remained and used in GWAS model as genomic file (Sadeqi et al, 2023). The SNPs with MAF ≤ 0.05, ≤ 0.01, ≤ 0.005 and ≤ 0.001, respectively were removed due to monomorphism in the marker (Fadista et al, 2016). To detect the regions that might be involved in LD, based on the pruned marker information, neighbored LD between adjacent SNPs (D') within 200Mb with promised to physical position of SNPs on each chromosome and genetic correlation between two loci (r^2) with MAF value (Joiret et al, 2022), was calculated using R/*Synbreed*.

GWAS models and adjusted *p-value*(s) generation:

Within single locus association model with R/*rrBLUP*, was fitted in the adjusted form the mixed linear model as $y = X\beta + Zu + e$, where y is the trait vector, X is the fixed effects matrix, β is the vector of coefficients including principal components and population structure, Z the matrix of random SNP effects coded as (-1, 0 and 1), $V(u) = K\sigma_g^2$, where K is the GRM as kinship matrix and σ_g^2 is additive genetic variance with IBS basis. It was removed from the model due to convergence of N × Y to zero. GWAS multi-locus association model with R/*mlmm.gwas*, in form $y_{i=1}^{n} = \mu + \sum_{j=1}^{m} M_{.j}\, \beta_j + e$, where $y_{i=1}^{n}$ is NUE vector with n genotypes, m is total number of SNPs, M_{ij} is the matrix of random SNP effects coded as (0, 1 and 2) and β_j is the vector of SNP effects and H_0 denoted in form of $\beta = \sigma_g^2 = 0$. Based on both GWAS models, after rejection of $H_0 = \beta = 0$ for each SNP, the vector *–log $_{10}$(raw p-value(s))* was retrieved. Due to high type I error and high number of false positives in the raw *p-value*(s) produced by the given GWAS model, it is uninformative and does not provide a reliable vector regarding FDR thresholding approaches. To enhance reliability and reproducibility of FDR approaches it is necessary to adjusting the raw *p-value*(s). Therefore, proportion of null raw *p-value*(s) with 2000 times bootstrap random SNPs (with replacement) from the original raw *p-value*(s) vector was estimated using R/*fdrtool.* Based on bootstrapping technique, tuning hyper-parameter $(\lambda = 0.05)$ to use in linear threshold approaches, and based on smoothing technique, the tuning hyper-parameter $(\lambda = 0.01)$ to use in non-linear threshold approaches, was taken respectively. Then based on given GWAS model (*rrBLUP* or *mlmm*) and through distinguished linear FDR thresholding approaches the adjusted *p-value*(s) vector at low and high N levels, was generated (Boca and Leek, 2018; Jafari and Ansari-Pour, 2019).

FDR thresholding based on linear approaches:

For the FDR thresholding based on linear approaches, six common methods will be considered include following adjustments using R/*FDRestimation*:

1- *Bonferroni* correction is defined by the following function (Nakagawa 2004):

$$Bonf_{p-value} = \sum_{i=1}^{m} ML(\pi_i \leq \frac{\alpha}{m})$$

Where, $Bonf_{p-value}$ is corresponded for Bonferroni correction *p-value* generated from hypothesis tests, m is number of hypothesis tests, ML is maximum likelihood of multi comparisons when H_0 is rejected and a type I error is produced, π_i is linear threshold coefficient factor for ith paired comparisons and α is confidence interval of paired comparisons (usually 0.05). Bonferroni correction only accounts for number of H_0 tests, separately while, definition does not have a component to cover the relation between hypothesis tests and FWER, simultaneously.

2- *Holm* threshold adjustment utilizes the same function which is in the Bonferroni correction, only instead of upper bound $\frac{\alpha}{m}$ the equality is in the definition following (Giacalone et al, 2018):

$$Holm_{p-value} = \sum_{i=1}^{m} ML(\pi_i = \frac{\alpha}{m})$$

With this equality the type II error increases lower than Bonferroni correction, but the other norms are getting same explanation.

3- *Hochberg* adjustment is defined with the function (Chen et al, 2017):

$$Hoch_{p-value} = \sum_{i=1}^{m} ML(\pi_i \leq \frac{\alpha}{m-i+1})$$

The Hochberg threshold adjustment conducts statistical inference of hypothesis by starting with the maximum *p-value* from H_m to H_i and then to H_0. Moreover, due to upper bound $\frac{\alpha}{m-i+1}$ the *p-value*(s) are taking weights. This weights vector minimize the bias in the FWER.

4- *Benjamini-Hochberg* procedure controls the FDR using function (Benjamini and Hochberg, 1995):

$$Ben - Hoch_{p-value} = \sum_{i=1}^{m} ML(\pi_i \leq \frac{\alpha(m+1)}{2m})$$

Due to upper bound $\frac{\alpha(m+1)}{2m}$ in the function, the ratio false discoveries to all discoveries at 0.01 significance leads to estimate the non-false discovery rate (NFDR) in the procedure. However, the procedure has been constructed for adjusted *p-value*(s).

5- *Benjamini-Yekutieli* procedure controls the FDR using function (Benjamini and Yekutieli, 2001):

$$Ben - Yeku_{p-value} = \sum_{i=1}^{m} ML(\pi_i \leq \frac{\alpha(m+1)}{2m[\ln(m)+1]})$$

Due to logit bound $\frac{\alpha(m+1)}{2m[\ln(m)+1]}$ in the function, and based on adjusted *p-value*(s), the ratio of false positives to all discoveries at 0.01 significance might to find optimal value for threshold line.

6- *Sidak* adjustment is defined with the function (Chen et al, 2017):

$$Sidak_{p-value} = \sum_{i=1}^{m} ML(\pi_i \leq 1 - \frac{\alpha}{m})$$

In the function $1 - \frac{\alpha}{m}$ is complementary event to minimize the bias in the FWER.

FDR thresholding based on non-linear approaches:

For the FDR thresholding based on non-linear approaches, two distinguished methods will be considered including:

1- *q-value* threshold was utilized to minimize the error variance of threshold in the posterior distribution of adjusted *p-value*(s) using the following function (Storey and Tibshirani, 2003) using Bioconductor/*q-value*:

$$\widehat{qFDR} = minP(\pi_0(\lambda) = 0 \,|z \in [c, \infty))$$

where, $qFDR$ is corresponded to the *q-value* generated from function, P is posterior probability of type I errors when $\pi_0(\lambda) = 0$ and H_0 is rejected, $f(c)$ is proper cumulative distribution of given adjusted *p-value*(s) when $P(Z \geq c)$ and $f(z)$ is the distribution of null *z-value*(s).

2- *Local FDR* was utilized to concentrate on tail-area calculations over adjusted *p-value*(s) with Bayesian inference, using the following function (Efron and Tibshirani 2002) using Bioconductor/*twilight*:

$$LFDR = P((\pi_0(\lambda) = 0 \,|Z = z_i) \text{ if } \lambda = \frac{\pi_0 f(z_0)}{f(z_i)}$$

where, $LFDR$ is corresponded for FDR value generated from function, P is posterior probability of type I errors when $\pi_0(\lambda) = 0$ and H_0 is rejected, z_i is a t-statistic comparing pairwise SNP associations with *N* (0, 1) distribution under null hypothesis, λ is proportion of true positives to all hypotheses and $f(z_i)$ calculated based on EBayes rule is the posterior distribution estimate at median point.

FDR threshold optimization:

Due to check generalization performance in FDR thresholding, the optimization method including regularization and penalization, is designed to determine significant threshold with high accuracy. In the given threshold approach, sparsity assumption has been measured with regularization parameter and scaling assumption with penalty parameter. In the FDR thresholding based on linear approaches, ML and upper bound of each function including $(\alpha\ , m)$ were taken as regularization parameter and penalty parameter, respectively. Also, for the FDR thresholding based on non-linear approaches, $\pi_0(\lambda)$ and $f(z)$ of each function were taken as regularization parameter and penalty parameter, respectively using Python/*Scikit-learn*. Then *EBayes* = 15000 *Gibbs* samples with $\theta_{Gibbs}(\hat{\mu}\ , \hat{k})$ were generated, which $\hat{\mu}$ implies to regularization and $\hat{k}$ to penalty parameters. The prior distribution were chosen to be uninformative to mildly informative.

2.3. Results

Phenotypic quality control:

Combined ANOVA by considering the N level, year and replication as fixed effect showed that there was a significant difference between low, middle and high N applications. Also, there was a significant difference between the genotypes as random factor within three years. However the effect of N × G was significant for NUE and its related traits include GY, GNY, SY and SNY (Table 2.1). Therefore, this not significant N × G × Y = 75 indicates, the NUE observations could be used in the given GWAS model as phenotypic vector.

Table 2.1. Combine analysis of variance of NUE and its agronomic related traits at low, middle and high N levels, during 2018, 2019 and 2020.

SOV	*df*	GY	GNY	SY	SNY	NUE
Y	2	2478000***	360***	1751***	0.0070***	1067***
N	2	1380000000***	3523***	211535***	1.4513***	130883***
G	441	225800000***	39***	69902***	0.6343**	112752***
N × G	882	411700 ^{ns}	10***	705 ^{ns}	0.0048***	430***
N × Y	2	170500000***	13***	144976***	0.1227***	45862***
G × Y	220	785200***	1.40 ^{ns}	1422***	0.0025 ^{ns}	104 ^{ns}
N × G × Y	440	562700 ^{ns}	1.41 ^{ns}	1107 ^{ns}	0.0029*	75 ^{ns}

Mean squares: ns: not significant, *, ** and ***: significant at level 0.05, 0.01 and 0.001, respectively.

Due to deal with outliers in the NUE vector belong to 221 bread wheat genotypes on three years, it has been averaged across three years by mean 24.181(%) and standard error 11.16 (%). In the histogram plot (Figure 2.1), the mean of Shapiro *p-value* = 0.0307 and the mean of Shapiro *p-value* = 0.0652 for low and high N, respectively across three years, which indicates the NUE vector is following normal distribution. We assumed that this test is not enough to decide on normality of NUE vector, so to focus more on the quality of observations, 2000 times the repeated random vectors with replacement from original NUE vector was generated. The mean Bayesian bootstrap *p-value* = 0.0263 and the mean Bayesian bootstrap *p-value* = 0.0641 for low and high N, is close to two-tailed 95% confidence interval (CI) of mean Shapiro *p-value*(s), which confirms the distribution of NUE vector is normal.

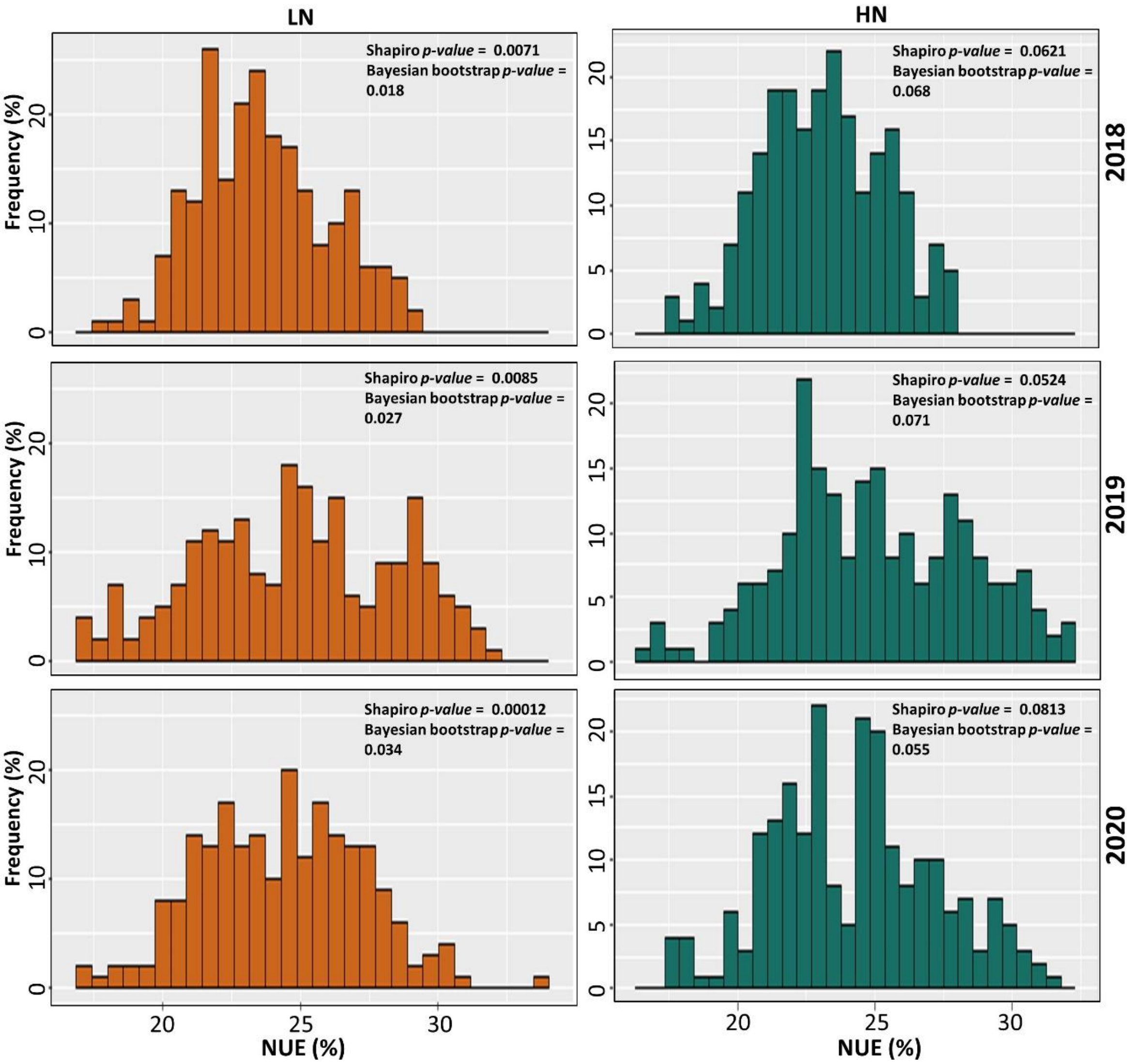


Figure 2.1: NUE vector quality control via Shapiro *p-value* and Bayesian bootstrap *p-value*, both *p-values* refers acceptable deal with outliers in the vector at low and high N levels, 2018, 2019 and 2020. The mean Bayesian bootstrap p-value = 0.0263 and the mean Bayesian bootstrap p-value = 0.0641 for low and high N, is close to 95% CI of mean Shapiro *p-value*(s), which verifies distribution of NUE vector is normal.

Genomic quality control:

Applying various fixed MAF values less than or equal to 0.001, 0.005, 0.01 and 0.05 to the platform of 150K affymetrix chip with 22489 SNP polymorphic among our 221 bread wheat genotypes, we found as expected at MAF ≤ 0.05 and as well at *–log₁₀ (p-value)* = 4.75, we would have the highest value of LD threshold (r^2 = 0.95) (Figure 2.2). While with the other MAF values the LD thresholds are still high (r^2 ≥ 0.80), it seems that genomic datasets with very low MAF have low heterogeneity and are then less informative. Therefore this genome wide LD threshold could be utilized in the given GWAS model and the linear FDR approaches such as Bonferroni correction for 150K SNP variants with MAF ≥ 0.05 and the FDR line *–log₁₀ (p-value)* ≅ 5.

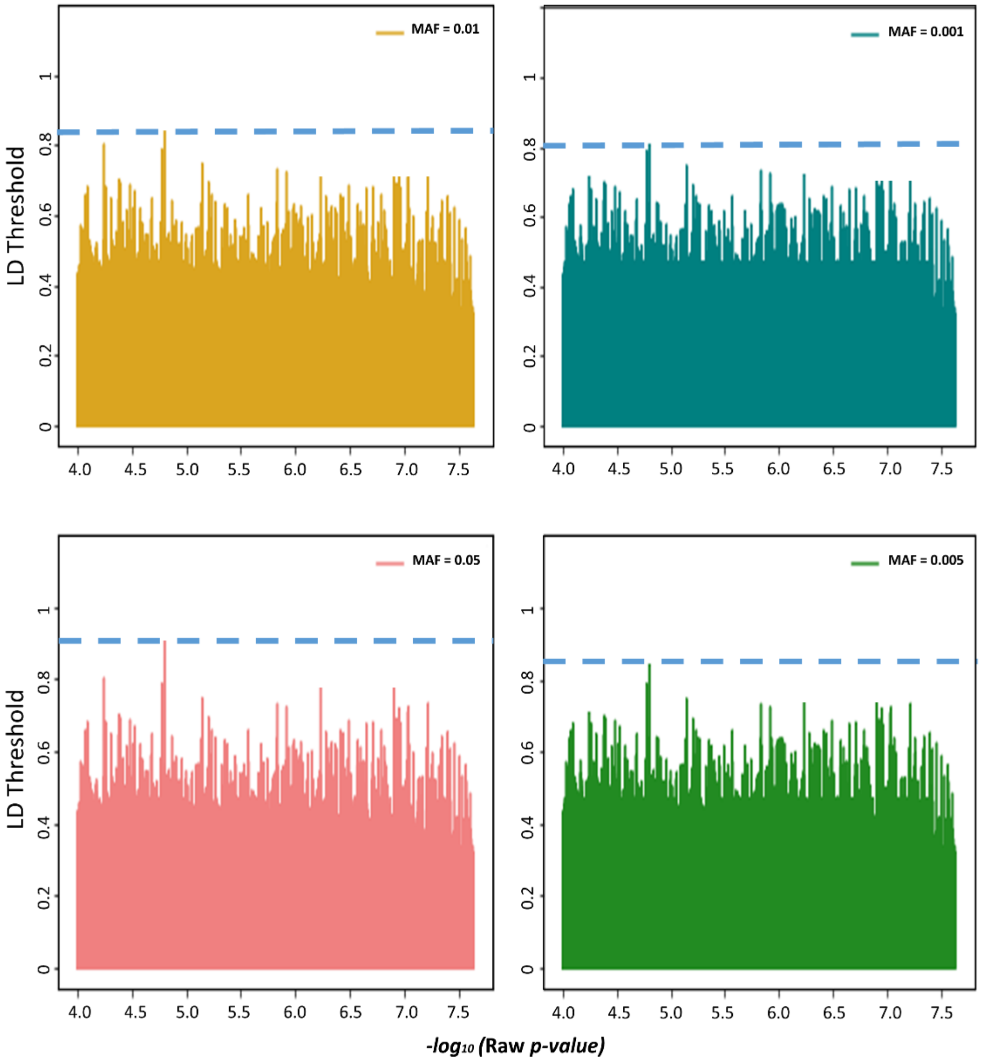


Figure 2.2: Visualization of fixed MAF values less than or equal to 0.001, 0.005, 0.01 and 0.05 to the platform of 150K affymetrix chip with 22489 SNP polymorphic among 221 bread wheat genotypes. LD threshold was specified as dash line for each MAF values. With MAF ≤ 0.05 and–log$_{10}$ (*p-value*) = 4.75, the highest value of LD threshold was received.

Generation of Adjusted *p-value*(s):

The impact of raw *p-value*(s) for NUE vector, received from single locus (*rrBLUP*) and multi locus associations (*mlmm*) GWAS models, under low and high N was revealed (Figure 2.3). Our findings on adjusted *p-value*(s) show, based on *rrBLUP* model, the likelihood of NUE vector for all linear FDR thresholding approaches is equal to 0.06 at low and high N levels. Due to type I error and high number of false positives in the raw *p-value*(s) produced by the *rrBLUP* GWAS model, the high bias in the adjusted *p-value*(s) vector is observed. Therefore it is provide this evidence that the raw *p-value*(s) received from this GWAS model could not be prone to use in the FDR thresholding approaches. In contrast, based on *mlmm* model, the likelihood of NUE vector is different depend to linear FDR thresholding approach. The *Bonferroni* correction and *Benjamini-Hochberg* approaches might not cover all range of raw *p-value*(s) received from *mlmm* GWAS model. In contrast, both *Hoch* and *Sidak* approaches have upper bounds $\frac{\alpha}{m-i+1}$ and $1-\frac{\alpha}{m}$, respectively, which minimize the bias in the FWER estimation. In addition this linear approaches cover very well all range of raw *p-value*(s) received from *mlmm* GWAS model, which implies to acceptable approaches to adjust the raw *p-value*(s) at low and high N levels.

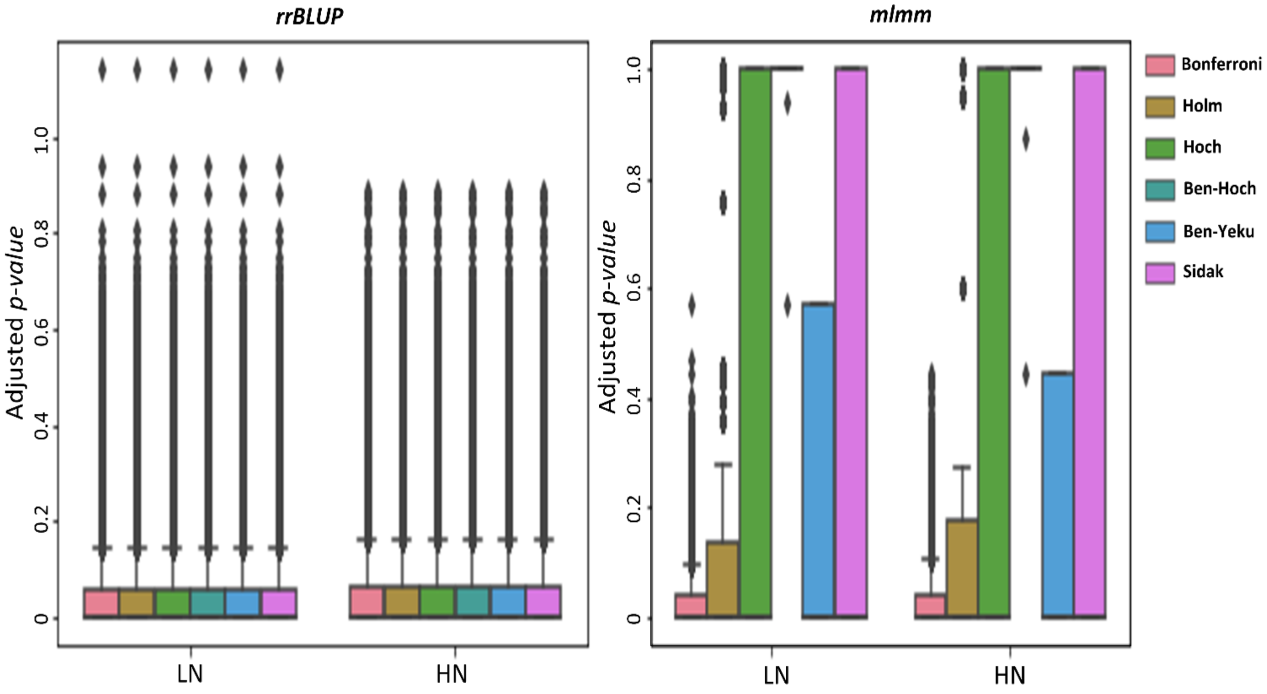


Figure 2.3: The impact of raw *p-value*(s) for NUE vector, received from left: single locus (*rrBLUP*) and right: multi locus associations (*mlmm*) GWAS models, under low and high N for *Bonferroni* correction, *Holm*, *Hoch*, *Benjamini-Hochberg*, *Benjamini-Yekuteili* and *Sidak* adjustments were investigated respectively. Both *Hoch* and *Sidak* approaches cover very well all range of raw *p-value*(s) for NUE vector received from *mlmm* GWAS model, which implies to acceptable approaches to adjust the raw *p-value*(s) at low and high N levels.

FDR thresholding based on linear approaches:

The performance of each FDR linear approach was highly dependent on the characteristics of the upper and lower bunds of their functions (Figure 2.4). The results based on the six distinguished FDR linear approaches demonstrated that FDR threshold line 0.05 could not be true value to make decision on false positives, which is common value in GWAS models. In addition the results show that the scaled version of adjusted *p-value* (s) doesn't follow easily distribution of raw *p-value*(s) and it is risky to accept the results of this linear approaches. However, in all approaches at low and high N levels, ML of null hypotheses was defined as regularization parameter in the lower bund to detect false positives. In contrast in the upper bound only the *Hoch* and *Sidak* linear adjustments could clearly provide false positives and we could assume that both related upper bounds $\frac{\alpha}{m-i+1}$ (in the *Hoch* adjustment) and $1-\frac{\alpha}{m}$ (in the *Sidak* adjustment) were defined proper to receive reliable and reproducible results in aim to FDR thresholding. The number of false positives in upper bound of both *Hoch* and *Sidak* adjustments are the same range. But when comparisons were extended to the lower bounds, the distribution of false positives were different.

FDR thresholding based on non-linear approaches:

We evaluated the performance of *q-value*(s) and *LFDR* as distinguished non-linear in the context of thresholding (Figure 2.5). The interested finding is that the density of false positives in both methods using the scaled version of adjusted *p-value* (s) are same at low and high N levels, which is significantly reduced the risk of soft thresholding against FDR linear approaches. Based on definition, for both non-linear approaches in order to identify prior probability of associations hypotheses, the ($\pi_0(\lambda)$) as regularization parameter, were estimated. Due to coincidence in the distribution tails, there is no significant difference in performance between *q-value*(s) and $\pi_0(\lambda)$ = 0.388 and 0.375 at low and high N levels, respectively. Due to genetic parameters and hyper-parameters such as MAF, LD and SNP high density, for the *LFDR* thresholding curve against $\pi_0(\lambda)$ = 0.388 and 0.375 at low and high N, there is no quite coincidence seems. In spite of no coincidence in the LFDR to $\pi_0(\lambda)$ distribution, $f(z_i)$ function in the definition could be estimated using EBayes algorithm. Therefore *LFDR* in the upper tail area at low and high N levels shows coherent behave for the FDR threshold, which make it reasonable accurate upon large scale simultaneous problem.

