International Journal of
Molecular Sciences

MDPI

Article

Deciphering Pleiotropic Signatures of Regulatory SNPs in *Zea mays* L. Using Multi-Omics Data and Machine Learning Algorithms

Ataul Haleem [1,2], Selina Klees [1,3], Armin Otto Schmitt [1,3] and Mehmet Gültas [2,3,*]

[1] Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; ataul.haleem@uni-goettingen.de (A.H.); selina.klees@uni-goettingen.de (S.K.); armin.schmitt@uni-goettingen.de (A.O.S.)
[2] Faculty of Agriculture, South Westphalia University of Applied Sciences, Lübecker Ring 2, 59494 Soest, Germany
[3] Center for Integrated Breeding Research (CiBreed), Georg-August University, Carl-Sprengel-Weg 1, 37075 Göttingen, Germany
* Correspondence: gueltas.mehmet@fh-swf.de

Abstract: Maize is one of the most widely grown cereals in the world. However, to address the challenges in maize breeding arising from climatic anomalies, there is a need for developing novel strategies to harness the power of multi-omics technologies. In this regard, pleiotropy is an important genetic phenomenon that can be utilized to simultaneously enhance multiple agronomic phenotypes in maize. In addition to pleiotropy, another aspect is the consideration of the regulatory SNPs (rSNPs) that are likely to have causal effects in phenotypic development. By incorporating both aspects in our study, we performed a systematic analysis based on multi-omics data to reveal the novel pleiotropic signatures of rSNPs in a global maize population. For this purpose, we first applied Random Forests and then Markov clustering algorithms to decipher the pleiotropic signatures of rSNPs, based on which hierarchical network models are constructed to elucidate the complex interplay among transcription factors, rSNPs, and phenotypes. The results obtained in our study could help to understand the genetic programs orchestrating multiple phenotypes and thus could provide novel breeding targets for the simultaneous improvement of several agronomic traits.

Keywords: multi-omics; regulatory SNPs; incremental feature selection; random forest; markov clustering; hierarchical network model; gene expression profiles

Citation: Haleem, A.; Klees, S.; Schmitt, A.O.; Gültas, M. Deciphering Pleiotropic Signatures of Regulatory SNPs in *Zea mays* L. Using Multi-Omics Data and Machine Learning Algorithms. *Int. J. Mol. Sci.* **2022**, *23*, 5121. https://doi.org/10.3390/ijms23095121

Academic Editor: Shaozhen He

Received: 30 March 2022
Accepted: 2 May 2022
Published: 4 May 2022

1. Introduction

Maize is an exceptional source of food, feed and fuel. It has become one of the most important cereal crops that feed the world by contributing 30% of the food calories for 4.5 billion people [1]. Over the past decade, maize production has increased remarkably by more than 1.16 million tons (FAOSTAT), but there is still a need for further yield increases to offset the food insecurity caused by the exponential increase in the world's population. However, the dramatic fluctuations in global mean temperatures observed over the past decades pose a serious threat to sustainable crop production and demand better strategies of crop improvement.

To deal with above-mentioned issues, several biochemical, physiological and morphological traits of maize (such as grain yield and biomass) have been of individual importance in breeding research [2–13]. For this purpose, several association studies have been conducted for marker-assisted selection of superior genotypes by applying conventional genome-wide association studies (GWAS), which could provide essential information about the genetic architecture of genotype × phenotype interactions [14]. Despite the rich literature on GWAS and their application in plant breeding, they are still criticized for

high false positive rates [15–17], requirement of large sample sizes for the detection of rare alleles [18] and missing heritability [19]. In order to overcome these limitations to certain extents, machine learning (ML) approaches, like Random Forests (RFs) or convolutional neural networks (CNNs), have been successfully applied to large genomic data sets, which employ non-parametric methods to decipher genotype × phenotype interactions [17,20–26]. Especially, recent studies [16,17,27,28] have demonstrated the utility of RF based-models for the analysis of a large number of loci and the identification of promising SNP candidates having strong associations with phenotypes. For example, Klees et al. [20] recently applied the RF approach to identify associations between the rapeseed oil content and regulatory SNPs (rSNPS), which are located in the promoter regions of genes and have a strong impact on the binding sites of transcription factors (TFs), thus affecting the development of phenotypes.

Another fundamental aspect of single-SNP-based studies (including rSNPs) is their utility to detect associations between a SNP and multiple phenotypes. Such type of associations of a single-SNP or a gene is referred to as pleiotropy [29–31], which, by definition, is a phenomenon of having a single genetic variant responsible for multiple phenotypes. Given the importance of pleiotropy, the aspect of investigating rSNPs could be quintessential for understanding the influence of variation in quantitative traits and their improvement against biotic and abiotic stresses. However, a limited number of studies have reported pleiotropic effects in maize [32–40]. The lack of such type of studies in maize can be compensated using modern multi-omics technologies, which have enabled the researchers to rapidly sequence large breeding populations and measure several phenotypes [41], facilitating pleiotropic studies. In particular, transcriptomics, proteomics, genomics, and phenomics are increasingly being used in plant sciences to gain a comprehensive understanding of complex genetic traits [20,42].

Recently, Liu et al. [43] generated a comprehensive multi-omics dataset of global maize germplasm, comprising genomic, transcriptomic and multiple phenotypic data of 368 maize inbred lines representing stiff-stock, non-stiff-stock, tropical, semi-tropical and mixed backgrounds [44] to identify genome-wide associations between individual SNPs and phenotypes [45–49].

Leveraging these multi-omics data, the main objectives of our study are to identify pleiotropic signatures of rSNPs in a systematic analysis and to construct hierarchical network models that could lead to new hypotheses to determine the crucial role of TFs controlling the development of different phenotypes in maize breeding research. Our analysis pipeline primarily consists of four distinct phases. In the first phase, rSNPs are identified, while in the second phase, the RF algorithm is used to identify relative importance of individual rSNP in phenotype associations [17,20]. Based on these association results, we identify pleiotropic rSNPs in the third phase using the Markov clustering algorithm [50], which is followed by the construction of hierarchical network models in the fourth phase that elucidate the complex interplay among TFs, rSNPs, and multiple phenotypes. Our findings demonstrate that systematic analysis of multi-omics data of global maize populations: (i) enables the identification of pleiotropic signatures of rSNPs along with their consequences on TF binding sites and; (ii) provides new insights into the genetic architecture and new breeding targets for the corresponding multiple phenotypes in maize.

2. Materials and Methods

In this section, we describe the multi-omics dataset analyzed and the methods applied in this study. Our analysis framework is structured as shown in the Figure 1. In particular, we start with the preprocessing of multi-omics data of 368 maize inbred lines and their systematic analysis towards the identification of pleiotropic signatures of rSNPs. For this purpose, we first identified the rSNPs from the genotype dataset by applying the MATCH™ algorithm [51] together with a non-redundant plant position weight matrix (PWM) library obtained from the TRANSFAC database [52]. Second, the Random forest algorithm [53]

was applied together with its specific feature selection wrapper, the Boruta algorithm [54], to assess the relative importance of each rSNP in terms of its involvement in the characterization of 20 agronomic phenotypes under study. This step was followed by the incremental feature selection (IFS) procedure [55,56] to find the optimal list of associated rSNPs for each phenotype. Next, using the Markov clustering algorithm (MCL) algorithm [50] the pleiotropic relationship signatures of rSNPs were uncovered, as suggested in [57]. Finally, we constructed hierarchical network models to elucidate the complex interplay among TFs, rSNPs, and multiple phenotypes by incorporating the corresponding transcriptome dataset to evaluate the importance of pleiotropic rSNPs. Detailed information on these analysis steps are given in Section 2.2 from Phase 1 to 4.

