Machine Learning-GWAS reveals the role of *WSD1* gene for cuticular wax ester biosynthesis and key genomic regions controlling early maturity in bread wheat.

Honoré Tekeu[1,2,3], Martine Jean[1,2], Eddy L. M. Ngonkeu[3], François Belzile[1,2]*

Corresponding author: francois.belzile@fsaa.ulaval.ca

[1]Département de Phytologie, Université Laval, Quebec City, QC, Canada

[2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

[3]Institute of Agricultural Research for Development, Yaoundé, Cameroon.

Abstract

This study employed Machine Learning-Genome-Wide Association Study (ML-GWAS) to identify genomic regions linked to cuticular wax ester biosynthesis (SW) and early maturity (DM) in wheat. Using a dataset with 170 wheat accessions and 74K SNPs, four GWAS tools (MLM, CMLM, FarmCPU, and BLINK) and five machine learning techniques (RF, ANN, SVR, CNN, and SVM) were applied. A highly significant SW association was found on chromosome 1A, with the peak SNP (chr1A:556842331) explaining 50% of the phenotypic variation. A promising candidate gene, *TraesCS1A01G385500*, was identified as an ortholog of Arabidopsis thaliana's *WSD1* gene, which plays a crucial role in very long-chain (VLC) wax ester biosynthesis. For DM, four QTLs were detected on chromosomes 4B (two QTLs), 2A, and 5A. Haplotype analysis revealed that alleles TT significantly contribute to cuticular wax ester biosynthesis and early maturity in wheat varieties. The study underscores the superior performance of ML models, especially when combined with advanced multi-locus GWAS models like BLINK and FarmCPU, with significantly lower p-values for identifying relevant QTLs compared to traditional methods. ML approaches hold potential for revolutionizing the study of complex genetic traits, offering insights to enhance wheat crops' resilience and quality. ML-GWAS emerges as a compelling tool for genomic-based breeding, enabling breeders to develop improved wheat varieties with greater precision and efficiency.

Keywords: International wheat collection, Genotyping-by-Sequencing, population structure, Genome-Wide Association Study, Machine learning, spike waxiness, number of days-to-maturity.

Introduction

Wheat (*Triticum aestivum* L.), a globally essential staple crop, faces a multitude of environmental challenges, from drought and salinity to extreme temperatures and pest pressures (He et al., 2022). A vital component of its adaptive response to these challenges is the hydrophobic cuticle, composed of cutin and cuticular waxes (Wang and Chang, 2022). This lipidic shield defends against non-stomatal transpiration, UV radiation, pathogens, and insect invasions while maintaining the integrity of adjacent plant organs (Ingram and Nawrath, 2017; Martin and Rose, 2014).

The cuticle consists of two primary constituents: cutin, an insoluble polyester, and cuticular waxes, encompassing very-long-chain (VLC) fatty acids, aldehydes, ketones, esters, alcohols, alkanes, and other compounds (Kunst and Samuels, 2009). Alkanes, a significant component of cuticular waxes, play a critical role in enhancing plant drought tolerance (Kosma et al., 2009; Seo et al., 2011). In the realm of wheat, genes related to wax biosynthesis, including TaFARs for primary alcohols and the W1 locus for β-diketones, have been identified (Hen-Avivi et al., 2016; Y. Wang et al., 2015a, 2015b). One pivotal gene, TaCER1-1A, has been recognized for its involvement in alkane accumulation in wheat (Li et al., 2019). In a recent study by (He et al., 2022), attention is drawn to TaCER1-6A, another key gene involved in alkane biosynthesis in wheat, with investigations extending to overexpression and CRISPR/Cas9-mediated gene editing.

To date, no study has pinpointed a gene responsible for the biosynthesis of wax VLC esters, which play a crucial role in mitigating leaf water loss, particularly under drought conditions. The journey of these wax constituents from the Golgi and trans-Golgi network (TGN) to the plasma membrane and onward to the cuticle involves pathways coordinated by ABCG subfamily half transporters and lipid transfer proteins (LTPs) (DeBono et al., 2009; Ichino and Yazaki, 2022; Pighin et al., 2004; Wang and Chang, 2022).

Additionally, the cultivation of early-maturing wheat varieties holds critical importance in regions characterized by short growing seasons and extended daylight, exemplified by the Northern Great Plains of Canada and the USA (A. Kamran et al., 2013). Early maturation not only enhances crop yields but also acts as a safeguard against frost damage, a threat that can significantly compromise grain quality and overall agricultural productivity (Iqbal et al., 2007). The precise timing of wheat's flowering is intricately regulated by a complex interplay of genes that dictate growth patterns and earliness. These genetic regulators encompass vernalization

(Vrn), photoperiod (Ppd), and earliness per se (Eps) genes, shaping when wheat plants initiate flowering and influencing their growth habits (Atif Kamran et al., 2013).

Adding to this complexity, certain genetic factors, such as dwarfing genes, subtly affect the timing of heading, flowering, and maturity, introducing further intricacies in the regulation of these vital agricultural traits (Chen et al., 2018; Daoura Goudia et al., 2014). Earliness per se genes also play a role in enhancing the adaptability of wheat plants, contributing to their resilience in varying environments (Snape et al., 2001). Recent studies, including one by Semagn et al. (2021), have delved into the intricate genetic mapping of Quantitative Trait Loci (QTLs) associated with days to maturity, particularly in wheat varieties evaluated under both conventional and organic farming practices. These studies have identified key QTLs on chromosome 4B, shedding light on the genetic mechanisms governing maturity. Furthermore, earlier research by authors such as (Zou et al., 2017a, 2017b) employing extensive genetic mapping using 1203 markers in RIL populations like 'Attila' and 'CDC Go' has uncovered a shared genomic region linked to maturity, situated on both chromosome 4B and 5A.

Among the array of genetic factors at play, certain dwarfing genes, including Rht-B1, Rht5, Rht8, and Rht12, have been identified as contributing factors, subtly influencing the timing of heading, flowering, and maturity in wheat varieties. These genetic elements add an additional layer of complexity to the intricate regulation of these pivotal traits (Chen et al., 2018; Daoura Goudia et al., 2014).

While molecular markers have facilitated characterizing genetic diversity, phenotypic assessments have primarily determined the utility of these genetic resources in breeding (Belzile et al., 2020). With the availability of high-density SNP markers, Genome-Wide Association Studies (GWAS) have become a powerful tool for identifying and mapping loci contributing to phenotypic variation among diverse genetic materials that have undergone extensive recombination (Yu and Buckler, 2006). Recent applications of highly reproducible GBS-derived SNPs have uncovered candidate genes influencing grain size in bread wheat (Tekeu et al., 2021). GWAS has become a standard approach across species for identifying genes associated with critical traits (Ashkenazy et al., 2022).

However, there remain challenges with conventional GWAS techniques, including the "large p, small n" issue when the number of markers surpasses the number of genotypes (Kaler et al., 2020; Mohammadi et al., 2020). Conventional GWAS methods are better suited for identifying common SNPs with substantial main effects, while the distinction between causal variants and

correlated genes linked by linkage disequilibrium remains problematic (Enoma et al., 2022; Nicholls et al., 2020). Moreover, conventional GWAS approaches lack the power to uncover minor-effect SNPs associated with specific traits (Zhou et al., 2019). Consequently, machine learning (ML) techniques offer an opportunity to address these limitations and gain insights into the complex genetic basis of traits, as demonstrated in other crop species (Ashkenazy et al., 2022; Kwon et al., 2022).

Machine learning models for GWAS vary in complexity, from simple logistic regression to sophisticated ensemble models such as random forests, gradient boosting, and neural networks. These ML algorithms focus on maximizing prediction accuracy and excel at capturing multi-locus SNP interactions better than conventional methods. Support Vector Regression (SVR) is one such machine learning technique that has shown promise in predicting important agricultural traits (Yoosefzadeh Najafabadi et al., 2021). While SVR has found application in various crop studies, the potential of other ML techniques, such as Random Forest (RF), Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), remains largely untapped when compared to the more conventional GWAS tools like Mixed Linear Model (MLM), Compressed Mixed Linear Model (CMLM), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK).

This study aims to address this gap and provide valuable insights to bolster crop resilience. By employing a diverse array of advanced ML techniques in GWAS analysis, we seek to identify the genomic regions associated with cuticular wax ester biosynthesis (SW) and early maturity (DM) in wheat. Our approach promises to shed light on the intricate genetic mechanisms governing these vital traits and contribute to the advancement of crop breeding efforts for improved wheat varieties. Our research hypotheses revolve around specific genomic regions influencing SW and DM in a diverse global collection of bread wheat accessions. Furthermore, it postulates that ML-GWAS approaches will outperform traditional GWAS methods, in identifying Quantitative Trait Loci (QTLs) relevant to SW and DM traits in wheat. The present study aims to decipher the genetic underpinnings of SW and DM using ML-GWAS approaches.

Materials and methods

Plant materials

In this study, an international collection of 170 accessions was employed for genome-wide association analyses. These cultivars were obtained from various international wheat breeding programs. The South African accessions consisted of spring wheat lines from the Western Cape region, along with some winter bread wheat lines from other parts of the country. The East African spring-type accessions were gathered in Kenya and Ethiopia. The Mexican accessions were obtained via the International Maize and Wheat Improvement Center (CIMMYT), and they included spring accessions from Mexicali and Baja California. The Central African accessions were provided by the Institute of Agricultural Research for Development (IRAD) and farmers (Tekeu et al., 2017). The French accessions were winter lines, and those from North Africa were composed of spring lines acquired from the International Center for Agricultural Research in the Dry Areas (ICARDA).

Phenotyping

A panel of 170 accessions of bread wheat was phenotyped and used for genome-wide association analyses. Field trials were conducted in two different locations in the bimodal humid forest zone of Cameroon, during the 2015-2016 season in Munt Mbankolo (1057 m above sea level) and during 2016-2017 in Nkolbisson (650 m a. s. l.). At each trial site, an incomplete alpha-lattice design with two replications was used and each accession was planted, as previously reported by (Tekeu et al., 2021). Then, fields trials were managed in accordance with the technical recommendations and standard agricultural practices for wheat (Pask et al., 2012). Spike waxiness (SW; 0: Absent, 2: Almost none, 3: Very little, 4: Little, 5: Intermediate, 6: Some, 7: Much, 8: Very much) and DM (days-to-maturity) were assessed when 50% of spikes had turned yellow (Zadoks et al., 1974).