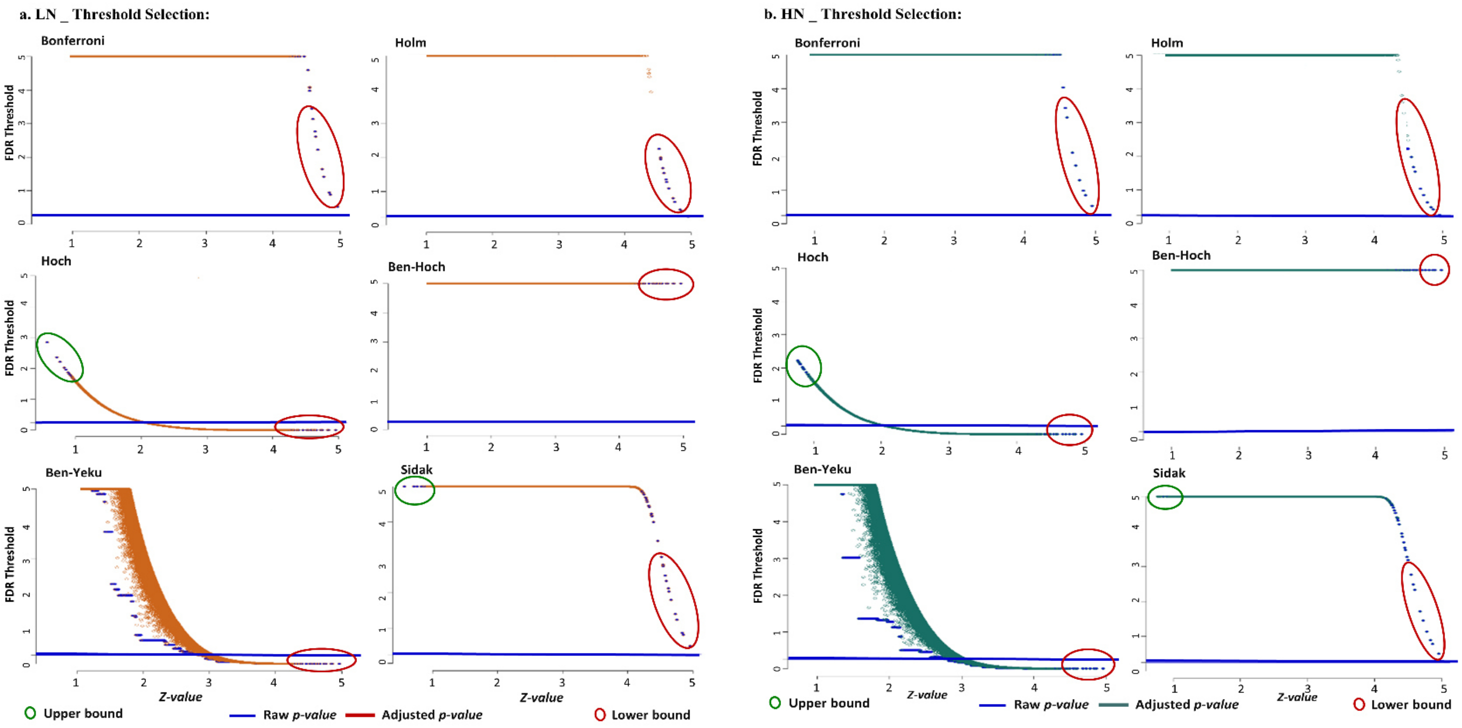


Figure 2.4: FDR thresholding based on the *Bonferroni* correction, *Holm*, *Hoch*, *Benjamini-Hochberg*, *Benjamini-Yekuteili* and *Sidak* linear approaches among adjust and raw *p-value*(s) at low (a) and high (b) N levels. The performance of each FDR linear approach was highly dependent on the characteristics of the upper and lower bunds of their functions. Only the *Hoch* and *Sidak* linear adjustments in the upper and lower bounds could clearly provide false positives.

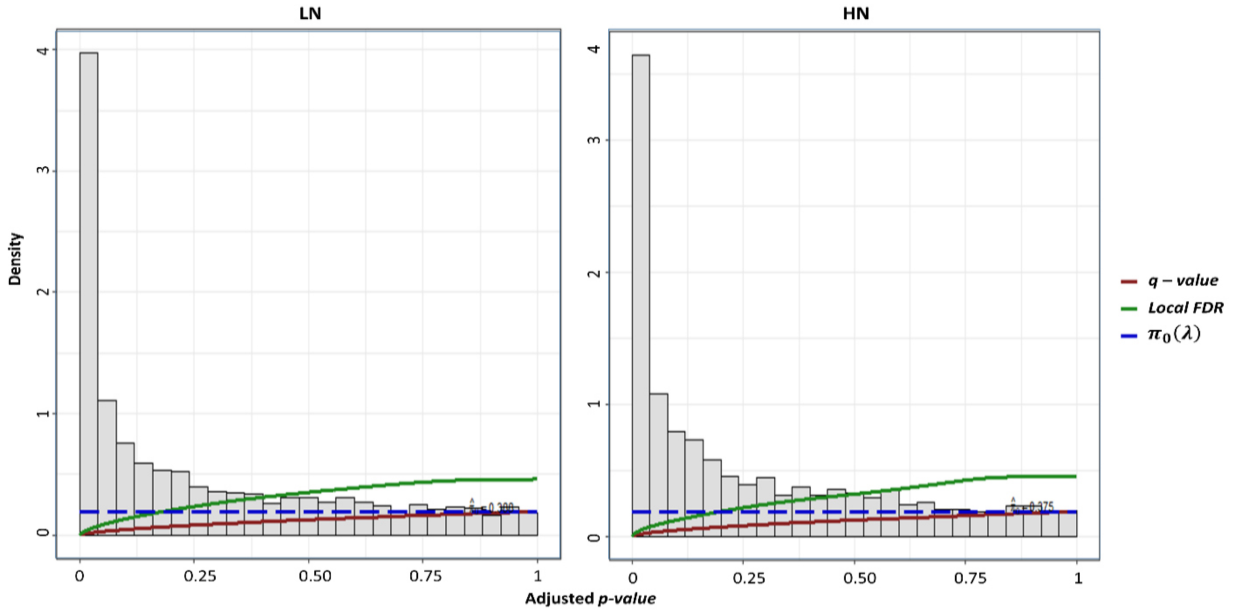


Figure 2.5. The performance of *q-value*(s) and *LFDR* as distinguished non-linear in the context of thresholding. Density of false positives in both methods using the adjusted *p-value* (s) are same at low and high N levels, which is significantly reduced the risk of soft thresholding. Due to coincidence in the distribution tails, there is no significant difference in performance between *q-value*(s) and $\pi_0(\lambda)$. *LFDR* in the upper tail area at low and high N levels shows coherent behave for the FDR threshold, which make it reasonable accurate upon large scale simultaneous problem.

FDR threshold optimization and selection:

To optimize the FDR threshold in the linear approaches, ML was estimated while it was faced with bias in the results. Even bias-variance analysis could not minimized this negative effect on the heavily tail simultaneous distributions at all. However, this problem was observed in the *Benjamini-Yekuteili* approach clearly and in the other linear approaches more or less. Consequently, the estimation of ML as lower bund was with bias and upper bounds with high variance. Therefore, the bounds could not estimate the significant FDR threshold and they were removed from optimization. Then, we evaluated the results of FDR non-linear approaches include *q-value*(s) and *LFDR* using *EBayes* FDR as index with uninformative to mildly informative in the prior distribution of false positives (Figure 2.6). The *EBayes* false positives histogram was obtained with 15000 Gibbs samples in two dimensions include regularization parameter ($\hat{\mu} = 5.8$) for sparsity and penalty parameter ($k = 0.25$) for scaling optimization. The results show that posterior distribution in the *LFDR* might be a good demonstration in deal with false positives especially in heavily tail areas. Because *LFDR* approach has cumulative distribution function of given adjusted *p-value*(s) in the tails as prior distribution. This cumulative function optimizes the sparsity and the scale of false positives very well in contrast to the *q-value*(s).

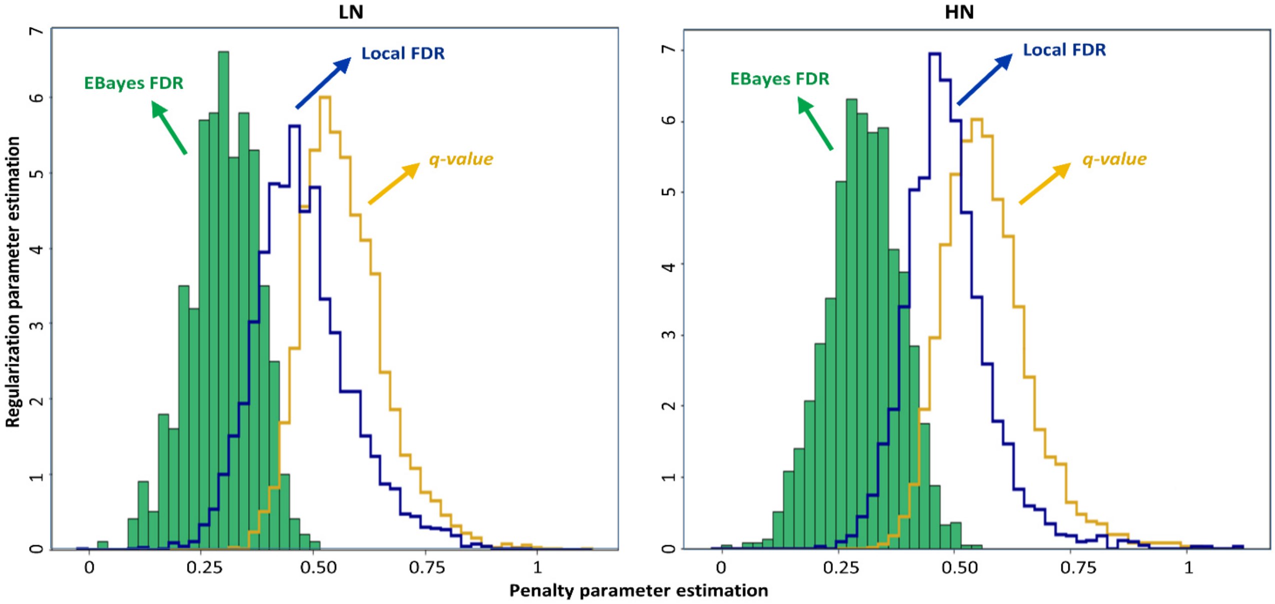


Figure 2.6: Posterior distribution estimation for *q-value*(s) and Local FDR thresholding approaches at low and high N using sparsity (with regularization parameter) and scaling (with penalty parameter) assumptions. In the solid area histogram *EBayes* = 15000 *Gibbs* samples with $\hat{\mu} = 5.8$ and $k = 0.25$ at low N and $\hat{\mu} = 5.8$ and $k = 0.25$ at high N were generated. For the Local FDR with dark blue line histogram: $\hat{\mu} = 6.0$ and $k = 0.50$ at low N and $\hat{\mu} = 6.9$ and $k = 0.50$ at high N. For the *q-value*(s) with orange line histogram: $\hat{\mu} = 5.9$ and $k = 0.60$ at low N and $\hat{\mu} = 5.8$ and $k = 0.60$ at high N.

2.4. Discussion

In GWAS, genetic parameters such as population structure, genomic relationship matrix, marker density, sample size and MAF have large effects on the performance and accuracy of the model (Wang et al., 2016). However, they play an important role in estimating the significant FDR threshold. In GWAS, one side of the model is assigned to the complex feature such as NUE. To reduce the multicollinearity and heteroscedasticity in the NUE observations, model evaluation and error measurement are necessary. A combined ANOVA including N level, year and replication as a fixed effect showed that there was a significant interaction between genotype and environment at low and high N levels with respect to NUE. Therefore, NUE is considered a complex trait for breeding and is controlled by many genes with minor effects or small p-value(s). Moreover, in the GWAS model, the null hypothesis often states that there is no association between a particular genetic variant and the trait of interest. The selection of the significance threshold is a crucial parameter to determine reliable associations between genetic variants and complex traits. The significance threshold is set to control and minimize the type I error rate, i.e. the probability that a true null hypothesis is falsely rejected. It helps researchers to distinguish true positive associations from random false positive associations with reasonable accuracy. In addition, the reliability and reproducibility of significant associations will be robust and applicable to different populations. The choice of significance threshold depends on various factors, such as MAF, LD threshold, crude or adjusted p-value(s) and the desired balance between sensitivity and specificity of the FDR approach. The crude p-value or crude p-values represent the probability that a test statistic will be as extreme as the one obtained in the study, assuming that the null hypothesis is true or that there is no association between the genetic variant and the complex trait. The FDR thresholding based on raw *p-value*(s) is affected by noises on false positives. Therefore to improve the robustness and validity of the identified FDR threshold, we generated the adjusted *p-value*(s) vector from the raw *p-value*(s) using bootstrap technique. To address this issue we applied different FDR thresholding in linear form including *Bonferroni* correction, *Holm*, *Hoch*, *Benjamini-Hochberg*, *Benjamini-Yekuteili* and *Sidak* adjustment, or non-linear form including *q-value*(s) and local FDR in platform of empirical Bayes estimation. While a few reports have shown in general that FDR thresholding is a non-linear concept (Emmert-Streib and Dehmer, 2019; Cao et al, 2023) we believe that sparsity and scale of false positives in the large simultaneous hypothesis tests are two important features in the FDR thresholding approach. Sparsity refers to the proportion of true null hypotheses in a set of hypotheses being tested. When there is high sparsity, meaning that a large proportion of hypotheses are true nulls, the FDR threshold could be accurate. Our results confirm that linear approaches are accompany with less stringent FDR threshold, it might be not appropriate because the overall rate of false discoveries is controlled even if a higher proportion of individual discoveries turn out to be false positives. However, in the all FDR linear approaches at low and high N levels, ML of null hypotheses was defined as regularization parameter in the lower bund to detect false positives. But ML estimation was accompanied with bias in the results. Even bias-variance analysis could not minimized this negative effect on the heavily tail simultaneous distributions at all. Whereas, both *q-value*(s) adjustment and *LFDR* are computed based on proportion of false positives through the function $\pi_0(\lambda) = 1$. Scaling refers to adjusting the significance threshold based on the number of null hypotheses, the coincidence behave between distribution of adjusted *p-value*(s) and raw *p-value*(s) and overall complexity of GWAS model. However, adjusting the FDR threshold based on scaling assumption is crucial when dealing with a large number of hypotheses to control the overall rate of false positives. In the context of FDR thresholding, *EBayes* index could be applied to predict the performance *q-value*(s) and *LFDR*, which incorporate the prior information that a given test statistic corresponds to a null hypothesis given the genomic data. The *EBayes* usually involve shrinkage or smoothing of estimates, which could improve the stability of *LFDR* estimates, especially when dealing with sparse data or low scaled where individual estimates may be unreliable due to small sample sizes. Heterogeneity often occurs in the distribution of true and null effects across different genomic features. However, the results of FDR thresholding are affected with bias in the estimation. *EBayes* algorithm could be designed to account for this heterogeneity, leading to more accurate *LFDR* estimates in the heavily tails of specific genomic regions. However, we evaluated the results of FDR non-linear approaches include *q-value*(s) and *LFDR* using *EBayes* algorithm as powerful index with uninformative to mildly informative in the prior distribution of false positives. In summary, *q-value*(s) and *LFDR* serve similar purposes in controlling false positives in high-dimensional data but have different focuses and applications. The choice between them depends on the approach definition, goals of the analysis and the characteristics of the data being examined. The *q-value*(s) offer a global control for false positives, while *LFDR* provides more reasonable information at the genotype test level. Therefore in our study, in spite of no coincidence in the $LFDR$ to $\pi_0(\lambda)$, the $f(z_i)$ function in the definition could be estimated with $EBayes$ algorithm. Moreover, $LFDR$ shows coherent behavior for the FDR threshold in the upper tail region at low and high N levels, making it a reasonable approach, while other common fits point to the large-scale simultaneity problem. In addition, the interpretation of false positives in GWAS requires some caution when using linear FDR adjustments in sparse and large genomic data. We encountered some limitations in defining the FDR threshold, particularly in the upper bounds of linear and nonlinear approaches. We emphasize that empirical null distributions based on permutation methods might be useful when the assumption of linear or parametric FDR approaches does not hold. Nevertheless, we believe that it is necessary to use modern statistical optimization techniques to evaluate the stability and performance of our results and to select a significant FDR threshold. By incorporating the neural network algorithm, it is possible to improve the reliability of the FDR threshold and increase the probability of identifying true genetic associations while minimizing the risk of false positives in GWAS.

Acknowledgments:

The authors would like to acknowledge the kind support of DAAD who generously provided us the opportunity to pursue our research and financing the study. Also, the authors are grateful to the Institute for Crop Science and Resource Conservation (INRES)-Plant Breeding, University of Bonn, Germany, for responsible supervision and execution plan during the study.

Author contribution:

M.B.S. problem statement, performing the experiments , data collection and analysis, writing original draft; A.B. supervision of the experiments, draft editing; S.D. resources, provided the improved marker data with the physical map; N.S. and A.P.K. supported data collection; M.K. supported data collection and editing manuscript draft; J.L. data analysis and interpretation, study supervision and funding acquisition.

Data availability statement:

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of interest:

The authors declare that they have no conflict of interest.

2.5. References

Asif, H., Alliey-Rodriguez, N., Keedy, S., Tamminga, C. A., Sweeney, J. A., Pearlson, G., Clementz, B. A., Keshavan, M. S., Buckley, P., Liu, C., Neale, B., and Gershon, E. S. (2020). Gwas significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Molecular Psychiatry*, 26(6), 2048–2055. https://doi.org/10.1038/s41380-020-0670-3

Benjamini, Y., and Bogomolov, M. (2013). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1), 297–318. https://doi.org/10.1111/rssb.12028

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4). https://doi.org/10.1214/aos/1013699998

Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds, C. F., &amp; Butters, M. A. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology, 23(2), 255–264. https://doi.org/10.1037/a0012850

Boca, S. M., and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6. https://doi.org/10.7717/peerj.6035

Brinster, R., Köttgen, A., Tayo, B. O., Schumacher, M., and Sekula, P. (2018). Control procedures and estimators of the false discovery rate and their application in low-dimensional settings: An empirical investigation. *BMC Bioinformatics*, 19(1). https://doi.org/10.1186/s12859-018-2081-x

Cao, Y., Sun, X., and Yao, Y. (2023). Controlling the false discovery rate in transformational sparsity: Split knockoffs. *Journal of the Royal Statistical Society Series B*: Statistical Methodology. https://doi.org/10.1093/jrsssb/qkad126

Chen, D., Liu, C., and Xie, J. (2016). Multi-locus test and correction for confounding effects in genome-wide association studies. *The International Journal of Biostatistics*, 12(2). https://doi.org/10.1515/ijb-2015-0091

Chen, S.-Y., Feng, Z., and Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*, 9(6), 1725–1729. https://doi.org/10.21037/jtd.2017.05.34

Chen, X., Robinson, D. G., and Storey, J. D. (2019). The functional false discovery rate with applications to genomics. Biostatistics, 22(1), 68–81. https://doi.org/10.1093/biostatistics/kxz010

Cook, J. P., Mahajan, A., and Morris, A. P. (2016). Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *European Journal of Human Genetics*, 25(2), 240–245. https://doi.org/10.1038/ejhg.2016.150

Cormier, F., Foulkes, J., Hirel, B., Gouache, D., Moënne-Loccoz, Y., & Le Gouis, J. (2016). Breeding for increased nitrogen-use efficiency: A review for wheat (*Triticum. aestivum* L.). *Plant Breeding*, 135(3), 255–278. https://doi.org/10.1111/pbr.12371

Efron, B. (2013). Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press, ISBN-13:978-1107619678

Efron, B., and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1), 70–86. https://doi.org/10.1002/gepi.1124

Emmert-Streib, F., and Dehmer, M. (2019). Large-scale simultaneous inference with hypothesis testing: Multiple testing procedures in practice. *Machine Learning and Knowledge Extraction*, 1(2), 653–683. https://doi.org/10.3390/make1020039

Fadista, J., Manning, A. K., Florez, J. C., and Groop, L. (2016). The (in)famous GWAS *p-value* threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8), 1202–1205. https://doi.org/10.1038/ejhg.2015.269

Furmańczyk, K. (2013). Some remarks on the control of false discovery rate under dependence. *Applicationes Mathematicae*, 40(3), 297–307. https://doi.org/10.4064/am40-3-3

Giacalone, M., Agata, Z., Cozzucoli, P. C., and Alibrandi, A. (2018). Bonferroni-Holm and permutation tests to compare health data: Methodological and applicative issues. *BMC Medical Research Methodology*, 18(1). https://doi.org/10.1186/s12874-018-0540-8

Hastie, T., Tibshirani, R., and Wainwright, M. (2020). Statistical learning with sparsity: The lasso and generalizations. *CRC Press*, Taylor &amp; Francis Group, ISBN-13:978-1498712163

Haynes, W. (2013). Bonferroni Correction. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds) *Encyclopedia of Systems Biology*. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_1213

Jafari M, Ansari-Pour N. (2019). Why, When and How to Adjust Your *p-values*? *Cell J.* 20(4):604-607. https://doi.org/10.22074/cellj.2019.5992.

Joiret, M., Mahachie John, J. M., Gusareva, E. S., and Van Steen, K. (2022). Correction: Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Mining*, 15(1). https://doi.org/10.1186/s13040-022-00296-9

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1716-1

Krzywinski, M., and Altman, N. (2014). Comparing samples—part II. *Nature Methods*, 11(4), 355–356. https://doi.org/10.1038/nmeth.2900

Lai, Y. (2017). A statistical method for the conservative adjustment of False Discovery Rate (Q-value). *BMC Bioinformatics*, 18(S3). https://doi.org/10.1186/s12859-017-1474-6

Lee, S., and Lee, D. K. (2020). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*, 73(6), 572–572. https://doi.org/10.4097/kja.d.18.00242.e1

Lozano, A. C., Ding, H., Abe, N., and Lipka, A. E. (2023). Regularized multi-trait multi-locus linear mixed models for genome-wide association studies and genomic selection in crops. *BMC Bioinformatics*, 24(1). https://doi.org/10.1186/s12859-023-05519-2

McCaw, Z. R., Colthurst, T., Yun, T., Furlotte, N. A., Carroll, A., Alipanahi, B., McLean, C. Y., and Hormozdiari, F. (2022). Deepnull models non-linear covariate effects to improve phenotypic prediction and association power. *Nature Communications*, 13(1). https://doi.org/10.1038/s41467-021-27930-0

Menyhart, O., Weltz, B., and Győrffy, B. (2022). Correction: Multipletesting.com: A tool for life science researchers for multiple hypothesis testing correction. *PLOS ONE*, 17(9). https://doi.org/10.1371/journal.pone.0274662

Midway, S., Robertson, M., Flinn, S., and Kaller, M. (2020). Comparing multiple comparisons: Practical guidance for choosing the best multiple comparisons test. PeerJ, 8. https://doi.org/10.7717/peerj.10387

Moll, R. H., Kamprath, E. J., & Jackson, W. A. (1982). Analysis and interpretation of factors which contribute to efficiency of nitrogen utilization. *Agronomy Journal*, 74(3), 562–564. https://doi.org/10.2134/agronj1982.00021962007400030037x

Nakagawa, S. (2004). A Farewell to Bonferroni: The Problems of Low Statistical Power and publication bias. *Behavioral Ecology*, 15(6), 1044–1045. https://doi.org/10.1093/beheco/arh107

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, 316(7139), 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236

Rathan, N. D., Krishna, H., Ellur, R. K., Sehgal, D., Govindan, V., Ahlawat, A. K., Krishnappa, G., Jaiswal, J. P., Singh, J. B., SV, S., Ambati, D., Singh, S. K., Bajpai, K., and Mahendru-Singh, A. (2022). Genome-wide association study identifies loci and candidate genes for grain micronutrients and quality traits in wheat (*Triticum aestivum* L.). *Scientific Reports*, 12(1). https://doi.org/10.1038/s41598-022-10618-w

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes&nbsp; usingfalse discovery rate controlling procedures. *Bioinformatics*, 19(3), 368–375. https://doi.org/10.1093/bioinformatics/btf877

Sadeqi, M. B., Ballvora, A., and Léon, J. (2023). Local and bayesian survival FDR estimations to identify reliable associations in whole genome of bread wheat. *International Journal of Molecular Sciences*, 24(18), 14011. https://doi.org/10.3390/ijms241814011

Salim, N., & Raza, (2019). A. Nutrient use efficiency (NUE) for sustainable wheat production: A Review. *Journal of Plant Nutrition*,; 43(2), 297–315. https://doi.org/10.1080/01904167.2019.1676907

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44(7), 825–830. https://doi.org/10.1038/ng.2314

Sørensen, E. S., Jansen, C., Windju, S., Crossa, J., Sonesson, A. K., Lillemo, M., and Alsheikh, M. (2022). Evaluation of strategies to optimize training populations for genomic prediction in oat (avena sativa). *Plant Breeding*, 142(1), 41–53. https://doi.org/10.1111/pbr.13061

Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B: Statistical Methodology, 64(3), 479–498. https://doi.org/10.1111/1467-9868.00346

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genome wide studies. *Proceedings of the National Academy of Sciences,* 100(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

Tan, Y.-D., and Xu, H. (2014). A general method for accurate estimation of false discovery rates in identification of differentially expressed genes. *Bioinformatics*, 30(14), 2018–2025. https://doi.org/10.1093/bioinformatics/btu124

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1). https://doi.org/10.1038/s43586-021-00056-9

VanderWeele, T. J., and Mathur, M. B. (2018). Some desirable properties of the Bonferroni correction: Is the Bonferroni Correction really so bad? *American Journal of Epidemiology*, 188(3), 617–618. https://doi.org/10.1093/aje/kwy250

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J. M., Xu, S., and Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports*, 6(1). https://doi.org/10.1038/srep19444

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., Wang, S.-B., Dunwell, J. M., Zhang, Y.-M., and Wu, R. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings in Bioinformatics*, 18(5), 906–906. https://doi.org/10.1093/bib/bbx028

Wilson, D. J. (2019). The harmonic mean *p-value* for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4), 1195–1200. https://doi.org/10.1073/pnas.1814092116

Zabaneh, D., and Mackay, I. J. (2003). Genome-wide linkage scan on estimated breeding values for a quantitative trait. *BMC Genetics*, 4(1). https://doi.org/10.1186/1471-2156-4-s1-s61

Software description:

R/*agricolae:*

Original idea was presented in the thesis "A statistical analysis tool for agricultural research" to obtain the degree of Master on science, National Engineering University (UNI), Lima-Peru. Some experimental data for the examples come from the CIP and others research. *Agricolae* offers extensive functionality on experimental design especially for agricultural and plant breeding experiments, which can also be useful for other purposes. It supports planning of lattice, Alpha, Cyclic, Complete Block, Latin Square, Graeco-Latin Squares, augmented block, factorial, split and strip plot designs. There are also various analysis facilities for experimental data, e.g. treatment comparison procedures and several non-parametric tests comparison, biodiversity indexes and consensus cluster (the canonical link https://CRAN.R-project.org/package=agricolae to use).

R/*bootstrap:*

Software (bootstrap, cross-validation, jackknife) and data for the book "An Introduction to the Bootstrap" by B. Efron and R. Tibshirani, 1993, Chapman and Hall. This package is primarily provided for projects already based on it, and for support of the book. New projects should preferentially use the recommended package *bootstrap* (the URL link https://gitlab.com/scottkosty/bootstrap to use).

R/*rrBLUP:*

Software for genomic prediction with the RR-BLUP mixed model (Endelman 2011, <doi:10.3835/plantgenome2011.08.0024>). One application is to estimate marker effects by ridge regression; alternatively, BLUPs can be calculated based on an additive relationship matrix or a Gaussian kernel.

R/*mlmm.gwas:*

Pipeline for Genome-Wide Association Study using Multi-Locus Mixed Model from Segura V, Vilhjálmsson BJ et al. (2012) <doi:10.1038/ng.2314>. The pipeline include detection of associated SNPs with MLMM, model selection by lowest eBIC and raw *p-value* threshold, estimation of the effects of the SNPs in the selected model and graphical functions.

R/*Synbreed*:

This package provides a framework for the analysis of genomic prediction data (Genomic Selection, GWAS, QTL-mapping) within an open source software (the URL link https://synbreed.r-forge.r-project.org/ to use).

R/*fdrtool*:

Estimates both tail area-based false discovery rates (FDR) as well as local false discovery rates (fdr) for a variety of null models (p-values, z-scores, correlation coefficients, t-scores). The proportion of null values and the parameters of the null distribution are adaptively estimated from the data. In addition, the package contains functions for non-parametric density estimation (Grenander estimator), for monotone regression (isotonic regression and antitonic regression with weights), for computing the greatest convex minorant (GCM) and the least concave majorant (LCM), for the half-normal and correlation distributions, and for computing empirical higher criticism (HC) scores and the corresponding decision threshold (the canonical link https://CRAN.R-project.org/package=fdrtool to use).