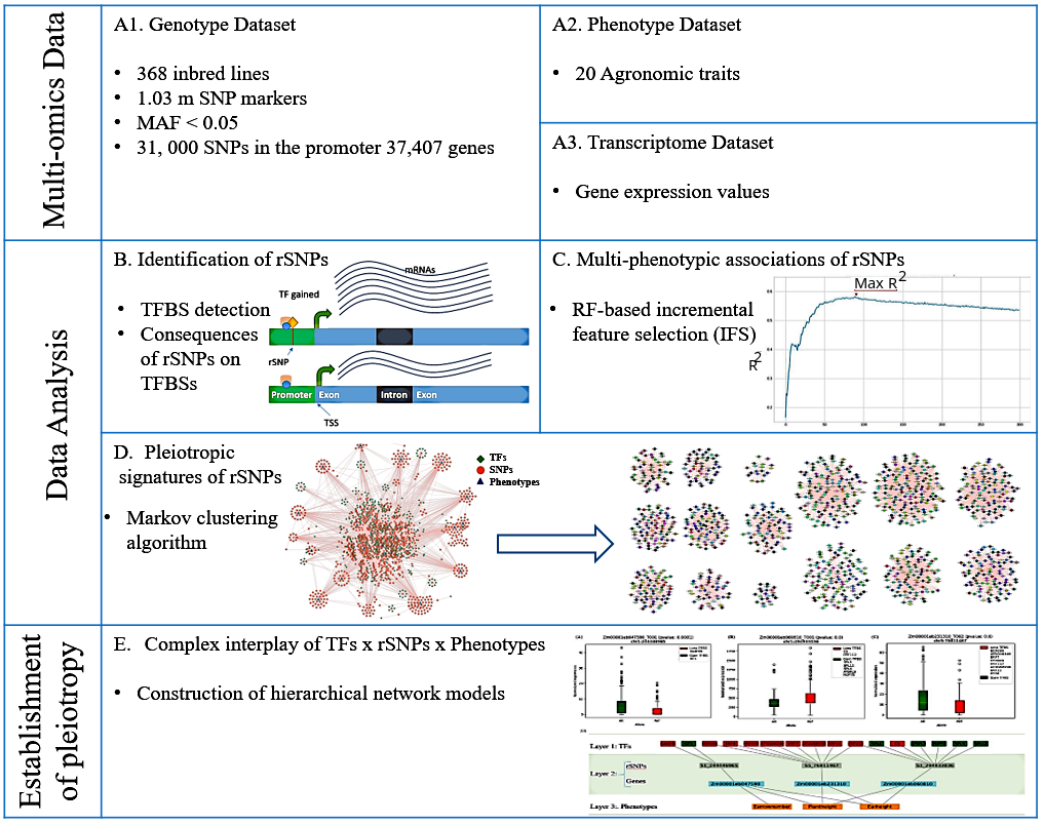

Figure 1. Overview of the analyses pipeline highlighting key machine learning algorithms for the identification of pleiotropic signatures of regulatory SNPs (rSNPs) to establish complex interplay of transcription factors (TFs), rSNPs and multiple phenotypes. The genotypic data (**A1**), consisting of 1.03 m SNP markers was filtered for MAF (<0.05) and 31,000 SNPs found within promoter regions of 37,407 maize genes were considered for association analysis with 20 quantitative agronomic traits (**A2**). RNA-seq (**A3**) dataset was utilized for the validation of pleiotropic rSNPs on the underlying gene expression. As of first step in the data analysis, rSNPs were identified (**B**) for their impact on the gain or loss of TFBSs, after which their association with multiple phenotypes was determined using random forest (RF) using the Boruta algorithm and incremental feature selection (IFS) technique (**C**). Pleiotropic signatures of rSNPs were then established by pruning weaker connections in the overall network into smaller non-overlapping fully connected clusters, using Markov clustering (MCL) algorithm (**D**) which provided the basis for the construction of hierarchical network models with three distinct layers modelling the complex interplay of TFs, rSNPs and multiple phenotype (**B**). Further, the boxplots show the impact of pleiotropic rSNPs at gene expression level as a function of gain or loss of TFBSs (**E**).

2.1. Multi-Omics Data

2.1.1. Genotype Dataset

The genotypic data of 368 inbred lines was obtained from: (i) CIMMYT; (ii) the Germplasm Enhancement of Maize (GEM) project in the USA; (iii) temperate and tropical/subtropical breeding programs in China. The inbred lines represent non-stiff stock, stiff stock, mixed, tropical and semi-tropical backgrounds. The genotyping has been performed using the MaizeSNP50 BeadChip with 56,110 SNP markers [58] which was further incremented to 1.03 million SNP markers using deep RNA-Seq data [43]. Similar to previous studies [47,59], SNPs with an MAF <0.05 are discarded, and genomic coordinates for SNP markers are lifted over from reference genome V2 to V5 using CrossMap [60]. Consequently, after filtering, the genotype dataset comprises 31,934 SNPs for 368 maize lines which are located on the chromosomes 1 to 10 including 37,407 genes.

2.1.2. Phenotype Dataset

The corresponding phenotypic data of the maize lines was collected in 2009 and 2010 in five different environments in China. The experimental units were completely randomized with a row length of 3 m, 11 plants per row, 25 cm spacing between plants, and 60 cm spacing between rows. The mean observed value of five randomly selected plants for 20 agronomic (quantitative) traits was taken, which additionally were converted into the best linear unbiased predictions (BLUPs) [40,48,49].

2.1.3. Transcriptome Dataset

Paired-end deep RNA-seq data for the 368 inbred lines, generated by Liu et al. [43], was retrieved from the European Nucleotide Archive (ENA) browser (study accession: PRJNA208608). Raw sequencing data were adapter and quality trimmed using Trim Galore [61]. High-quality reads were then mapped to the *Zea mays* L. reference genome, V5 (available at https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/Zm-B73-REFERENCE-NAM-5.0.fa.gz) (accessed on 16 April 2021), using STAR (v2.7.3a) [62]. Raw read counts for each transcript were obtained using HTSeq [63] and normalized using the median-of-ratios normalization method implemented in the R package DESeq2 [64].

2.2. Data Analysis

Our analysis framework consists of four phases to decipher the complex interplay among transcription factors (TFs), regulatory SNPs (rSNPs) and multiple phenotypes using the multi-omics dataset under study.

Phase 1: We identified rSNPs from the genotype dataset by applying our analysis pipeline introduced in [14], which consists of the following steps. First, considering the promoter region of genes (−500 bp to +100 bp relative to the transcription start site), all SNPs in these regions were selected. Second, we extracted the flanking sequence of each selected SNP, which covers ±25 bp relative to the SNP position. In total, each sequence is 51 bp long and the SNP is located in the central position. Then, two copies of the extracted sequences were constructed: while the first copy contains the reference allele at the SNP position, the second has the alternate allele. Next, by employing the MATCH™ program [51], we scanned each sequence to predict binding sites of transcription factors with their affinity scores $\in [0, 1]$. For the application of the MATCH™ program, we used a non-redundant plant position weight matrix (PWM) library obtained from the TRANSFAC database [52]. Finally, mainly focusing on the alterations of transcription factor binding sites (TFBSs) in the sequences of each SNP, we collected its potential consequence as: (i) ”Loss of TFBS”: only the sequence with the reference allele contains the TFBS of a specific transcription factor (TF), but the same TFBS is not found in the sequence with alternate allele; and (ii) “Gain of TFBS”: the TFBS can be found only in the sequence with the alternate allele; and (iii) “No Strong Effect”: indicating that the SNP consequence has either no effect or entails a slight change in binding affinity of TFs. As suggested in the

previous studies [14,20,65], we consider in our following analysis a SNP as an rSNP if it leads to a "Gain of TFBS" or a "Loss of TFBS" for at least one TF.

Phase 2: Following the association analysis strategy described in [16,17,20], we used a Random Forest (RF)-based feature selection approach to assess the relative importance of each rSNP in predicting the response variable (phenotype) of interest. To this end, we applied the Boruta algorithm [54], a powerful wrapper designed specifically for the RF-based feature selection technique, to rank the importance of variables (in this case rSNPs). Consequently, by constructing multiple decision trees based on random subsets of features, the Boruta algorithm calculates an importance score for each rSNP and thus provides a ranking.

Using the ranked rSNPs determined by the Boruta algorithm, we further performed the incremental feature selection (IFS) procedure to retrieve the optimal list of features as suggested in [55,56]. During the IFS application, the rSNPs were incrementally added from higher to lower ranks in the ordered feature set, based on which an RF classifier was constructed. The predictive performance of RF was examined based on the R^2 values. This enabled us to determine the optimal numbers of associated rSNPs for a certain phenotype of interest (see Figure 2).