Analysis of phenotypic data

We conducted the analysis of variance for each trait using PROC MIXED in SAS 9.4. In this analysis, each cultivar was considered a fixed effect, while replications and environments were treated as random effects. Pearson correlation coefficients between pairs of phenotypic traits were computed using Pearson's correlation in SPSS 20.0. To assess the heritability of each trait, we utilized the broad-sense heritability (h^2) formula: $h^2 = VG / (VG + VGE + Ve)$, where VG

represents genetic variance, VGE is the genetic-environment interaction variance, and Ve is the error variance.

DNA isolation, GBS library construction and sequencing

To extract genomic DNA from dried young leaf tissue (~ 5 mg) of all accessions, we used a CTAB DNA isolation method (Doyle and Doyle, 1990). The extracted DNA was quantified using a Quant-iT™ PicoGreen kit (ThermoFisher Scientific, Canada), and concentrations were normalized to 20 ng/µl for library preparation. We constructed three 96-plex *PstI-MspI* GBS libraries as described by (Elshire et al., 2011). Subsequently, each library was sequenced on three P1 chips using an Ion Torrent PGM sequencer at the Plate-forme d'Analyses Génomiques of the Institut de Biologie Intégrative et des Systèmes (Université Laval, Québec, Canada).

Single nucleotide polymorphism calling and bioinformatics analysis.

Genomic DNA sequences of wheat samples, with an average of 2.4 million reads per wheat line, were analyzed using the FastGBS pipeline (Torkamaneh et al., 2017). The reads were aligned to the wheat reference genome (Chinese Spring v1.0), and SNPs were called using FastGBS. Standard filtration steps were applied to the FastGBS results, as previously described by (Tekeu et al., 2021). Additional filtration steps were carried out on this subset to retain only SNPs with a minor allele frequency (MAF) of at least 0.05.

Machine Learning-Genome-Wide Association Study

We conducted a genome-wide association study (GWAS) to identify genomic regions associated with variation in SW and DM using a dataset comprising 170 accessions and 74K single nucleotide polymorphisms (SNPs). We employed a comprehensive approach that integrated four GWAS analytical methods, namely the Mixed Linear Model (MLM), Compressed Mixed Linear Model (CMLM), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK). In addition, we harnessed the power of five machine learning algorithms, which included Random Forest (RF), Support Vector Regression (SVR), Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). This integrated approach allowed us to assess the association between SNP markers and estimated genotypic values (BLUEs) for each trait.

For MLM, CMLM, FarmCPU, and BLINK methods, we made use of the Genomic Association and Prediction Integrated Tool (GAPIT) version 2 (Lipka et al., 2012) in conjunction with the rMVP packages (Yin et al., 2021). Our association analyses were performed while correcting for both population structure and relationships among individuals, with the incorporation of either the Q+K matrices. The K matrix was computed using the Van Raden method (Lipka et al., 2012). The significance threshold for genome-wide association was determined based on a false discovery rate (FDR-adjusted $p < 0.05$).

In the case of machine learning algorithms, we utilized a scaled method (ranging from 0 to 100) to estimate the importance of each SNP associated with the traits of interest. To integrate the machine learning approach into GWAS, we implemented a five-fold cross-validation strategy with ten repetitions to estimate the variable importance of each SNP, following (Siegmann and Jarmer, 2015). Therefore, we applied a global empirical threshold, as proposed by (Churchill and Doerge, 1994; Doerge and Churchill, 1996). This threshold was determined by fitting the ML algorithm, recording SNPs with the highest variable importance scores, repeating the process 1000 times, and selecting associated SNPs based on $\alpha = 0.5$. The machine learning methods were executed using the Caret package (Kuhn et al., 2020) in R software version 4.2.2. Throughout these analyses, we ensured that association analysis was conducted while correcting for both population structure and relationships among individuals, using a combination of the Q + K matrices. The p-value threshold for significance in the genome-wide association was determined based on a false discovery rate (FDR-adjusted $p \leq 0.05$).

Identification of candidate genes and haplotype analysis

To identify candidate genes contributing to SW and DM, we defined haplotype blocks containing the peak SNP. Each region with the peak SNP was visually explored for its LD structure and for genes located in such regions, and the annotated genes within each interval were screened thanks to the annotated and ordered reference genome sequence in place by (International Wheat Genome Sequencing Consortium (IWGSC), 2018). Candidate genes potentially involved in each trait were further investigated. The function of these genes was also inferred by a BLAST of their sequences to the UniProt reference protein database (http://www.uniprot.org/blast/). To further provide more information about potential candidate genes, we used RNA-seq data of (Ramírez-González et al., 2018), based on the electronic fluorescent pictograph (eFP) at bar.utoronto.ca/eplant (by (Waese et al., 2017) to identify in what tissues and at which developmental stages candidate genes were expressed in wheat.

To better define the possible alleles in a strong candidate gene and trait, we defined haplotypes around the peak SNP. For each haplotype, we calculated the trait mean for lines sharing the same haplotype using the R ggpubr program.

Results

Phenotypic characterization

In order to delve into the traits of SW and DM in wheat, we meticulously assessed their phenotypes over the span of two years at two distinct sites. As summarized in Table 1, the observed means (± standard deviation) for these traits were as follows: 5.35 (±1.56) for SW and 98.06 days (±4.65) for DM. The broad-sense heritability estimates were robust, measuring 55.4% for SW and 50% for DM. An analysis of variance uncovered noteworthy differences attributable to genotypes (G) for all traits, and, in the case of SW and DM, the interaction between genotype and environment (GxE) also emerged as a significant factor. A correlation analysis unveiled a highly significant positive correlation between SW and DM ($r = 0.273$; $p < 0.01$).

Upon scrutinizing the relationship between SW and DM using bagplots analysis with the 170 accessions in our collection, no outliers were detected when considering the interplay between these two traits (Supplementary Figure S1). Consequently, for subsequent analyses including those involving population structure and genome-wide association studies (GWAS), all accessions were retained. The distribution of phenotypic traits appeared to approximate a normal distribution and exhibited characteristics of quantitative inheritance (Figure 1).

Table 1. Descriptive statistics, broad-sense heritability (h^2), and F-values from variance analysis for two key agronomic traits in a cohort of 170 wheat lines.

Traits	R^2	CV	Range		Mean±SD	h^2	F-values		
			Min	Max			Genotype (G)	Environment (E)	G x E
SW	0.665	29.181	0	8	5.35±1.56	55.4	1.45***	66.87	0.14
DM	0.852	4.744	74	125	98.06±4.65	50	3.39***	310.38**	3.39**

SD Standard deviation, h^2 Broad sense heritability, R^2: R-squared; CV: Coefficient of variation; ***, ** and *: significant at $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively.

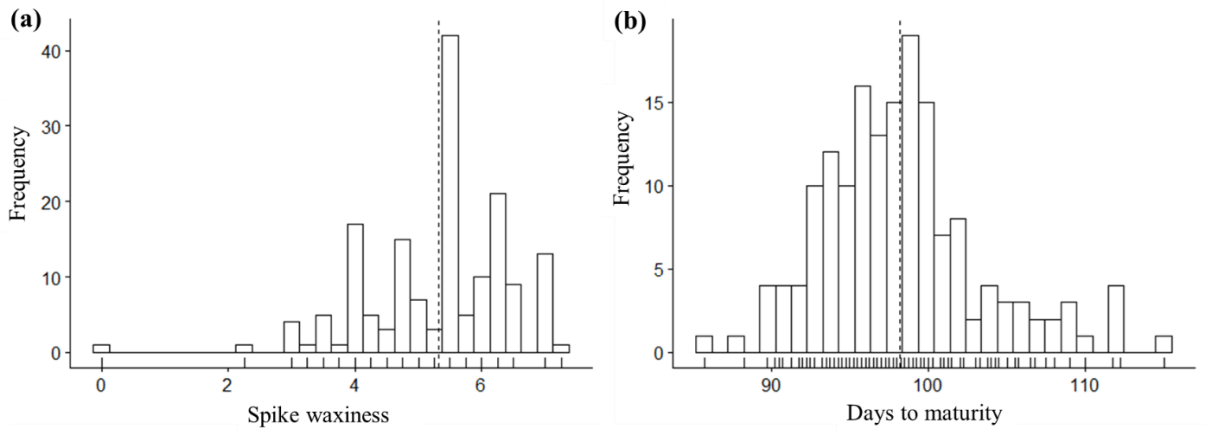


Figure 1. Distribution of phenotypes for spike waxiness (a) and days-to-maturity (b). Histograms are based on the average trait value of each wheat line across the different environments. The bars under the histograms represent the density of individuals.

Genome coverage and population structure

Our comprehensive analysis revealed a total of 73,784 polymorphic SNP markers that spanned across the 21 chromosomes of the wheat genome, as depicted in Figure 2. As previously reported in our prior study, the examination of population structure within the accessions of this association panel revealed that K=6 provided the optimal representation of population structure within this set of accessions. These clusters notably aligned with the geographic regions of origin. The distribution of wheat accessions among these six subpopulations ranged from 6 to 43, with the largest number of accessions hailing from northwestern Baja California, Mexico, specifically represented by Mexico 1 (43). Conversely, the smallest subpopulation was observed in East and Central Africa, encompassing just 6 accessions.

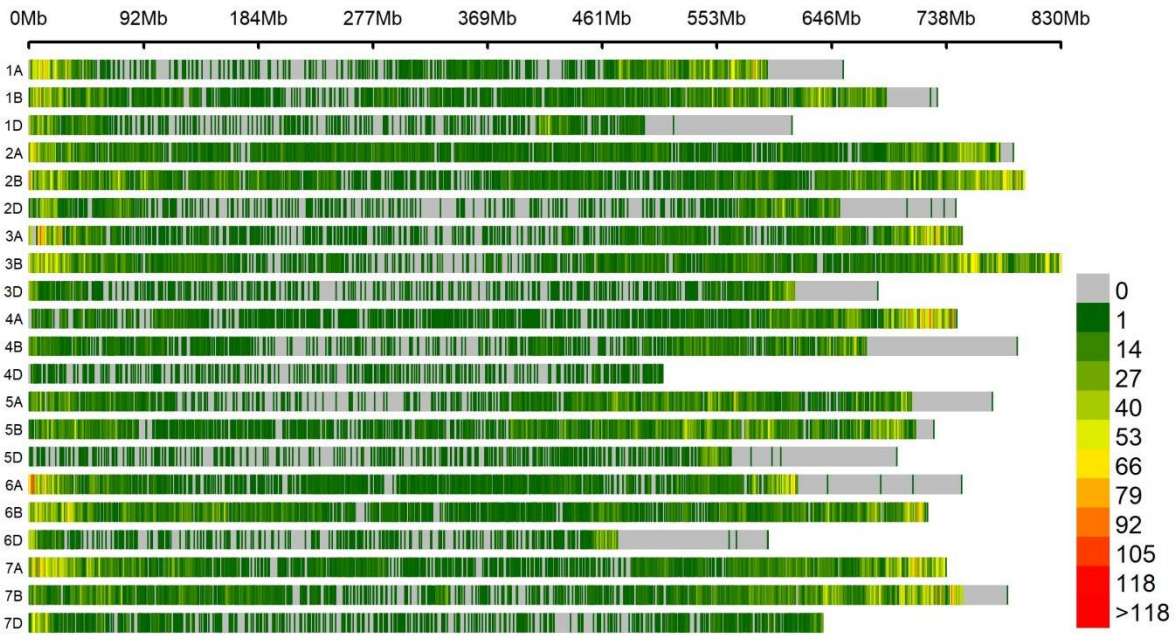


Fig 2. Genome coverage of polymorphic SNP markers over the physical map of the 21 chromosomes of the hexaploid wheat lines. The color reflects the density of SNP markers (i.e. number of SNPs within a sliding 1-Mb window).