R/*fdrestimation*:

The user can directly compute and display false discovery rates from inputted *p-values* or z-*scores* under a variety of assumptions. p.fdr() computes FDRs, adjusted *p-values* and decision reject vectors from inputted *p-values* or z-*values*. get.pi0() estimates the proportion of data that are truly null. plot.p.fdr() plots the FDRs, adjusted p-values, and the raw *p-values* points against their rejection threshold lines (the canonical link https://CRAN.R-project.org/package=FDRestimation to use).

R/Bioconductor/*qvalue*:

This package takes a list of p-values resulting from the simultaneous testing of many hypotheses and estimates their *q-values* and local FDR values. The *q-value* of a test measures the proportion of false positives incurred (called the false discovery rate) when that particular test is called significant. The local FDR measures the posterior probability the null hypothesis is true given the test's *p-value*. Various plots are automatically generated, allowing one to make sensible significance cut-offs. Several mathematical results have recently been shown on the conservative accuracy of the estimated *q-values* from this software. The software can be applied to problems in genomics, brain imaging, astrophysics, and data mining (the URL link http://github.com/jdstorey/qvalue to use).

R/Bioconductor/*twilight*:

In a typical microarray setting with gene expression data observed under two conditions, the local false discovery rate describes the probability that a gene is not differentially expressed between the two conditions given its corresponding observed score or *p-value* level. The resulting curve of *p-values* versus local false discovery rate offers an insight into the twilight zone between clear differential and clear non-differential gene expression. Package 'twilight' contains two main functions: Function twilight.pval performs a two-condition test on differences in means for a given input matrix or expression set and computes permutation based *p-values*. Function twilight performs a stochastic downhill search to estimate local false discovery rates and effect size distributions. The package further provides means to filter for permutations that describe the null distribution correctly. Using filtered permutations, the influence of hidden confounders could be diminished (the URL link http://compdiag.molgen.mpg.de/software/twilight.shtml to use).

Python/*Scikit-learn*:

Model selection and evaluation (the URL link https://scikit-learn.org/stable/model_selection.html to use).

International Journal of
Molecular Sciences

MDPI

Article

Local and Bayesian Survival FDR Estimations to Identify Reliable Associations in Whole Genome of Bread Wheat

Mohammad Bahman Sadeqi, Agim Ballvora * and Jens Léon

INRES-Plant Breeding, Rheinische Friedrich-Wilhelms-Universität Bonn, 53113 Bonn, Germany; mbsadeghi1@gmail.com (M.B.S.); j.leon@uni-bonn.de (J.L.)
* Correspondence: ballvora@uni-bonn.de

Abstract: Estimating the FDR significance threshold in genome-wide association studies remains a major challenge in distinguishing true positive hypotheses from false positive and negative errors. Several comparative methods for multiple testing comparison have been developed to determine the significance threshold; however, these methods may be overly conservative and lead to an increase in false negative results. The local FDR approach is suitable for testing many associations simultaneously based on the empirical Bayes perspective. In the local FDR, the maximum likelihood estimator is sensitive to bias when the GWAS model contains two or more explanatory variables as genetic parameters simultaneously. The main criticism of local FDR is that it focuses only locally on the effects of single nucleotide polymorphism (SNP) in tails of distribution, whereas the signal associations are distributed across the whole genome. The advantage of the Bayesian perspective is that knowledge of prior distribution comes from other genetic parameters included in the GWAS model, such as linkage disequilibrium (LD) analysis, minor allele frequency (MAF) and call rate of significant associations. We also proposed Bayesian survival FDR to solve the multi-collinearity and large-scale problems, respectively, in grain yield (GY) vector in bread wheat with large-scale SNP information. The objective of this study was to obtain a short list of SNPs that are reliably associated with GY under low and high levels of nitrogen (N) in the population. The five top significant SNPs were compared with different Bayesian models. Based on the time to events in the Bayesian survival analysis, the differentiation between minor and major alleles within the association panel can be identified.

Keywords: GWAS; local FDR; Bayesian survival analysis; wheat genome; grain yield

check for updates

Citation: Sadeqi, M.B.; Ballvora, A.; Léon, J. Local and Bayesian Survival FDR Estimations to Identify Reliable Associations in Whole Genome of Bread Wheat. *Int. J. Mol. Sci.* **2023**, *24*, 14011. https://doi.org/10.3390/ijms241814011

Academic Editor: Xiaojun Nie

Received: 3 August 2023
Revised: 2 September 2023
Accepted: 7 September 2023
Published: 12 September 2023

1. Introduction

Bread wheat (*Triticum aestivum* L.) is among one of the major crops for food security worldwide which alone contributes 20% of the protein and calories in our daily diet. Increasing Grain yield (GY) is the main goal of wheat breeding programs. GY is a complex trait controlled by many genes with small effects. Under field conditions, both genetic and environmental factors, and interaction between genotype and environment (G × E) affect GY [1]. However it is observed that the analysis of GY related parameters presents challenge. Because the observations should be independent and identical and the residuals must follow normal distribution. In practice, they are affected by genetic and environmental factors and the distribution of data is far from normal [2]. There is only one mean vector obtained from the observations, which is not sufficiently informative to confirm whether the distribution is normal or not. One of the most popular tools to compute true confidence interval for the mean and standard error is bootstrap [3], which is a resampling statistical device to measure the accuracy of observed bias and variance [4] in complex traits like GY. In field experiments, genotypes have random effects, which lead to some outliers among the observations. The best linear unbiased predictors (BLUPs) is utilized in phenotypic mixed linear models to estimate random effects. BLUPs with frequentist perspective can be an analogy to

Bayesian inference in dealing with outliers [5]. In parallel, genetic estimated breeding value (GEBV) describes the individual genetic merit to produce superior offspring, and correlation between GEBVs and BLUPs is a common approach to check the quality of GY vector [6]. GEBVs of complex traits can be computed by BLUPs via genome wide association study (GWAS) models. So, GWAS is a successful and fast strategy for genetic dissection of complex traits in plants [7]. The objective of GWAS is to determine the presence of a significant relationship between genotypes and traits of interest. Therefore, understanding variations in rare phenotypes among complex traits and associations between large number of markers (typically SNPs) and a given trait are of great interest [8]. It is particularly useful when little information on the genetic control of a quantitative complex trait is available. The main challenge is to find significant reliable associations in GWAS results. Population structure and genomic relationship (kinship) matrix are two main components involved in the GWAS model that can reduce structural and systematic effects. Furthermore, there are some genetic hyper-parameters such as marker effect, minor allele frequency (MAF), number of call rate for each marker [9] and epistasis effects between loci [10], which are not presented in usual GWAS models, but can improve the power and accuracy of models significantly. Moreover, various GWAS models, such as single locus association, multi-locus association [11,12], Bayesian whole genome regression and whole genome variance, have potential to attain reliable significant associations. In the single locus association model, population structure and environmental factors are taken as fixed effects while phenotypic values and markers are considered as random effects [13]. It further requires Bonferroni correction with false discovery rate (FDR) of 0.05 for pairwise tests between the complex trait and SNPs. At this threshold level, although the risk of a false positive has been reduced, the chance of producing true positive associations is very low. As such, this model has led to many pseudo associations in the results of some studies [14]. Moreover, high over or low under-fitting in single locus association model due to high-density SNPs makes optimization difficult. Meanwhile, the multi-locus association model based on the restricted maximum likelihood (REML) method applies better FDR correction to reduce selection criterion, but due to pairwise comparisons, collinearity in this model remains high. It has also led to over- or under-fitting in the results of GWAS model [15]. Particularly, when the epistasis effects between loci are high, due to low bias and high variance in the model parameters, many pseudo associations arise in the results [16–18]. The whole genome variance model provides the analysis of genetic variance based on whole genome regression [19]. The probability of receiving reliable significant associations in this model is still low, due to multicollinearity and heteroscedasticity in model components. The single, multi-locus and whole genome variance models are based on linear regression algorithm, which has low efficiency for big genomic datasets with large-*p-value*-small effects. However, the Bayesian whole genome regression model based on prior and posterior probability distributions is a convenient algorithm to deal with these difficulties [20]. Therefore, the model's accuracy can be higher than three previous frequentist models, because the complex trait and SNPs are taking prior distribution [21]. In this model, SNP effects and dimensions of genomic dataset are interpreted with Bayes factor (BF), which is easier and more reliable against SNP *p-values* [22]. BF is a summary measure that provides an alternative to the *p-value* for ranking significant associations [23]. In addition, in this model, false coverage rate (FCR) is considered analogous to FDR [24]. FCR covers more dimensions among the selected intervals in the genomic file [25]. Nevertheless, determining the actual prior distribution for the GY vector and SNPs represents a difficult part of computation [26]. Genomic prediction (GP) as a promising technique in molecular breeding provides the possibility to consider the performance of GWAS models. Bayesian survival FDR analysis is a robust strategy based on probability theory to determine the true prior distribution in large-scale genomic datasets [17,27]. Moreover, this analysis is a regularization approach that can be applied as a GP model with high performance [28]. It attempts to estimate all genomic effects among GY vector, while the pseudo effects of covariates are reduced to zero within SNPs [29].

2. Local and Bayesian Survival FDR Analysis in GWAS

In the last decade, GWAS has been made feasible by high-throughput genomic technologies based on next-generation sequencing (NGS) techniques. When multiple pairwise association tests are performed between case and control individuals, it becomes difficult to minimize type I error (false positive error) in hypothesis testing simultaneously. Some efforts to control the family wise error rate (FWER) versus multiple pairwise errors were proposed by Bonferroni, then Benjamini [30], who proposed FDR control as an alternative approach to overall type I error. GWAS is very sensitive in identifying the genetic basis of complex traits such as GY with many agronomic components with small effects. The reasons for this sensitivity could be the many genetic variations with small effects among individuals and the variation in genetic structure of sub population. Thus, the power of common GWAS models, such as genome regression models, to control for the FDR threshold is very low to detect non null hypotheses for associations. Efron and Tibshirani developed an extension of FDR for large-scale simultaneous hypothesis testing, called local FDR based on MLE in scale of z-value versus *p-value* [31]. Large-scale genetic association studies are conducted in the hope of discovering signal SNPs involved in the association of complex traits. Identifying the correct FDR to decide between signal associations is the critical key of GWAS. Based on frequent perspective, the final extension of the correct threshold in GWAS is the local FDR approach proposed by Efron [32]. The local FDR approach is suitable for simultaneously testing many associations, based on the empirical Bayes perspective.

In the local FDR framework, the FDR is a measurement of posterior signal SNPs with rule:

$$P(z_i(\pi_0 f_0 + \pi_1 f_1 + \cdots + \pi_n f_n) > (1-\alpha)$$

where z_i is a t-statistic comparing pairwise SNP associations with $N(0,1)$ distribution under null hypothesis, π is proportion of true null hypothesis, f the mixture density estimate at midpoint of bin, calculated based on empirical Bayes rule of z_i with degree of freedom (df) for fitting the MLE of z_i, α is significant level of association test. The local FDR (*locFDR*) is then defined as:

$$locFDR(z_i) = \frac{z_i(\pi_0 f_0)}{\sum z_i(\pi_i f_i)}$$

Thus in *locFDR*, maximum likelihood estimator is sensitive to bias when the GWAS model includes two or more explanatory variables as genetic parameters. The results are based on maximum likelihood of marker effects, which is only a point estimation and there is no further information about prior distribution of the signal and followers of the signal associations.

However, the advantage of the Bayesian perspective is that the knowledge of prior distribution comes from other genetic parameters included in the GWAS model, such as linkage disequilibrium (LD) analysis, MAF and call rate of significant associations. In the current study, we propose to solve the multicollinearity and large-scale problems in the GY vector as a complex trait in bread wheat and large-scale SNP information with Bayesian survival analysis. The aim of this study is to obtain a short list of SNPs that are reliably associated to GY vector among three levels of N. The hypothesis to test, for each SNP = SNP_1,. . ., SNP_p when pairwise comparisons between censored and relapsed observations, are conditionally independent, $x_{SNP} \sim P(\theta_{SNP})$, so x_{SNP} is GY values for each genotype. We suppose the SNP effect as parameter space of $\theta_{SNP} \in \Theta$, where Θ is an unobserved scalar parameter that can be partitioned in two parts (Θ_0, Θ_1) such that:

H_0: $\theta_{SNP} \in \Theta_0$ or SNP_i is not significantly associated with GY among population.

H_a: $\theta_{SNP} \in \Theta_1$ or SNP_i is significantly associated with GY among population.

In Bayesian survival framework, the FDR is a measurement of posterior signal SNPs with rule:

$$lP(\theta_{SNP} \in \Theta_1 \, |x_{SNP}) > (1-\alpha)$$

Where α takes MAF of SNP in the pairwise comparisons of censored and returned observations.

The Bayesian survival FDR (bsFDR) is then defined:

$$bsFDR(\lambda_{SNP}) = \frac{\sum_{SNP} P(\theta_{SNP} \in \Theta_1 \,|\lambda_{SNP}.\theta_{SNP})}{\sum_{SNP}(\lambda_{SNP}.\,\theta_{SNP})}$$

where λ_{SNP} is Bayes survived factor that indicate SNP effect based on MAF value for GY of each genotype.

Estimating a significant FDR threshold in GWAS is still a major challenge to distinguish true positive hypotheses from false positives and negative errors. Several comparative methods for multiple testing have been developed to determine the significance threshold; however, these methods may be overly conservative and lead to an increase in false negative errors. Here, we developed two empirical methods to determine the statistical significance threshold based on SNP- heritability of GY as a target trait. To test the locFDR and bsFDR methods for significance threshold under different distinguished GWAS models, we used the mean of results from three years of field experiments on GY vector in bread wheat under low N (LN) and high N (HN) applications.

3. Results

3.1. Grain Yield Quality Control

The GY observations obtained from 221 bread wheat genotypes under HN treatment were averaged across three years with a mean value of 6400 and standard division of 145.16 (gr/m^2). In the histogram plot, the Shapiro *p-value* = 0.0891, which indicates the GY vector follows a normal distribution. However, we considered this test inadequate, so to focus more on the quality of observations, the random vectors were repeated 2000 times with replacement from original GY vector was generated and Bayesian bootstrap *p-value* = 0.00325, which shows the Shapiro *p-value* is accurate (Figure 1b). Consequently, a confidence interval of 95% and standard error in simulated samples are close to original observations. Correlation between BLUPs and BLUEs as analog for EBVs is high (r = 0.814), showing that the effects of genotypes are distributed randomly and the quality of GY vector is acceptable to use in GWAS and GP models. Moreover, under LN treatment after removing outliers the Bayesian bootstrap *p-value* = 0.3048, which shows the Shapiro *p-value* = 0.2462 follows a normal distribution and is ready to apply in GWAS models (Figure 1a).

3.2. Population Structure Analysis

In Figure 2, based on the population of 221 genotypes, genetic data are divided into three clusters which implies these genotypes differ from each other in genetic content.

The variance is explained by PCA being normally distributed in ten PCs and the first three PCs with explained variance 20.5, 10.8 and 9.7%, respectively, are considerable as PC number in the GWAS analysis. The allocation of genotypes in related clusters could be attributed to their genetic descent and common ancestry. The F_{st} plot based on 150 K SNPs with MAF greater than 0.05, indicates how heterozygosity varies in sub-populations in comparison with the whole population. It is thus considered a threshold of calculated proportion of homozygosity in whole genome, which is clearly observed in Figure 2d and the proportion of homozygosity in the population is acceptable.

3.3. GWAS and GP Model Selection

To determine the best GWAS model, local FDR analysis on SNP *p-values* for each model was run.

In Figure 3a, the results under low N show that in the variance component model (*sommer*) the standard error is minimum (sigma = 0.048), but Bayesian whole genome regression (*NAM*) exhibits the highest Delta (μ) = 0.589 and highest Proportion (H$_0$) = 0.625, which shows that the false positive rate to reject H$_0$ via *sommer* under LN treatment is not high, so this minimum sigma = 0.048 is not reliable. For all GWAS models, the local FDR plot was revealed in one tail.

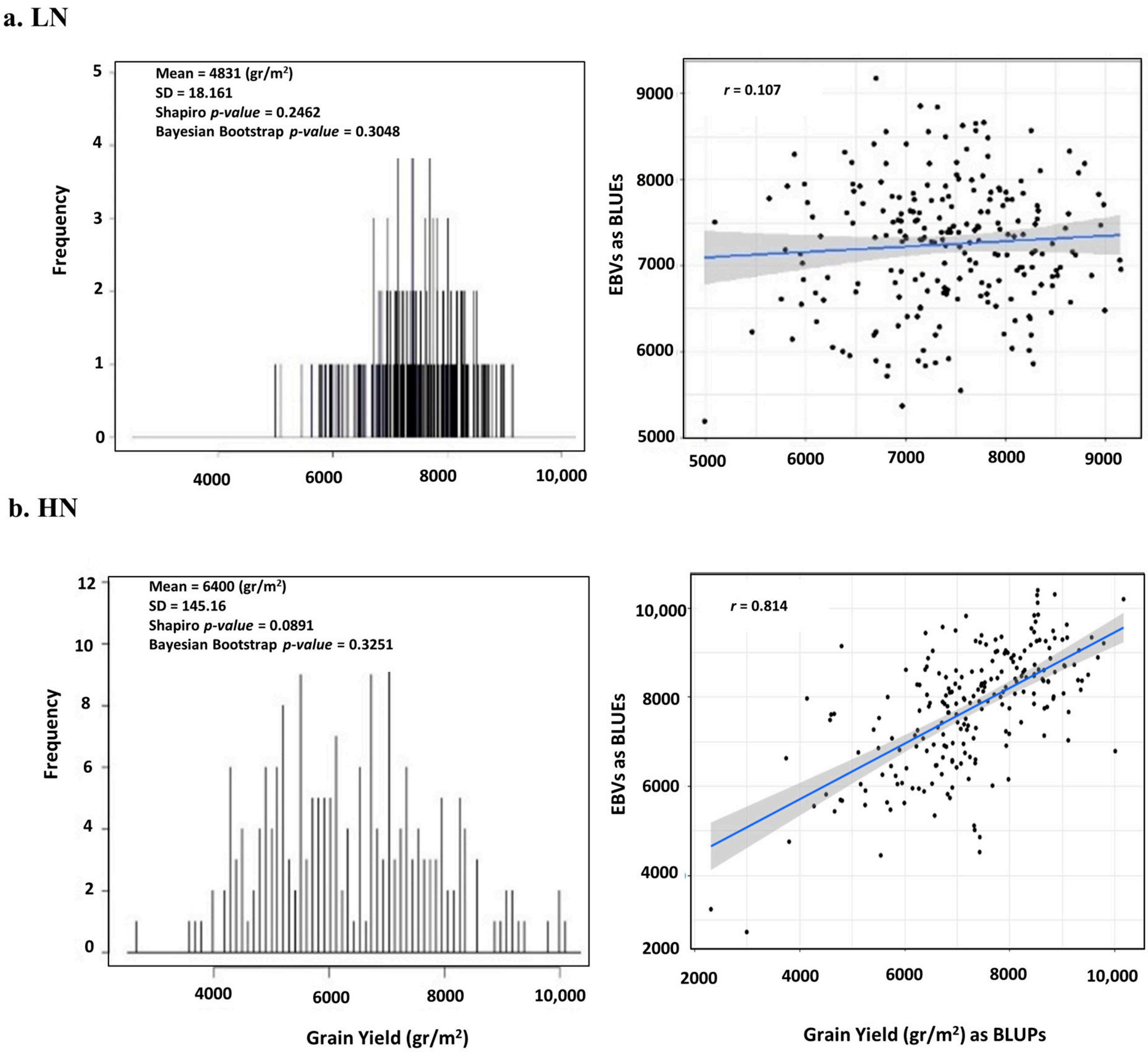


Figure 1. (**a**) Left, GY vector under LN treatment after removing outliers the Bayesian bootstrap *p-value* = 0.3048, which shows the Shapiro *p-value* = 0.2462 is following normal distribution, but on the right, the correlation between BLUPs and BLUEs as analog for EBVs is low (r = 0.107). (**b**) Left, GY vector under HN treatment, mean of three agronomic years 2018, 2019 and 2020, quality control via Shapiro *p-value* and Bayesian bootstrap *p-value*, both *p*-values represent acceptable management of outliers in the vector; right, the correlation between BLUPs and BLUEs as analog for EBVs is high (r = 0.814).

In Figure 3b, the results under high N show that for Bayesian whole genome regression (*NAM*) the standard error is minimum (sigma = 0.062), the Delta is highest (μ) = 0.661 and Proportion is highest (H_0) = 1.034, in front in variance component model, sigma = 0.664 is highest and the higher sigma with lower Delta (μ) = 0.358, and lower Proportion (H_0) = 0.433, representing lower model accuracy, so the *sommer* model was removed from further analysis.

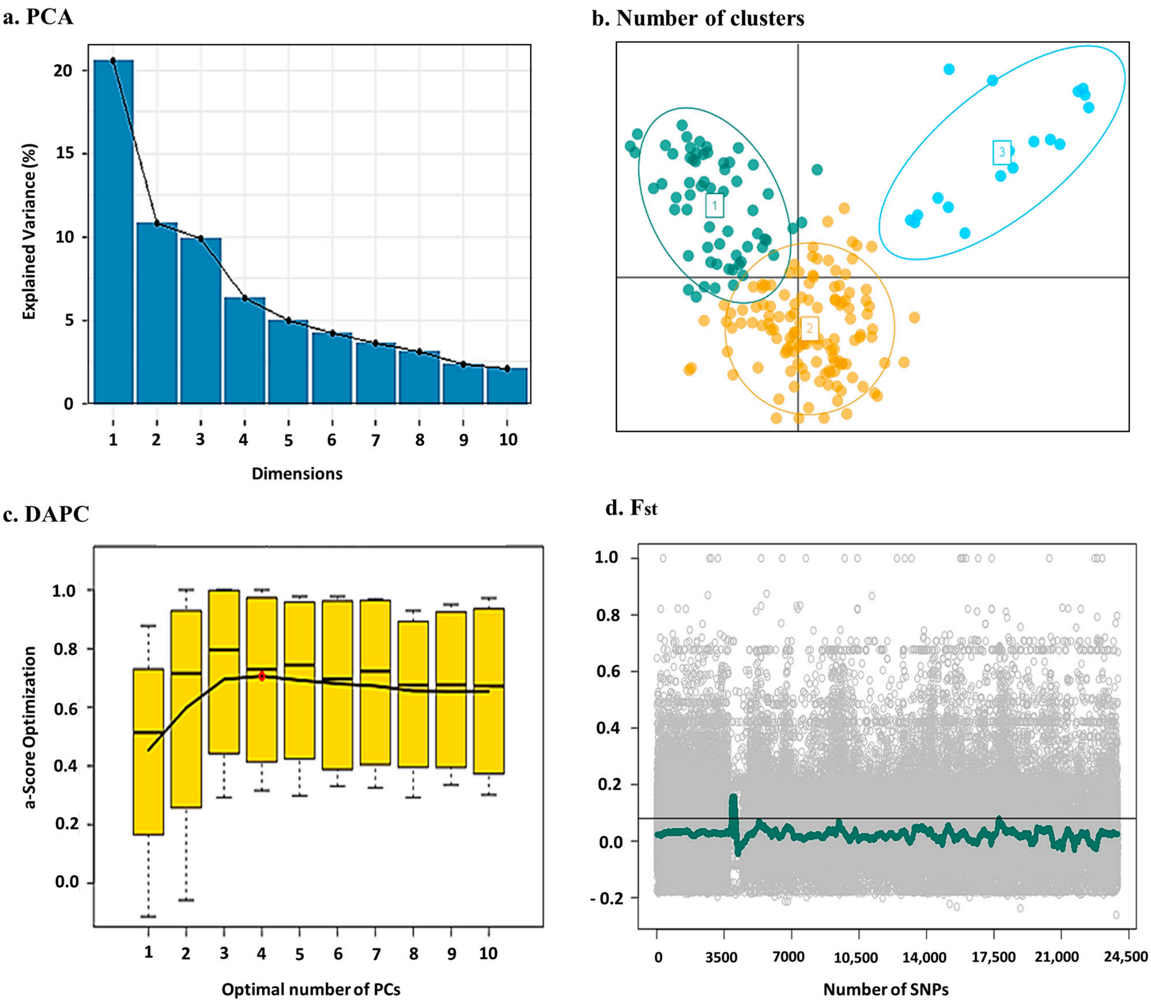


Figure 2. (**a**,**b**) PCA-explained variance and number of clusters by PCs based on 150 K SNP Chip on 221 bread wheat genotypes, (**c**) discriminate analysis of principle components with alpha-score based on genetic diversity between clusters, (**d**) fixation index using 150 K SNP Chips of the genetic variation among and within population.

For *rrBLUP* and *NAM*, the local FDR plot was one tail but in *mlmm* model the two tails of significant SNPs was revealed. To check the performance of the remaining three GWAS models, the Bayesian survival analysis based on highest significant association was performed for GY vector under LN and HN treatments separately, and the mean and standard error of survived SNPs ($S(SNP_i)$) were calculated. To measure the accuracy of each model, Akaike information criterion (AIC) and Bayesian information criterion (BIC) were calculated and the results are shown in Figure 4. Under LN, GY vector in the *rrBLUP* and *mlmm* models was the only major allele that survived with mean of $S(SNP_i)$ 0.547 and 0.481. However, in the NAM model both major and minor alleles during survival analysis have remained. Moreover, the NAM displayed the minimum standard error (SE $S(SNP_i)$ = 0.189) in comparison to the other two models. The AIC and BIC were 150,408 and 150,867, respectively, which are lower than *rrBLUP* and *mlmm*. For GY vector under HN condition, the bootstrapping method was applied and in the *mlmm* model only the major allele survived with mean of $S(SNP_i)$ 0.597. However, in the *rrBLUP* both major and minor alleles and in the NAM all major, minor, heterozygous and missing (NA) alleles during survival analysis remained. In addition, the NAM exhibited minimum SE $S(SNP_i)$ = 0.104 in comparison to the other two models. Here, the AIC and BIC values were also less than those against *rrBLUP* and *mlmm*. Therefore, both local FDR and Bayesian survival analyses have confirmed NAM to be the best GWAS model with minimum residual errors on the

SNPs. The Manhattan and QQ-plots for NAM model for GY vector under HN are shown in Figure 5a,b.