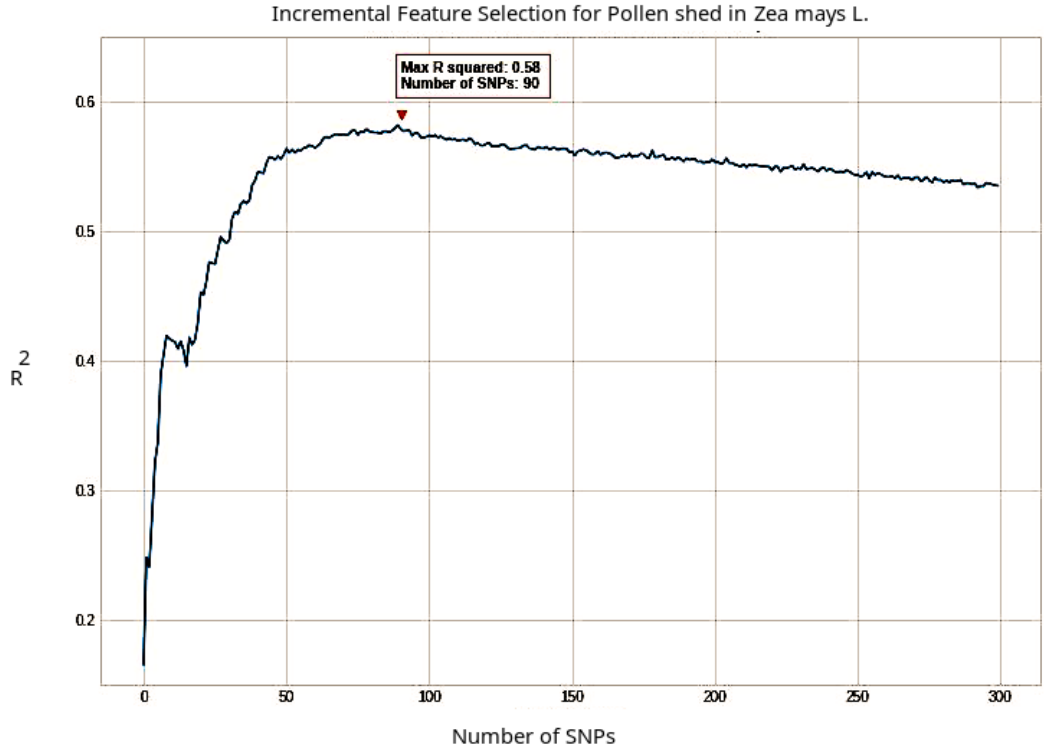

Figure 2. A plot to show the change of R^2-values versus the number of rSNPs in association with the phenotype pollen shed. The incremental feature selection (IFS) curves were drawn using the ranking of rSNPs. The R^2- value reached a peak when considering the first 90 rSNPs. These rSNPs were used for the further analysis of this phenotype.

These analyses were repeated for each of the 20 phenotypes to identify the optimal numbers of the associated rSNPs, that are are given in Table 1.

Phase 3: To reveal the unique pleiotropic relationship signatures of rSNPs and thus decipher their complex interplay with TFs and with multiple phenotypes, we applied the Markov clustering algorithm (MCL). MCL is a very effective network-based clustering algorithm that detects distinct groups in a network by eliminating negligible connections (edges) based on their weights [50].

Table 1. Phenotypes and the optimal numbers of their associated rSNPs determined by incremental feature selection (IFS) procedure.

Phenotype	Max R^2	#rSNPs
Leaf number above ear	0.490740	89
Ear leaf width	0.484029	70
Cob diameter	0.445720	64
Ear height	0.509523	109
Kernel width	0.418115	172
Ear leaf length	0.553292	112
Tassel main axis length	0.498562	96
Pollen shed	0.581765	90
Heading date	0.537987	49
Ear length	0.434011	82
Silking time	0.506520	122
Ear diameter	0.481445	110
Cob weight	0.460850	37
X100 grain weight	0.389332	51
Tassel branch number	0.507112	142
Ear row number	0.491663	46
Kernel number per row	0.350717	27
Plant height	0.532837	72
kernel length	0.580691	168
Kernel thickness	0.437589	64

For the detection of pleiotropic relationships, Weighill et al. [57] have successfully applied the MCL algorithm by constructing a profile matrix which represents the SNP × phenotype associations. Following this idea and thus the main concept of MCL, we first created such a profile matrix $\mathcal{M}$, where rows correspond to rSNPs determined by the IFS procedure and columns refer to names of both the phenotypes and TFs. The entry of $\mathcal{M}$ at position (i, j) , $\mathcal{M}_{ij}$ is defined as:

$$\mathcal{M}_{ij} = \begin{cases} 1 & \text{if rSNP}_i \text{ is associated with phenotype } j \\ 1 & \text{if the consequence of rSNP}_i \text{ is “Gain” or “Loss” for TF } j \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{M}$ was then converted into an rSNP association matrix, $\mathcal{A}_{n\times n}$ (n is the number of rSNPs (rows) in $\mathcal{M}$), using the Proportional Similarity Index [66]. The entry of $\mathcal{A}$ at a position (k, l) is calculated between the rSNPs (rows) k and l in $\mathcal{M}$ as:

$$\mathcal{A}_{kl} = 2 \cdot \frac{\sum_j \min(\mathcal{M}_{kj},\ \mathcal{M}_{lj})}{\sum_j (\mathcal{M}_{kj} + \mathcal{M}_{lj})}$$

Next, we employed MCL [50] using the matrix $\mathcal{A}$ to cluster rSNPs in subgroups based on their similar relationship signatures. Consequently, each of the resulting clusters reflects a collection of rSNPs and their complex interplay with TFs and phenotypes, based on which we designed a hierarchical network model using three layers. These layers are: (i) TFs whose binding site is lost or gained due to the rSNPs; (ii) rSNPs located in the promoter of the genes; and (iii) phenotypes whose development is strongly connected to the expression level of the corresponding genes. In a final step, we carefully removed the rSNPs from the

hierarchical network models if they were associated with only one phenotype for ensuring the pleiotropy in the clusters [57].

Phase 4: For the assessment of the consequences of pleiotropic rSNPs on TF-binding activities, which in turn affect the regulation of gene expression and thus the development of the phenotype, we evaluated their potential effects using the corresponding RNA-seq data. For this purpose, we focused only on genes with pleiotropic rSNPs in their promoters. By considering these genes, we divided the 368 maize lines into two groups for each pleiotropic rSNP: While the plants in the first group have the reference allele at the corresponding genomic position, the plants in the second group contain an alternate allele of the rSNP under study. As a result, we compared the gene expression values between those two groups using the Wilcoxon test in order to determine whether the consequence of a pleiotropic rSNP on a certain TF-binding activity ("Gain" or "Loss") leads to a significant alteration in the expression of the corresponding gene.

In our following analysis, we further pruned the hierarchical network models by removing the rSNPs and corresponding TFs whose binding activities do not result in a significant change in the related gene expression values.

3. Results and Discussion

In this study, we systematically analyzed a multi-omics dataset of 368 maize inbred lines to decipher the complex interplay among TFs, rSNPs and multiple phenotypes. For this purpose, we first identified the rSNPs from genotype dataset and then applied the Boruta algorithm followed by IFS procedure to determine the rSNPs having a strong association with the phenotypes of interest. The number of associated rSNPs along with corresponding phenotypes is given in Table 1 and Figure 3. Next, considering the multi-phenotypic associations of the rSNPs as well as their consequences on TF binding, we employed the MCL algorithm to cluster the rSNPs, which is additionally used to construct hierarchical network models to elucidate the relationship signatures between TFs and rSNPs together with those between rSNPs and multiple phenotypes, indicating their pleiotropic functions.