Marker-trait associations

To uncover the genomic regions responsible for the variation in SW and DM, we conducted an association analysis (GWAS) on a subset of accessions with phenotypic data (170 accessions and 73,784 SNPs). In this analysis, we employed four GWAS analytical tools (MLM, CMLM, FarmCPU, and BLINK), complemented by five machine learning techniques (RF, ANN, SVR, CNN, and SVM). Notably, the quantile-quantile (QQ) plots in Figure 3 demonstrated the effective control of confounding effects related to population structure and relatedness by all conventional GWAS and machine learning models. Deviations from the diagonal were observed only for the most extreme p-values, indicating a well-controlled analysis for both traits.

For the SW trait, the results of the association analyses are visualized in the Manhattan plots presented in Figure 3. Using a threshold for false discovery rate (FDR) of ≤ 0.05 (as detailed in Supplementary Figure S2, marked by the green horizontal line), we identified four QTLs. Remarkably, only one QTL was co-identified by at least two models (Figure 4). The most robust and consistent association, located on chromosome 1A, is summarized in Table 2.

This particular QTL was defined by its peak SNP, marked as chr1A:556842331, and was identified by the multi-locus models (FarmCPU and BLINK) as well as all five machine learning algorithms (RF, ANN, SVR, CNN, and SVM). Notably, this QTL explained a substantial 50% of the phenotypic variation observed in SW. The minor allele frequency (MAF) at this locus was 0.09, and it exhibited an allelic effect of 0.66.

These findings highlight a significant and consistent genetic association with SW on chromosome 1A, showcasing the power of both traditional GWAS and machine learning approaches in identifying key genomic regions influencing this trait.

Turning our attention to the DM trait, our investigation unveiled a total of eight genomic regions that displayed significant associations. The results of these association analyses are visualized in the Manhattan plots featured in Figure 3, with a stringent threshold for false discovery rate (FDR) of ≤ 0.05, as outlined in Supplementary Figure S2 and highlighted by the green horizontal line. However, we noted the co-identification of only four Quantitative Trait Loci (QTLs) by at least two models (Figure 3). Among these, the most robust associations, localized on chromosomes 4B, 5A, and 2A, are thoughtfully summarized in Table 2.

Of noteworthy mention is chr4B:666048201, which emerged as the peak SNP and was jointly identified by both multi-locus GWAS models (FarmCPU and BLINK) and four machine learning algorithms (RF, ANN, SVR, and CNN). These markers formed a robust linkage block, with all markers exhibiting perfect linkage disequilibrium (LD) ($r^2 = 1$), as detailed in Supplementary Table S1. This discovery delineated a single QTL, with the peak SNP accounting for a substantial 19.3% of the phenotypic variation associated with DM. The minor allele frequency (MAF) at this locus was observed to be 0.08, while the allelic effect amounted to 3.84 days (Table 2).

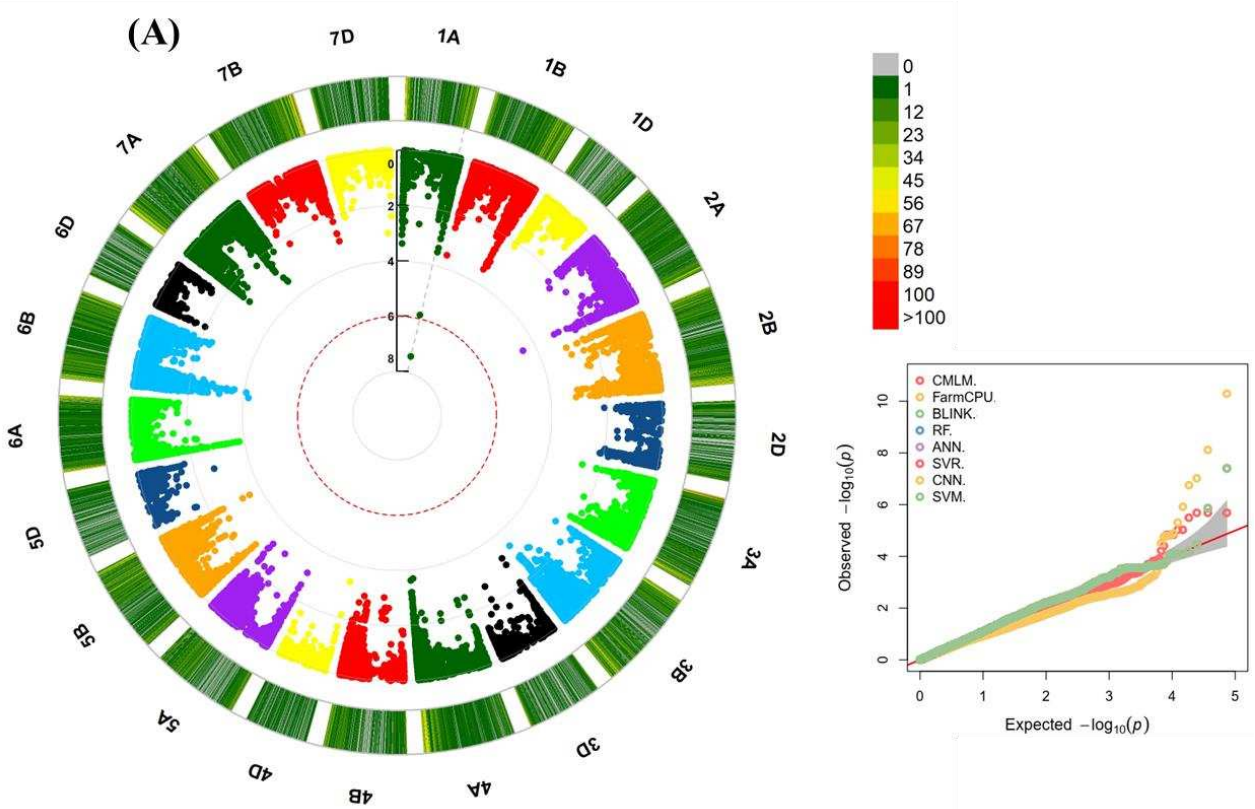
In addition, another noteworthy association with DM on chromosome 4B was unveiled, defined by the peak SNP chr4B:37907825. This association was identified by the GWAS model BLINK and all five machine learning methods (RF, ANN, SVR, CNN, and SVM). It explained 18.44% of the phenotypic variation for DM, with a MAF of 0.09 and an allelic effect of -2.76 days.

Moreover, a QTL residing on chromosome 2A was brought to light, marked by the peak SNP chr2A:605016602, which explained 8.65% of the phenotypic variation for DM. This QTL was detected using both multi-locus GWAS models (FarmCPU and BLINK) and all five machine learning methods (RF, ANN, SVR, CNN, and SVM).

Furthermore, an additional QTL on chromosome 5A, characterized by the peak SNP chr5A:580797118, was identified through the BLINK model and all five machine learning methods (RF, ANN, SVR, CNN, and SVM). This QTL contributed to 0.72% of the phenotypic variation associated with DM.

These findings underscore the efficacy of our approach in uncovering key genomic regions associated with DM and highlight the potential of both traditional GWAS and machine learning techniques in unraveling the genetic underpinnings of complex traits.

Overall, the GWAS and ML methods successfully mitigated the confounding effects of population structure and relatedness and identified multiple genomic regions associated with spike waxiness and Days to maturity in wheat. These findings can provide insights into the genetic architecture of these traits and aid plant breeders in developing new bread wheat varieties with improved SW and maturity.



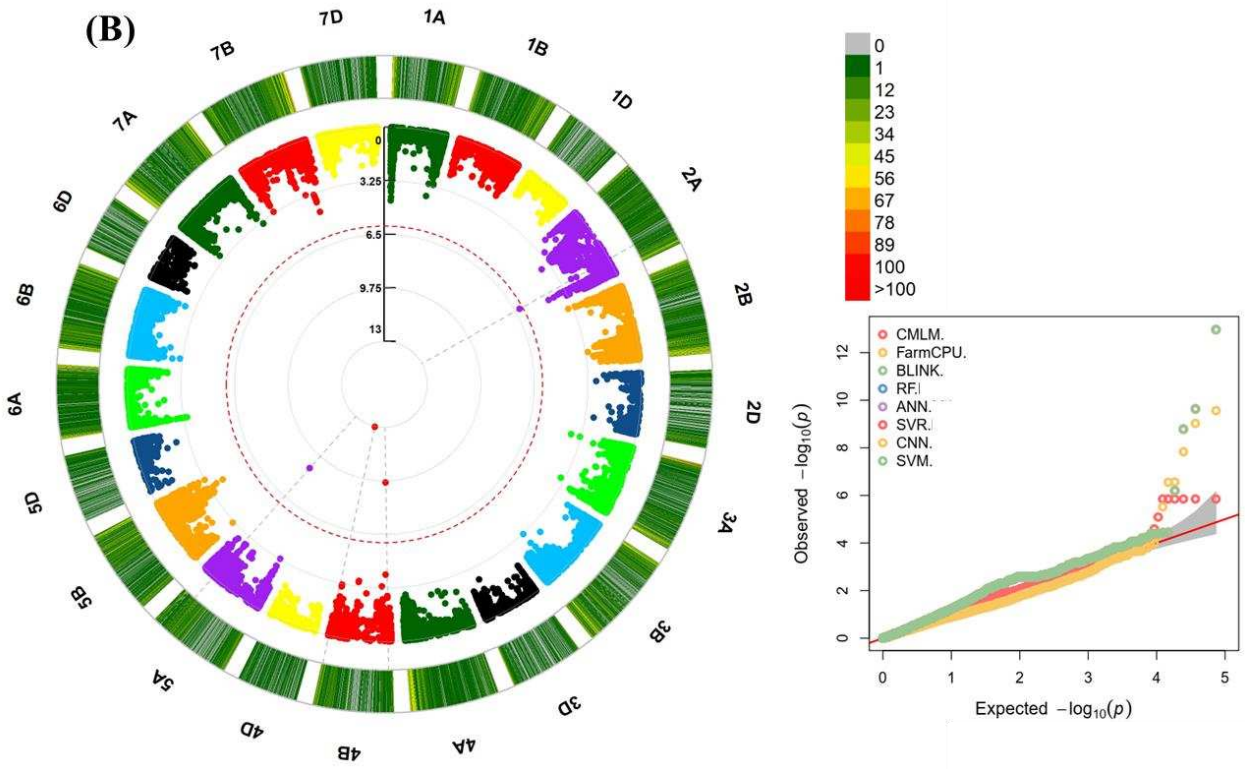


Figure 3. Genome-wide association analysis of 170 hexaploid wheat cultivars. Manhattan and Q-Q plots for all models shows the degree of association between SNPs and SW (A) and DM (B). In both cases, associations are declared significant at an FDR ⩽0.05. One marker (see the red circle) displayed significant associations with the SW trait. Four SNP markers (see the red circle) displayed significant associations with the DM.