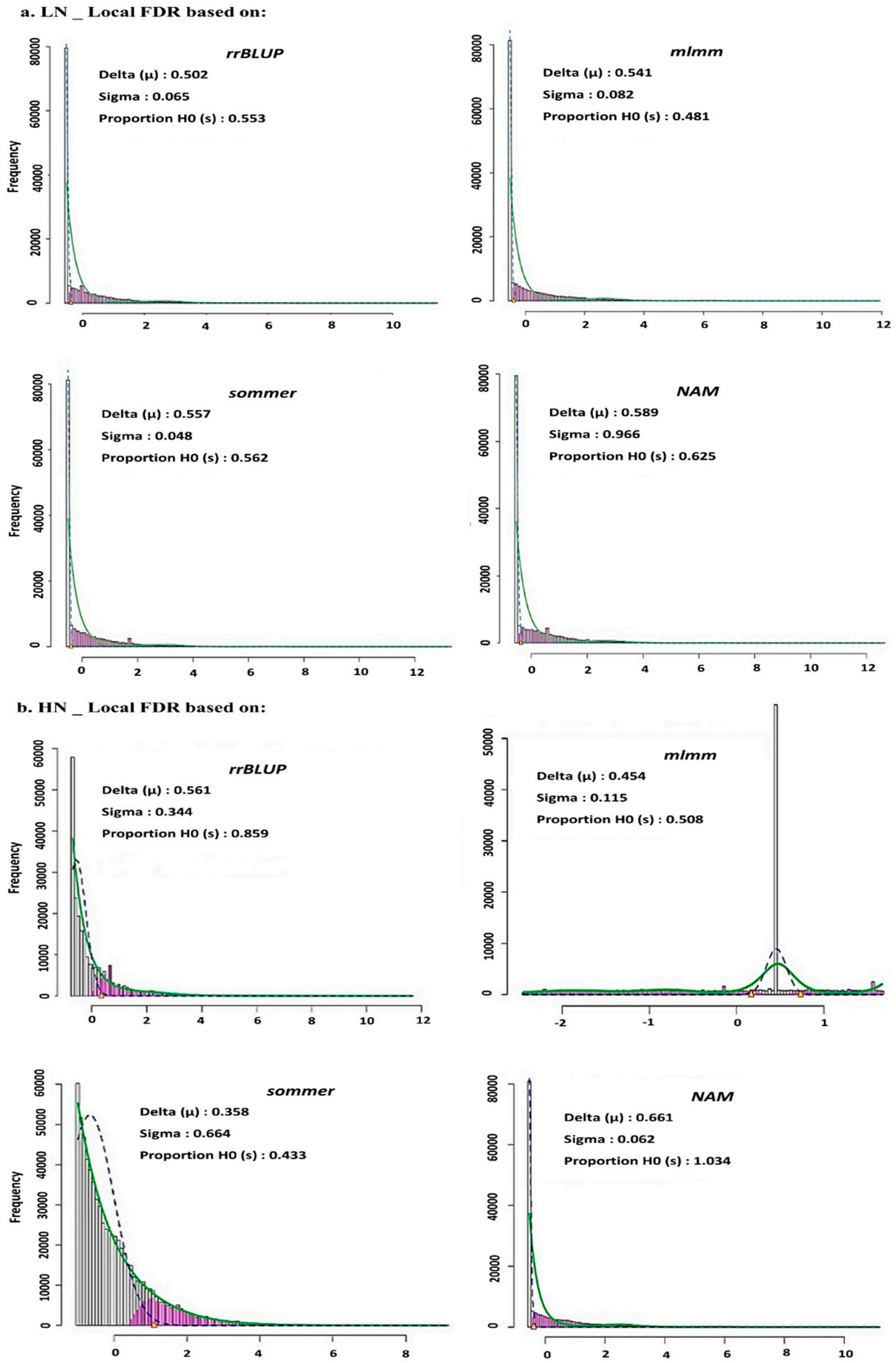


Figure 3. (**a**) GWAS model selection based on local FDR approach, in each model: Delta shows maximum likelihood for mean estimation, Sigma is standard error and Proportion of H_0 refers to proportion of false rejection hypotheses. All three estimators are important to determine the accuracy and performance of GWAS model for grain yield under low N fertilizer. (**b**) GWAS model selection based on local FDR approach, in each model: Delta shows maximum likelihood for mean estimation, Sigma is standard error and Proportion of H_0 refers proportion of false rejection hypotheses. All three estimators are important to determine the accuracy and performance of GWAS model for grain yield under high N fertilizer.

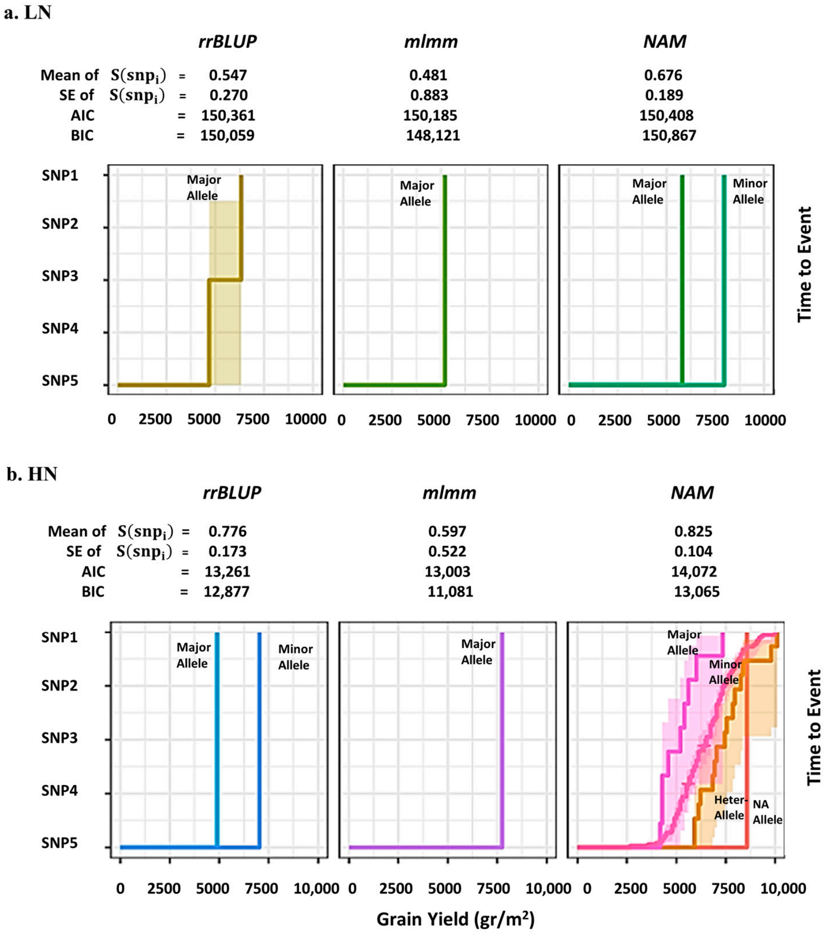


Figure 4. GWAS model selection based on Bayesian survival analysis, the minor and major alleles in each SNP were involved in computation based on time to events, which is necessary to attain expected SNP prior distribution. SNP1: AX-158576714, SNP2: AX-109898892, SNP3: AX-110385692, SNP4: AX-109470057 and SNP5: BS00094057_51. (**a**) Under LN, GY vector in the *rrBLUP* and *mlmm* models only major allele survived with mean of $S(SNP_i)$ 0.547 and 0.481, but in the NAM model both major and minor alleles during survival analysis remained. (**b**) Under HN condition, in the *mlmm* model only the major allele survived with mean of $S(SNP_i)$ 0.597, but in the *rrBLUP* both major and minor alleles and in the NAM all major, minor, heterozygous and missing (NA) alleles during survival analysis remained.

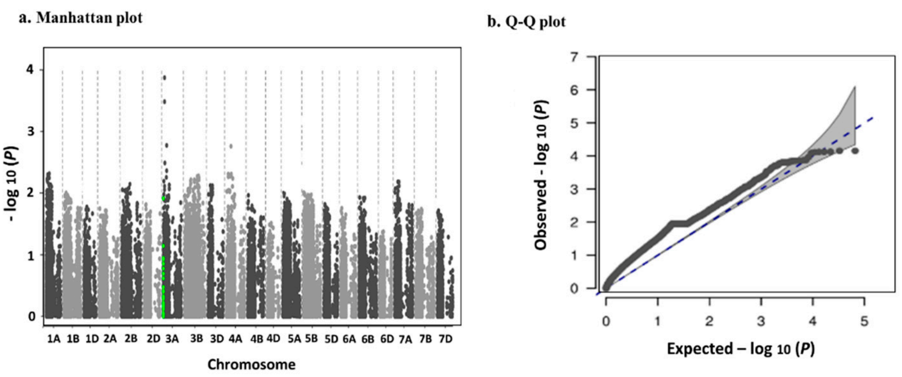


Figure 5. (**a**) Manhattan upon Bayesian whole genome regression GWAS model, the top five significant SNPs include Ax-110385692 in chr. 3A, Ax-158547970 in chr. 3A, Ax-158538619 in chr. 4A, Ax-158522989 in chr. 3A and Ax-109470057 in chr. 3A, were plotted with NAM package. (**b**) In QQ plot, all top five significant SNPs closely follow the Chi-square association line trend.

The association points well above the gray area in QQ-plot correspond to GWAS hits and in Manhattan plot by threshold $-log10$ (p = 4), the continuous points on chromosome region 3A specify the reliable SNPs, which show association with investigated trait (Tables 1 and 2).

Table 1. GWAS results using rrBLUP, mlmm, Sommer and NAM models for GY vector of 221 wheat genotypes under low N treatment.

GWAS Model	N Level	SNP	Chr [a]	Pos [b]	Maj. Allele [c]	MAF [d]	$-\log(p\text{-}Value)$ [e]	ML ($locFDR(z_i)$) [f]	P ($bsFDR(\lambda_{SNP})$) [g]
rrBLUP	LN	AX-158596644	2A	136,652,249	NA	---	4.456407631	0.502	0.195
	LN	AX-158568302	7D	55,688,624	NA	---	4.256507384		0.179
	LN	AX-109865927	6D	160,078,273	AA	0.176	3.669168314		0.235
	LN	AX-158576714	3A	69,099,202	CC	0.181	3.288885		0.237
	LN	AX-109898892	3A	73,588,989	CC	0.127	3.128732		0.261
mlmm	LN	AX-108915613	4A	79,319,176	NA	---	3.189661	0.541	0.112
	LN	AX-110600522	6A	111,630,421	CC	0.064	3.099851		0.178
	LN	AX-158544205	7D	24,335,511	CC	0.064	2.900164		0.181
	LN	AX-158612323	3A	3,508,832	NA	---	2.778897		0.233
	LN	AX-109299894	5A	97,047,183	TT	0.288	2.683941		0.163
Sommer	LN	AX-158544205	6A	24,335,511	CC	0.064	5.3026811	0.557	0.188
	LN	AX-108741100	6B	67,958,573	NA	---	5.1678802		0.175
	LN	AX-108915613	4A	79,319,176	NA	---	4.6011748		0.110
	LN	AX-109299894	5A	97,047,183	CC	0.108	3.6739810		0.122
	LN	AX-110600522	6A	121,657,421	NA	---	3.5461284		0.135
NAM	LN	AX-110385692	3A	47,067,572	TT	0.115	3.541840	0.589	0.288
	LN	AX-158547970	3A	46,623,733	AA	0.112	3.174208		0.276
	LN	AX-158608942	2A	64,279,945	GG	0.224	2.110374		0.234
	LN	AX-86178623	2A	248,384,372	TT	0.213	2.028563		0.260
	LN	wsnp_Ex_c351_689415	7B	706,808,922	CC	0.115	2.016566		0.251

[a] Chr: Chromosome, [b] Pos: Position (bp), [c] Maj. Allele: Major Allele, [d] MAF: Minor Allele Frequency, [e] $-\log(p$ *Value*) for five top significant SNPs received from different GWAS models, [f] ML $locFDR(z_i)$: maximum likelihood of five significant SNPs based on local FDR threshold in the GWAS model, [g] P ($bsFDR(\lambda_{SNP})$): posterior probability of SNP when alternative hypothesis is true.

Table 2. GWAS results using rrBLUP, mlmm, Sommer and NAM models for GY vector of 221 wheat genotypes under high N treatment.

GWAS Model	N Level	SNP	Chr [a]	Pos [b]	Maj. Allele [c]	MAF [d]	$-\log(p\text{-}Value)$ [e]	ML ($locFDR(z_i)$) [f]	*P* ($bsFDR(\lambda_{SNP})$) [g]
rrBLUP	HN	AX-158576714	3A	69,099,202	CC	0.181	3.288885	0.561	0.389
	HN	AX-109898892	3A	73,588,989	CC	0.127	3.128732		0.391
	HN	AX-111563200	1B	15,596	TT	0.112	2.764430		0.269
	HN	AX-110038979	5B	13,056	NA	---	2.173367		0.266
	HN	AX-108852922	5B	14,184	NA	---	2.039261		0.266
mlmm	HN	AX-109898892	3A	73,588,989	CC	0.127	3.288885	0.454	0.372
	HN	AX-158544205	7D	24,335,511	CC	0.064	2.900164		0.284
	HN	AX-158612323	3A	3,508,832	NA	---	2.778897		0.357
	HN	AX-108915613	4A	79,319,176	NA	---	2.153261		0.251
	HN	AX-109299894	5A	97,047,183	TT	0.288	2.073941		0.272
Sommer	HN	AX-108915613	4A	79,319,176	NA	---	4.6011748	0.661	0.248
	HN	AX-109299894	5A	97,047,183	CC	0.108	3.6739810		0.223
	HN	AX-158544205	6A	24,335,511	CC	0.064	3.211453		0.217
	HN	AX-158576714	3A	69,099,202	CC	0.181	3.288885		0.281
	HN	IACX2540	5A	619,684,824	CC	0.213	2.78284		0.272
NAM	HN	AX-158576714	3A	69,099,202	CC	0.181	3.288885	0.558	0.415
	HN	AX-109898892	3A	73,588,989	CC	0.127	3.2845915		0.427
	HN	AX-110385692	3A	47,067,572	TT	0.115	3.1852291		0.414
	HN	AX-109470057	3A	61,095,990	CC	0.162	2.5953862		0.427
	HN	BS00094057_51	3A	7,437,103	TT	0.063	1.938022		0.412

[a] Chr: Chromosome, [b] Pos: Position (bp), [c] Maj. Allele: Major Allele, [d] MAF: Minor Allele Frequency, [e] $-\log(p$ Value) for top five significant SNPs received from different GWAS models, [f] ML $locFDR(z_i)$: maximum likelihood of five significant SNPs based on local FDR threshold in the GWAS model, [g] P ($bsFDR(\lambda_{SNP})$): posterior probability of SNP when alternative hypothesis is true.

3.4. SNP Effect Estimation

In this study, we were also interested in examining the impact of prior distribution on the SNP *p-values* of NAM as the best model of GWAS and estimation accuracy. Therefore, the GP models were assessed with six different Bayesian prior distributions: A, B, C, LASSO, ridge regression and survival. For all prior distributions, the markers with null effects were removed from analysis. For Bayesian survival prior distribution, $\lambda_0 = 1$ and $\exp(\beta_{GY}) = 0.001$, was assigned to obtain flat prior. To calculate $\lambda(SNP_i)$ and *p-values*, a Markov chain Monte Carlo (MCMC) of length 150,000 iterations was considered after a burning bootstrap period of 2000 iterations. The *p-values* generated from all prior distributions were converted to z-values for easier interpretation and minimizing the effect of the SNP dimension. By comparing the box plots of the SNP Z-scores generated from six different Bayesian prior distributions (Figure 6a), the Bayesian LASSO and Bayesian survival models exhibit highest accuracy by correlation between predicted genetic values $(\hat{g})$ with true breeding values (g), $r_{Bayes_L}(\hat{g}, g)$ = 0.8722 and $r_{Bayes_{Surv}}(\hat{g}, g)$ = 0.8876. The top five significant SNPs received from NAM model (Figure 6a) were compared with different Bayesian models. Only the Bayesian survival analysis can identify the differentials within them due to time to events among minor and major alleles within association panel. With this approach, the first two signal SNPs show a trend, and share reliable associations. However, for other Bayesian platforms there is no trend among SNP effects (Figure 6b).

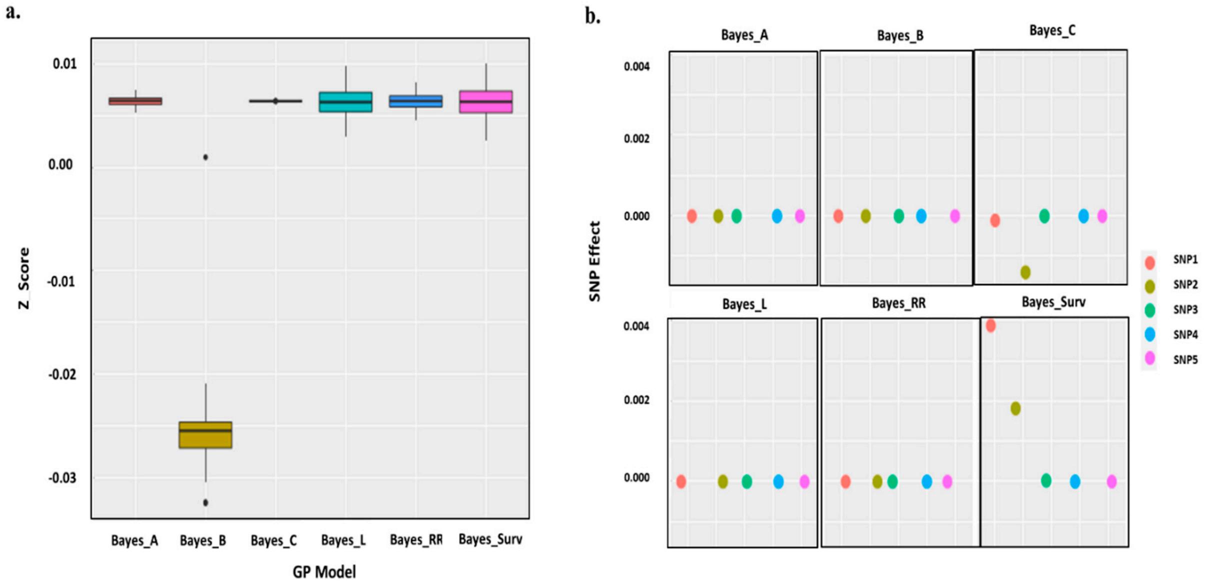


Figure 6. (**a**) Accuracy of GP models based on Bayesian inference, Z-score was used due to increase mean deviation in genetic value for each genotype. (**b**) The top five significant SNPs were compared with different Bayesian models. Only the Bayesian survival analysis can identify the differentials within them due to time to events among minor and major alleles within association panel. SNP1: AX-158576714, SNP2: AX-109898892, SNP3: AX-110385692, SNP4: AX-109470057 and SNP5: BS00094057_51.

4. Discussion

Nitrogen is the main nutrient for canopy growth and photosynthesis, which is responsible for GY and quality. Allelic variation for GY under low and high N could be high due to large mutations in the signal SNPs within candidate genes. Moreover, considerable effort is required to identify all involved variants among complex traits. Therefore, the main challenge in GWAS is to find significant and reliable associations related to a given complex trait. To address this dilemma, local FDR correction and Bayesian survival analysis, two precise and efficient computational approaches, serve as different filters to determine the best GWAS and GP models and consequently obtain a reliable association in the result of the best selected model. Mixed populations and variants with small effect sizes or rare alleles in kinship coefficients derived from marker information or outliers in the phenotypic vector, may lead to causal signals and false association between marker and complex trait. In parallel with these genetic parameters, there are some genetic hyper-parameters such as panel size, number of markers, MAF and number of call rates for each marker, that are not represented in the given model, but which have effects on the performance of the model and accuracy of results. In the GWAS and GP, one side of the model is allocated to the complex trait. To reduce the multicollinearity and heteroscedasticity in phenotypic observations, a tradeoff between bias and variance is required. Outliers in the GY vector are the first cause of bias results in GWAS. We used bootstrapping as a powerful and well-known tool to deal with outliers in the GY vector. However, in this parametric resampling method, the GY vector is simulated to obtain the confidence interval of the mean and standard error of the phenotypic observations. However, randomization with global replacement in parametric bootstrapping leads to underestimation of the variation of the complex trait. In this study, in addition to dealing with outliers with usual procedures, we applied Bayesian bootstrap using MCMC state resampling, which can be a good alternative to control underestimation of variance. Because the observations appeared with a certain probability with bias and variance tradeoff on the results. This probability based on prior distribution offers more precise estimation. In the results, Bayesian bootstrap *p-value* shows more significance with lower sampling error among the GY vector (Figure 1a). Therefore, the correlation between BLUPs and BLUEs is higher than 0.8 among prior distribution of the GY vector. However, the low replicability and reliability of results in the linear GWAS models, such as single-locus, multi-locus association and whole-genome variance models, pose a major challenge. Basically, in self-pollinated wheat crops, there are some non-additive relationships between loci, such as codominant and epistasis, due to heterogeneity within the locus. However, linear GWAS models do not account for this problem, and the genomic relationship matrix is estimated based only on additive fixed effects between SNPs. Consequently, the false positive rate and type I error are high, and many SNPs with true non-zero effects behave like null effects and, conversely, many SNPs with null effects show significant associations in the results. In contrast, in Bayesian whole genome regression, SNPs are considered as random effects, which introduces MAF as the main genetic hyper-parameter in the model. Additionally, the effects of minor and major alleles are accounted for via the $XZ_{\alpha\beta}$ component in the model, which reveals the allelic effects from the interaction between population structure and kinship matrix. All these different types of information, including genetic parameters and hyper-parameters, bear implications for assessing reliable associations in the GWAS results. With the Bayesian framework, it is possible to include all this information in both sides of the model, which greatly reduces the type I and II errors due to the posterior specifications for each component of the model. In the present work, we have studied two precise and efficient computational approaches as different filters to determine the best GWAS and GP models and consequently obtain reliable association in the results of the best selected model.

Local FDR as frequentist inference with Fisher information background was applied to check the distribution of SNPs especially in the tails as critical regions of the distribution. This method, based on maximum likelihood estimation, is more precise for determining the effect size and strength of a large-scale genomic file. As can be observed in the results, the expected mean likelihoods based on SNP information are acceptable, but the standard error in the models with a scale of $-log10$ (*SNP*) is still high. The main criticism of this approach is that it locally focuses only on SNP effects in tails, whereas signaling associations were distributed throughout the genome. Moreover, the magnitude of SNP effects alone may be insufficient to inform a conclusion about the performance of the model. For example, the degree of freedom (df) to fit the estimated density among SNPs is affected by the MAF and the quality of SNP chip (e.g., NAs and imputation). A larger df is required for sharper tails in $f(zz)$, but in the *mlmm* model df is low, and consequently the sigma is partially low, which may lead to underestimation or overestimation of the model. With this limitation, we also considered whether we should use this approach because the proportion of false rejection hypotheses is larger in this approach than in the family wise error rate (FWER) approach. The most interesting result of this approach is that in the variance component GWAS model, proportion of H_0 (sigma/mean) = 1.034 is the highest due to the highest sigma (SNP standard error), thus this model can clearly be removed due to its very low power. In the Bayesian whole genome regression with a low proportion of H_0 (sigma/mean) = 0.633, the putative signal SNPs in the tails were very well covered and the model produced the lowest error (sigma = 0.062).

Bayesian survival FDR analysis as a modern inference based on posterior estimation of SNP effects is commonly used for large-scale genomic data with small effects. In the survival part with the $(-\Delta(SNP_{i=0}^{n})$ component of the approach, we included the minor and major alleles in each SNP based on the time to events, which is necessary to achieve the expected prior distribution. Therefore, the MAF and the quality of SNP chip were considered in the computation. In GWAS with the original SNP information among GY vector, only in the *NAM* model were both minor and major alleles revealed in the results with the highest AIC and BIC criteria versus two models. In GWAS with SNP simulation, all allelic effects including major, minor, heterozygous and missing were revealed in the same model, which may present good evidence supporting this model as the best performing. Thus, we found that this is in good agreement with [29] to minimize the type I error at the significance level of MAF. As shown in Figures 5 and 6b, the relationship between effect size and MAF is not strong for SNPs with higher MAF value. The difference is claimed for SNPs with low MAFs (from SNP3: AX-110385692 to SNP5: BS00094057_51). In this case, it indicates that the informative nature of $locFDR(z_i)$ is lower than $bsFDR(\lambda_{SNP})$ to determine reliable associations, which closely aligns with the results of [23,33]. Binary MCMC sampling is a renewable technique in survival Bayesian inference to generate large-scale genomic datasets; however, it is affected by long computation times in the linear algorithms. In the Bayesian part, we solved this technical difficulty using the exponential form $\exp(-\Delta(SNP_{i=0}^{n})$ to predict the SNP prior distribution. For both the local FDR and Bayesian survival approaches, the standard error of the *NAM* model was lower than the others. To compare the performance of the Bayesian survival model with other commonly used Bayesian platforms, we proposed $\lambda(SNP_i)$ based on covariate effects between SNPs. These covariate effects have negative implications for the expected prior distribution in all Bayesian approaches, including survival. To solve this problem, β_{GY} should take an exponential form based on λ_0 which allows the GP algorithms to examine SNP effects with minimal bias and variance. To compare the accuracy of GP models, we used Z-score because it increases the mean and variance of genetic value for each genotype. However, it makes interpretation easier and more accurate than *p-value* and Bayes factor with a smaller comparison interval. The five top significant SNPs were compared with different Bayesian models. Only Bayesian survival analysis can identify the difference between minor and major alleles based on time to events. In $bsFDR(\lambda_{SNP})$, λ_{SNP} is a semi-parametric statistic because it is based on MAF and allelic scores of genotypes simultaneously [34]. It seems that the semi-parametric empirical Bayes factor better controls both false positive and false negative errors in GWAS to identify allelic variations and find templates for significant SNPs. Therefore, we propose the utilization of different GWAS models at more N levels to increase the replicability of the posterior probability [35] of the identified signal associations.

5. Materials and Methods

5.1. Plant Materials and Field Experiment

In this study, a set of 221 bread wheat genotypes from Breeding Innovations In Wheat for resilient Cropping Systems (BRIWECS) project were cultivated as split-plots design, in three cropping seasons 2018, 2019 and 2020 at the agricultural research station Campus Klein-Altendorf, University of Bonn, Germany. We recorded the GY value of each genotype under high and low N fertilizer annually. Then, the average GY values of the genotypes were used as GY vector for subsequent analysis. The outliers in the GY vector were checked. To deal with outliers, they were kept since they reflected the actual field values across all years. The distribution of the residuals was checked using Shapiro normality test. To improve the quality of the vector, 2000 repeated random samples with replacement were generated from the original GY vector using R/*bootstrap*, and then Bayesian bootstrap *p-value* was calculated. To also control genotype random effect, the EBVs and BLUPs were based on broad sense heritability [36] plotted.

5.2. SNP Quality Control

A platform of 150 K affymetrix SNP Chip was used to apply the GY vector in the GWAS and GP models. The SNPs with MAF ≤ 0.05 were removed due to monomorphism in the marker. After checking SNPs that deviated from the Hardy–Weinberg equilibrium (HWE), only 22,489 polymorphic SNP markers remained, which were used in GWAS and GP analyses.

5.3. Population Structure

Cumulative variance explained by the eigenvalues of the principal components was calculated, and then discriminant analysis of principal components (DAPC), which minimizes genetic variation within clusters, was also conducted using R/*adegenet* for assessing the population structure. To determine the actual number of PCs retained in the DAPC, cross validation (CV) was performed to simulate the low and high numbers of PCs in the model. To control outliers in the population clusters, information from the identical-by-state (IBS), which measures differences in allelic states, and the fixation index (F_{st}), which refers to the amount of heterozygosity at different levels of population structure, were calculated based on covariance estimation [37]. To detect the regions that might be involved in the linkage pattern of the population, the whole genome wide plot of F_{st} was constructed.

5.4. Construction of Genomic Relationship Matrix (GRM)

Basically, the GRM as kinship matrix is used in GWAS and GP models. Based on the method suggested by (Mathew et al. [38]), the covariance between individuals g_i and g_j can be equal by covariance of SNP_{ij}, therefore the GRM was calculated using $G = \frac{\Sigma_{k=1}^{L}(g_{ik}-p_k)^2}{\Sigma_{k=1}^{L}p_k(1-p_k)}$, and $p_k = \frac{1}{n}\Sigma_{i=1}^{n} g_{ik}$ where, L is number of loci, p_k is MAF for the locus k and g_i.