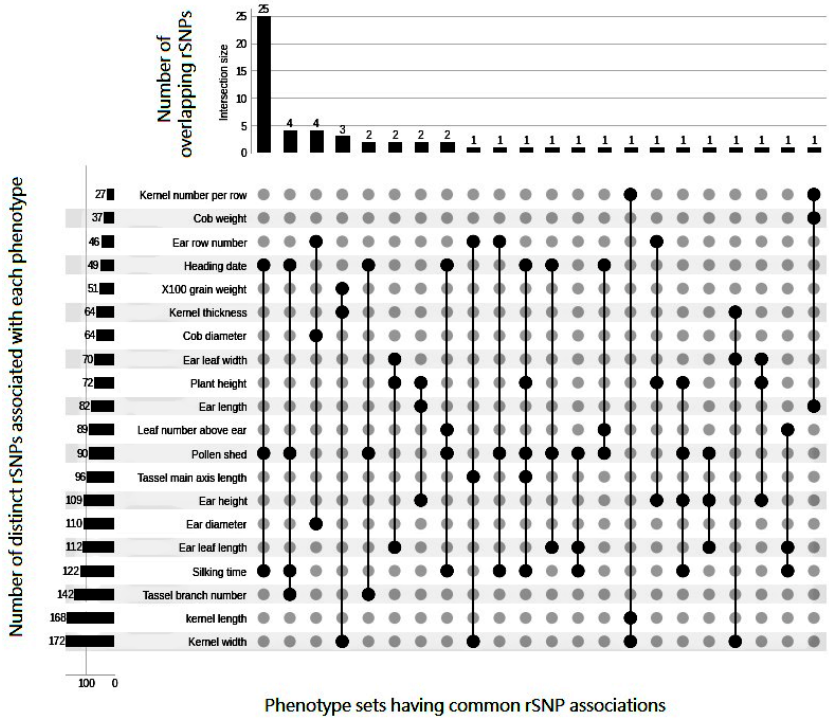

Figure 3. Number of associated rSNPs determined by the incremental feature selection (IFS) procedure for each phenotype and their overlap represented in matrix layouts using the UpSet technique [67]. Black circles in the matrix layout are related to the phenotypes that are part of the intersection. For the sake of clarity, not all intersections are displayed.

3.1. Pleiotropic Association Signatures of rSNPs

The application of the MCL algorithm to the rSNPs determined by the IFS procedure (Table 1) results in eleven clusters that reveal the unique pleiotropic relationship signatures of rSNPs arising from their complex interplay with TFs and multiple associated phenotypes. A brief description of the clusters is given in Table 2 and additional information about rSNPs and genes are provided for each cluster in Supplementary File S1.

Table 2. Result of Markov clustering algorithm (MCL) including the numbers of rSNPs together with their related genes and their associated multiple phenotypes.

Cluster	Numbers of Pleiotropic		Phenotypes
	rSNPs	Genes	
Cluster-1	15	10	Heading date, Pollen shed, Silking time and Ear height
Cluster-2	9	7	Cob weight, Heading date, Pollen shed, Tassel main axis length, Ear leaf length, Plant height, Ear leaf width, Ear row number and Ear height
Cluster-3	7	6	Kernel length, Kernel thickness, Kernel number per row, Ear diameter and X100 grain weight
Cluster-4	6	5	Ear diameter, Cob diameter and Ear row number
Cluster-5	6	4	Heading date, Pollen shed, Silking time, Ear height and Tassel branch number
Cluster-6	5	3	Ear diameter, Cob diameter and Ear row number
Cluster-7	3	3	Ear height, Plant height and Ear row number
Cluster-8	3	3	Kernel width, Kernel length, Kernel thickness and X100 grain weight
Cluster-9	2	2	Ear leaf length, Leaf number above ear, Kernel length
Cluster-10	2	2	Ear length and Kernel number per row
Cluster-11	2	2	Ear diameter, Tassel main axis length and Cob weight

As shown in Table 2, the individual clusters are quite different regarding the numbers of pleiotropic rSNPs as well as their related phenotypes. The largest cluster contains ten genes (Zm00001eb354560, Zm00001eb403000, Zm00001eb328980, Zm00001eb137870, Zm00001eb366710, Zm00001eb190550, Zm00001eb364380, Zm00001eb156700, Zm00001eb337030, Zm00001eb397560) each associated with four phenotypes. Among these genes, Zm00001eb137870, which encodes the VIN3-like protein 1, is associated with heading date, silking time and pollen shed. Its homologue in arabdopsis, AT3G24440, is known as the Vernalization Insensitive 3-like 1 (VIL1) gene that regulates expression of the Flowering Locus C (FLC:floral repressors) and the Flowering Locus M (FLM) in response to vernalization. VIL1 and VIN3 (Vernalization Insensitive 3) are essential for epigenetic modification of the FLC and FLM loci [68,69]. The VIL1 gene acts upstream of many biological processes including administration and regulation of histone methylation, vernalization response and regulation of flower development. Additionally, it is a negative regulator of gene expression, which is achieved by its role in the positive regulation of histone H3-K27 methylation [70]. VIL1 is also involved in photoperiodism, flowering, vernalization response and response to cold [71]. Other genes in this cluster are known for their involvement in protein metabolism (Zm00001eb354560, Zm00001eb403000, Zm00001eb364380) and fat metabolism (Zm00001eb328980), however, the functionality of the remaining genes in this cluster is currently not known.

Another interesting instance of pleiotropy in action can be seen in Cluster-2, where most of the phenotypes are clustered together, representing a highly complex and coordinated developmental program of the disparate phenotypes in this cluster. The cluster contains genes belonging to protein metabolism (Zm00001eb131510, Zm00001eb288200), carbohydrate and fat metabolism (Zm00001eb372200, Zm00001eb418690), along with flowering

time genes (Zm00001eb188340, Zm00001eb418690). Among these genes, Zm00001eb131510 encodes an intracellular protein transporter and is associated with ear leaf length and cob weight. This protein transporter is known for its role in exocytosis, golgi to plasma membrane transport and intracellular protein transport [70]. Its arabidopsis homologue, AT4G02350, is vital for pollen tube growth, pollen germination and acceptance in arabidopsis [72]. Such a gene may play an important role in the source-sink relationship between the developing maize kernels and the flag leaf for translocation of proteins. The arabidopsis homologue, AT1G66430, of the oil content related gene, Zm00001eb372200 is associated with ear row number and ear leaf height in this cluster and is also known for its role in carbohydrate biosynthetic, fatty acid biosynthetic, and fructose metabolic processes [73]. Several genes in this cluster indicates the co-regulation of floral transition and seed development in maize.

A closer look at Cluster-4 reveals that it contains six pleiotropic rSNPs found within five genes (Zm00001eb068110, Zm00001eb140090, Zm00001eb424640, Zm00001eb049750, Zm00001eb057150), associated with only ear traits (ear diameter, cob diameter and ear row number). The genes in this cluster are involved in riboflavin (Zm00001eb068110) [74] and the fatty acid metabolism (Zm00001eb049750) [47], protein relocalization to mitochondrion (Zm00001eb140090) [75] as well as acy carrier activity (Zm00001eb049750) [76]. The gene Zm00001eb049750 encoding the acyl carrier protein is known for its association with maize kernel oil contents [47], while the fax1 (AT3G57280) mutants, a homologue of Zm00001eb424640, are characterized by a decrease in biomass, plant height, stem thickness, reduced male sterility and defective pollen cell wall biosynthesis [77].

Among the smallest clusters are the Clusters 9-11, containing two pleiotropic rSNPs and two corresponding genes each. Additionally there are within the Cluster-10 another two kernel and ear development genes (Zm00001eb213840, Zm00001eb349520). The arabidopsis homologue (AT4G07960.1) of Zm00001eb213840 is known to encode XyG glucan synthase. Arabidopsis mutants of this gene have smaller rosettes and inflorescence stems, weak inflorescence stems and a reduced number of pollen tubes after pollination [78]. Moreover, the product of Zm00001eb349520 is known as a aspartic acid proteinase inhibitor. Today it is well known that high levels of aspartic acids are observed in maize cobs during early reproductive development [79]. Aspartic acids accumulate in kernels as N-assimilates, suggesting its role in kernel growth.