In order to gain a deeper understanding of the relationship between the peak SNP (chr1A:556842331) and SW, we delved into the realm of SNP haplotypes. Through a thorough analysis of haplotypes encompassing this peak marker, we unveiled two distinct haplotypes (Figure 4). Remarkably, we observed a notable divergence in phenotypic outcomes between these haplotypes. Haplotype TT displayed significantly higher values (5.481) compared to the values associated with haplotype CC (3.642). This revelation suggests that SNP markers flanking this gene could serve as valuable tools in marker-assisted breeding programs. By selecting alleles that contribute to drought-resistant wheat varieties, these programs hold the potential to enhance wheat productivity and bolster its resilience in the face of water scarcity.

To further refine our understanding of the association between the peak SNP (chr4B:666048201) and DM, we embarked on an exploration of SNP haplotypes. This investigation uncovered two distinct haplotypes encircling the peak SNP. Notably, our scrutiny of these haplotypes revealed a substantial difference in phenotypic outcomes (Figure 4). Haplotype TT was linked to significantly lower values (97.41) in comparison to haplotype CC (107.36). This observation posits that SNP markers flanking this gene have the potential to be valuable assets in marker-assisted breeding programs. By selecting alleles conducive to the development of short-season wheat varieties, these programs can contribute to the improvement of wheat productivity and the creation of cultivars better equipped to thrive in varying environmental conditions. More details are provided in Supplementary Table S2.

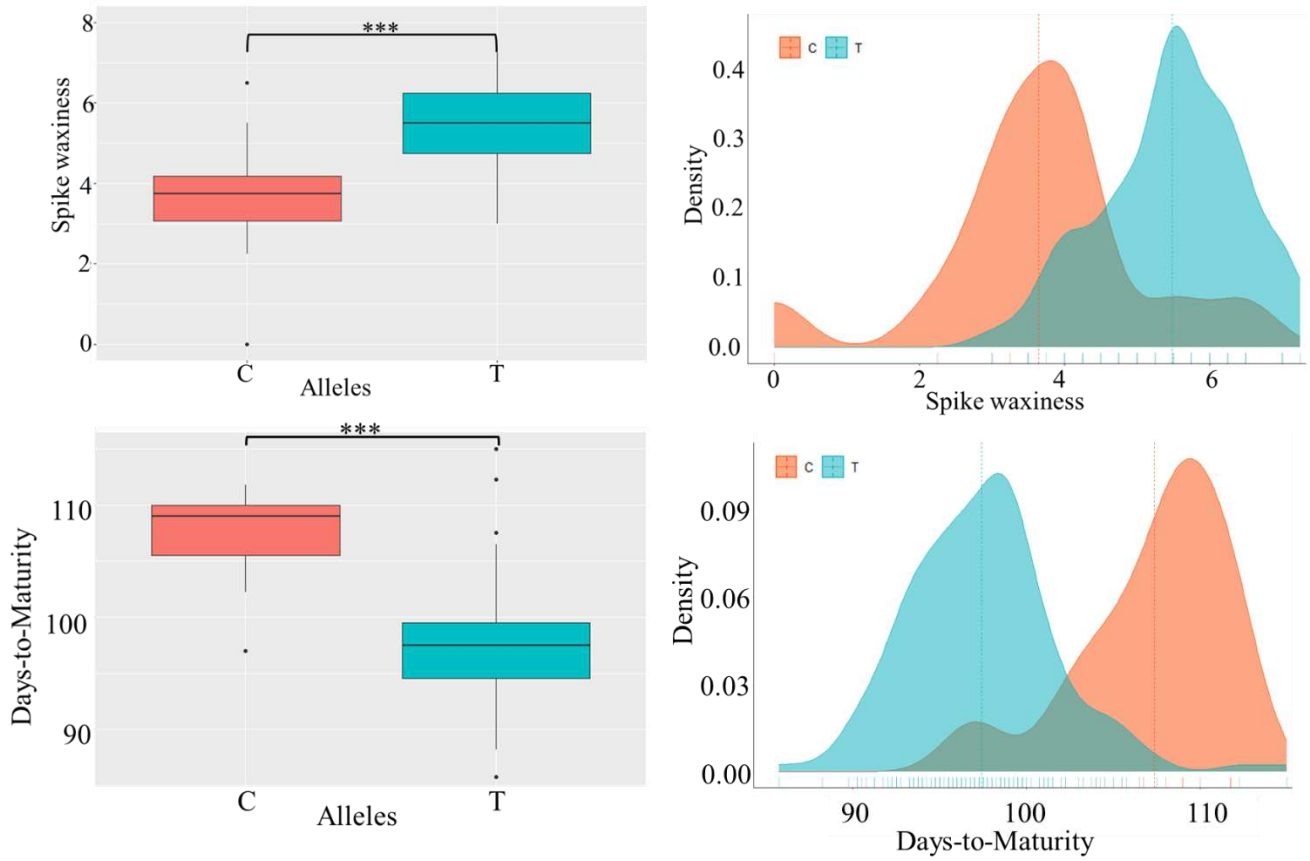


Figure 4. Boxplots (lef) and bimodal distribution (right) of Spike Waxiness and Days-to-maturity are represented for each haplotype. ***: signifcant at $P < 0.001$

Table 2 Details of loci associated with phenotypic traits identified by at least two methods in wheat.

Traits	Loci	P.value	MAF	Allelic effect	PVE(%)	Alleles (Maj/Min)	Models
SW	Chr1A:556842331	3.95E-08	0.09	0.66	50.00	T/C	BLINK/FarmCPU/RF/ANN/SVR/CNN/SVM
DM	Chr4B:666048201	2.73E-10	0.08	-3.84	19.3	T/C	FarmCPU/BLINK/RF/ANN/SVR/CNN
	Chr4B:37907825	2.32E-10	0.09	-2.76	18.44	T/C	BLINK/RF/ANN/SVR/CNN/SVM
	Chr2A:605016602	6.36E-07	0.06	2.39	8.65	G/A	FarmCPU/BLINK/RF/ANN/SVR/CNN/SVM
	Chr5A:580797118	1.64E-09	0.09	-2.28	10.72	C/T	BLINK/RF/ANN/SVR/CNN/SVM

SW: Spike Waxiness; DM: Days to Maturity; MAF: Minor Allele Frequency; PVE: Phenotype_Variance_Explained (%).

Three conventional GWAS analytical tools, including CMLM (Compressed Mixed Linear Model); FarmCPU (Fixed and random model Circulating Probability Unification) and BLINK (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway), completed by five machine learning algorithms, which included RF (Random Forest), SVR (Support Vector Regression), SNN (Convolutional Neural Networks), ANN (Artificial Neural Networks), and SVM (Support Vector Machines) were used. The most highly associated SNP within each QTL is indicated along with the associated statistics of RF model.

Identification of candidate genes

To pinpoint the candidate genes that potentially govern SW and DM in our diverse wheat collection, we conducted an analysis of the genes located within the same linkage block as the peak SNP for each QTL.

In the genomic interval encompassing the QTL that contributes the most to the phenotypic variation in SW (50%) specifically, the region from 1A_555 to 557 Mb, surrounding the peak SNP (chr1A:556842331), we identified a total of 24 high-confidence genes. Upon a detailed examination of the gene annotations and expression profiles, one gene, TraesCS1A01G385500 on chromosome 1A, emerged as the most promising candidate. TraesCS1A01G385500 is an ortholog of the *Arabidopsis Thaliana* O-acyltransferase gene, commonly known as WSD1, a bifunctional wax ester synthase/diacylglycerol acyltransferase, involved in cuticular wax biosynthesis and essential for reducing leaf water loss, particularly during drought conditions. WSD1 has also been associated with the biosynthesis of very long-chain (VLC) wax esters, contributing to drought tolerance in Arabidopsis. This gene exhibits the highest expression levels in spike, roots, and shoot axis tissues (Figure 5). More details are provided in Supplementary Table S3.

In our quest to identify potential candidate genes governing DM trait in our diverse wheat collection, we performed a meticulous analysis of the genes residing within the same linkage block as the peak SNP for each QTL. Within the genomic interval encompassing the QTL that makes the most substantial contribution to the phenotypic variation in DM (comprising 19.3% of the total variation), specifically spanning from 4B_666 to 668 Mb and surrounding the peak SNP (chr4B:666048201), we pinpointed a total of 27 high-confidence genes. Through an in-depth examination of gene annotations and expression profiles, no one emerged as the most promising candidate. More details are provided in Supplementary Table S3.