5.5. GWAS and GP Models

A total of four different GWAS models, including single-locus association with R/*rrBLUP* [39], multi-locus association with R/*mlmm.gwas* [11], variance component association with R/*sommer* [40] and Bayesian whole genome regression with R/*NAM* [41] at low and high N levels were performed to detect the association between SNPs and GY as a target trait. The single-locus association model was fitted in the adjusted form the mixed linear model as $y = X\beta + Zu + e$, where y is the trait vector, X is the fixed effects matrix, β is the vector of coefficients including principal components and population structure, Z the matrix of random SNP effects coded as (-1, 0 and 1), $V(u) = K\sigma_g^2$, where K is the GRM as kinship matrix and σ_g^2 is additive genetic variance with IBS basis. It was removed from the model due to convergence of N × Y to zero. Multi-locus association model, form $y_{i=1}^{n} = \mu + \sum_{j=1}^{m} M_{.j}\beta_j + e$, where $y_{i=1}^{n}$ is trait vector with n genotypes, m is total number of SNPs, M_{ij} is the matrix of random SNP effects coded as (0, 1 and 2) and β_j is the vector of SNP effects and H_0 is given in terms of $\beta = \sigma_g^2 = 0$. Once the $-log$ *(p-value)* is above the FDR threshold line, H_0 is rejected. In the variance component model, $y = \sum_{i=1}^{c} K_i\sigma_g^2 + e$, where c is non-overlapping classes of SNPs, K_i is the class of kinship matrix based on genomic data and the components of genetic variance are from REML of the mean of SNP information. Bayesian whole genome regression model was fitted to $y = \mu + X\alpha + Z\beta + XZ_{\alpha\beta} + e$, where y is trait vector, X is matrix of genotypes and SNPs, α is corresponding vector of SNP effects that captures small effects of all SNPs, β is vector that captures additional effects of SNPs with large effect based on Bayes factor.

GP models were performed with Bayesian whole genome regression based on the following model: $y = \mu + X\beta + (Z_1Z_2)u + e$, where, y is the GY vector, X, is SNP information for the genotypes, β is regression coefficient for each SNP, indicating the SNP effect in the model, Z_1 is effect of major allele, Z_2 is effect of minor allele, u is vector of Bayes factor for SNP matrix and e is implies to residual term [42]. Bayesian models including Bayes A, B, C, LASSO, RR and Bayesian survival analysis, were then applied to predict genomic estimated breeding values (GEBVs). The CV approach was used to evaluate the accuracy of each model. To determine the best model, correlations $\rho(\check{y}EBV,\ \check{y}GEBV)$ were estimated for each GP model.

5.6. SNP Effect Estimation under locFDR and bsFDR

To check the distribution of SNPs *p-value*, for each GWAS model under low and high N levels, zz vector was created under the null hypothesis *N* (0, 1), which is necessary, when SNPs *p-value* is very far from normal distribution. To fit the density of *f (zz)* with heavy tails, degree of freedom (*df*) was determined based on SNP sample size. Then, empirical null hypotheses were used to estimate the parameters of *f (zz)* by maximum likelihood, indicating the accuracy of each GWAS model, and the results were visualized using R/*locfdr* [32]. Based on $locFDR(z_i)$ function, the significant associations above FDR threshold line of 0.05 were selected for the GWAS models. The survival function was represented as a cumulative hazard function using R/*survival* [43], $S(SNP_i) = \exp\left(-\Delta\left(SNP_{i=0}^{n}\right)\right)$, where, $S(snp)$ is survived coefficient for SNP, which estimates the tendency between uniformity in the bottom of SNPs *p-value* distribution and the peaks, while there were no other observations in the bottom of distribution $S(y_{SNP})$. Based on the MAF and missing values of each SNP, the prior distribution of significant associations were calculated using the function $bsFDR(\lambda_{SNP})$, for the GWAS models. To check for covariate effects within SNPs in the GP models, the Bayesian survival function was applied in the form of $\lambda(SNP_i) = \lambda_0(SNP)\exp(\beta_{GY})$, where, $\lambda(SNP_i)$ is the Bayes survived factor indicating the SNP effect, λ_0 is the baseline for function and β_{GY} is regression coefficient of the whole genome under the GY vector.

Author Contributions: M.B.S. literature and problem statement, methodology, formal analysis, writing original draft; A.B. validation, review, editing, funding acquisition; J.L. editing and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to acknowledge the kind support of DAAD who generously provided us the opportunity to pursue our research and financing the study. Also, the authors are grateful to the Institute for Crop Science and Resource Conservation (INRES)-Plant Breeding, University of Bonn, Germany, for responsible supervision and execution plan during the study.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Eltaher, S.; Baenziger, P.S.; Belamkar, V.; Emara, H.A.; Nower, A.A.; Salem, K.F.; Alqudah, A.M.; Sallam, A. GWAS revealed effect of genotype × environment interactions for grain yield of Nebraska winter wheat. *BMC Genom.* **2021**, *22*, 2. [CrossRef]
2. Nehe, A.; Akin, B.; Sanal, T.; Evlice, A.K.; Ünsal, R.; Dinçer, N.; Demir, L.; Geren, H.; Sevim, I.; Orhan, Ş.; et al. Genotype x environment interaction and genetic gain for grain yield and grain quality traits in Turkish spring wheat released between 1964 and 2010. *PLoS ONE* **2019**, *14*, e0219432. [CrossRef]
3. Efron, B.; Hastie, T.J. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*; Cambridge University Press: Cambridge, UK, 2019.
4. Chernick, M.R.; LaBudde, R.A. *An Introduction to Bootstrap Methods with Applications to R*; Wiley-Blackwell: Hoboken, NJ, USA, 2011.
5. Gianola, D.; Cecchinato, A.; Naya, H.; Schön, C.C. Prediction of Complex Traits: Robust Alternatives to Best Linear Unbiased Prediction. *Front. Genet.* **2018**, *9*, 195. [CrossRef] [PubMed]
6. Ma, X.; Christensen, O.F.; Gao, H.; Huang, R.; Nielsen, B.; Madsen, P.; Jensen, J.; Ostersen, T.; Li, P.; Shirali, M.; et al. Prediction of breeding values for group-recorded traits including genomic information and an individually recorded correlated trait. *Heredity* **2020**, *126*, 206–217. [CrossRef] [PubMed]

7. Deja-Muylle, A.; Parizot, B.; Motte, H.; Beeckman, T. Exploiting natural variation in root system architecture via genome-wide association studies. *J. Exp. Bot.* **2020**, *71*, 2379–2389. [CrossRef]
8. Gondro, C.; Van der Werf, J.; Hayes, B. *Genome-Wide Association Studies and Genomic Prediction*; Humana Press: Totowa, NJ, USA, 2017.
9. Maldonado, C.; Mora-Poblete, F.; Contreras-Soto, R.I.; Ahmar, S.; Chen, J.T.; do Amaral Júnior, A.T.; Scapim, C.A. Genome-Wide Prediction of Complex Traits in Two Outcrossing Plant Species Through Deep Learning and Bayesian Regularized Neural Network. *Front. Plant Sci.* **2020**, *11*, 593897. [CrossRef]
10. Wei, W.H.; Hemani, G.; Haley, C.S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **2014**, *15*, 722–733. [CrossRef]
11. Segura, V.; Vilhjálmsson, B.J.; Platt, A.; Korte, A.; Seren, Ü.; Long, Q.; Nordborg, M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **2012**, *44*, 825–830. [CrossRef]
12. Würschum, T.; Kraft, T. Evaluation of multi-locus models for genome-wide association studies: A case study in sugar beet. *Heredity* **2014**, *114*, 281–290. [CrossRef]
13. Müller, B.U.; Stich, B.; Piepho, H.P. A general method for controlling the genome-wide type I error rate in linkage and association mapping experiments in plants. *Heredity* **2010**, *106*, 825–831. [CrossRef]
14. Mutshinda, C.M.; Sillanpää, M.J. Swift block-updating EM and pseudo-EM procedures for Bayesian shrinkage analysis of quantitative trait loci. *Theor. Appl. Genet.* **2012**, *125*, 1575–1587. [CrossRef]
15. Wang, S.B.; Feng, J.Y.; Ren, W.L.; Huang, B.; Zhou, L.; Wen, Y.J.; Zhang, J.; Dunwell, J.M.; Xu, S.; Zhang, Y.M. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **2016**, *6*, 19444. [CrossRef] [PubMed]
16. Mathew, B.; Léon, J.; Sannemann, W.; Sillanpää, M.J. Detection of Epistasis for Flowering Time Using Bayesian Multilocus Estimation in a Barley MAGIC Population. *Genetics* **2018**, *208*, 525–536. [CrossRef] [PubMed]
17. Kärkkäinen, H.P.; Li, Z.; Sillanpää, M.J. An Efficient Genome-Wide Multi locus Epistasis Search. *Genetics* **2015**, *201*, 865–870. [CrossRef]
18. Li, Z.; Sillanpää, M.J. Estimation of Quantitative Trait Locus Effects with Epistasis by Variation Bayes Algorithms. *Genetics* **2012**, *190*, 231–249. [CrossRef]
19. Svishcheva, G.R.; Axenovich, T.I.; Belonogova, N.M.; van Duijn, C.M.; Aulchenko, Y.S. Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.* **2012**, *44*, 1166–1170. [PubMed]
20. Banerjee, S.; Zeng, L.; Schunkert, H.; Söding, J. Bayesian multiple logistic regression for case-control GWAS. *PLoS Genet.* **2018**, *14*, e1007856. [CrossRef]
21. Chen, C.; Steibel, J.P.; Tempelman, R.J. Genome Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods. *Genetics* **2017**, *206*, 1791–1806. [CrossRef]
22. He, L.; Kulminski, A.M. Fast Algorithms for Conducting Large-Scale GWAS of Age-at-Onset Traits Using Cox Mixed-Effects Models. *Genetics* **2020**, *215*, 41–58. [CrossRef] [PubMed]
23. Wakefield, J. Bayes factors for genome-wide association studies: Comparison with *p*-values. *Genet. Epidemiol.* **2009**, *33*, 79–86. [CrossRef]
24. Lee, Y.; Luca, F.; Pique-Regi, R.; Wen, X. Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics. *bioRxiv* **2018**. [CrossRef]
25. McDaid, A.F.; Joshi, P.K.; Porcu, E.; Komljenovic, A.; Li, H.; Sorrentino, V.; Litovchenko, M.; Bevers, R.P.; Rüeger, S.; Reymond, A.; et al. Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nat. Commun.* **2017**, *8*, 15842. [CrossRef]
26. Hughey, J.J.; Rhoades, S.D.; Fu, D.Y.; Bastarache, L.; Denny, J.C.; Chen, Q. Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC Genom.* **2019**, *20*, 805. [CrossRef] [PubMed]
27. Theodoratou, E.; Farrington, S.M.; Timofeeva, M.; Din, F.V.N.; Svinti, V.; Tenesa, A.; Liu, T.; Lindblom, A.; Gallinger, S.; Campbell, H.; et al. Genome-wide scan of the effect of common nsSNPs on colorectal cancer survival outcome. *Br. J. Cancer* **2018**, *119*, 988–993. [CrossRef] [PubMed]
28. Habier, D.; Fernando, R.L.; Kizilkaya, K.; Garrick, D.J. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinform.* **2011**, *12*, 186. [CrossRef]
29. Bi, W.; Fritsche, L.G.; Mukherjee, B.; Kim, S.; Lee, S. A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank. *Am. J. Hum. Genet.* **2020**, *107*, 222–233. [CrossRef] [PubMed]
30. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
31. Efron, B.; Tibshirani, R. Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **2002**, *23*, 70–86. [CrossRef]
32. Efron, B. Correlation and Large-Scale Simultaneous Significance Testing. *J. Am. Stat. Assoc.* **2007**, *102*, 93–103. [CrossRef]
33. Ouyang, H. Bayesian Approach for Nonlinear Dynamic System and Genome-Wide Association Study. Ph.D. Thesis, North Carolina State University, Raleigh, NC, USA, 2010.
34. Morisawa, J.; Otani, T.; Nishino, J.; Emoto, R.; Takahashi, K.; Matsui, S. Semi-parametric empirical Bayes factor for genome-wide association studies. *Eur. J. Hum. Genet.* **2021**, *29*, 800–807. [CrossRef]

35. McGuire, D.; Jiang, Y.; Liu, M.; Weissenkampen, J.D.; Eckert, S.; Yang, L.; Chen, F.; Berg, A.; Vrieze, S.; Jiang, B.; et al. Model-based assessment of replicability for genome-wide association meta-analysis. *Nat. Commun.* **2021**, *12*, 1964. [CrossRef] [PubMed]
36. Schmidt, P.; Hartung, J.; Bennewitz, J.; Piepho, H.P. Heritability in Plant Breeding on a Genotype-Difference Basis. *Genetics* **2019**, *212*, 991–1008. [CrossRef] [PubMed]
37. Dadshani, S.; Mathew, B.; Ballvora, A.; Mason, A.S.; Léon, J. Detection of breeding signatures in wheat using a linkage disequilibrium-corrected mapping approach. *Sci. Rep.* **2021**, *11*, 5527. [CrossRef]
38. Mathew, B.; Léon, J.; and Sillanpää, M.J. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. *Heredity* **2017**, *120*, 356–368. [CrossRef]
39. Endelman, J.B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* **2011**, *4*, 250–255. [CrossRef]
40. Covarrubias-Pazaran, G. Software update: Moving the R package sommer to multivariate mixed models for genome-assisted prediction. *bioRxiv* **2018**, 354639. [CrossRef]
41. Xavier, A.; Xu, S.; Muir, W.M.; Rainey, K.M. NAM: Association studies in multiple populations. *Bioinformatics* **2015**, *31*, 3862–3864. [CrossRef]
42. de los Campos, G.; Hickey, J.M.; Pong-Wong, R.; Daetwyler, H.D.; Calus, M.P. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* **2013**, *193*, 327–345. [CrossRef]
43. Rizvi, A.A.; Karaesmen, E.; Morgan, M.; Preus, L.; Wang, J.; Sovic, M.; Hahn, T.; Sucheston-Campbell, L.E. gwasurvivr: An R package for genome wide survival analysis. *Bioinformatics* **2018**, *35*, 1968–1970. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of
Molecular Sciences

MDPI

Article

Genetic Parameter and Hyper-Parameter Estimation Underlie Nitrogen Use Efficiency in Bread Wheat

Mohammad Bahman Sadeqi [1], Agim Ballvora [1,*], Said Dadshani [2] and Jens Léon [1]

[1] INRES-Plant Breeding, Rheinische Friedrich-Wilhelms-Universität Bonn, 53113 Bonn, Germany; mbsadeghi1@gmail.com (M.B.S.); j.leon@uni-bonn.de (J.L.)

[2] INRES-Plant Nutrition, Rheinische Friedrich-Wilhelms-Universität Bonn, 53113 Bonn, Germany; dadshani@uni-bonn.de

* Correspondence: ballvora@uni-bonn.de

Abstract: Estimation and prediction play a key role in breeding programs. Currently, phenotyping of complex traits such as nitrogen use efficiency (NUE) in wheat is still expensive, requires high-throughput technologies and is very time consuming compared to genotyping. Therefore, researchers are trying to predict phenotypes based on marker information. Genetic parameters such as population structure, genomic relationship matrix, marker density and sample size are major factors that increase the performance and accuracy of a model. However, they play an important role in adjusting the statistically significant false discovery rate (FDR) threshold in estimation. In parallel, there are many genetic hyper-parameters that are hidden and not represented in the given genomic selection (GS) model but have significant effects on the results, such as panel size, number of markers, minor allele frequency, number of call rates for each marker, number of cross validations and batch size in the training set of the genomic file. The main challenge is to ensure the reliability and accuracy of predicted breeding values (BVs) as results. Our study has confirmed the results of bias–variance tradeoff and adaptive prediction error for the ensemble-learning-based model STACK, which has the highest performance when estimating genetic parameters and hyper-parameters in a given GS model compared to other models.

Keywords: genetic parameter; hyper-parameter; genomic selection model; estimation; nitrogen use efficiency and wheat

check for updates

Citation: Sadeqi, M.B.; Ballvora, A.; Dadshani, S.; Léon, J. Genetic Parameter and Hyper-Parameter Estimation Underlie Nitrogen Use Efficiency in Bread Wheat. *Int. J. Mol. Sci.* **2023**, *24*, 14275. https://doi.org/10.3390/ijms241814275

Academic Editor: Ivan Kreft

Received: 8 August 2023
Revised: 7 September 2023
Accepted: 14 September 2023
Published: 19 September 2023

1. Introduction

Estimation and prediction play a key role in breeding programs. For a long time, breeders have tried to predict better genetic performance, genotypic value (GV) or breeding value (BV) from observations of a phenotype of interest by using estimators of genetic and phenotypic variance. This ratio is usually the heritability based on the line mean [1]. Single nucleotide polymorphisms (SNP) with a microarray platform has become the most popular high-throughput genotyping system in recent decades, and has been extensively used for quantitative trait loci (QTL) and experimental population analysis [2–5]. Thousands of QTLs inheriting simple traits of agronomic importance have been identified in major crops, and these can be used to accelerate marker-assisted selection (MAS). However, the genetic improvement of complex quantitative traits by using QTL-associated markers or MAS is not very efficient in practical breeding programs due to QTL × environment interactions or variation in genetic population structure. MAS is effective only for alleles with large effects on quality traits. However, it cannot improve polygenic traits, and many important traits in plant breeding are polygenic [6,7]. However, genomic selection (GS) uses high-density markers to predict the genetic values of genotypes, which is different from QTL analysis, MAS and association mapping (AM). With the availability of cheap and abundant molecular markers, genomic selection (GS) is becoming an efficient method for selection in both animal and plant breeding programs. Currently, phenotyping is still expensive, requires high-throughput technologies and is a very time-consuming process compared to genotyping. Therefore, researchers are trying to predict phenotypes based on marker information. GS consists of genomic file (SNPs) and phenotype file (individuals) in the reference (training) population and predicts the phenotypes or breeding values (BVs) of candidates for selection in the test (validation) population using statistical machine learning models [8]. The training set combines genomic data as independent variables with the agronomic trait of interest as dependent variables. The density of markers is very high and contains enough information to train the GS model with greater accuracy. However, the test set contains only the genomic data of some individuals and predicts the BVs of individuals according to the GS model. A higher correlation between the predicted BVs and the true phenotypic values of the individuals implies higher accuracy and performance. Basically, there are two type of features in the GS model including genetic parameters with random effects and hyper-parameters with fixed effects, which determine the results. Genetic parameters such as population structure, genomic relationship matrix, marker density and sample size are the major factors that increase the power and accuracy of the model. However, they play an important role in adjusting the statistically significant false discovery rate (FDR) threshold in estimation. For example, mixed populations [9,10], copy numbers of variants with small effect sizes [11], rare alleles in linkage disequilibrium (LD) decay derived from marker information [12] and outliers in phenotypic observations of interesting complex traits can lead to casual signals and pseudo association between marker and trait. In parallel, there are many genetic hyper-parameters that are hidden and are not represented on a given GS model, but have significant effects on the results, such as panel size, number of markers, minor allele frequency (MAF), number of call rates for each marker, number of cross validations (CV), and batch size in the training set of the genomic file [13,14]. However, GS models face a range of practical and theoretical problems in estimating the genetic parameters and hyper-parameters of the model. The main challenge is being sure of the reliability and accuracy of the predicted BVs as results. GS via linear mixed regression is based on conventional point estimators such as maximum likelihood (ML) and restricted ML (REML), which are generally susceptible to estimating genetic parameters in the whole genomic dataset because of high collinearity in the model. Therefore, the model introduces a strong estimation bias. Consequently, QTLs with small effects are completely missed in the results [2,15]. Moreover, high variance due to high convergence under marker density is a problem that often occurs when implementing complex mixed linear GS models [16]. Recent developments in shrinkage estimation [17] and the utilization of Markov Chain Monte Carlo (MCMC) sampling methods have made GS based on Bayesian whole genome regression feasible. Nevertheless, MCMC sampling algorithms can suffer from slow convergence rates and poor mixing of sample chains [18], especially when non-additive genetic random effects are included in the model [19]. In GS, the use of high-density markers requires the application of advanced feature selection algorithms. In Bayesian whole genome regression, even shrinkage algorithms cannot provide an acceptable tradeoff between bias and variance due to the high convergence rate under high marker density [20]. Using modern statistical models could be a logical solution to this challenge. In recent decades, new technologies such as sensors, robotics and satellite data have led to a high throughput of phenotypic data. In parallel, next generation sequencing (NGS) techniques have made it possible to simultaneously generate a large training dataset for traits of interest. Thus, GS with large genomic datasets and high-density markers as features in the model, requires statistical machine learning methods with more computational power, especially for complex traits such as nitrogen use efficiency (NUE) in wheat [14,21].

1.1. GS Model Definition

In the rrBLUP model, $GEBV = X\hat{g}$, which can be considered as a regularization parameter. So,

$$\begin{bmatrix}\hat{\mu}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}1_n * 1_n & 1_n * X\\ X^* 1_n & X^* X + I\frac{\sigma_e^2}{\sigma_g^2}\end{bmatrix}^{-1}\begin{bmatrix}1_n * y\\ X^* y\end{bmatrix}$$

The $I\frac{\sigma_e^2}{\sigma_g^2}$ component in the rrBLUP matrix occurs in theory, but in practice it is required to be adjusted to a more accurate parameter such as the $G^{-1}\frac{\sigma_e^2}{\sigma_v^2}$ component in the gBLUP matrix:

$$\begin{bmatrix}\hat{\mu}\\ \hat{v}\end{bmatrix} = \begin{bmatrix}1_n * 1_n & 1_n * Z\\ Z^* 1_n & Z^* Z + G^{-1}\frac{\sigma_e^2}{\sigma_v^2}\end{bmatrix}^{-1}\begin{bmatrix}1_n * y\\ Z^* y\end{bmatrix}$$

In the LASSO model, the regularization parameter is constrained by minimum ordinary least squares (OLS) of the large genomic dataset. The GS models rrBLUP and gBLUP face incredible biases in *GEBV* calculation due to collinearity. To obtain a solution for LASSO, the $X^* X + \lambda I$ component is updated so that either marker effect with addition or subtraction can be computed per each iteration. The SNP effects were calculated as

$$\begin{bmatrix}\hat{\mu}\\ \hat{\beta}\end{bmatrix} = \begin{bmatrix}1_n * 1_n & 1_n * X\\ X^* 1_n & X^* X + \lambda I\end{bmatrix}^{-1}\begin{bmatrix}1_n * y\\ X^* y\end{bmatrix}$$

*GEBV*s for each individual BGLR matrix are estimated as follows:

$$\begin{bmatrix}\hat{\mu}\\ \hat{g_1}\\ \cdot\\ \cdot\\ \cdot\\ \hat{g_p}\end{bmatrix} = [u]\begin{bmatrix}1_n * 1_n & \cdots & 1_n * X_1\\ \vdots & \ddots & \vdots\\ X_p{}^* 1_n & \cdots & X_P{}^* X_P + I\frac{\sigma_e^2}{\sigma_{gp}^2}\end{bmatrix}^{-1}\begin{bmatrix}1_n * y\\ \cdot\\ \cdot\\ \cdot\\ X_P{}^* y\end{bmatrix}$$

In the BGLR model, the coefficients of marker effects were tested under the following hypotheses:

$$\mathrm{H_0}:\ g_1 = g_2 = g_n = 0$$

$$\mathrm{H_a}:\ g_1 \neq g_2 \neq g_n \neq 0$$

Thus, the u value, as Bayes factor and regularization parameter, is an index for testing these hypotheses. When the number of features (p) is much higher than the number of observations (n), the challenge is to find optimal hyper-parameters for the given GS model that play a dominant role in GS. However, in these four methods, SNP selection is performed via a nonlinear transformation, typically achieved using the kernel trick. The kernel (RKHS and SVM) or ensemble (boosting and bagging) approaches allow feature selection in the n-dimensional space of SNPs with a random number of predictors p [22].

In the RKHS model, if $g(x_i)$ is a nonparametric function of large marker density, it can be written as

$$f(g(x_i)|\lambda) = \frac{1}{2}[y - \mathrm{W}\theta - g(x_i)]R^{*-1} + \frac{\lambda}{2}|g(x_i)|_{H^*}^{2}]$$

where y is the *NUE* vector, W is the incidence matrix of parametric SNP effects θ on y, R is the residual covariance matrix, $g(x_i)$ is the vector of genotypes (SNPs), λ is the regularization parameter under squared norm of H^* as Hilbert space. Thus, λ controls the tradeoff between goodness and complexity of the model [23].

The SVM model can be interpreted as a class of kernel algorithms, which is very similar to the RKHS model. It is written as

$$f(g(x_i)|\lambda) = \frac{1}{2}V[y, g(x_i)]R^{*-1} + \frac{\lambda}{2}|g(x_i)|^{2}_{H^*}]$$

where $V[y, g(x_i)]$ is the insensitive loss function of support vectors for y (NUE vector) under covariance matrix (R), and λ is the regularization parameter under the squared norm of H^* as Hilbert space.

In the Boosting model, the $GEBV$ for NUE among each genotype is represented as follows:

$$GEBV = \sum_{L=1}^{L_i} vh_z\left(g_{ens}\left(x_p\right)\middle|W\right) + Bias\left(g_{ens}\left(x_p\right)\right) + e$$

where L_i is the learning rate of predictors, v is the regularization parameter, h_z is the accuracy mean of the GS model, W is the incidence matrix of SNP effects on the ensemble estimator of $g_{ens}\left(x_p\right)$ with the expected function and e is the residuals with independent and identity distribution (IID).

In the Bagging model, the BLUE estimator is represented as follows:

$$GEBV = \frac{1}{L}\;\hat{g}_{Bag}(ij)\mathrm{E}[\hat{g}_{ens}(i.)] + Var(\hat{g}_{ens}(.j))$$

where L_i is the learning rate of predictors, $\hat{g}_{Bag}(ij)$ is the bagged estimator based on allele frequency of the genotype, $\hat{g}_{ens}(..)$ is the ensemble estimator of the model that can be considered as a regularization parameter, $\mathrm{E}[\hat{g}_{ens}(i.)]$ is the bootstrap mean as an estimation of bias of $\hat{g}_{ens}(..)$.

The stacking model combines all GS models through meta-learning. It is presented as follows:

$$Loss_{GEBV} = \begin{cases} 0 & if\ f(x) < \varepsilon \\ |f(x)| - \varepsilon & if\ f(x) = \varepsilon \\ |f(x) + (Bias \pm Var)| & if\ f(x) > \varepsilon \end{cases}$$

So, $f(x) = \sum_{L=1}^{L_i}(m_{.j}\alpha_{.j})$ and $(x)\ \sim \hat{y}$.

$f(x)\ \sim \hat{y}$ is the predicted NUE value among each genotype, ε is the residual error of $GEBV$ estimation, L_i is the learning rate of predictors in a given GS model, $(Bias \pm Var)$ is the tradeoff between bias and variance of the $GEBV$ estimation, $allel_{.j}$ is the $.j$th MAF for SNP and $\alpha_{.j}$ is $.j$th SNP effect based on the given GS model.