3.2. Construction of Hierarchical Network Models

To further elaborate the complex interplay between pleiotropic rSNPs and TFs by assessing their potential impact on the expression level of the respective genes, we constructed a hierarchical network model for each cluster found by the MCL algorithm. These network models help explain the potential biological functions of TFs in regulating gene expression and, hence, assess the importance of individual pleiotropic rSNPs. They also provide with new hypotheses to advance our knowledge of why the consideration of the TFs could play a crucial role in maize breeding research to understand the genetic mechanisms underlying the development of different phenotypes. An example of our hierarchical network model is presented in Figure 4, showing the complex regulatory circuitry of Cluster-7 phenotypes. The phenotype set in this cluster represented by layer 3 (Figure 4D) comprises ear height, plant height and ear row number, whereas the layer 2 lists three associated rSNPs, three corresponding genes and their pleiotropic associations. Finally, in the first layer the TFs and the change in their binding activity is highlighted, showing loss of binding for ten TFs (in red) and gain of six (in green) TFs. In this cluster, a pleiotropic rSNP in Zm00001eb047590 (opf6 gene) results in the loss of a TFBS for GAMYB but in the gain for a GT1 binding site. The transcription factor GAMYB in arabidopsis is known for its role in Gibberelic acid (GA) signalling regarding floral transition [80,81], whereas the factor GT1 is vital to the regulation of stress tolerance in rice as well as response to light [82,83]. GT1 binds upstream of the light-responsive, rbcS-3A (RUBISCO) gene in peas, hence plays a crucial role in the regulation of photosynthesis [84]. The functional

analysis of GT1 in arabidopsis shows its regulation in the target promoters that may have a repressive function in transcription activity [85]. Our findings indicate that the replacement of GAMYB with GT1 at the opf6 promoter (a transcription repressor gene) results in its significantly higher levels of gene expression (Figure 4A). This variation in gene expression may have an impact on the regulation of the GA pathway. For their role in flower development and flowering time [86], GA dynamic of the cells may contribute to the development of associated phenotypes (plant height, ear height and ear row number). Another pleiotropic rSNP in the bzip41 (Zm00001eb060810) promoter in this cluster results in loss of TFBSs for several SQUAMOSA-promoter binding protein-like TFs (SPL family TFs), whereas a TFBS is gained for C1 TF, resulting in significant change in gene expression (Figure 4B). C1 is known for regulating pigmentation in the aleurone layer of the maize kernels [87], whereas ERF112 is a potential negative regulator of the JA-responsive gene expression [88]. SPL proteins are a plant specific family of TFs and are known as potential candidates for the genetic improvement of agronomic traits due to their role in the physiological and reproductive development of plants [89]. For example, SPL4 is required for developmental transition and plays an important role in the determination of flowering time [90,91], whereas SPL12 is known to be expressed during plant development [92] and is responsive to abscisic acid biosynthesis in heat stress [93]. The association of this rSNP and recruitment of respective TFs is in line with the development of associated phenotypes. The pleiotropic rSNP (chr5:76811467) results in the loss of TFBSs for several ethylene responsive transcription factors (ERFs), which significantly increases the gene expression (Figure 4C). ERF TFs in this cluster mainly act as inhibitor of transcription [88,94], hence loss of their TFBS is translated as higher promoter activity on this gene, which may further contribute to the development of the associated phenotypes. Our results also show the importance of hierarchical network models in explaining the impact of pleiotropic rSNP on the expression of corresponding genes, which may directly influence all the associated phenotypes. TF−rSNPs(gene)−Phenotype network models for all other clusters are given in the Supplementary File S2.

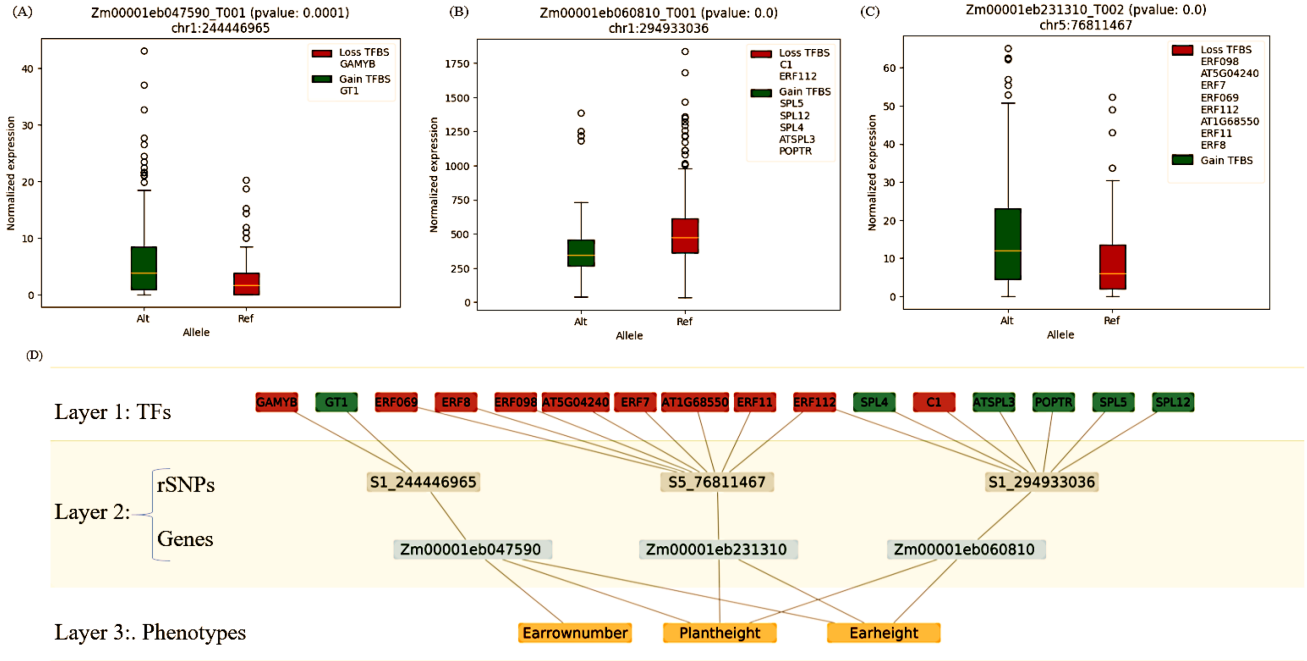

Figure 4. Hierarchical network model constructed using Cluster-7 to elucidate the complex interplay among TFs−rSNPs(genes)−Phenotypes. (**A**–**C**) show the significant changes in the gene expression values resulting from the consequences of pleiotropic rSNPs. (**D**) Hierarchical network model with three layers.

4. Conclusions

Pleiotropy and rSNPs are the two main concepts in the field of genetics by providing novel targets for the acceleration of plant breeding strategies. Therefore we considered in this study both of these concepts together and established hierarchical network models for elucidating the complex interplay among TFs−rSNPs−Phenotypes using multi-omics data. Our results show that most of the identified TFs and genes play essential roles for the development of multiple phenotypes. Our findings further suggest common genetic mechanisms underlying several interrelated phenotypes found in Clusters-1 or -8 as well as disparate phenotypes, like plant height and kernel traits, found as in Clusters-2 or -5. To the best of our knowledge, by mainly focusing on the important role of rSNPs and their consequences on TFs, this is the first study which provides the pleiotropic relations of several agronomically important phenotypes of maize. The outcomes of our analysis could be highly relevant for the understanding of the genetic programs governing the development of multiple phenotypes. Therefore, further molecular biology progress is needed not only to assess the potential role of these rSNP candidates, but also to gain a deeper insight into the genetic mechanisms underlying biological processes in maize.

5. Future Directions

Unraveling the genetic architecture of complex traits is a key component for improving plants against biotic and biotic stresses. In this context, plant breeding strategies focus on developing QTL maps for different types of stresses, accounting for linkage disequilibrium and high false positive rate, standardizing genome-wide polygenetic scores [95,96], and incorporating epistatic effects into genome-wide association studies to explain missing heritability and pleiotropy using traditional marker-assisted selection (MAS) and genomic prediction (GS) strategies as well as ML approaches [97–99]. As the desired and undesired traits could share a pleiotropic relationship, consideration of pleiotropy is essential to direct breeding programs. Additionally consideration of pleiotropic signatures of rSNPs in the analysis of large genomic datasets with regarding MAS and GS, polygenetic scores, and polygenetic biotic interactions can impact the outcome of the analysis. We suggest that the incorporation of pleiotropic effects of rSNPs in the analysis of genomic data could improve outcome of MAS as well as GS studies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23095121/s1.