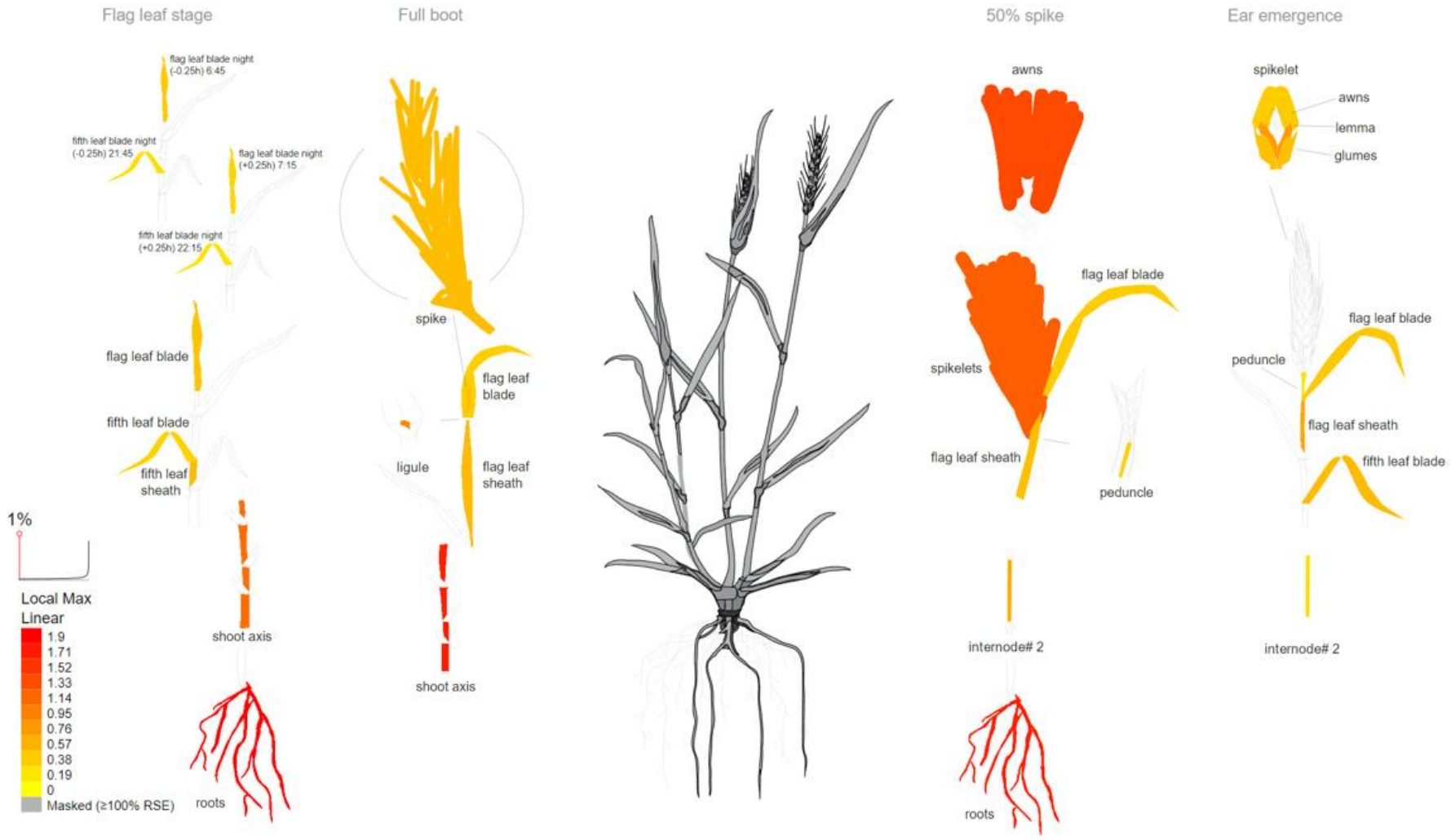


Figure 5. Expression profile of *TraesCS1A01G385500* gene based on transcriptomic analysis in wheat. As shown, this gene is most expressed in spike, roots and shoot axis and the image was generated with the eFP (RNA-Seq data) at http://bar.utoronto.ca/eplant/ by Waese et al.51. The legend at bottom lef presents the expression levels, coded by colors (yellow=low, red=high).

Comparison of ML and conventional GWAS methods for identifying genomic regions

In our study, the peak SNP position on chr1A:556842331 exhibited the highest allelic effect (0.66) and explained a substantial phenotypic variance (50%) among all the identified SNPs associated with wheat SW trait. These associations were successfully detected using both GWAS models (BLINK and FarmCPU) and five machine learning algorithms (RF, ANN, SVR, CNN, and SVM). Importantly, conventional methods (MLM and CMLM) failed to identify this peak SNP linked to the SW trait (Supplementary Table S4). Furthermore, the machine learning models produced significantly lower p-values (CNN, RF, SVM, SVR, and ANN, all with p-values of 3.95E-08) compared to conventional GWAS models (CMLM with a p-value of 5.97E-05, MLM with a p-value of 5.97E-05, FarmCPU with a p-value of 6.34E-03, and BLINK with a p-value of 3.95E-08).

Shifting our focus to the DM trait, our investigation revealed the co-identification of four QTLs on chromosomes 4B, 5A, and 2A (with peak SNPs chr4B:666048201, chr4B:37907825, chr5A:580797118, and chr2A:605016602) by both GWAS models (BLINK and FarmCPU) and the five machine learning algorithms (RF, ANN, SVR, CNN, and SVM). The details of these robust associations are succinctly presented in Table 2. Regrettably, conventional methods (MLM and CMLM) were unable to detect these SNP peaks associated with the DM trait. Additionally, the machine learning models yielded significantly lower p-values for each of the associated markers compared to conventional GWAS models (Supplementary Table S4).

For the SNP chr4B:666048201, the machine learning models generated markedly lower p-values (CNN, RF, SVM, SVR, and ANN, all with p-values of 1.09E-13) compared to conventional GWAS models (CMLM with a p-value of 1.42E-06, MLM with a p-value of 1.42E-06, FarmCPU with a p-value of 2.73E-10, and BLINK with a p-value of 1.09E-13).

For the SNP chr4B:37907825, the machine learning models produced substantially lower p-values (CNN, RF, SVM, SVR, and ANN, all with p-values of 2.32E-10) compared to conventional GWAS models (CMLM with a p-value of 2.65E-04, MLM with a p-value of 2.65E-04, FarmCPU with a p-value of 4.81E-04, and BLINK with a p-value of 2.32E-10).

For the SNP chr5A:580797118, the machine learning models yielded notably lower p-values (CNN, RF, SVM, SVR, and ANN, all with p-values of 1.64E-09) compared to conventional GWAS models (CMLM with a p-value of 4.72E-04, MLM with a p-value of 4.72E-04, and FarmCPU with a p-value of 1, and BLINK with a p-value of 1.64E-09).

For the SNP chr2A:605016602, the machine learning models generated significantly lower p-values (CNN, RF, SVM, SVR, and ANN, all with p-values of 1.64E-09) compared to

conventional GWAS models (CMLM with a p-value of 1.38E-01, MLM with a p-value of 1.38E-01, and FarmCPU with a p-value of 8.63E-02, and BLINK with a p-value of 6.36E-07).

In essence, these traditional methods used for genetic data analysis and the establishment of associations between genetic variations and traits have limitations when it comes to identifying subtle or minor effects of certain SNPs on SW and DM characteristics in wheat. On the contrary, machine learning algorithms, particularly in conjunction with the recent multi-locus GWAS models (BLINK and FarmCPU), exhibited superior performance in identifying relevant QTLs when compared to traditional MLM and CMLM methods.

Discussion

In this study, we employed four GWAS models and five machine learning algorithms to investigate the genomic regions associated with spike waxiness and days to maturity within a dataset consisting of 170 accessions and 74K SNPs. Our analyses consistently identified a robust QTL located on chromosome 1A, demonstrating significance across both conventional GWAS models (FarmCPU and BLINK) and a variety of machine learning models (BLINK, RF, ANN, SVR, CNN, and SVM). Notably, the peak SNP (chr1A:556842331) within this QTL explained a substantial portion, 50%, of the phenotypic variation observed. Within the genomic interval encompassed by this QTL (1A_555 to 557 Mb) and centered around the peak SNP (chr1A:556842331), we identified a total of 24 high-confidence genes. Upon closer examination of gene annotations and expression profiles, one candidate gene, *TraesCS1A01G385500* on chromosome 1A, stood out as particularly promising. This gene exhibits high expression levels in spike, roots, and shoot axis tissues and shares orthology with the *Arabidopsis Thaliana* O-acyltransferase gene, widely known as the *WSD1* gene. Previous research has highlighted the significance of the *WSD1* gene, which serves as a bifunctional wax ester synthase/diacylglycerol acyltransferase (Li et al., 2008; Patwari et al., 2019). Its involvement in cuticular wax biosynthesis is well-documented, and it plays a pivotal role in reducing leaf water loss, particularly during drought conditions (Li et al., 2008; Patwari et al., 2019). The WSD1 gene has also been associated with the biosynthesis of very long chain (VLC) wax esters, contributing to drought tolerance in Arabidopsis (Patwari et al., 2019). VLC primary alcohols and acyl-CoAs serve as precursors for wax ester biosynthesis, catalyzed by the bifunctional wax ester synthase/diacylglycerol acyltransferase WSD1 (Li et al., 2008; Patwari et al., 2019). These wax components, including VLC fatty acids, aldehydes, alkanes, alcohols, ketones, and esters, undergo trafficking through the Golgi and trans-Golgi network (TGN)

pathways to the plasma membrane (PM). From there, they are exported to the cuticle via ABCG subfamily half transporters and lipid transfer proteins (LTPs) (DeBono et al., 2009; Ichino and Yazaki, 2022; Pighin et al., 2004; Wang and Chang, 2022).

Moreover, prior studies have revealed the role of AtCER1 in VLC alkane biosynthesis in Arabidopsis (Aarts et al., 1995; Bourdenx et al., 2011; Sakuradani et al., 2013). Recently, He et al. (2022) identified a homologous gene of AtCER1 in wheat, named TaCER1-6A, which shares 55% amino acid identity with AtCER1. Similar to previously reported AtCER1 orthologs, including rice OsCER1 (Ni et al., 2018), wheat TaCER1-1A (Li et al., 2019), Brachypodium BdCER1-8 (Wu et al., 2019), cucumber CsCER1 (W. Wang et al., 2015), and P. pratensis PpCER1 (Wang et al., 2021), TaCER1-6A also contains three specific His-rich motifs essential for VLC alkane biosynthesis (Bernard et al., 2012). Therefore, (He et al., 2022) speculated that TaCER1-6A likely plays a similar role in VLC alkane biosynthesis in wheat. Notably, we observed that alleles associated with higher wax content were more prevalent in lines originating from East African spring-type accessions (Kenya and Ethiopia) and North Africa. These accessions primarily consist of spring lines cultivated in arid regions and were acquired from the International Center for Agricultural Research in the Dry Areas (ICARDA).

Ultimately, our study has unveiled a promising candidate gene, TraesCS1A01G385500, linked to spike waxiness, with implications for cuticular wax biosynthesis and its role in drought tolerance, as established in previous research. This discovery sheds light on the genetic mechanisms underpinning spike waxiness in bread wheat, offering valuable insights for future breeding and crop improvement efforts.