1.2. Feature Selection in GS Model

Modern statistical genomics algorithms utilize high-dimensional genomic data to perform customized and accurate genomic prediction and selection. In these algorithms, feature selection is the key step for analyzing high-dimensional genomic and phenotypic data simultaneously. Recent advances include next generation sequencing (NGS) and high-throughput phenotyping techniques that generate a large number of variables in different types of GS models. This development of complex data structures leads to structured identification of important features in the model. This big dataset can be considered as a matrix, with columns corresponding to variables such as SNPs and explanatory phenotypic traits, and rows corresponding to individuals. Since the number of measured features is much larger than the number of individuals, this high dimensionality of the dataset has led to heterogeneous feature selection [24,25]. The question arises of how to identify the important features among the trait of interest from this large-scale data [26]. Breeders often measure multiple variables in each genotype simultaneously. Therefore, multivariate data are very common due to the facility of data collection. Therefore, for complex features, the relationship between individuals is nonlinear. Defining a complex model is usually the solution against poor accuracy, especially for the GS model, which is inherently nonlinear, with parametric and some semi-parametric estimators. Feature selection summarizes the variables into a small subset. Two complementary items include predictive performance and stability of the selected features, which makes up an acceptable feature selection method. Simple feature selection algorithms rely only on univariate GS models such as single locus prediction. However, in practice, most genomic datasets are multivariate with different classes and probability distributions [27]. Thus, one challenge is to identify the subset of variables that are useful in a given prediction GS model. The interpretability of selected variables is important for the stability of the given model, which is related to the heterogeneous nature of genomic data [28,29]. Indeed, marker information with minor or major alleles, heterozygous and missing data are highlighted at the categorical scale, while agronomic (phenotypic) traits may be on continuous or categorical scales. Currently, there are few feature selection methods that directly handle both continuous and categorical variables. In general, kernel methods and tree ensemble approaches are common for semiparametric GS models with both continuous and categorical variables [30]. A mixed linear GS model with high marker density leads to nonlinear regression in the surrounding space. This is typically achieved via the kernel method, which allows the computations to be performed in the n dimensional space of variables for any number of predictors p. The kernel method is actually kernel smoothing of the mean, which is a set of tools used to perform non- or semi-parametric estimation [31]. Multi-trait reproducing kernel Hilbert space (RKHS) methods can be modeled with marker–environment interaction; therefore, they are suitable for continuous response variables such as NUE with unknown prior distribution [32]. Tree ensemble approaches such as random forest (RF) provide a better generalization performance because, in these approaches, the errors of the estimators (e.g., BLUE) are distributed across different decision trees [1,33]. In general, the tree strategy attempts to minimize covariate error when approximating the true class distribution while shrinking the effects of known factors to null [34].

1.3. Regularization of GS Model

The main objective of genomic prediction (GP) approaches is to estimate genotypic values among unobserved true phenotypic values. However, determining the relevant predictive genomic estimated breeding values ($GEBV$s) based on high-density marker information is a fundamental problem in GS models, especially for complex quantitative trait with large number of SNPs (n) and very small p-values [35,36]. SNP heritability estimation takes the role of regularization, but how can one tell whether these estimated BVs are good or bad? Before providing an answer to this question, it should be clarified that GP is different from genomic inference. GP can be empirically calibrated, but inference cannot. Therefore, the regularization approach is an important step towards correct inference [37]. Deep learning (DL) algorithms attempt to estimate any minor or major genetic effects. The utilization of DL algorithms in GS provides the opportunity to obtain a meta-picture of the whole GS performance [38]. Even when using DL algorithms as a modern statistical perspective, GS models suffer from under- or over-fitted results. Therefore, regularization techniques provide a balance between under- or over-fitting in the GS model. That is, regularization provides a set of tools to find a logical tradeoff between bias and variance of the parameters and semi-parameters of the estimated genetic gain in the GS model. Other GS models with kernel, Bayesian or DL roots have been derived from Equation (1). Genetic gain in specific or estimated breeding value (EBV) in general can be defined by the following formula [39]:

$$\Delta G = g_{si}\beta_i\sigma_e \tag{1}$$

where ΔG is the genetic gain based on marker information, g_{si} is the genetic selection density in the population, β_i is the power of the equation based on the FDR threshold of significant markers and σ_e is the genetic standard error (SE) or the residuals of the equation. Once high-density markers are used in the GS model, computing g_{si} with many pairwise hypothesis tests at the same time has a high bias and high variance. Therefore, modern feature selection such as kernel or ensemble tree approaches could be a solution to this challenge. The β_i can be defined as the matrix below:

$$\sum \beta_i = \log \begin{bmatrix} X'X & X'Z \\ Z'X & 1 + \frac{1^{G^{*-1}}}{\sigma_e^2} \end{bmatrix} \tag{2}$$

where X is the incidence matrix for the proportion of individuals in the population structure ($\mathrm{n_{ps}}$) × marker (m) with fixed effects, X' is a transformation of X, Z is a designed matrix for the effect of genotype (n) × marker effect (p), including all random effects, Z' is the transformed Z, G^{*-1} is the inverted matrix of the genomic relationship matrix (GRM) when the effect of non-associated markers has shrunk toward zero with $N(0, \sigma_e^2)$ and σ_e^2 is the covariate error of the GS model in the form of BLUEs. β_i is the power of the GS model, and it can be considered as a regularization parameter when fitting a neural network (NN) GS model. The regularization parameter is a function of tradeoff between bias and variance. This clearly indicates that the GS model is over-fitting or under-fitting the training set. Other parameters of the model are controlled by the regularization parameter. In the network, an input layer with large weights can lead to large changes in the output layer as a result of the model [39]. In this situation, parameter estimation and model performance (β_i) are likely to be poor for new data. Therefore, when using modern statistical algorithms such as kernel or tree ensemble, the weights in the input layers are kept small [40]. Feature selection and regularization are two useful concepts in situations where, in the GS model, the number of genetic parameters is much larger than the number of observations. Feature selection is required to prevent over-fitting or under-fitting of selection or classification models, and it minimizes the computation time and loss function error of the model [1,41]. Regularization attempts to account for genetic hyper-parameters such as number of clusters in the population structure, the effect of MAF on genomic relationship matrix (GRM), LD decay and SNP covariate effects [42]. Thus, this approach leads to a higher performance in estimation, when the regularization parameter is very well defined in the given GS model. In this study, we evaluate different GS models and their predictive ability to improve prediction accuracy in the context of a phenotypic and genotypic NUE dataset of 221 bread wheat genotypes among three classes of regression learning methods, kernel and ensemble algorithms. We emphasize that the focus of this study is to compare GS methods based on ensemble learning algorithms and regression approaches with the aim of (i) optimizing genetics parameters and hyper-parameters of the population in the given GS model, (ii) identifying an appropriate regularization parameter for a given specific GS problem and (iii) demonstrating the performance of the best GS model through bias–variance analysis and error measurement.

2. Results

2.1. Genetic Parameters and Hyper-Parameter Estimation

Genetic parameter estimates and their 95% CL bootstraps for the *NUE* vector of 221 bread wheat genotypes based on all GS models separately. SNP heritability estimates were derived from random SNP effects in the given GS model. The highest and lowest SNP heritability estimates were related to the STACK model (0.62) and the gBLUP model (0.28), respectively, at low N levels. At the HN level, the highest SNP heritability was found with the STACK model (0.71) and the lowest with the gBLUP (0.30) and BGLR (0.30) models. In both the training phase with $\mathrm{CV_{K\text{-}fold}} = 10$ and the testing phase with $\mathrm{CV_{K\text{-}fold}} = 5$, the STACK model, which is based on ensemble learning inference, had the highest *GEBV* mean under low and high N levels, at 0.69 and 0.76, respectively. For the rrBLUP, gBLUP, BGLR, RKHS and SVM models, differences in hypothesis testing between the *GEBV* means in the training and testing phases were significant at low and high N levels (p-value with $\alpha = 0.05$), indicating that the performance of these models is worse than against other inferences, such as ensemble learning GS models. Genetic hyper-parameter estimates including learning rate, number of iterations and batch size for the *NUE* vector of

221 bread wheat genotypes based on all GS models separately are shown in Table 1. Based on the regularization parameters of each model, the minimum learning rate of 0.01 was computed for the rrBLUP, LASSO, RKHS and BOOST models, using the rule $\alpha_j = \frac{100\alpha_0}{100+j}$, where α_0 is the initial learning rate 1 and j is a counter of epochs up to 9900. For the remaining GS models, the minimum learning rate of 0.001 was calculated. The comparison between accuracy (%) in both the training phase with $\text{CV}_{\text{K-fold}}$ = 10 and the testing phase with $\text{CV}_{\text{K-fold}}$ = 5, among all ensemble learning algorithms, including BOOST, BAGG and STACK, was not significant, indicating that the accuracy of these models is higher than other GS models with kernel and linear algorithms.

Table 1. Genetic parameter estimation for the *NUE* vector under low and high N levels using different genomic selection models.

Inference	Model	N Level	SNP-h^{2a}	Training Set ($\text{CV}_{\text{K\_fold}}$ = 10)				Test Set ($\text{CV}_{\text{K\_fold}}$ = 5)				*p*-Value [f]
				*GEBV*s Mean [b]	Boot. *GEBV*s Mean [c]	Vg [d]	Ve [e]	*GEBV*s Mean	Boot. *GEBV*s Mean	Vg	Ve	
Frequentist	rrBLUP	LN	0.30	0.65	0.66	27.49	64.15	0.64	0.66	28.77	67.13	0.05 *
		HN	0.33	0.73	0.70	9.79	19.88	0.73	0.70	8.77	17.81	
	gBLUP	LN	0.28	0.64	0.61	26.78	68.87	0.64	0.61	22.61	54.14	0.044 *
		HN	0.30	0.71	0.69	4.93	11.51	0.71	0.68	4.57	10.61	
Bayesian	LASSO	LN	0.31	0.62	0.61	32.40	72.12	0.62	0.60	32.03	71.12	0.144 ns
		HN	0.31	0.70	0.68	7.94	17.69	0.69	0.65	8.39	18.69	
	BGLR	LN	0.32	0.63	0.68	35.49	75.43	0.65	0.65	35.35	75.12	0.042 *
		HN	0.30	0.72	0.69	9.70	22.65	0.74	0.65	26.24	61.24	
Kernel	RKHS	LN	0.45	0.64	0.64	57.11	69.81	0.64	0.64	53.29	65.14	0.031 *
		HN	0.61	0.72	0.72	28.16	18.01	0.72	0.71	42.71	27.21	
	SVM	LN	0.38	0.13	0.22	24.30	39.66	0.18	0.32	24.64	40.21	0.048 *
		HN	0.57	0.18	0.29	73.09	55.14	0.18	0.33	59.66	45.01	
Ensemble	BOOST	LN	0.48	0.61	0.61	62.84	67.08	0.61	0.65	61.92	67.08	0.164 ns
		HN	0.62	0.68	0.69	28.73	17.61	0.68	0.72	27.76	17.02	
	BAGG	LN	0.55	0.60	0.68	71.03	58.12	0.61	0.71	69.82	57.13	0.679 ns
		HN	0.55	0.64	0.71	39.80	32.57	0.64	0.74	38.12	31.19	
	STACK	LN	0.62	0.69	0.78	49.33	30.24	0.69	0.79	50.85	31.17	0.0924 ns
		HN	0.71	0.76	0.78	72.98	29.81	0.76	0.79	73.49	30.02	

a—SNP-h^2: SNP-Heritability, b—*GEBV*s mean: mean of genomic estimated breeding values, c—Boot. *GEBV*s mean: Bootstrap mean of genomic estimated breeding values, d—V_g: Genetic variance, e—V_e: Error variance, f—*p*-value: H0: there is no significant difference between the means of genomic estimated breeding values (*GEBV*s) in training and testing sets under low and high N levels ($\alpha = 0.05$), significance levels of *p*-value * $p \leq 0.01$, ns = not significant.

2.2. *Bias–Variance Tradeoff in GS Models*

The bias–variance tradeoff analysis for the *NUE* vector generated from the 150 K affymetrix SNP Chip, under low and high N is shown in (Tables 2 and 3). As can be seen, the loss value of a given GS model using the Scikit-learn algorithm indicates an irreducible error that is constant at both low and high N levels, and it is possible to minimize and control the effect of hyper-parameters in the model that were not defined in the given model. The effects of the main genetic parameters depending on the definition of the given model are clearly shown in Figure 1a,b. Based on the genetic structure and kinship of the population after model regularization, K = 3 was determined as the optimal number for model complexity analysis. At low N levels, both SVM and RKHS models with kernel inference exhibited the highest and lowest bias and variance, respectively. Thus, these models may represent upper and lower thresholds for the bias–variance tradeoff. At high N levels, the SVM model with kernel inference and rrBLUP with frequentist linear inference had the highest and lowest bias and variance, respectively. It can be concluded that the interaction between N level and wheat genotypes based on SNP information is significant. At both low and high N levels, the behavior of the BOOST and BAGG models with ensemble learning inference showed a moderate tradeoff between under and upper fit. Thus, after mean comparison (LSD (0.05)), we can conclude that the bias, variance and average of expected loss between the GS models overall resulted in statistically significantly different means. This difference indicates that the performance of some models to predict *GEBV*s is higher than others.

Table 2. Genetic hyper-parameter estimations for *NUE* vector under low and high N levels using different genomic selection models.

Inference	Model	N Level	Training Set (CV_{K_fold} = 10)				Test Set (CV_{K_fold} = 5)				*p*-Value [b]
			LR [a]	No. of Iteration	No. of Batch Size	Accuracy (%)	LR	No. of Iteration	No. of Batch Size	Accuracy (%)	
Frequentist	rrBLUP	LN HN	0.01 0.01	27 27	10 100	81.12 81.09	0.01 0.01	27 27	5 25	92.22 92.21	0.007 **
	gBLUP	LN HN	0.001 0.001	27 27	10 100	80.88 80.41	0.001 0.001	27 27	5 25	92.21 92.22	0.001 **
Bayesian	LASSO	LN HN	0.01 0.01	468 468	10 100	85.51 85.12	0.01 0.01	468 468	5 25	92.43 92.44	0.0024 *
	BGLR	LN HN	0.001 0.001	468 468	10 100	84.47 85.77	0.001 0.001	468 468	5 25	92.74 92.76	0.0031 *
Kernel	RKHS	LN HN	0.01 0.01	2050 2050	100 1000	85.11 85.33	0.01 0.01	2050 2050	25 250	95.64 95.37	0.001 **
	SVM	LN HN	0.001 0.001	2050 2050	100 1000	84.04 85.77	0.001 0.001	2050 2050	25 250	95.33 95.18	0.0014 **
Ensemble	BOOST	LN HN	0.01 0.01	5520 5520	100 1000	91.12 92.11	0.01 0.01	5520 5520	25 250	96.01 96.12	0.098 ns
	BAGG	LN HN	0.001 0.001	5520 5520	100 1000	92.31 92.16	0.001 0.001	5520 5520	25 250	96.48 97.22	0.1445 ns
	STACK	LN HN	0.001 0.001	5520 5520	100 1000	93.58 93.79	0.001 0.001	5520 5520	25 250	97.54 97.84	0.0905 ns

a—LR: Learning rate of given GS model, b—*p*-value: H0: there is no significant difference between model accuracy in train and test sets under low and high N levels ($\alpha = 0.05$), significance levels of *p*-value * $p \leq 0.01$, ** $p \leq 0.001$, ns = not significant.

Table 3. Bias–variance tradeoff analysis for *NUE* vector under low and high N levels using different genomic selection models.

Inference	Model	Low N				High N			
		$\overline{Bias}$	$\overline{Var}$	$\overline{Exp.Loss}$	$\overline{Sk.Loss}$	$\overline{Bias}$	$\overline{Var}$	$\overline{Exp.Loss}$	$\overline{Sk.Loss}$
Frequentist	rrBLUP	9.1×10^5	6.1×10^7	6.2×10^7	0.0009	0.2×10^6	0.9×10^6	1.1×10^6	0.0009
	gBLUP	3.1×10^6	0.8×10^6	1.2×10^6	0.0009	0.22×10^6	3.7×10^6	4.0×10^6	0.0009
Bayesian	LASSO	3.1×10^6	0.82×10^6	1.2×10^6	0.0009	0.2×10^6	0.8×10^6	1.2×10^6	0.0009
	BGLR	3.1×10^6	5.57×10^6	6.4×10^6	0.0009	0.2×10^6	5.2×10^6	5.5×10^6	0.0009
Kernel	RKHS	25×10^4	49×10^4	55×10^4	0.00081	0.3×10^6	1.1×10^6	1.5×10^6	0.00081
	SVM	0.1×10^{12}	$2.5 \times 1\ 0^{12}$	2.7×10^{12}	0.00080	0.15×10^{10}	1.5×10^{10}	1.6×10^{10}	0.00080
Ensemble	BOOST	4.1×10^6	4.2×10^6	5.1×10^6	0.00007	0.09×10^7	1.1×10^7	1.19×10^7	0.00007
	BAGG	0.75×10^7	0.5×10^7	0.8×10^7	0.00007	0.11×10^7	1.6×10^7	1.75×10^7	0.00007
	STACK	0.41×10^6	0.6×10^6	1.2×10^6	0.00002	0.5×10^6	5.1×10^6	5.8×10^6	0.00002
Mean		12.3×10^9	30.0×10^9	33×10^9	0.00059	1.67×10^8	1.68×10^9	1.78×10^9	0.00059
LSD (0.05)		0.92×10^9	0.48×10^9	1.25×10^9	2.11	0.34×10^8	0.70×10^9	1.03×10^9	2.11

$\overline{Bias}$: average of bias, $\overline{Var}$: average of variance, $\overline{Exp.Loss}$: average expected loss that is equal to mean square error (MSE) of GS model by bias–variance analysis simultaneously using k-nearest neighbor algorithm and $\overline{Sk.Loss}$: mean irreducible error (IE) of GS model by bias–variance analysis simultaneously using *Scikit-learn* algorithm. LSD (0.05) for comparison of GS models performances.

2.3. Error Measurement of GS Models

Measurement error in *GEBV* predictors causes bias in estimated genetic parameters in GS models. Thus, error measurement of GS models clarifies the origin of this error, which is related to model definition or variables in the model, such as marker information and phenotypic values. After performing the bias–variance analysis to obtain an overall view of the performance of GS models, the adaptive standard error of prediction for each model under low and high N levels, were pairwise compared (Figure 2). Under low N levels, the BAGG model, and at high N levels, the STACK model, showed the lowest error in predicting the genomic parameters. The lowest error in prediction confirms the result of the bias–variance analysis on these two models with ensemble learning inference. Probably due to collinearity in the whole genomic regression analysis, rrBLUP has the highest error in the prediction at both low and high N levels compared to other models (Figure S1a,b). Therefore, STACK can be selected as the best GS model with a high performance for predicting *GEBV*s.

a. LN

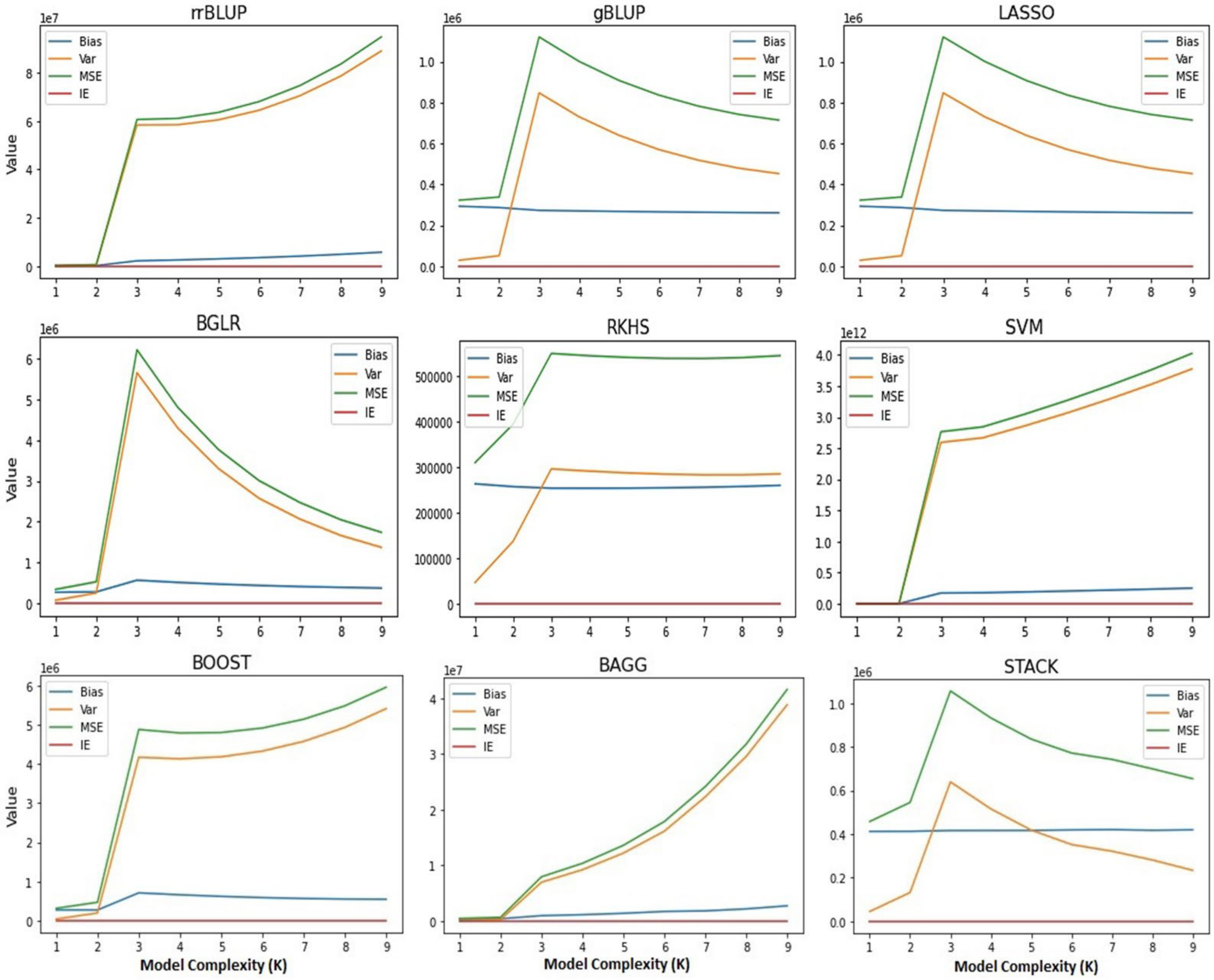


Figure 1. *Cont.*

b. HN

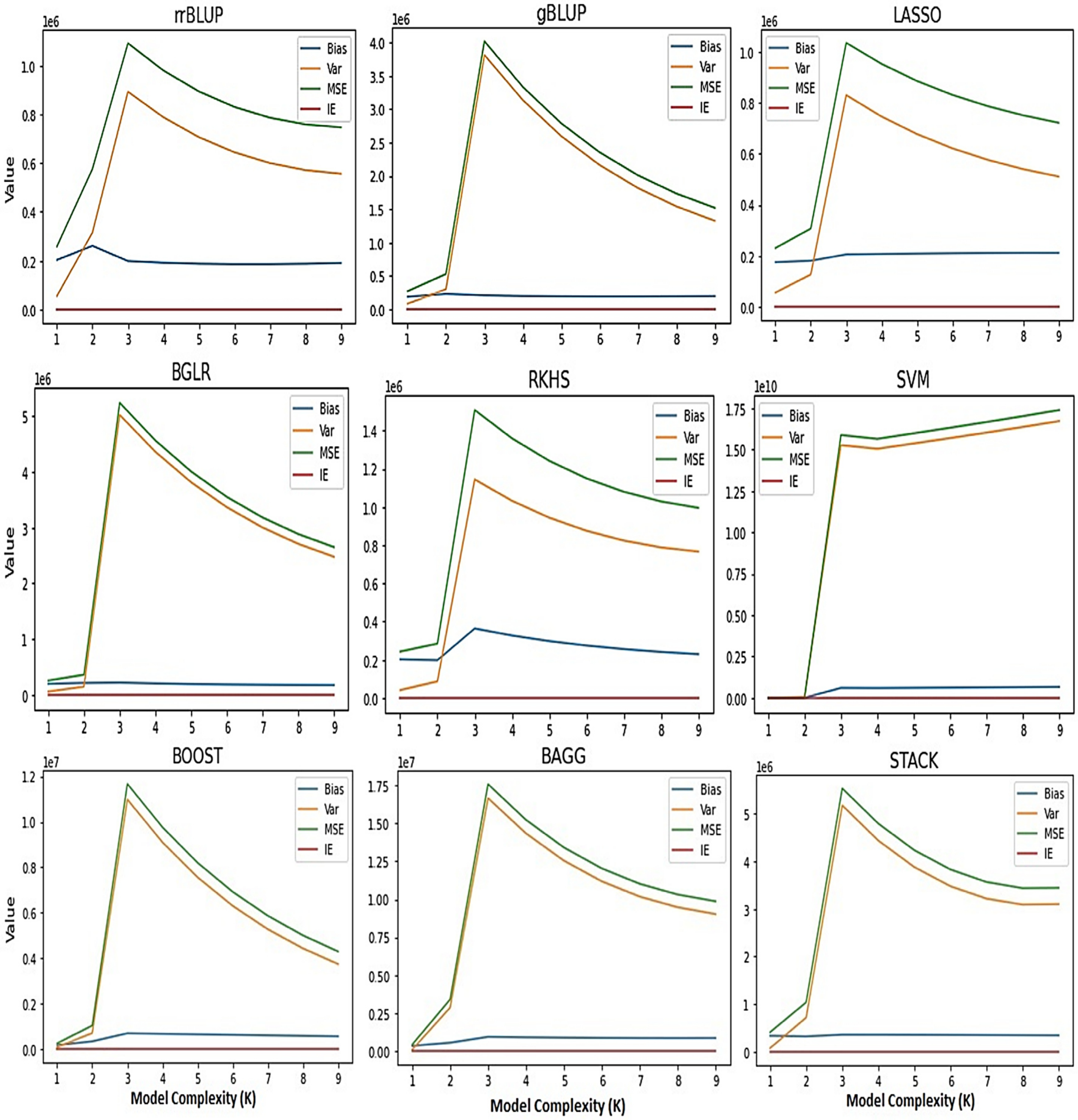


Figure 1. (**a**) Model complexity analysis using k-nearest neighbor (KNN) algorithm for the *NUE* vector under low N levels using different genomic selection models. X-axis represents value of bias (blue line), variance (orange line), MSE (green line) and irreducible error (red line); y-axis shows GS model complexity. le (number): ×10 $^{\text{number}}$. (**b**) Model complexity analysis using k-nearest neighbor (KNN) algorithm for the *NUE* vector under high N levels using different genomic selection models. X-axis shows the value of bias (blue line), variance (orange line), MSE (green line) and irreducible error (red line); y-axis shows GS model complexity, le (number): ×10 $^{\text{number}}$.