Author Contributions: M.G. designed and supervised the research. A.H. together with M.G. participated in the design of the study. Further, A.H. conducted computational analyses, prepared the data sets, implemented the framework and performed the literature survey. S.K. involved in the rSNP analysis and interpreted the results with A.O.S., A.H. and M.G.; A.H. and M.G. wrote the final version of the manuscript. M.G. conceived of and managed the project. All authors read and approved the final manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University. We would like to thank our collegues Abirami Rajavel and Felix Heinrich for providing helpful advice and discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shiferaw, B.; Prasanna, B.M.; Hellin, J.; Bänziger, M. Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Secur.* **2011**, *3*, 307–327. [CrossRef]
2. Prasanna, B.M.; Palacios-Rojas, N.; Hossain, F.; Muthusamy, V.; Menkir, A.; Dhliwayo, T.; Ndhlela, T.; San Vicente, F.; Nair, S.K.; Vivek, B.S.; et al. Molecular breeding for nutritionally enriched maize: Status and prospects. *Front. Genet.* **2020**, *10*, 1392. [CrossRef] [PubMed]
3. Ortiz-Monasterio, J.I.; Palacios-Rojas, N.; Meng, E.; Pixley, K.; Trethowan, R.; Pena, R. Enhancing the mineral and vitamin content of wheat and maize through plant breeding. *J. Cereal Sci.* **2007**, *46*, 293–307. [CrossRef]
4. Bänziger, M.; Betrán, F.; Lafitte, H. Efficiency of high-nitrogen selection environments for improving maize for low-nitrogen target environments. *Crop. Sci.* **1997**, *37*, 1103–1109. [CrossRef]
5. Suwarno, W.B.; Pixley, K.V.; Palacios-Rojas, N.; Kaeppler, S.M.; Babu, R. Genome-wide association analysis reveals new targets for carotenoid biofortification in maize. *Theor. Appl. Genet.* **2015**, *128*, 851–864. [CrossRef] [PubMed]
6. Wu, J.; Lawit, S.J.; Weers, B.; Sun, J.; Mongar, N.; Van Hemert, J.; Melo, R.; Meng, X.; Rupe, M.; Clapp, J.; et al. Overexpression of zmm28 increases maize grain yield in the field. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 23850–23858. [CrossRef]
7. Boćanski, J.; Srećkov, Z.; Nastasić, A. Genetic and phenotypic relationship between grain yield and components of grain yield of maize (*Zea mays* L.). *Genetika* **2009**, *41*, 145–154. [CrossRef]
8. Veldboom, L.R.; Lee, M. Genetic mapping of quantitative trait loci in maize in stress and nonstress environments: I. Grain yield and yield components. *Crop. Sci.* **1996**, *36*, 1310–1319. [CrossRef]
9. Betran, F.; Beck, D.; Bänziger, M.; Edmeades, G. Genetic analysis of inbred and hybrid grain yield under stress and nonstress environments in tropical maize. *Crop. Sci.* **2003**, *43*, 807–817. [CrossRef]
10. Dhugga, K.S. Maize biomass yield and composition for biofuels. *Crop. Sci.* **2007**, *47*, 2211–2227. [CrossRef]
11. Fernandez, M.G.S.; Becraft, P.W.; Yin, Y.; Lübberstedt, T. From dwarves to giants? Plant height manipulation for biomass yield. *Trends Plant Sci.* **2009**, *14*, 454–461. [CrossRef]
12. Xue, J.; Gao, S.; Fan, Y.; Li, L.; Ming, B.; Wang, K.; Xie, R.; Hou, P.; Li, S. Traits of plant morphology, stalk mechanical strength, and biomass accumulation in the selection of lodging-resistant maize cultivars. *Eur. J. Agron.* **2020**, *117*, 126073. [CrossRef]
13. Mazaheri, M.; Heckwolf, M.; Vaillancourt, B.; Gage, J.L.; Burdo, B.; Heckwolf, S.; Barry, K.; Lipzen, A.; Ribeiro, C.B.; Kono, T.J.; et al. Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biol.* **2019**, *19*, 45. [CrossRef] [PubMed]
14. Heinrich, F.; Wutke, M.; Das, P.P.; Kamp, M.; Gültas, M.; Link, W.; Schmitt, A.O. Identification of regulatory SNPs associated with vicine and convicine content of Vicia faba based on genotyping by sequencing data using deep learning. *Genes* **2020**, *11*, 614. [CrossRef]
15. Pearson, T.A.; Manolio, T.A. How to interpret a genome-wide association study. *JAMA* **2008**, *299*, 1335–1344. [CrossRef]
16. Ramzan, F.; Gültas, M.; Bertram, H.; Cavero, D.; Schmitt, A.O. Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations. *Genes* **2020**, *11*, 892. [CrossRef] [PubMed]
17. Ramzan, F.; Klees, S.; Schmitt, A.O.; Cavero, D.; Gültas, M. Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forests. *Genes* **2020**, *11*, 464. [CrossRef]
18. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [CrossRef]
19. Patron, J.; Serra-Cayuela, A.; Han, B.; Li, C.; Wishart, D.S. Assessing the performance of genome-wide association studies for predicting disease risk. *PLoS ONE* **2019**, *14*, e0220215. [CrossRef]
20. Klees, S.; Lange, T.M.; Bertram, H.; Rajavel, A.; Schlüter, J.S.; Lu, K.; Schmitt, A.O.; Gültas, M. In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in Brassica napus L. Using Multi-Omics Data. *Int. J. Mol. Sci.* **2021**, *22*, 789. [CrossRef]
21. Liu, Y.; Wang, D.; He, F.; Wang, J.; Joshi, T.; Xu, D. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet.* **2019**, *10*, 1091. [CrossRef] [PubMed]
22. Nguyen, T.T.; Huang, J.Z.; Wu, Q.; Nguyen, T.T.; Li, M.J. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genom.* **2015**, *16*, S5. [CrossRef] [PubMed]
23. Zhao, Y.; Chen, F.; Zhai, R.; Lin, X.; Wang, Z.; Su, L.; Christiani, D.C. Correction for population stratification in random forest analysis. *Int. J. Epidemiol.* **2012**, *41*, 1798–1806. [CrossRef]
24. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef]
25. Schrider, D.R.; Kern, A.D. Supervised machine learning for population genetics: A new paradigm. *Trends Genet.* **2018**, *34*, 301–312. [CrossRef] [PubMed]
26. Cortés, A.J.; López-Hernández, F.; Osorio-Rodriguez, D. Predicting thermal adaptation by looking into populations' genomic past. *Front. Genet.* **2020**, *11*, 1093. [CrossRef]
27. Jansen, S.; Baulain, U.; Habig, C.; Ramzan, F.; Schauer, J.; Schmitt, A.O.; Scholz, A.M.; Sharifi, A.R.; Weigend, A.; Weigend, S. Identification and Functional Annotation of Genes Related to Bone Stability in Laying Hens Using Random Forests. *Genes* **2021**, *12*, 702. [CrossRef]