Regarding DM, we identified four strong genomic regions significantly associated with the trait on chromosomes 4B, 2A and 5A. Our results were consistent with those of (Semagn et al., 2021), who performed QTL mapping in four RIL populations evaluated under conventional and organic management systems and reported two QTLs associated with days to maturity on chromosome 4B (explaining 20.8% of the phenotypic variances), where one (QMat.dms-4B.2) at chr4B:569184188-599613837 is located on the extremity of long chromosome 4B arm with the peak SNP chr4B:666048201 (explaining 19.3% of the phenotypic variation) that was jointly identified by both multi-locus GWAS models and four ML algorithms (RF, ANN, SVR, and CNN) in the present study. QTL mapping conducted by previous authors (Zou et al., 2017a, 2017b) in the ‘Attila’ and ‘CDC Go’ RIL populations using genetic maps of 1203 markers identified a coincident genomic region associated with maturity under both management systems on chromosome 4B and 5A. We also identified QTLs on chromosomes 4B and 5A,

with peak SNPs chr4B:666048201, chr4B:37907825, and chr5A:580797118 explaining 19.3%, 18.44%, and 10.72%, of the phenotypic variation, respectively. The favorable alleles for those QTLs on 4B and 5A were most originated from the accessions of North Africa, including spring lines (Attila) acquired from the International Center for Agricultural Research in the Dry Areas (ICARDA). Interesting, previous works also identified two QTLs for maturity on chromosome 4B (QMat.dms-4B) and chromosome 5A (QMat.dms-5A.2), which individually explained 15.9% and 14.0% of the phenotypic variance, respectively, and together accounted for 29.9% of the phenotypic variance across seven environments (Zou et al., 2017a, 2017b). The favorable alleles for QMat.dms-4B and QMat.dms-5A.2 originated from 'Attila' and 'CDC Go', respectively. (Chen et al., 2020) also identified a QTL associated with maturity on chromosome 4B (4B_s4991673- 4B_d1258252) using a linkage map of 4439 markers produced by DArTseq technology and phenotype data from 'Peace' and 'Carberry' RIL populations assessed for two years under organic management and conventional systems, consistent with our results.

Our investigation into candidate genes associated with maturity in wheat led us to a genomic interval spanning the QTL that contributes significantly to the phenotypic variation in Days-to-Maturity (19.3% of the variation). This region, located between 4B_666 and 668 Mb and centered around the peak SNP (chr4B:666048201), contained a total of 27 high-confidence genes. Our findings align with prior research that has identified genomic regions on chromosome 4B associated with maturity and housing candidate genes related to flower-promoting factors. Notably, certain dwarfing genes, such as Rht-B1, Rht5, Rht8, and Rht12, have been reported to exert slight delays in heading, flowering, and/or maturity time in wheat. These genetic factors add complexity to the regulation of these traits (Chen et al., 2018; Daoura Goudia et al., 2014). The discovery in the present study contributes to our understanding of the genetic factors underpinning wheat maturity and sets the stage for future research aimed at elucidating the molecular mechanisms involved.

The results our study highlight the remarkable superiority of machine learning (ML) models in identifying significant genetic associations compared to traditional Genome-Wide Association Study (GWAS) methods, as demonstrated through substantially lower p-values. For SW, the peak SNP was efficiently identified by both GWAS models (BLINK and FarmCPU) and the five ML algorithms, emphasizing their robustness. Notably, the ML models, including CNN, RF, SVM, SVR, and ANN, produced significantly lower p-values (3.95E-08) compared to the traditional GWAS models, which had p-values ranging from 5.97E-05 to 6.34E-03. Traditional methods (MLM and CMLM) failed to detect this critical SNP, showcasing their limitations in capturing minor genetic effects. Shifting the focus to DM, the robust associations identified by

GWAS models and ML algorithms demonstrated that conventional methods (MLM and CMLM) were less effective, failing to detect these essential SNP peaks. Once again, ML models consistently delivered significantly lower p-values, underscoring their increased sensitivity and accuracy in identifying genetic markers linked to DM. The differences in p-values were substantial, with ML models consistently outperforming the traditional GWAS methods.

These findings reveal that traditional GWAS methods face limitations in detecting minor genetic effects on SW and DM traits in wheat. Conversely, ML models, especially when coupled with advanced multi-locus GWAS models like BLINK and FarmCPU, exhibited a superior performance characterized by significantly lower p-values. This work demonstrates the potential of ML approaches to revolutionize the study of complex genetic traits, offering valuable insights for crop improvement and stress resilience in bread wheat. Our hypotheses (1) regarding the presence of specific genomic regions associated with SW and DM in a diverse global collection of bread wheat accessions and (2) the superior performance of Machine Learning-Genome-Wide Association Study (ML-GWAS) approaches over traditional GWAS methods in identifying relevant genomic regions associated with SW and DM traits in bread wheat have been confirmed. Our study has provided evidence that conventional GWAS approaches, such as MLM, and CMLM, lack the ability to effectively detect SNPs with minor effects underlying specific traits. In other words, these traditional methods used to analyze genetic data and establish associations between genetic variations and traits are not sensitive enough to identify subtle or minor effects of certain SNPs on the characteristics of SW and DM in wheat. These findings align with previous research conducted by (Yoosefzadeh-Najafabadi et al., 2023; Zhou et al., 2019), which also highlighted the limited power of conventional GWAS approaches in detecting SNPs with minor effects on specific traits.

However, our study has revealed the effectiveness of an alternative approach, utilizing machine learning algorithms in GWAS. By employing this method, we were able to overcome the limitations of traditional GWAS and more accurately identify SNPs with smaller yet significant effects on SW and DM traits in wheat. Additionally, the most robust associations identified by modern GWAS methodologies (BLINK and FarmCPU) were reaffirmed by machine learning techniques. These results are consistent with the studies conducted by (Yoosefzadeh-Najafabadi et al., 2023) and (Zhou et al., 2019), which compared SVR and RF algorithms, respectively, to conventional GWAS methods in soybean. They reported that machine learning algorithms are more accurate and sensitive in detecting subtle or minor effects of certain SNPs on traits of interest. Additionally, our study demonstrated the effectiveness of the new GWAS model, BLINK and FarmCPU, in accurately and efficiently detecting SNPs with smaller but significant

589 effects on SW and DM traits in wheat. In fact, both real and simulated data analyses have shown
590 that BLINK significantly improves statistical power compared to FarmCPU, while also
591 reducing computing time (Huang et al., 2019).

592 The FarmCPU, developed by (Liu et al., 2016), represents an iterative method that addresses
593 the issue of false positive control and confounding between testing markers and cofactors
594 simultaneously. As FarmCPU tests markers in a fixed-effect model, it is computationally more
595 efficient than methods that test markers in a random-effect model, such as MLM, CMLM,
596 ECMLM, SUPER, and MLMM (Liu et al., 2016). Studies have demonstrated the order of the
597 statistical power of these methods: BLINK > FarmCPU > CMLM > MLM (Huang et al., 2019;
598 Liu et al., 2016; Zhang et al., 2010).

599 The utilization of machine learning algorithms (RF, ANN, SVR, CNN, and SVM), along with
600 the recent multi-locus GWAS model, BLINK, and FarmCPU has enabled a more sensitive and
601 precise identification of genetic factors influencing specific traits. This opens up new
602 opportunities for wheat improvement and selection. Indeed, ML algorithms are focused on
603 maximizing prediction accuracy at the individual subject level and have been shown to capture
604 multi-locus SNP interactions better than univariate association studies (Okser et al., 2014,
605 2013). Additionally, ML techniques provide an opportunity to better understand multi-locus
606 genetic variants and their interactions in predicting complex traits (Ashkenazy et al., 2022;
607 Kwon et al., 2022). This approach provides a more sophisticated and reliable means of
608 discovering genetic markers associated with SW and DM traits, which can have significant
609 implications for agriculture, varietal selection, and understanding the genetic mechanisms
610 governing crop characteristics.

611 Overall, both GWAS and machine learning methods have successfully addressed the
612 confounding effects of population structure and relatedness, allowing us to identify multiple
613 genomic regions associated with SW and DM traits in wheat. These findings shed light on the
614 genetic architecture of these traits and offer valuable insights to plant breeders in their efforts
615 to develop new bread wheat varieties with improved SW and DM.

616
617
618
619
620
621

Conclusion

In this study, our primary objective was to identify the genomic regions associated with SW and DM using state-of-the-art Machine Learning-Genome-Wide Association Study (ML-GWAS) techniques. Our findings provide a deep understanding of the genetic landscape governing these critical traits, delivering valuable insights that can significantly inform wheat breeding and crop improvement strategies. Leveraging ML-GWAS, we successfully identified a robust QTL significantly associated with SW on chromosome 1A, represented by the peak SNPs chr1A:556842331, explaining an impressive 50% of the phenotypic variation. Additionally, we detected four strong genomic regions significantly associated with DM on chromosomes 4B, 2A, and 5A, employing the same cutting-edge methods. Notably, our study unveiled a candidate gene linked to the QTLs for SW. *TraesCS1A01G385500*, an ortholog of the *Arabidopsis Thaliana* O-acyltransferase gene *WSD1*, plays a pivotal role in cuticular wax biosynthesis. It is essential for reducing leaf water loss, particularly during drought conditions, and contributes to drought tolerance through the biosynthesis of very long-chain (VLC) wax esters. Our study also shows that, ML models, especially when coupled with advanced multi-locus GWAS models like BLINK and FarmCPU, exhibited a superior performance characterized by significantly lower p-values in identifying relevant QTLs compared to traditional methods like MLM and CMLM. This work demonstrates the potential of ML approaches to revolutionize the study of complex genetic traits, offering valuable insights for crop improvement and stress resilience in bread wheat. ML-GWAS emerges as a compelling approach for genomic-based breeding strategies, providing breeders with more accurate and efficient tools to develop improved wheat varieties. Our research significantly advances the precision and effectiveness of GWAS, emphasizing the importance of incorporating advanced computational methods into crop breeding studies. The insights into the genetic architecture of SW and DM traits in wheat offer essential knowledge for designing targeted crop improvement strategies. Moreover, the versatility and effectiveness demonstrated by the ML-GWAS approach extend its applicability beyond wheat and can be harnessed to address other crop traits, thus enhancing progress in crop genetics research and breeding efforts on a broader scale. Overall, the integration of machine learning techniques with GWAS stands as a potent tool for dissecting complex traits in crop genetics research. The findings of our study hold great promise for the field of wheat breeding and crop improvement strategies, making substantial contributions to enhancing agricultural productivity and ensuring food security in the face of evolving global challenges.

Additional Information Supplementary information for this paper is available at:

Competing interests: The authors declare that they have no competing interests.