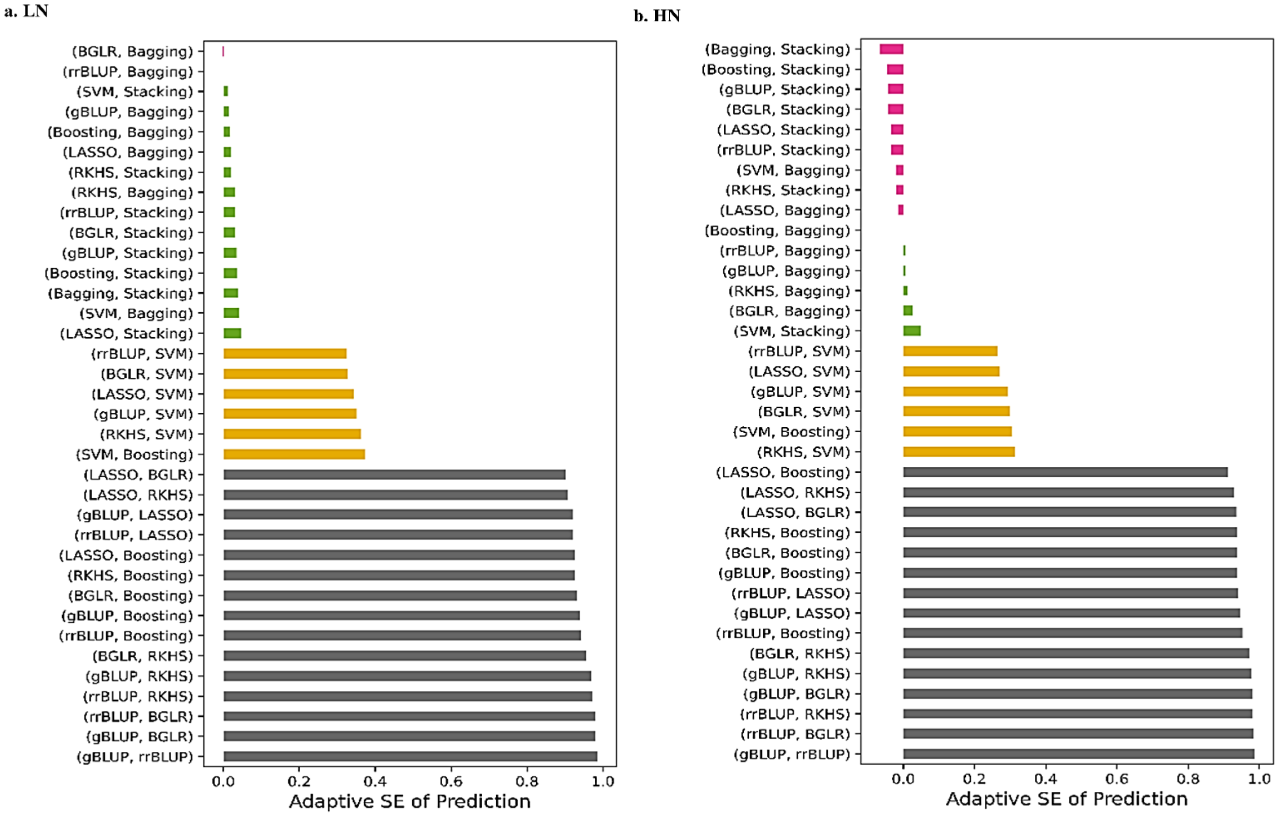


Figure 2. Adaptive standard error (SE) of prediction for pairwise comparison of GS models under low and high N levels. The ranges of adaptive SE values were normalized from −0.2 to 1. Pink color: the adaptive SE of prediction for the pairwise comparisons, is less than zero, green color: the adaptive SE of prediction for the pairwise comparisons, is equal by zero, orange color: the adaptive SE of prediction for the pairwise comparisons, is almost equal by 0.3 and gray color: the adaptive SE of prediction for the pairwise comparisons, is equal by 1.

2.4. Genetic Selection Gain Estimation Based on Selected Model

Genetic selection gain was estimated using the genetic value of genotypes in the GRM and STACK models. For this purpose, the wheat population was divided into training set (75%) and testing set (25%) genotypes. As can be observed in Figure 3a, the *NUE* values (%) predicted based on *GEBV*s are positively correlated with the actual *NUE* values, as the regression slope in the training set is positive at both low and high N levels. Additionally, the distribution of genotypes around the regression line is almost the same in the training and testing sets, indicating that the accuracy of the STACK is a good fit to the predicted values of *NUE* (%) at both low and high N levels. In Figure 3b, the RE (%) of the top ten genotypes with the highest *GEBV*s based on the STACK model is shown. Six genotypes with high RE (%) were replicated at both low and high N levels. Therefore, allelic variation in these top six genotypes was plotted against the minor and major alleles of the entire population (Figure 4). As can be observed, there is a significant difference in allelic content between the average of the top six genotypes and the minor alleles in the whole population, but there is no difference in the major alleles, and they are in the same group. Since the regularization parameter in this model is MAF with bias and variance, it could find more than 50% of the top genotypes with high accuracy.

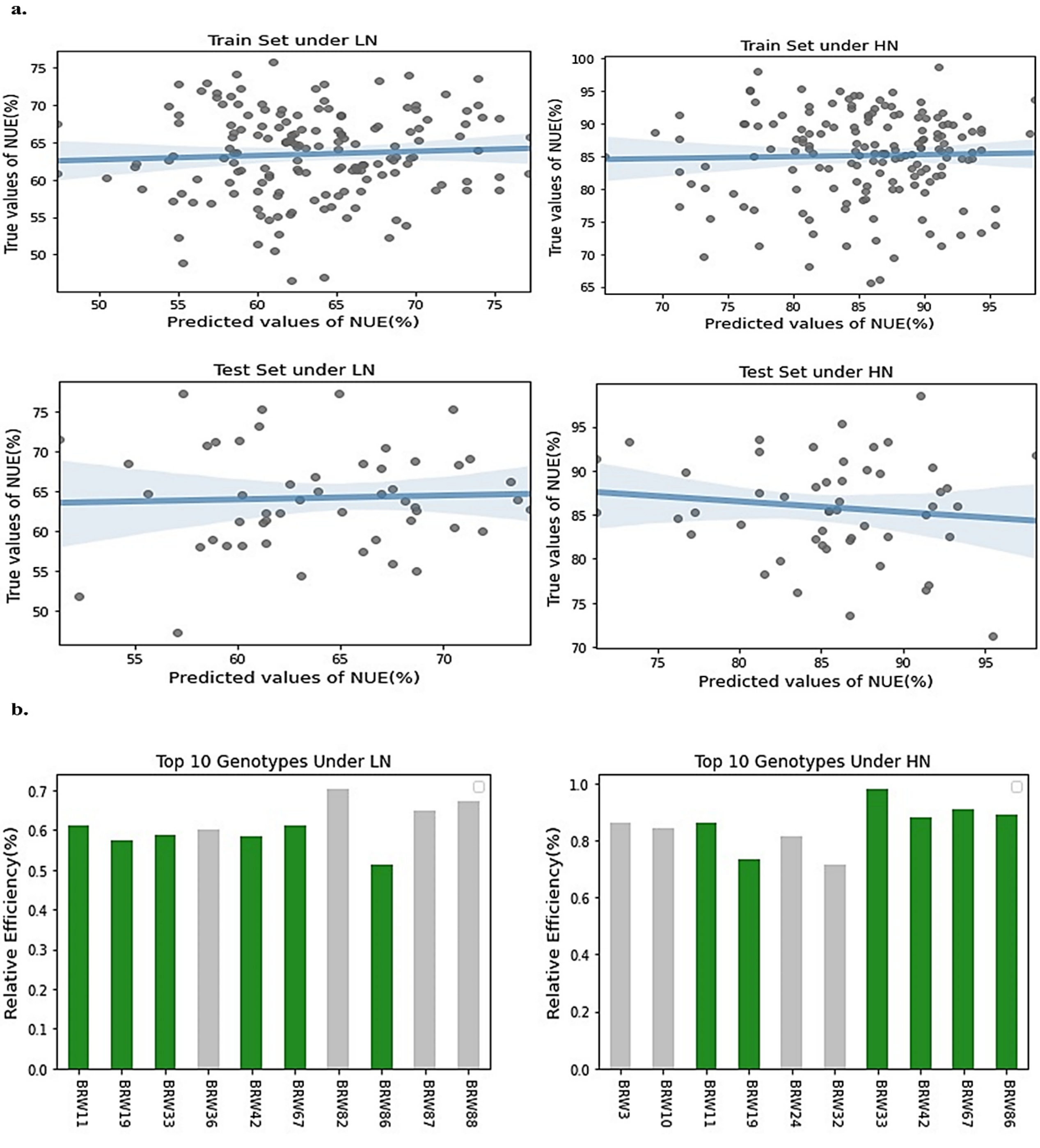
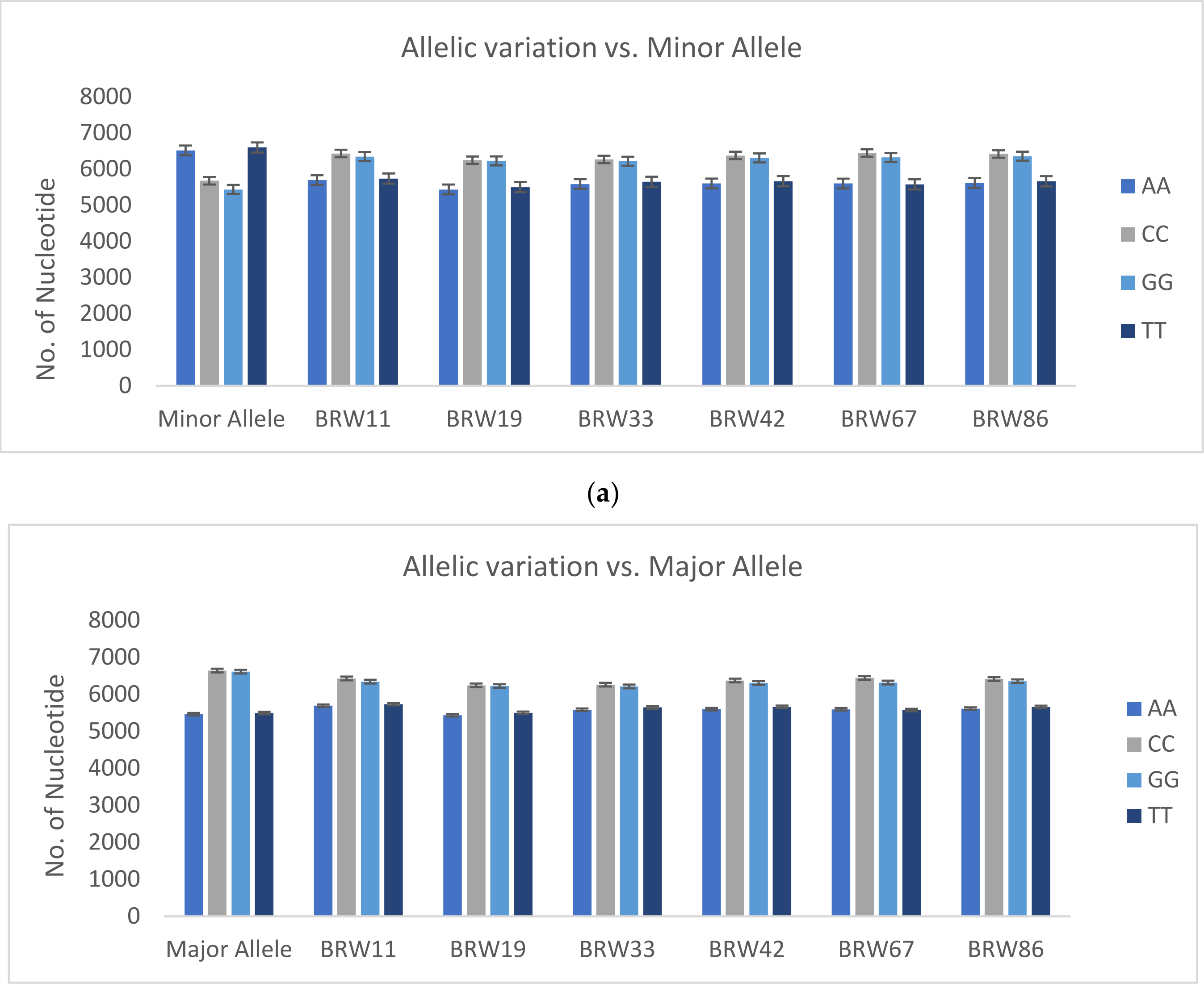


Figure 3. (**a**) Bias–variance analysis and adaptive SE of prediction indicate that the STACK model shows the best performance and accuracy. The predicted values of *NUE* (%) for both training and test sets under low and high N levels are displayed on the x-axis, while the true values from the dataset are displayed on the y-axis. (**b**) GS gain in the form of relative efficiency (%) for the STACK model among the 221 wheat genotypes and the top 10 genotypes under low and high N levels has been specified. Genotypes with green color under both low and high N levels are duplicate. Genotypes with gray color are distinct at each N level.



(**a**)

(**b**)

Figure 4. (**a**) Mean comparison between allele content of six top genotypes with highest GS gain for *NUE* (%) vector versus minor allele content in whole population. (**b**) Mean comparison between allele content of six top genotypes with highest GS gain for *NUE* (%) vector versus major allele content in whole population.

3. Discussion

Currently, phenotyping is still expensive, requires high-throughput technologies, and is a very time-consuming process compared to genotyping. The use of modern statistical models could be a logical solution to this challenge. In recent decades, the use of new technologies such as sensors, robotics and satellite data has led to high-throughput phenotypic data. In parallel, new techniques such as next generation sequencing (NGS) have made it possible to create a big training dataset for the trait of interest. Thus, GS with a big genomic dataset and high-density markers as features in the model requires statistical machine learning methods with more computational power, especially for complex traits such as nitrogen use efficiency (NUE) in wheat [14,21]. Many machine learning methods have been adapted and developed for GP and GS [1,43]. In particular, several important parametric models, such as neural networks, kernel regression and ensemble learning algorithms have been generalized to handle high-density SNP data [38,44]. They provide GS studies with more comprehensive and flexible methods to estimate *GEBV*s with high accuracy. For the

*GEBV*s to achieve high accuracy, all the assumptions underlying the *GEBV* equations are required to be met. These assumptions are numerous and relate to several factors, such as the extent of coverage of genetic variability at QTLs by markers, which depends on the number and placement of markers in the genome. In addition, the estimation of the quality of markers is critical, as it is influenced by allele frequencies and the degree of linkage disequilibrium with QTLs, and may vary between populations, especially between reference and selection populations. Another crucial factor is the absence of nonadditive effects. The model rrBLUP is based on the additive relationship matrix (A) formed by the estimation of IBS as a marker-associated trait. Ridge regression is the main core of rrBLUP and is used to analyze genotypic data when they suffer from multicollinearity. With increasing marker density on genetic maps, the concept of multicollinearity is limited to strong LD, which is unusual for all genotypes between mono- and polymorphic SNPs. Therefore, even with ridge regression, unbiased least squares and large variance are observed in the GP results. Thus, the accuracy of model is affected by the training set, number of markers and heritability. Consequently, predicted genetic values are far from the actual values. The model gBLUP is based on the matrix A mating formed by the estimation of IBD and its relatives between the complex trait of interest and associated loci. The unavoidable and desired presence of LD between causal and/or marker loci modifies the simple interpretation of the matrices. The variability of LD in the genome due to heterogeneity within loci may lead to bias in the calculation of SNP heritability based on this genomic matrix. Both rrBLUP and gBLUP models are linear systems with an REML approach to predicting GS. Thus, the question of how to estimate genetic parameters such as variance–covariance components to calculate the fixed and random effects of high-density SNPs remains unanswered, and only point estimation of likelihood for *GEBV*s is available. However, GS linear models based on the REML approach can be specified without having to worry too much about the assumptions. However, Bayesian inference is more flexible when it comes to assumptions, and it has a range of answers that may change with each run [45]. The main challenge with BGLR and LASSO is to ensure that the distribution of statistical estimators follows the genetic parameters of the population. Therefore, it is often observed that kernel methods perform better compared to linear models. Kernel inference provides a linear solution in the feature space while being nonlinear in the input space, and it is a combination of linear and nonlinear parameters by definition. RKHS has partial similarity to rrBLUP and gBLUP models and utilizes a kernel matrix that represents Euclidean distances between focal points in Hilbert space. This kernel matrix in RKHS optimizes a more general structure of covariance between individuals compared to the GRM used to measure similarities in genetic values related to individuals. This allows greater flexibility in capturing complex relationships between individuals and improves the accuracy of predictions in GS. SVM can be interpreted as part of the class of kernel approaches. It is a method that is traditionally applied to classification problems, but has also been used for regression (prediction or selection). In SVM, optimization of the hyper-parameters can be carried out via various methods, but the most commonly used and convenient method is the grid search. The grid search method evaluates all possible combinations of hyper-parameters, allowing an exhaustive search in the hyper-parameter space. The BOOST, BAGG and STACK models are based on the DL algorithm with an ensemble method used to jointly solve a complex problem. A number of algorithms, generally nonparametric, are combined to improve the prediction or classification ability of the assembled model [46]. Our study has revealed that the definition and optimization of the regularization parameter is crucial to demonstrate the performance and accuracy of the GS model, which has not been sufficiently addressed in previous GS studies. However, in the BOOST model, the regularization parameter is only used to control the bias of the model. By adjusting the regularization parameter, the model can be made less complex and less prone to over-fitting. In the BAGG model, the regularization parameter is used to control the variance of the model. By reducing the variance, the model can be made more stable and less sensitive to noise. In STACK, both bias and variance are considered by adjusting the regularization parameter to find a balance between model complexity and stability. Thus, our study confirmed that the results of bias–variance tradeoff and adaptive prediction error for the STACK model were intermediate compared with other models. This remarkable result for the STACK model is consistent with previous results [23,47]. In all ensemble models, especially in the STACK, the number of epochs and batches of hyper-parameters need to be specified along with the activation process. The number of epochs determines how often the weights of model are updated, and it is affected by LR on the training dataset. It is important to carefully tune these hyper-parameters to optimize the performance of the model and avoid over-fitting or under-fitting. Therefore, a smaller LR in the training data set and a batch size of 1000 produces maximum SNP heritability and *GEBV* means. Thus, the epoch number indicates the forward and backward runs of the whole training data through the model. However, epoch number can be computationally intensive for the memory of computation, so these are divided into small batch sizes.

4. Materials and Method

4.1. Phenotypic Data

In this study, a set of 221 bread wheat genotypes from the Breeding Innovations in Wheat for Resilient Cropping Systems (BRIWECS) project were grown at the agricultural research station, Campus Klein-Altendorf, University of Bonn, Germany, in three cropping seasons during 2018, 2019 and 2020, in a split-plot design. The *NUE* value of each genotype under low-N (LN) and high-N (HN) fertilizers was calculated using the following formula:

$$NUE = \frac{GY}{N_s} = (\frac{N_t}{N_s})(\frac{GY}{N_t}) \tag{3}$$

where GY is grain yield (gr/m^2), N_s is the nutrient supplied and N_t is the total aboveground plant nutrient at maturity [48].

4.2. Genotypic Data

In order to characterize the *NUE* vector among the bread wheat population, a platform of 150K affymetrix SNP Chip at TraitGenetics GmbH (SGS GmbH Gatersleben, Germany), was used. To minimize monomorphism in the Chip, the SNPs with MAF ≤ 0.05 were removed. After checking for SNPs that deviated from the Hardy–Weinberg equilibrium (HWE), only 22,489 polymorphic SNP markers, were remained and were used in GS models.

4.3. Construction of GRM

Basically, the GRM is used as a kinship matrix in genome-wide association studies (GWASs) and GS models. For the rrBLUP model, based on the method suggested by [49], the covariance between individuals g_i and g_j can be equal to the covariance of SNP_{ij}; therefore, the GRM was calculated using $G = \frac{\Sigma_{k=1}^{L}(g_{ik}-p_k)^2}{\Sigma_{k=1}^{L}p_k(1-p_k)}$ and $p_k = \frac{1}{n}\Sigma_{i=1}^{n}g_{ik}$, where L is the number of loci, p_k is the MAF for the loci k and g_i. In the rrBLUP model, $y = \mu + \sum_{i=1}^{p} X_i g_i + e$, where y is the *NUE* vector with n genotypes, p is the total number of SNPs, X_i is the matrix of random SNP effects coded as (-1, 0 and 1), g_j is the main diameter of GRM and H_0 denoted in the form of $\sigma_g^2 = 0$ based on identity by state (IBS) only, and e is the residuals of the model. In the gBLUP model, $y = \mu + Zu + e$, where y is the *NUE* vector, Z is the matrix of genetic values for an individual, $V(u) = K\sigma_g^2$, where K is the GRM as a kinship matrix and σ_g^2 is the additive genetic variance with IBS or identity by descent (IBD). In the LASSO model, marker effects were calculated as $\hat{\beta} = (X^*X - \lambda I)^{-1}X^*y$, where $[\lambda] = \frac{\sigma_e^2}{\sigma_\beta^2}$ is taken as a kinship matrix and I is an identity matrix. σ_β^2 and σ_e^2 are computed from SNP heritability and phenotypic variance, respectively [50,51]. The BGLR model with a Bayes factor was fitted to $y = \mu + X\alpha + Z\beta + XZ_{\alpha g}u + e$, where y is the *NUE* vector, X is a matrix with genotypes and SNPs, α is the corresponding vector of SNP effects which captures small effects of all SNPs, g is a vector that captures prior distribution of SNP effects with prior distribution of $g \sim N\left(0,\ G^2_{gu}\right)$, where G is a marker-derived genomic relationship matrix, u is a vector of Bayes factors for the SNP matrix and e is the residual term [52,53]. In the RKHS, SVM, Boosting, Bagging and Stacking GS models, the objective is to classify SNPs with higher accuracy. Typically, in these models, the GRM with a linear scale has been replaced by a distance matrix with a Euclidean family scale, so the regularization parameter in any given GS model is a prerequisite for obtaining the kinship matrix.

4.4. Genomic Selection Models

1—In order to carry out GS based on frequentist and Bayesian perspectives, models including ridge regression best linear unbiased prediction (rrBLUP) with R/*rrBLUP* [54], genomic best linear unbiased prediction (gBLUP) with R/*BGLR* [55,56], least absolute shrinkage and selection operator (LASSO) with R/*glmnet* [45] and Bayesian generalized linear regression (BGLR) [48] with R/*BGLR* were applied. 2—To perform GS, models based on Kernel whole genome regression reproducing kernel Hilbert spaces (RKHS) with R/*BWGS* [57–59] and support vector machine (SVM) with R/*kernlab* [8,60] were utilized. 3—From tree ensemble algorithms, Boosting with R/*gbm* [61,62] and Bagging [33,63] with R/*ipred* were used. To run the Stacking GS model, [63,64] Python/*Scikit-learn* was utilized. All three category of GS models were run among the *NUE* vector of 221 wheat genotypes under low and high N levels.

4.5. Genetic Parameters and Hyper-Parameters Estimation

To solve the fixed effects coming from environmental factors or population structure, and random effects coming from SNPs, the residual maximum likelihood (REML) in all GS models were minimized through the BLUP vector as additive genomic variance and the BLUE vector as residual variances, and then SNP effects were estimated. The h^2, based on SNP information, was estimated as $h^2 = \frac{\sigma^2_a}{\sigma^2_a+\sigma^2_e}$, where σ^2_a and σ^2_e are additive genomic and residual variances, respectively. In the CV strategy with the training set (K-fold = 10) and test set (K-fold = 5), genetic parameters, including *GEBV*s mean, bootstrap of *GEBV*s mean, genetic variance and error variance were applied to assess accuracy in the given GS model. In the CV approach, accuracy is measured using the correlation coefficient $r(\check{y}Obs., \check{y}GEBV)$ between observed values (OBVs) and genomic estimated breeding values (*GEBV*s), which were estimated to divide the population into validation and training sets. The genotypes assigned to the validation set were used as predicted breeding values, and the remaining genotypes were used for the training set. Therefore, besides the genetic parameters in the definition of the GS models, before starting the training and testing processes, hyper-parameters were determined. They included learning rate (LR), number of hidden layers, number of iterations and batch size per one epoch computed. Based on the regularization parameters of a given model, the minimum learning rate was optimized with the rule $\alpha_j = \frac{100\alpha_0}{100+j}$, where α_0 is the initial learning rate 1 and j is a counter of epochs to 9900 [65,66]. They were also defined based on SNP effects (p-values) generated from the *NUE* vector under low and high N content.

4.6. Bias–Variance Tradeoff in GS Models

Model evaluation was carried out for all GS models via analysis of bias and variance. The *NUE* vector was considered as a target trait at low and high N contents among all 221 bread wheat genotypes. Bias was taken as the difference between *GEBV*s and the true *NUE* vector. Variance was measured to identify the difference between parameters of a given model and its training set. In order to deal with model over- and under-fitting, the mean squared error (MSE) based on k-nearest neighbor (KNN) algorithm was calculated.

To reduce model complexity after dealing with outlier values in the *NUE* vector, irreducible error (IE) was measured.

4.7. Error Measurement between GS Models

To evaluate the GS models, additional error measurements were performed. Models based on *GEBV*s generated from the *NUE* vector under low and high N content into two pairwise groups were compared. To check the distribution of SNPs *p*-value, a matrix of standard errors (SE) was generated from each GS model under low and high N levels, after randomly sampling SNP effects 2000 times with replacement, under the null hypothesis *N* (0, 1), which is necessary when SNPs' *p*-values are near to the normal distribution. For SNPs with *p*-values far from the normal distribution, the adaptive standard error of the prediction matrix was calculated with an adjusted FDR threshold of 0.05.

4.8. Genetic Selection Gain Estimation Based on the Selected Model

The equation $R = \frac{ir(\check{y}Obs.,\check{y}GEBV)}{y}\, log \begin{bmatrix} X'X & X'Z \\ Z'X & 1+\frac{1^{G^{*-1}}}{\sigma_e^2} \end{bmatrix}$ was utilized to estimate the expected genetic selection gain (R). i is the selection intensity, $r(\check{y}Obs.,\check{y}GEBV)$ is the selection accuracy, y is the number of years. The component $log \begin{bmatrix} X'X & X'Z \\ Z'X & 1+\frac{1^{G^{*-1}}}{\sigma_e^2} \end{bmatrix}$ is equal to β_i which is the power of the given GS model. In this component, X is the incidence matrix for the proportion of individuals in the population structure (n_{ps}) × marker (m) with fixed effect, X' is the transformed X, Z is a designed matrix for the effect of genotype (n) × marker effect (p), including all random effects, Z' is the transformed Z, G^{*-1} is an invert matrix of the genomic relationship matrix (GRM), when the effect of non-associated markers are shrunken toward null with $N(0, \sigma_e^2)$, and σ_e^2 is the covariate error of the GS model in the form of BLUEs. To clarify genetic gain from the GS model against gain from phenotypic selection, a relative efficiency (RE) index was developed. The RE of indirect GS under low and high N levels was calculated with $RE_{per\ N\ level} = \frac{r(\check{y}Obs.,\check{y}GEBV)}{\sqrt{R^2}}$.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms241814275/s1.