28. Brieuc, M.S.; Waters, C.D.; Drinan, D.P.; Naish, K.A. A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Mol. Ecol. Resour.* **2018**, *18*, 755–766. [CrossRef]
29. Pendergrass, S.A.; Brown-Gentry, K.; Dudek, S.; Frase, A.; Torstenson, E.S.; Goodloe, R.; Ambite, J.L.; Avery, C.L.; Buyske, S.; Bžková, P.; et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* **2013**, *9*, e1003087. [CrossRef]
30. Pendergrass, S.; Brown-Gentry, K.; Dudek, S.; Torstenson, E.; Ambite, J.; Avery, C.; Buyske, S.; Cai, C.; Fesinmeyer, M.; Haiman, C.; et al. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* **2011**, *35*, 410–422. [CrossRef]
31. Solovieff, N.; Cotsapas, C.; Lee, P.H.; Purcell, S.M.; Smoller, J.W. Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genets.* **2013**, *14*, 483–495. [CrossRef]
32. Mayfield, S.; Nelson, T.; Taylor, W.; Malkin, R. Carotenoid synthesis and pleiotropic effects in carotenoid-deficient seedlings of maize. *Planta* **1986**, *169*, 23–32. [CrossRef]
33. Pilu, R.; Landoni, M.; Cassani, E.; Doria, E.; Nielsen, E. The maize lpa241 mutation causes a remarkable variability of expression and some pleiotropic effects. *Crop. Sci.* **2005**, *45*, 2096–2105. [CrossRef]
34. Wen, L.; Chase, C.D. Pleiotropic effects of a nuclear restorer-of-fertility locus on mitochondrial transcripts in male-fertile and S male-sterile maize. *Curr. Genet.* **1999**, *35*, 521–526. [CrossRef]
35. Bomblies, K.; Doebley, J.F. Pleiotropic effects of the duplicate maize FLORICAULA/LEAFY genes zfl1 and zfl2 on traits under selection during maize domestication. *Genetics* **2006**, *172*, 519–531. [CrossRef]
36. Asakura, Y.; Hirohashi, T.; Kikuchi, S.; Belcher, S.; Osborne, E.; Yano, S.; Terashima, I.; Barkan, A.; Nakai, M. Maize mutants lacking chloroplast FtsY exhibit pleiotropic defects in the biogenesis of thylakoid membranes. *Plant Cell* **2004**, *16*, 201–214. [CrossRef]
37. Chourey, P.S.; Li, Q.B.; Cevallos-Cevallos, J. Pleiotropy and its dissection through a metabolic gene Miniature1 (Mn1) that encodes a cell wall invertase in developing seeds of maize. *Plant Sci.* **2012**, *184*, 45–53. [CrossRef]
38. Clark, R.M.; Wagler, T.N.; Quijada, P.; Doebley, J. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet.* **2006**, *38*, 594–597. [CrossRef]
39. Wisser, R.J.; Kolkman, J.M.; Patzoldt, M.E.; Holland, J.B.; Yu, J.; Krakowsky, M.; Nelson, R.J.; Balint-Kurti, P.J. Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7339–7344. [CrossRef]
40. Brown, P.J.; Upadyayula, N.; Mahone, G.S.; Tian, F.; Bradbury, P.J.; Myles, S.; Holland, J.B.; Flint-Garcia, S.; McMullen, M.D.; Buckler, E.S.; et al. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet.* **2011**, *7*, e1002383. [CrossRef]
41. Houle, D.; Govindaraju, D.R.; Omholt, S. Phenomics: The next challenge. *Nat. Rev. Genet.* **2010**, *11*, 855–866. [CrossRef]
42. Rajavel, A.; Klees, S.; Schlüter, J.S.; Bertram, H.; Lu, K.; Schmitt, A.O.; Gültas, M. Unravelling the Complex Interplay of Transcription Factors Orchestrating Seed Oil Content in *Brassica napus* L. *Int. J. Mol. Sci.* **2021**, *22*, 1033. [CrossRef]
43. Liu, H.; Wang, F.; Xiao, Y.; Tian, Z.; Wen, W.; Zhang, X.; Chen, X.; Liu, N.; Li, W.; Liu, L.; et al. MODEM: Multi-omics data envelopment and mining in maize. *Database* **2016**, *2016*, baw117. [CrossRef]
44. Yang, X.; Gao, S.; Xu, S.; Zhang, Z.; Prasanna, B.M.; Li, L.; Li, J.; Yan, J. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol. Breed.* **2011**, *28*, 511–526. [CrossRef]
45. Wen, W.; Araus, J.L.; Shah, T.; Cairns, J.; Mahuku, G.; Bänziger, M.; Torres, J.L.; Sánchez, C.; Yan, J. Molecular characterization of a diverse maize inbred line collection and its potential utilization for stress tolerance improvement. *Crop. Sci.* **2011**, *51*, 2569–2581. [CrossRef]
46. Fu, J.; Cheng, Y.; Linghu, J.; Yang, X.; Kang, L.; Zhang, Z.; Zhang, J.; He, C.; Du, X.; Peng, Z.; et al. RNA sequencing reveals the complex regulatory network in the maize kernel. *Nat. Commun.* **2013**, *4*, 1–12. [CrossRef]
47. Li, H.; Peng, Z.; Yang, X.; Wang, W.; Fu, J.; Wang, J.; Han, Y.; Chai, Y.; Guo, T.; Yang, N.; et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* **2013**, *45*, 43–50. [CrossRef]
48. Wen, W.; Li, D.; Li, X.; Gao, Y.; Li, W.; Li, H.; Liu, J.; Liu, H.; Chen, W.; Luo, J.; et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. Commun.* **2014**, *5*, 1–10. [CrossRef]
49. Yang, N.; Lu, Y.; Yang, X.; Huang, J.; Zhou, Y.; Ali, F.; Wen, W.; Liu, J.; Li, J.; Yan, J. Genome Wide Association Studies Using a New Nonparametric Model Reveal the Genetic Architecture of 17 Agronomic Traits in an Enlarged Maize Association Panel. *PLoS Genet.* **2014**, *10*, e1004573. [CrossRef]
50. Van Dongen, S. Graph Clustering by Flow Simulation. Ph.D. Thesis, University of Utrecht, Utrecht, The Netherlands, 2000.
51. Kel, A.E.; Gössling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O.V.; Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**, *31*, 3576–3579. [CrossRef]
52. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* **2008**, *9*, 326–332. [CrossRef] [PubMed]
53. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
54. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
55. Li, B.Q.; Hu, L.L.; Chen, L.; Feng, K.Y.; Cai, Y.D.; Chou, K.C. Prediction of Protein Domain with mRMR Feature Selection and Analysis. *PLoS ONE* **2012**, *7*, e39308. [CrossRef]