Acknowledgements:

Authors are grateful to the International Maize and Wheat Improvement Center (CIMMYT), the International Center for Agricultural Research in the Dry Areas (ICARDA) and the Plant Breeding Laboratory (Department of Genetics, Stellenbosch University) for their technical supports and wheat varieties collection. We would like to thank Dr Wuletaw Tadesse, Mr Tsimi Patrick, Mr Charly Mam for their technical support. We are grateful to Dr Amina Abed and Dr Jérôme Laroche for their technical support during bioinformatics analyses.

References

Aarts, M.G., Keijzer, C.J., Stiekema, W.J., Pereira, A., 1995. Molecular characterization of the CER1 gene of arabidopsis involved in epicuticular wax biosynthesis and pollen fertility. Plant Cell 7, 2115–2127. https://doi.org/10.1105/tpc.7.12.2115

Ashkenazy, N., Feder, M., Shir, O.M., Hübner, S., 2022. GWANN: Implementing deep learning in genome wide association studies. https://doi.org/10.1101/2022.06.01.494275

Belzile, F., Abed, A., Torkamaneh, D., 2020. Time for a paradigm shift in the use of plant genetic resources. Genome 63, 189–194. https://doi.org/10.1139/gen-2019-0141

Bernard, A., Domergue, F., Pascal, S., Jetter, R., Renne, C., Faure, J.-D., Haslam, R.P., Napier, J.A., Lessire, R., Joubès, J., 2012. Reconstitution of plant alkane biosynthesis in yeast demonstrates that Arabidopsis ECERIFERUM1 and ECERIFERUM3 are core components of a very-long-chain alkane synthesis complex. Plant Cell 24, 3106–3118. https://doi.org/10.1105/tpc.112.099796

Bourdenx, B., Bernard, A., Domergue, F., Pascal, S., Léger, A., Roby, D., Pervent, M., Vile, D., Haslam, R.P., Napier, J.A., Lessire, R., Joubès, J., 2011. Overexpression of Arabidopsis ECERIFERUM1 promotes wax very-long-chain alkane biosynthesis and influences plant response to biotic and abiotic stresses. Plant Physiol. 156, 29–45. https://doi.org/10.1104/pp.111.172320

Chen, H., Bemister, D.H., Iqbal, M., Strelkov, S.E., Spaner, D.M., 2020. Mapping genomic regions controlling agronomic traits in spring wheat under conventional and organic managements. Crop Sci. 60, 2038–2052. https://doi.org/10.1002/csc2.20157

Chen, L., Du, Y., Lu, Q., Chen, H., Meng, R., Cui, C., Lu, S., Yang, Y., Chai, Y., Li, J., Liu, L., Qi, X., Li, H., Mishina, K., Yu, F., Hu, Y.-G., 2018. The Photoperiod-Insensitive Allele Ppd-D1a Promotes Earlier Flowering in Rht12 Dwarf Plants of Bread Wheat. Front. Plant Sci. 9.

Churchill, G.A., Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics 138, 963–971. https://doi.org/10.1093/genetics/138.3.963

Daoura Goudia, B., Chen, L., Yingying, D., Yingang, H., 2014. Genetic effects of dwarfing gene Rht-5 on agronomic traits in common wheat (Triticum aestivum L.) and QTL analysis on its linked traits. Field Crops Res. 156, 22–29. https://doi.org/10.1016/j.fcr.2013.10.007

DeBono, A., Yeats, T.H., Rose, J.K.C., Bird, D., Jetter, R., Kunst, L., Samuels, L., 2009. Arabidopsis LTPG Is a Glycosylphosphatidylinositol-Anchored Lipid Transfer Protein Required for Export of Lipids to the Plant Surface. Plant Cell 21, 1230–1238. https://doi.org/10.1105/tpc.108.064451

Doerge, R.W., Churchill, G.A., 1996. Permutation Tests for Multiple Loci Affecting a Quantitative Character. Genetics 142, 285–294. https://doi.org/10.1093/genetics/142.1.285

Doyle, J.J., Doyle, J.L., 1990. Isolation ofplant DNA from fresh tissue. Focus 12, 39–40.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6, e19379. https://doi.org/10.1371/journal.pone.0019379

Enoma, D.O., Bishung, J., Abiodun, T., Ogunlana, O., Osamor, V.C., 2022. Machine learning approaches to genome-wide association studies. J. King Saud Univ. - Sci. 34, 101847. https://doi.org/10.1016/j.jksus.2022.101847

He, J., Li, C., Hu, N., Zhu, Y., He, Z., Sun, Y., Wang, Z., Wang, Y., 2022. ECERIFERUM1-6A is required for the synthesis of cuticular wax alkanes and promotes drought tolerance in wheat. Plant Physiol. 190, 1640–1657. https://doi.org/10.1093/plphys/kiac394

Hen-Avivi, S., Savin, O., Racovita, R.C., Lee, W.-S., Adamski, N.M., Malitsky, S., Almekias-Siegl, E., Levy, M., Vautrin, S., Bergès, H., Friedlander, G., Kartvelishvily, E., Ben-Zvi, G., Alkan, N., Uauy, C., Kanyuka, K., Jetter, R., Distelfeld, A., Aharoni, A., 2016. A Metabolic Gene Cluster in the Wheat W1 and the Barley Cer-cqu Loci Determines β-Diketone Biosynthesis and Glaucousness. Plant Cell 28, 1440–1460. https://doi.org/10.1105/tpc.16.00197

Huang, M., Liu, X., Zhou, Y., Summers, R.M., Zhang, Z., 2019. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience 8, giy154. https://doi.org/10.1093/gigascience/giy154

Ichino, T., Yazaki, K., 2022. Modes of secretion of plant lipophilic metabolites via ABCG transporter-dependent transport and vesicle-mediated trafficking. Curr. Opin. Plant Biol. 66, 102184. https://doi.org/10.1016/j.pbi.2022.102184

Ingram, G., Nawrath, C., 2017. The roles of the cuticle in plant development: organ adhesions and beyond. J. Exp. Bot. 68, 5307–5321. https://doi.org/10.1093/jxb/erx313

International Wheat Genome Sequencing Consortium (IWGSC), 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaar7191. https://doi.org/10.1126/science.aar7191

Iqbal, M., Navabi, A., Salmon, D.F., Yang, R.-C., Murdoch, B.M., Moore, S.S., Spaner, D., 2007. Genetic analysis of flowering and maturity time in high latitude spring wheat. Euphytica 154, 207–218. https://doi.org/10.1007/s10681-006-9289-y

Kaler, A.S., Gillman, J.D., Beissinger, T., Purcell, L.C., 2020. Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. Front. Plant Sci. 10, 1794. https://doi.org/10.3389/fpls.2019.01794

Kamran, A., Iqbal, M., Navabi, A., Randhawa, H., Pozniak, C., Spaner, D., 2013. Earliness per se QTLs and their interaction with the photoperiod insensitive allele Ppd-D1a in the Cutler × AC Barrie spring wheat population. Theor. Appl. Genet. 126, 1965–1976. https://doi.org/10.1007/s00122-013-2110-0

Kamran, Atif, Randhawa, H.S., Pozniak, C., Spaner, D., 2013. Phenotypic Effects of the Flowering Gene Complex in Canadian Spring Wheat Germplasm. Crop Sci. 53, 84–94. https://doi.org/10.2135/cropsci2012.05.0313

Kosma, D.K., Bourdenx, B., Bernard, A., Parsons, E.P., Lü, S., Joubès, J., Jenks, M.A., 2009. The impact of water deficiency on leaf cuticle lipids of Arabidopsis. Plant Physiol. 151, 1918–1929. https://doi.org/10.1104/pp.109.141911

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R.C., 2020. Package 'caret.' R J. 223.

Kunst, L., Samuels, L., 2009. Plant cuticles shine: advances in wax biosynthesis and export. Curr. Opin. Plant Biol. 12, 721–727. https://doi.org/10.1016/j.pbi.2009.09.009

Kwon, O.-S., Hong, M., Kim, T.-H., Hwang, I., Shim, J., Choi, E.-K., Lim, H.E., Yu, H.T., Uhm, J.-S., Joung, B., Oh, S., Lee, M.-H., Kim, Y.-H., Pak, H.-N., 2022. Genome-wide association study-based prediction of atrial fibrillation using artificial intelligence. Open Heart 9, e001898. https://doi.org/10.1136/openhrt-2021-001898

Li, F., Wu, X., Lam, P., Bird, D., Zheng, H., Samuels, L., Jetter, R., Kunst, L., 2008. Identification of the Wax Ester Synthase/Acyl-Coenzyme A:Diacylglycerol Acyltransferase WSD1 Required for Stem Wax Ester Biosynthesis in Arabidopsis. Plant Physiol. 148, 97–107. https://doi.org/10.1104/pp.108.123471

Li, T., Sun, Y., Liu, T., Wu, H., An, P., Shui, Z., Wang, J., Zhu, Y., Li, C., Wang, Y., Jetter, R., Wang, Z., 2019. TaCER1-1A is involved in cuticular wax alkane biosynthesis in hexaploid wheat and responds to plant abiotic stresses. Plant Cell Environ. 42, 3077–3091. https://doi.org/10.1111/pce.13614

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., Zhang, Z., 2012. GAPIT: genome association and prediction integrated tool. Bioinforma. Oxf. Engl. 28, 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Liu, X., Huang, M., Fan, B., Buckler, E.S., Zhang, Z., 2016. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genet. 12, e1005767. https://doi.org/10.1371/journal.pgen.1005767

Martin, L.B.B., Rose, J.K.C., 2014. There's more than one way to skin a fruit: formation and functions of fruit cuticles. J. Exp. Bot. 65, 4639–4651. https://doi.org/10.1093/jxb/eru301

Mohammadi, M., Xavier, A., Beckett, T., Beyer, S., Chen, L., Chikssa, H., Cross, V., Freitas Moreira, F., French, E., Gaire, R., Griebel, S., Lopez, M.A., Prather, S., Russell, B., Wang, W., 2020. Identification, deployment, and transferability of quantitative trait loci from genome-wide association studies in plants. Curr. Plant Biol. 24, 100145. https://doi.org/10.1016/j.cpb.2020.100145

Ni, E., Zhou, L., Li, J., Jiang, D., Wang, Z., Zheng, S., Qi, H., Zhou, Y., Wang, C., Xiao, S., Liu, Z., Zhou, H., Zhuang, C., 2018. OsCER1 Plays a Pivotal Role in Very-Long-Chain Alkane Biosynthesis and Affects Plastid Development and Programmed Cell Death of Tapetum in Rice (Oryza sativa L.). Front. Plant Sci. 9, 1217. https://doi.org/10.3389/fpls.2018.01217

Nicholls, H.L., John, C.R., Watson, D.S., Munroe, P.B., Barnes, M.R., Cabrera, C.P., 2020. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. Front. Genet. 11.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., Aittokallio, T., 2014. Regularized Machine Learning in the Genetic Prediction of Complex Traits. PLOS Genet. 10, e1004754. https://doi.org/10.1371/journal.pgen.1004754

Okser, S., Pahikkala, T., Aittokallio, T., 2013. Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives. BioData Min. 6, 5. https://doi.org/10.1186/1756-0381-6-5

Pask, A.J.D., Pietragalla, J., Mullan, D.M., Reynolds, M.P., 2012. Physiological breeding II: a field guide to wheat phenotyping. Cimmyt.