Author Contributions: Conceptualization, M.B.S.; Methodology, M.B.S.; Formal analysis, M.B.S.; Investigation, A.B. and J.L.; Resources, S.D.; Writing—original draft, M.B.S.; Supervision, S.D. and J.L.; Project administration, A.B. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: Hereby the authors would like to acknowledge the kindly support of DAAD who gave us generous opportunity to pursue our research and financing the study. Additionally, the authors are grateful to the Institute for Crop Science and Resource Conservation (INRES)-Plant Breeding, University of Bonn, Germany, for responsible supervision and execution plan on the study.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Abdollahi-Arpanahi, R.; Gianola, D.; Peñagaricano, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evol.* **2020**, *52*, 1–15. [CrossRef]
2. Wang, X.; Xu, Y.; Hu, Z.; Xu, C. Genomic selection methods for crop improvement: Current status and prospects. *Crop J.* **2018**, *6*, 330–340. [CrossRef]
3. Crossa, J.; Pérez-Rodríguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; de los Campos, G.; Burgueño, J.; González-Camacho, J.M.; Pérez-Elizalde, S.; Beyene, Y.; et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* **2017**, *22*, 961–975. [CrossRef] [PubMed]

4. Bhat, J.A.; Ali, S.; Salgotra, R.K.; Mir, Z.A.; Dutta, S.; Jadon, V.; Tyagi, A.; Mushtaq, M.; Jain, N.; Singh, P.K.; et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* **2016**, *7*, 221. [CrossRef] [PubMed]
5. Bernardo, R. Bandwagons I, too, have known. *Theor. Appl. Genet.* **2016**, *129*, 2323–2332. [CrossRef] [PubMed]
6. Sneller, C.H.; Mather, D.E.; Crepieux, S. Analytical approaches and population types for finding and utilizing QTL in complex plant populations. *Crop Sci.* **2009**, *49*, 363–380. [CrossRef]
7. Schön, C.C.; Utz, H.F.; Groh, S.; Truberg, B.; Openshaw, S.; Melchinger, A.E. Quantitative trait locus mapping based on resampling in a vast maize Testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics* **2004**, *167*, 485–498. [CrossRef]
8. Montesinos-López, O.A.; Martín-Vallejo, J.; Crossa, J.; Gianola, D.; Hernández-Suárez, C.M.; Montesinos-López, A.; Juliana, P.; Singh, R. A benchmarking between deep learning, support Vector Machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3 Genes Genomes Genet.* **2019**, *9*, 601–618. [CrossRef]
9. Werner, C.R.; Gaynor, R.C.; Gorjanc, G.; Hickey, J.M.; Kox, T.; Abbadi, A.; Leckband, G.; Snowdon, R.J.; Stahl, A. How population structure impacts genomic selection accuracy in cross-validation: Implications for practical breeding. *Front. Plant Sci.* **2020**, *11*, 592977. [CrossRef]
10. Delfini, J.; Moda-Cirino, V.; dos Santos Neto, J.; Ruas, P.M.; Sant'ana, G.C.; Gepts, P.; Gonçalves, L.S. Population structure, genetic diversity and genomic selection signatures among a Brazilian common bean germplasm. *Sci. Rep.* **2021**, *11*, 1–12. [CrossRef]
11. Lyra, D.H.; Galli, G.; Alves, F.C.; Granato, Í.S.; Vidotti, M.S.; Bandeira e Sousa, M.; Morosini, J.S.; Crossa, J.; Fritsche-Neto, R. Modeling copy number variation in the genomic prediction of maize hybrids. *Theor. Appl. Genet.* **2018**, *132*, 273–288. [CrossRef] [PubMed]
12. Won, S.; Park, J.-E.; Son, J.-H.; Lee, S.-H.; Park, B.H.; Park, M.; Park, W.-C.; Chai, H.-H.; Kim, H.; Lee, J.; et al. Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. *Front. Genet.* **2020**, *11*, 134. [CrossRef] [PubMed]
13. Han, J.; Gondro, C.; Reid, K.; Steibel, J.P. Heuristic hyperparameter optimization of deep learning models for genomic prediction. *G3 Genes Genomes Genet.* **2021**, *11*, jkab032. [CrossRef]
14. Okut, H. Deep learning algorithms for complex traits genomic prediction. *Hayvan Bilim. ve Ürünleri Derg.* **2021**, *4*, 225–239. [CrossRef]
15. Jannink, J.-L.; Lorenz, A.J.; Iwata, H. Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genom.* **2010**, *9*, 166–177. [CrossRef] [PubMed]
16. Zhou, X.; Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **2014**, *11*, 407–409. [CrossRef] [PubMed]
17. Berger, S.; Pérez-Rodríguez, P.; Veturi, Y.; Simianer, H.; los Campos, G. Effectiveness of shrinkage and variable selection methods for the prediction of complex human traits using data from distantly related individuals. *Ann. Hum. Genet.* **2015**, *79*, 122–135. [CrossRef]
18. Guo, P.; Zhu, B.; Niu, H.; Wang, Z.; Liang, Y.; Chen, Y.; Zhang, L.; Ni, H.; Guo, Y.; Hay, E.H.; et al. Fast genomic prediction of breeding values using parallel Markov chain Monte Carlo with convergence diagnosis. *BMC Bioinform.* **2018**, *19*, 1–11. [CrossRef]
19. Zhang, H.; Yin, L.; Wang, M.; Yuan, X.; Liu, X. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front. Genet.* **2019**, *10*, 189. [CrossRef]
20. Shi, S.; Li, X.; Fang, L.; Liu, A.; Su, G.; Zhang, Y.; Luobu, B.; Ding, X.; Zhang, S. Genomic prediction using Bayesian regression models with global–local prior. *Front. Genet.* **2021**, *12*, 628205. [CrossRef]
21. Sandhu, K.S.; Lozada, D.N.; Zhang, Z.; Pumphrey, M.O.; Carter, A.H. Deep learning for predicting complex traits in spring wheat breeding program. *Front. Plant Sci.* **2021**, *11*, 613325. [CrossRef] [PubMed]
22. Morota, G.; Koyama, M.; M Rosa, G.J.; Weigel, K.A.; Gianola, D. Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. *Genet. Sel. Evol.* **2013**, *45*, 1–15. [CrossRef] [PubMed]
23. Reinoso-Peláez, E.L.; Gianola, D.; González-Recio, O. Genome-enabled prediction methods based on machine learning. *Methods Mol. Biol.* **2022**, *2467*, 189–218. [CrossRef] [PubMed]
24. Pal, R. Feature selection and extraction from heterogeneous genomic characterizations. In *Predictive Modeling of Drug Sensitivity*; Academic Press: Cambridge, MA, USA, 2017; pp. 45–81. [CrossRef]
25. Yu, L. Feature selection for Genomic Data Analysis. In *Computational Methods of Feature Selection*, 1st ed.; Liu, H., Motoda, H., Eds.; CRC: New York, NY, USA, 2007; pp. 337–353. [CrossRef]
26. Hastie, T.; Friedman, J.; Tisbshirani, R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2017.
27. Shi, L.; Westerhuis, J.A.; Rosén, J.; Landberg, R.; Brunius, C. Variable selection and validation in multivariate modelling. *Bioinformatics* **2018**, *35*, 972–980. [CrossRef] [PubMed]
28. Morillo-Salas, J.L.; Bolón-Canedo, V.; Alonso-Betanzos, A. Dealing with heterogeneity in the context of distributed feature selection for classification. *Knowl. Inf. Syst.* **2020**, *63*, 233–276. [CrossRef]
29. Paul, J. Feature Selection from Heterogeneous Biomedical Data: Semantic Scholar. Undefined. 2015. Available online: https://www.semanticscholar.org/paper/Feature-selection-from-heterogeneous-biomedical-Paul/47054794c57a8c57665d83bed606fd40b7ef011f (accessed on 29 October 2022).

30. Rustam, Z.; Kharis, S.A. Multiclass classification on brain cancer with multiple support Vector Machine and feature selection based on kernel function. *AIP Conf. Proc.* **2018**, *2023*, 020233. [CrossRef]
31. Efron, B.; Hastie, T. *Computer Age Statistical Inference*; Cambridge University Press: Cambridge UK, 2021.
32. Cuevas, J.; Crossa, J.; Soberanis, V.; Pérez-Elizalde, S.; Pérez-Rodríguez, P.; Campos, G.d.; Montesinos-López, O.A.; Burgueño, J. Genomic prediction of genotype × environment interaction kernel regression models. *Plant Genome* **2016**, *9*. [CrossRef]
33. Li, B.; Zhang, N.; Wang, Y.-G.; George, A.W.; Reverter, A.; Li, Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* **2018**, *9*, 237. [CrossRef]
34. Asafu-Adjei, J.K.; Sampson, A.R. Covariate adjusted classification trees. *Biostatistics* **2017**, *19*, 42–53. [CrossRef]
35. Sillanpää, M.J. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* **2010**, *106*, 511–519. [CrossRef]
36. Thavamanikumar, S.; Dolferus, R.; Thumma, B.R. Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. *G3 Genes Genomes Genet.* **2015**, *5*, 1991–1998. [CrossRef] [PubMed]
37. Martini, J.W.; Hearne, S.J.; Gardunia, B.; Wimmer, V.; Toledo, F.H. Editorial: Genomic selection: Lessons learned and Perspectives. *Front. Plant Sci.* **2022**, *13*, 890434. [CrossRef] [PubMed]
38. Montesinos-López, O.A.; Montesinos-López, A.; Pérez-Rodríguez, P.; Barrón-López, J.A.; Martini, J.W.; Fajardo-Flores, S.B.; Gaytan-Lugo, L.S.; Santana-Mancilla, P.C.; Crossa, J. A review of deep learning applications for Genomic Selection. *BMC Genom.* **2021**, *22*, 1–23. [CrossRef]
39. Shen, X.; De Jonge, J.; Forsberg, S.K.; Pettersson, M.E.; Sheng, Z.; Hennig, L.; Carlborg, Ö. Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS Genet.* **2014**, *10*, e1004842. [CrossRef] [PubMed]
40. Fujimoto, Y. Kernel regularization for low-frequency decay systems. In Proceedings of the 60th IEEE Conference on Decision and Control (CDC), Austin, TX, USA, 13–17 December 2021. [CrossRef]
41. Piles, M.; Bergsma, R.; Gianola, D.; Gilbert, H.; Tusell, L. Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. *Front. Genet.* **2021**, *12*, 611506. [CrossRef]
42. Liang, M.; An, B.; Li, K.; Du, L.; Deng, T.; Cao, S.; Du, Y.; Xu, L.; Gao, X.; Zhang, L.; et al. Improving genomic prediction with machine learning incorporating TPE for hyperparameters optimization. *Biology* **2022**, *11*, 1647. [CrossRef]
43. Antonio, M.L.O.; López, A.M.; Crossa, J. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*; Springer International Publishing AG: Berlin/Heidelberg, Germany, 2022. [CrossRef]
44. Ho, D.S.; Schierding, W.; Wake, M.; Saffery, R.; O'Sullivan, J. Machine learning SNP based prediction for Precision Medicine. *Front. Genet.* **2019**, *10*, 267. [CrossRef]
45. Mathew, B.; Sillanpää, M.J.; Léon, J. Advances in statistical methods to handle large data sets for genome-wide association mapping in crop breeding. In *Advances in Breeding Techniques for Cereal Crops*; Burleigh Dodds Science Publishing: Cambridge, UK, 2019; pp. 437–450. [CrossRef]
46. Chollet, F. *Deep Learning with Python*; Manning Publications Co.: London, UK, 2018.
47. Liang, M.; Chang, T.; An, B.; Duan, X.; Du, L.; Wang, X.; Miao, J.; Xu, L.; Gao, X.; Zhang, L.; et al. A Stacking Ensemble Learning Framework for genomic prediction. *Front. Genet.* **2021**, *12*, 600040. [CrossRef]
48. Moll, R.H.; Kamprath, E.J.; Jackson, W.A. Analysis and interpretation of factors which contribute to efficiency of nitrogen utilization 1. *Agron. J.* **1982**, *74*, 562–564. [CrossRef]
49. Mathew, B.; Léon, J.; Sillanpää, M.J. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. *Heredity* **2017**, *120*, 356–368. [CrossRef]
50. Usai, M.G.; Goddard, M.E.; Hayes, B.J. Lasso with cross-validation for Genomic Selection. *Genet. Res.* **2009**, *91*, 427–436. [CrossRef] [PubMed]
51. Foster, S.D.; Verbyla, A.P.; Pitchford, W.S. Incorporating lasso effects into a mixed model for quantitative trait loci detection. *J. Agric. Biol. Environ. Stat.* **2007**, *12*, 300–314. [CrossRef]
52. Chen, C.; Steibel, J.P.; Tempelman, R.J. Genome wide association analyses based on broadly different specifications for prior distributions, genomic windows, and Estimation Methods. *Genetics* **2017**, *206*, 1791–1806. [CrossRef] [PubMed]
53. Pérez, P.; de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **2014**, *198*, 483–495. [CrossRef]
54. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [CrossRef]
55. VanRaden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [CrossRef]
56. Habier, D.; Tetens, J.; Seefried, F.-R.; Lichtner, P.; Thaller, G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* **2010**, *42*, 1–12. [CrossRef]
57. Morota, G.; Gianola, D. Kernel-based whole-genome prediction of complex traits: A Review. *Front. Genet.* **2014**, *5*, 363. [CrossRef]
58. Campos, G.D.L.; Gianola, D.; Rosa, G.J.M.; Weigel, K.A.; Crossa, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* **2010**, *92*, 295–308. [CrossRef]
59. Gianola, D.; van Kaam, J.B. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of Quantitative Traits. *Genetics* **2008**, *178*, 2289–2303. [CrossRef]

60. Zhao, W.; Lai, X.; Liu, D.; Zhang, Z.; Ma, P.; Wang, Q.; Zhang, Z.; Pan, Y. Applications of support vector machine in genomic prediction in pig and maize populations. *Front. Genet.* **2020**, *11*, 598318. [CrossRef] [PubMed]
61. González-Recio, O.; Jiménez-Montero, J.A.; Alenda, R. The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *J. Dairy Sci.* **2013**, *96*, 614–624. [CrossRef] [PubMed]
62. Grinberg, N.F.; Orhobor, O.I.; King, R.D. An evaluation of machine-learning for predicting phenotype: Studies in yeast, Rice, and wheat. *Mach. Learn.* **2019**, *109*, 251–277. [CrossRef] [PubMed]
63. Perez, B.C.; Bink MC, A.M.; Churchill, G.A.; Svenson, K.L.; Calus MP, L. Prediction Performance of Linear Models and Gradient Boosting Machine on Complex Phenotypes in Outbred Mice. *bioRxiv* **2021**, *12*, 1–13. [CrossRef] [PubMed]
64. Nazzicari, N.; Biscarini, F. Stacked kinship CNN vs. GBLUP for genomic predictions of additive and complex continuous phenotypes. *Sci. Rep.* **2022**, *12*, 19889. [CrossRef] [PubMed]
65. Franchini, G.; Porta, F.; Ruggiero, V.; Trombini, I.; Zanni, L. Learning rate selection in stochastic gradient methods based on line search strategies. *Appl. Math. Sci. Eng.* **2023**, *31*, 2164000. [CrossRef]
66. Na, G.S. Efficient learning rate adaptation based on hierarchical optimization approach. *Neural Netw.* **2022**, *150*, 326–335. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Chapter 5:

General Discussion

Nitrogen (N) plays an important role in plant production. It is the main nutrient for canopy growth and photosynthesis, which determine grain yield and quality (Beres et al., 2018; Walsh et al., 2018 and Ondoua et al., 2019). N fertilizer as the most common application in cereal cultivation is necessary to increase the shoot biomass and dry matter of bread wheat (Saleem et al., 2021). To meet the needs of the growing population, wheat varieties require sustainable management and facilities that can cope with all environmental factors such as biotic and abiotic stress and maintain expected yields. Increased N fertilizer levels contribute significantly to yield stability in bread wheat (Zemichael et al. 2017), but soil and environmental pollution due to significant greenhouse gas emissions from N fertilizer production (Garnett et al., 2015), nitrate leaching (Pathak et al. 2011), volatilization, surface runoff and denitrification from the soil-plant system (Yadav et al. 2017) can be considered as major negative consequences of high N fertilizer use in wheat production. The recovery of nitrogen fertilizer in cereals is generally poor, and only 33% of the applied nitrogen is actually harvested in the grain, while the remaining proportion (67%) remains in the soil (Sharma and Bali, 2017; Doe, 2015).

5.1. Breeding for NUE in wheat

Basically, NUE is a complex trait determined by many other related agronomic traits, each of which is controlled by many genes in cooperation with low effects and many environmental factors. As a result, genetic progress and narrow sense heritability (h^2) are very low for this type of quantitative trait per year. The second reason is due to the type of GWAS and GP models used for the given trait. Most GWAS models for single loci and multiple loci exhibit collinearity and over- or under-fitting of results due to pairwise comparisons between single nucleotide polymorphisms (SNP). This problem will be intensified, especially, when the value of epistatic variance (V_I) in genetic variance (V_g) is high for the trait of interest (Wang et al, 2016, Kärkkäinen et al, 2015; Li and Sillanpää 2012). Allelic variation for NUE could be high due to the large mutation target size within candidate genes. In addition, it is of considerable importance to identify all involved variants between traits (Robinson et al., 2014). Therefore, the main challenge in GWAS is to find a significant and reliable association in a complex trait. To address this dilemma, in the second chapter, the Local FDR correction and Bayesian survival analysis were considered as different filters to determine the best GWAS and GP models and consequently obtaining the reliable association in the GWAS results.

5.2. Significant threshold selection for NUE in whole genome of bread wheat

The choice of GWAS model and the corresponding statistical inference (linear or non-linear FDR estimation) for adjusting *p-value*(s) can have a significant impact on the interpretation of results. In summary, in *rrBLUP* GWAS model (single locus association), the adjusted *p-value*(s) based on the *rrBLUP* model show a high bias due to type I error and a high number of false positives in the raw *p-value*(s). This suggests that the raw *p-value*(s) from the *rrBLUP* model may not be suitable in case genomic dataset with high bias, which is very usual in practice. Thus, we do not suggest the *rrBLUP* model for the FDR thresholding and false positives detection. In other hand, the linear FDR thresholding such as *Bonferroni* correction and *Benjamini-Hochberg* adjustments may not cover the entire range of raw *p-value*(s) received from the *mlmm* GWAS model (multi locus association). Our findings show that the *mlmm* model with specific FDR thresholding approaches may be more robust in this context compared to the *rrBLUP* model. The performance of each FDR linear approach is highly dependent on the characteristics of the upper and lower bounds of their functions. We demonstrated that the choice of FDR linear approach is crucial and relying on commonly used threshold values may not be appropriate. The *Hoch* and *Sidak* adjustments, with their well-defined upper bounds, seem to provide more reliable and reproducible results, particularly in detecting false positives. Additionally, the differences in the distribution of false positives between lower bounds highlight the importance of carefully considering both upper and lower bounds in the linear FDR thresholding approaches. In contrast for the non-linear approaches, we demonstrated that the density of false positives in both of *q-value*(s) and *LFDR* methods using the scaled version of adjusted *p-value*(s) are same at low and high N levels, which is significantly reducing the risk of soft thresholding against FDR linear approaches. Both approaches provide consistent and reliable results particularly when utilizing the scaled version of adjusted *p-value* (s). This scaling minimizes the risk of soft thresholding in the both approaches. But the estimation of the regularization parameter and the coherent behavior of *LFDR* in the upper tail area make it a reasonable and accurate approach for handling large-scale simultaneous problems, due to minimizes the risk of hard thresholding, as well.

5.3. Local FDR and Bayesian survival analyses to identify reliable associations for grain yield in bread wheat

In the GWAS, detection of significant and reliable associations related to the given complex trait, is still a challenge. In GWAS and GP, one side of the model is assigned to the complex trait. Outliers in the NUE vector could be the first cause of pseudo-marker trait associations (MTAs) in the GWAS results. In the first chapter of our study, the performance of *LFDR* approach was highest to handle the large-scale simultaneous problem. Therefore the *LFDR* as frequentist inference with Fisher information background was applied to check the distribution of SNPs especially in the tails as critical regions of the distribution. This method, based on maximum likelihood estimation handles the effect size of a large scale genomic file especially in the heavy tails. We found that the expected mean likelihoods based on SNP information are acceptable, but the standard error in the models with a scale of *–log10 (SNP)* is still high. The main criticism of *LFDR* approach is that it locally focuses only on SNP effects in tails, whereas signaling associations were distributed throughout the genome. Moreover, the magnitude of SNP effects alone may not be sufficient to make a decision about the performance of the model. Bayesian survival FDR analysis as a modern inference based on posterior estimation of SNP effects is commonly used for large scale genomic data with small effects. In the survival part with the $(-\Delta(SNP_{i=0}^{n})$ component of the approach, we included the minor and major alleles in each SNP based on the time to events, which is necessary to achieve the expected prior distribution. Moreover in the Bayesian part the semi-parametric empirical Bayes factor better controls both false positive and false negative errors in the GWAS to identify allelic variations and find templates for significant SNPs. Therefore, it is proposed to utilize different GWAS models at more N levels to increase the replicability of the posterior probability [35] of the identified signal associations.

5.4. Genetic parameter and hyper-parameter estimation underlie NUE in Bread Wheat

The estimation of breeding values (BVs) based on extensive genomic data for complex traits is the main goal in wheat breeding programs. Currently, phenotyping of complex traits such as NUE in wheat is still expensive, requires high throughput technologies and is very time consuming compared to genotyping. Therefore, breeding programs are trying to predict phenotypes based on high-density marker information. Genetic parameters such as population structure, genomic relationship matrix, marker density and sample size are important factors that increase the power and accuracy of the model. In parallel, there are many genetic hyper-parameters that are hidden and unrepresented in the given genomic selection (GS) model but have a significant impact on the results, e.g. panel size, number of markers, minor allele frequency, number of call rates for each marker, number of cross validations and batch size in the training set of the genomic file. The main challenge is to ensure the reliability and accuracy of genomic estimated BVs (GEBVs) as GS results. A number of algorithms, generally nonparametric, are combined to improve the predictive or classification ability of the compiled model. Our study has shown that the definition and optimization of the regularization parameter is crucial to demonstrate the performance and accuracy of the given GS model, which has not been sufficiently addressed in previous studies.

5.5. Conclusion

This study is one of the first to model NUE as complex trait using modern statistical genomics algorithms. To sum up, FDR thresholding based on non-linear approaches is more stringent than controlling the family wise error rate and they obtain adequate information to maintain the error rate in the large scale genomic hypotheses. We concluded that using Bayesian techniques, such as *EBayes* FDR, could be highlighted for effectiveness in the context of FDR thresholding optimization. Also, the *Bayesian survival* FDR analysis as modern statistical inference based on posterior estimation of SNP effects could be used to identify reliable associations in the GWAS and GP models. Two major challenges are appeared and require still attention in the future. The first challenge is how to tradeoff between bias and variance at the same time to minimize under or over-fitting performance of GWAS and GP models, which has significant impact to receive the reliable and reasonable results. The second challenge is how to define GWAS and GP models and how to optimize the hidden layers of genetic hyper-parameters in large scale of genomic data.

5.6. References

Beres, B., Graf, R., Irvine, R., O'Donovan, J., Harker, K., Johnson, E. Stevenson, F. (2018). Enhanced nitrogen management strategies for winter wheat production in the Canadian prairies. Canadian Journal of Plant Science, 98(3), 683-702. doi:10.1139/cjps-2017-0319

Doe, J. (2015). Ammonia and Nitrate Losses from Agriculture and Their Effect on Nitrogen Recovery in the EU and U.S. *CSA News*, 60(4), p.16

Garnett, T., Plett, D., Heuer, S., & Okamoto, M. (2015). Genetic approaches to enhancing nitrogen-use efficiency (NUE) in cereals: Challenges and future directions. *Functional Plant Biology, 42*(10), 921. doi:10.1071/fp15025

Kärkkäinen, H. P., Li, Z., & Sillanpää, M. J. (2015). An Efficient Genome-Wide Multilocus Epistasis Search. Genetics, 201(3), 865–870. https://doi.org/10.1534/genetics.115.182444

Li, Z., & Sillanpää, M. J. (2012). Estimation of Quantitative Trait Locus Effects with Epistasis by Variational Bayes Algorithms. Genetics, 190(1), 231–249. https://doi.org/10.1534/genetics.111.134866

McGuire D, Jiang Y, Liu M, Weissenkampen JD, Eckert S, Yang L, Chen F, Berg A, Vrieze S, Jiang B, Li Q, Liu DJ. (2021). Model-based assessment of replicability for genome-wide association meta-analysis. *Nature Communications. 12(1).* https://doi.org/10.1038/s41467-021-21226-z

Ondoua, R. N., & Walsh, O. (2017). Varietal differences in nitrogen use efficiency among spring wheat varieties in Montana. *Crops & Soils*, *50*(5), 40–42. https://doi.org/10.2134/cs2017.50.0505

Pathak, R., Lochab, S., & Raghuram, N. (2011). Improving Plant Nitrogen-Use Efficiency. *Comprehensive Biotechnology*, 209-218. doi:10.1016/b978-0-08-088504-9.00472-4

Robinson, M. R., Wray, N. R., & Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics*, *30*(4), 124–132. https://doi.org/10.1016/j.tig.2014.02.003

Saleem Kubar, M., Feng, M., Sayed, S., Hussain Shar, A., Ali Rind, N., Ullah, H., Ali Kalhoro, S., Xie, Y., Yang, C., Yang, W., Ali Kalhoro, F., Gasparovic, K., Barboricova, M., Brestic, M., El Askary, A., & El-Sharnouby, M. (2021). Agronomical traits associated with yield and yield components of winter wheat as affected by nitrogen managements. *Saudi Journal of Biological Sciences*, *28*(9), 4852–4858. https://doi.org/10.1016/j.sjbs.2021.07.027

Sharma, L., & Bali, S. (2017). A Review of Methods to Improve Nitrogen Use Efficiency in Agriculture. *Sustainability*, 10(2), 51. doi:10.3390/su10010051

Walsh, O., Shafian, S., & Christiaens, R. (2018). Nitrogen Fertilizer Management in Dryland Wheat Cropping Systems. *Plants, 7*(1), 9. doi:10.3390/plants7010009

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J. M., Xu, S., & Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Scientific Reports, 6(1). https://doi.org/10.1038/srep19444

Yadav, M. R., Kumar, R., Parihar, C. M., Yadav, R. K., Jat, S., Ram, H. Jat, M. L. (2017). Strategies for improving nitrogen use efficiency: A review. *Agricultural Reviews,* (OF). doi:10.18805/ag.v0iof.7306

Zemichael, B., Dechassa, N., & Abay, F. (2017). Yield and Nutrient Use Efficiency of Bread Wheat (*Triticum Aestivum* L.) as Influenced by Time and Rate of Nitrogen Application in Enderta, Tigray, Northern Ethiopia. *Open Agriculture*, 2(1). doi:10.1515/opag-2017-0065.

Publications:

1- **Sadeqi, M.B.,** Ballvora, A., Dadshani, S., Siddiqui, N., Kamruzzuman, M., Koua, A. P. and Léon, J. (2024) Significant threshold selection for nitrogen use efficiency in whole genome of bread wheat, manuscript under review by *Plant Direct*.

2- **Sadeqi, M. B.,** Ballvora, A., and Léon, J. (2023). Local and Bayesian Survival FDR Estimations to Identify Reliable Associations in Whole Genome of Bread Wheat. *International Journal of Molecular Sciences, 24(18), 14011*. https://doi.org/10.3390/ijms241814011

Author contribution:

M.B.S. literature and problem statement, methodology, formal analysis, writing original draft; A.B. validation, review, editing, funding acquisition; J.L. editing and supervision.

3- **Sadeqi, M. B.,** Ballvora, A., Dadshani, S. and Léon, J. (2023). Genetic Parameter and Hyper-Parameter Estimation Underlie Nitrogen Use Efficiency in Bread Wheat. *International Journal of Molecular Sciences, 24(18), 14275.* https://doi.org/10.3390/ijms241814275

Author contribution:

M.B.S.; conceptualization, methodology, formal analysis, investigation, writing—original draft, A.B. review, editing, project administration, S.D. resources, review and J.L. editing and supervision.