56. Li, B.Q.; Feng, K.Y.; Chen, L.; Huang, T.; Cai, Y.D. Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. *PLoS ONE* **2012**, *7*, e43927. [CrossRef]
57. Weighill, D.; Jones, P.; Bleker, C.; Ranjan, P.; Shah, M.; Zhao, N.; Martin, M.; DiFazio, S.; Macaya-Sanz, D.; Schmutz, J.; et al. Multi-phenotype association decomposition: Unraveling complex gene-phenotype relationships. *Front. Genet.* **2019**, *10*, 417. [CrossRef]
58. Ganal, M.W.; Durstewitz, G.; Polley, A.; Bérard, A.; Buckler, E.S.; Charcosset, A.; Clarke, J.D.; Graner, E.M.; Hansen, M.; Joets, J.; et al. A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* **2011**, *6*, e28334. [CrossRef]
59. Xu, J.; Chen, G.; Hermanson, P.J.; Xu, Q.; Sun, C.; Chen, W.; Kan, Q.; Li, M.; Crisp, P.A.; Yan, J.; et al. Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize. *Genome Biol.* **2019**, *20*, 1–16. [CrossRef]
60. Zhao, H.; Sun, Z.; Wang, J.; Huang, H.; Kocher, J.P.; Wang, L. CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **2014**, *30*, 1006–1007. [CrossRef]
61. Sun, K. Ktrim: An extra-fast and accurate adapter-and quality-trimmer for sequencing data. *Bioinformatics* **2020**, *36*, 3561–3562. [CrossRef]
62. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef] [PubMed]
63. Putri, G.H.; Anders, S.; Pyl, P.T.; Pimanda, J.E.; Zanini, F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *arXiv* **2021**, arXiv:2112.00939 .
64. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 1–21. [CrossRef] [PubMed]
65. Klees, S.; Heinrich, F.; Schmitt, A.O.; Gültas, M. agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species. *Biology* **2021**, *10*, 790. [CrossRef] [PubMed]
66. Bloom, S.A. Similarity indices in community studies: Potential pitfalls. *Mar. Ecol. Prog. Ser.* **1981**, *5*, 125–128. [CrossRef]
67. Conway, J.R.; Lex, A.; Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **2017**, *33*, 2938–2940. [CrossRef] [PubMed]
68. De Lucia, F.; Crevillen, P.; Jones, A.M.; Greb, T.; Dean, C. A PHD-polycomb repressive complex 2 triggers the epigenetic silencing of FLC during vernalization. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 16831–16836. [CrossRef]
69. Mylne, J.; Greb, T.; Lister, C.; Dean, C. Epigenetic regulation in the control of flowering. In *Proceedings of the Cold Spring Harbor Symposia on Quantitative Biology*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2004; Volume 69, pp. 457–464.
70. Berardini, T.Z.; Reiser, L.; Li, D.; Mezheritsky, Y.; Muller, R.; Strait, E.; Huala, E. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* **2015**, *53*, 474–485. [CrossRef]
71. Kim, D.H.; Sung, S. Role of VIN3-LIKE 2 in facultative photoperiodic flowering response in Arabidopsis. *Plant Signal. Behav.* **2010**, *5*, 1672–1673. [CrossRef]
72. Qi, H.; Jiang, Z.; Zhang, K.; Yang, S.; He, F.; Zhang, Z. PlaD: A transcriptomics database for plant defense responses to pathogens, providing new insights into plant immune system. *Genom. Proteom. Bioinform.* **2018**, *16*, 283–293. [CrossRef]
73. Stein, O.; Avin-Wittenberg, T.; Krahnert, I.; Zemach, H.; Bogol, V.; Daron, O.; Aloni, R.; Fernie, A.R.; Granot, D. Corrigendum: Arabidopsis fructokinases are important for seed oil accumulation and vascular development. *Front. Plant Sci.* **2017**, *8*, 303. [CrossRef] [PubMed]
74. Jiao, Y.; Peluso, P.; Shi, J.; Liang, T.; Stitzer, M.C.; Wang, B.; Campbell, M.S.; Stein, J.C.; Wei, X.; Chin, C.S.; et al. Improved maize reference genome with single-molecule technologies. *Nature* **2017**, *546*, 524–527. [CrossRef] [PubMed]
75. Baudisch, B.; Klösgen, R.B. Dual targeting of a processing peptidase into both endosymbiotic organelles mediated by a transport signal of unusual architecture. *Mol. Plant* **2012**, *5*, 494–503. [CrossRef] [PubMed]
76. Fu, X.; Guan, X.; Garlock, R.; Nikolau, B.J. Mitochondrial Fatty Acid Synthase Utilizes Multiple Acyl Carrier Protein Isoforms1[OPEN]. *Plant Physiol.* **2020**, *183*, 547–557. [CrossRef]
77. Li, N.; Gügel, I.L.; Giavalisco, P.; Zeisler, V.; Schreiber, L.; Soll, J.; Philippar, K. FAX1, a novel membrane protein mediating plastid fatty acid export. *PLoS Biol.* **2015**, *13*, e1002053. [CrossRef]
78. Kim, S.J.; Chandrasekar, B.; Rea, A.C.; Danhof, L.; Zemelis-Durfee, S.; Thrower, N.; Shepard, Z.S.; Pauly, M.; Brandizzi, F.; Keegstra, K. The synthesis of xyloglucan, an abundant plant cell wall polysaccharide, requires CSLC function. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 20316–20324. [CrossRef]
79. Seebauer, J.R.; Moose, S.P.; Fabbri, B.J.; Crossland, L.D.; Below, F.E. Amino acid metabolism in maize earshoots. Implications for assimilate preconditioning and nitrogen signaling. *Plant Physiol.* **2004**, *136*, 4326–4334. [CrossRef]
80. Gocal, G.F.; Sheldon, C.C.; Gubler, F.; Moritz, T.; Bagnall, D.J.; MacMillan, C.P.; Li, S.F.; Parish, R.W.; Dennis, E.S.; Weigel, D.; et al. GAMYB-like genes, flowering, and gibberellin signaling in Arabidopsis. *Plant Physiol.* **2001**, *127*, 1682–1693. [CrossRef]
81. Woodger, F.J.; Millar, A.; Murray, F.; Jacobsen, J.V.; Gubler, F. The role of GAMYB transcription factors in GA-regulated gene expression. *J. Plant Growth Regul.* **2003**, *22*, 176–184. [CrossRef]
82. Fang, Y.; Xie, K.; Hou, X.; Hu, H.; Xiong, L. Systematic analysis of GT factor family of rice reveals a novel subfamily involved in stress responses. *Mol. Genet. Genom.* **2010**, *283*, 157–169. [CrossRef]

83. Hiratsuka, K.; Wu, X.; Fukuzawa, H.; Chua, N.H. Molecular dissection of GT-1 from Arabidopsis. *Plant Cell* **1994**, *6*, 1805–1813. [PubMed]
84. Green, P.J.; Yong, M.H.; Cuozzo, M.; Kano-Murakami, Y.; Silverstein, P.; Chua, N. Binding site requirements for pea nuclear protein factor GT-1 correlate with sequences required for light-dependent transcriptional activation of the rbcS-3A gene. *EMBO J.* **1988**, *7*, 4035–4044. [CrossRef] [PubMed]
85. Le Gourrierec, J.; Delaporte, V.; Ayadi, M.; Li, Y.F.; Zhou, D.X. Functional analysis of Arabidopsis transcription factor GT-1 in the expression of light-regulated genes. *Genome Lett.* **2002**, *1*, 77–82. [CrossRef]
86. Cheng, H.; Qin, L.; Lee, S.; Fu, X.; Richards, D.E.; Cao, D.; Luo, D.; Harberd, N.P.; Peng, J. Gibberellin regulates Arabidopsis floral development via suppression of DELLA protein function. *Development* **2004**, *131*, 1055–1064. [CrossRef]
87. Cone, K.C.; Cocciolone, S.M.; Burr, F.A.; Burr, B. Maize anthocyanin regulatory gene pl is a duplicate of c1 that functions in the plant. *Plant Cell* **1993**, *5*, 1795–1805.
88. Caarls, L.; Van der Does, D.; Hickman, R.; Jansen, W.; Verk, M.C.V.; Proietti, S.; Lorenzo, O.; Solano, R.; Pieterse, C.M.; Van Wees, S. Assessing the role of ETHYLENE RESPONSE FACTOR transcriptional repressors in salicylic acid-mediated suppression of jasmonic acid-responsive genes. *Plant Cell Physiol.* **2017**, *58*, 266–278. [CrossRef]
89. Yu, N.; Yang, J.C.; Yin, G.T.; Li, R.S.; Zou, W.T. Genome-wide characterization of the SPL gene family involved in the age development of Jatropha curcas. *BMC Genom.* **2020**, *21*, 68. [CrossRef]
90. Jung, J.H.; Seo, P.J.; Kang, S.K.; Park, C.M. miR172 signals are incorporated into the miR156 signaling pathway at the SPL3/4/5 genes in Arabidopsis developmental transitions. *Plant Mol. Biol.* **2011**, *76*, 35–45. [CrossRef]
91. Jung, J.H.; Lee, H.J.; Ryu, J.Y.; Park, C.M. SPL3/4/5 integrate developmental aging and photoperiodic signals into the FT-FD module in Arabidopsis flowering. *Mol. Plant* **2016**, *9*, 1647–1659. [CrossRef]
92. Cardon, G.; Höhmann, S.; Klein, J.; Nettesheim, K.; Saedler, H.; Huijser, P. Molecular characterisation of the Arabidopsis SBP-box genes. *Gene* **1999**, *237*, 91–104. [CrossRef]
93. Chao, L.M.; Liu, Y.Q.; Chen, D.Y.; Xue, X.Y.; Mao, Y.B.; Chen, X.Y. Arabidopsis transcription factors SPL1 and SPL12 confer plant thermotolerance at reproductive stage. *Mol. Plant* **2017**, *10*, 735–748. [CrossRef] [PubMed]
94. Ohta, M.; Matsui, K.; Hiratsu, K.; Shinshi, H.; Ohme-Takagi, M. Repression domains of class II ERF transcriptional repressors share an essential motif for active repression. *Plant Cell* **2001**, *13*, 1959–1968. [CrossRef] [PubMed]
95. Cortés, A.J.; López-Hernández, F. Harnessing crop wild diversity for climate change adaptation. *Genes* **2021**, *12*, 783. [CrossRef] [PubMed]
96. Guevara-Escudero, M.; Osorio, A.N.; Cortés, A.J. Integrative pre-breeding for biotic resistance in forest trees. *Plants* **2021**, *10*, 2022. [CrossRef]
97. Ma, C.; Zhang, H.H.; Wang, X. Machine learning for big data analytics in plants. *Trends Plant Sci.* **2014**, *19*, 798–808. [CrossRef]
98. Cortés, A.J.; Restrepo-Montoya, M.; Bedoya-Canas, L.E. Modern strategies to assess and breed forest tree adaptation to changing climate. *Front. Plant Sci.* **2020**, *11*, 1606. [CrossRef]
99. Tong, H.; Nikoloski, Z. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J. Plant Physiol.* **2021**, *257*, 153354. [CrossRef]