795 Patwari, P., Salewski, V., Gutbrod, K., Kreszies, T., Dresen-Scholz, B., Peisker, H., Steiner,
796 U., Meyer, A.J., Schreiber, L., Dörmann, P., 2019. Surface wax esters contribute to
797 drought tolerance in Arabidopsis. Plant J. Cell Mol. Biol. 98, 727–744.
798 https://doi.org/10.1111/tpj.14269
799 Pighin, J.A., Zheng, H., Balakshin, L.J., Goodman, I.P., Western, T.L., Jetter, R., Kunst, L.,
800 Samuels, A.L., 2004. Plant cuticular lipid export requires an ABC transporter. Science
801 306, 702–704. https://doi.org/10.1126/science.1102331
802 Ramírez-González, R.H., Borrill, P., Lang, D., Harrington, S.A., Brinton, J., Venturini, L.,
803 Davey, M., Jacobs, J., van Ex, F., Pasha, A., Khedikar, Y., Robinson, S.J., Cory, A.T.,
804 Florio, T., Concia, L., Juery, C., Schoonbeek, H., Steuernagel, B., Xiang, D., Ridout,
805 C.J., Chalhoub, B., Mayer, K.F.X., Benhamed, M., Latrasse, D., Bendahmane, A.,
806 International Wheat Genome Sequencing Consortium, Wulff, B.B.H., Appels, R.,
807 Tiwari, V., Datla, R., Choulet, F., Pozniak, C.J., Provart, N.J., Sharpe, A.G., Paux, E.,
808 Spannagl, M., Bräutigam, A., Uauy, C., 2018. The transcriptional landscape of
809 polyploid wheat. Science 361, eaar6089. https://doi.org/10.1126/science.aar6089
810 Sakuradani, E., Zhao, L., Haslam, T.M., Kunst, L., 2013. The CER22 gene required for the
811 synthesis of cuticular wax alkanes in Arabidopsis thaliana is allelic to CER1. Planta
812 237, 731–738. https://doi.org/10.1007/s00425-012-1791-y
813 Semagn, K., Iqbal, M., Chen, H., Perez-Lara, E., Bemister, D.H., Xiang, R., Zou, J., Asif, M.,
814 Kamran, A., N'Diaye, A., Randhawa, H., Pozniak, C., Spaner, D., 2021. Physical
815 Mapping of QTL in Four Spring Wheat Populations under Conventional and Organic
816 Management Systems. I. Earliness. Plants 10, 853.
817 https://doi.org/10.3390/plants10050853
818 Seo, P.J., Lee, S.B., Suh, M.C., Park, M.-J., Go, Y.S., Park, C.-M., 2011. The MYB96
819 transcription factor regulates cuticular wax biosynthesis under drought conditions in
820 Arabidopsis. Plant Cell 23, 1138–1152. https://doi.org/10.1105/tpc.111.083485
821 Siegmann, B., Jarmer, T., 2015. Comparison of different regression models and validation
822 techniques for the assessment of wheat leaf area index from hyperspectral data. Int. J.
823 Remote Sens. 36, 4519–4534. https://doi.org/10.1080/01431161.2015.1084438
824 Snape, J., Butterworth, K., Whitechurch, E., Worland, A.J., 2001. Waiting for Fine Times:
825 Genetics of Flowering Time in Wheat, in: Bedö, Z., Láng, L. (Eds.), Wheat in a
826 Global Environment: Proceedings of the 6th International Wheat Conference, 5–9
827 June 2000, Budapest, Hungary, Developments in Plant Breeding. Springer
828 Netherlands, Dordrecht, pp. 67–74. https://doi.org/10.1007/978-94-017-3674-9_7
829 Tekeu, H., Ngonkeu, E.L.M., Bélanger, S., Djocgoué, P.F., Abed, A., Torkamaneh, D., Boyle,
830 B., Tsimi, P.M., Tadesse, W., Jean, M., Belzile, F., 2021. GWAS identifies an
831 ortholog of the rice D11 gene as a candidate gene for grain size in an international
832 collection of hexaploid wheat. Sci. Rep. 11, 19483. https://doi.org/10.1038/s41598-
833 021-98626-0
834 Tekeu, H., Ngonkeu, E.M.L., Djocgoue, F.P., Ellis, A., Lendzemo, V., Springfield, L.,
835 Moulin, L., Klonowska, A., Diouf, D., Botes, W.C., Bena, G., 2017. Genetic diversity
836 of Cameroonian bread wheat (Triticum aestivum L.) cultivars revealed by
837 microsatellite markers. Afr. J. Biotechnol. 16, 1832–1839.
838 https://doi.org/10.5897/AJB2017.16090
839 Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., Belzile, F., 2017. Fast-GBS: a new
840 pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-
841 sequencing data. BMC Bioinformatics 18, 1–7.
842 Waese, J., Fan, J., Pasha, A., Yu, H., Fucile, G., Shi, R., Cumming, M., Kelley, L.A.,
843 Sternberg, M.J., Krishnakumar, V., Ferlanti, E., Miller, J., Town, C., Stuerzlinger, W.,
844 Provart, N.J., 2017. ePlant: Visualizing and Exploring Multiple Levels of Data for

Hypothesis Generation in Plant Biology. Plant Cell 29, 1806–1821. https://doi.org/10.1105/tpc.17.00073

Wang, D., Ni, Y., Liao, L., Xiao, Y., Guo, Y., 2021. Poa pratensis ECERIFERUM1 (PpCER1) is involved in wax alkane biosynthesis and plant drought tolerance. Plant Physiol. Biochem. PPB 159, 312–321. https://doi.org/10.1016/j.plaphy.2020.12.032

Wang, W., Zhang, Y., Xu, C., Ren, J., Liu, X., Black, K., Gai, X., Wang, Q., Ren, H., 2015. Cucumber ECERIFERUM1 (CsCER1), which influences the cuticle properties and drought tolerance of cucumber, plays a key role in VLC alkanes biosynthesis. Plant Mol. Biol. 87, 219–233. https://doi.org/10.1007/s11103-014-0271-0

Wang, X., Chang, C., 2022. Exploring and exploiting cuticle biosynthesis for abiotic and biotic stress tolerance in wheat and barley. Front. Plant Sci. 13, 1064390. https://doi.org/10.3389/fpls.2022.1064390

Wang, Y., Wang, M., Sun, Y., Hegebarth, D., Li, T., Jetter, R., Wang, Z., 2015a. Molecular Characterization of TaFAR1 Involved in Primary Alcohol Biosynthesis of Cuticular Wax in Hexaploid Wheat. Plant Cell Physiol. 56, 1944–1961. https://doi.org/10.1093/pcp/pcv112

Wang, Y., Wang, M., Sun, Y., Wang, Yanting, Li, T., Chai, G., Jiang, W., Shan, L., Li, C., Xiao, E., Wang, Z., 2015b. FAR5, a fatty acyl-coenzyme A reductase, is involved in primary alcohol biosynthesis of the leaf blade cuticular wax in wheat (Triticum aestivum L.). J. Exp. Bot. 66, 1165–1178. https://doi.org/10.1093/jxb/eru457

Wu, H., Shi, S., Lu, X., Li, T., Wang, J., Liu, T., Zhang, Q., Sun, W., Li, C., Wang, Z., Chen, Y., Quan, L., 2019. Expression Analysis and Functional Characterization of CER1 Family Genes Involved in Very-Long-Chain Alkanes Biosynthesis in Brachypodium distachyon. Front. Plant Sci. 10, 1389. https://doi.org/10.3389/fpls.2019.01389

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., Liu, X., 2021. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. Genomics Proteomics Bioinformatics 19, 619–628. https://doi.org/10.1016/j.gpb.2020.10.007

Yoosefzadeh Najafabadi, M., Tulpan, D., Eskandari, M., 2021. Using Hybrid Artificial Intelligence and Evolutionary Optimization Algorithms for Estimating Soybean Yield and Fresh Biomass Using Hyperspectral Vegetation Indices. Remote Sens. 13, 2555. https://doi.org/10.3390/rs13132555

Yoosefzadeh-Najafabadi, M., Torabi, S., Tulpan, D., Rajcan, I., Eskandari, M., 2023. Application of SVR-Mediated GWAS for Identification of Durable Genetic Regions Associated with Soybean Seed Quality Traits. Plants 12, 2659. https://doi.org/10.3390/plants12142659

Yu, J., Buckler, E.S., 2006. Genetic association mapping and genome organization of maize. Curr. Opin. Biotechnol. 17, 155–160. https://doi.org/10.1016/j.copbio.2006.02.003

Zadoks, J.C., Chang, T.T., Konzak, C.F., 1974. A decimal code for the growth stages of cereals. Weed Res. 14, 415–421. https://doi.org/10.1111/j.1365-3180.1974.tb01084.x

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., Buckler, E.S., 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42, 355–360. https://doi.org/10.1038/ng.546

Zhou, W., Bellis, E.S., Stubblefield, J., Causey, J., Qualls, J., Walker, K., Huang, X., 2019. Minor QTLs mining through the combination of GWAS and machine learning feature selection. https://doi.org/10.1101/712190

Zou, J., Semagn, K., Iqbal, M., Chen, H., Asif, M., N'Diaye, A., Navabi, A., Perez-Lara, E., Pozniak, C., Yang, R.-C., Randhawa, H., Spaner, D., 2017a. QTLs associated with agronomic traits in the Attila × CDC Go spring wheat population evaluated under

conventional management. PLOS ONE 12, e0171528. https://doi.org/10.1371/journal.pone.0171528

Zou, J., Semagn, K., Iqbal, M., N'Diaye, A., Chen, H., Asif, M., Navabi, A., Perez-Lara, E., Pozniak, C., Yang, R., Randhawa, H., Spaner, D., 2017b. Mapping QTLs Controlling Agronomic Traits in the 'Attila' × 'CDC Go' Spring Wheat Population under Organic Management using 90K SNP Arra. Crop Sci. 57, 365–377. https://doi.org/10.2135/cropsci2016.06.0459