LEANDRO TONELLO ZUFFO

COMBINING GENOTYPING-BY-SEQUENCING DATA SETS FOR GWAS AND GWS STUDY FOR MAIZE ROOT SYSTEM ARCHITECTURE TRAITS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Fitotecnia, para obtenção do título de *Doctor Scientiae*.

Orientador: Rodrigo Oliveira de Lima.

VIÇOSA – MINAS GERAIS
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
Z94c
2019

Zuffo, Leandro Tonello, 1989-
Combining genotyping-by-sequencing data sets for GWAS and GWS study for maize root system architecture traits / Leandro Tonello Zuffo. – Viçosa, MG, 2019.
33 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Orientador: Rodrigo Oliveira de Lima.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 28-33.

1. *Zea mays. 2.* Milho - Seleção. 3. Genômica. 4. Miho - Mapeamento genômico. 5. Milho - Populações. I. Universidade Federal de Viçosa. Departamento de Fitotecnia. Programa de Pós-Graduação em Fitotecnia. II. Título.

CDD 22. ed. 633.152

LEANDRO TONELLO ZUFFO

COMBINING GENOTYPING-BY-SEQUENCING DATA SETS FOR GWAS AND GWS STUDY FOR MAIZE ROOT SYSTEM ARCHITECTURE TRAITS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Fitotecnia, para obtenção do título de *Doctor Scientiae*.

APROVADA: 31 de outubro de 2019.

Assentimento:

Leandro Tonello Zuffo
Autor

Rodrigo Oliveira de Lima
Orientador

AKNOWLEDGEMENTS

To God, for everything he always gave me and for being present in my life.

To my parents, Rogerio and Jusara, and my sister Jessica for their support, incentives and prayers.

To my advisor and friend Rodrigo Oliveira e Lima, for the advice, encouragement and confidence over these years.

To Dra. Eveline Teixeira Caixeta Moura, Dr. Kaio Olimpio das Graças Dias, and professors José Marcelo Soriano Viana and Pedro Crescêncio Souza Carneiro for their availability and suggestions for enriching the thesis.

To my advisor, Dr. Thomas Lubberstedt, professor at Department of Agronomy at Iowa State University, for the opportunity to do an internship in his laboratory and for the knowledge passed in this period.

To the team Programa Milho-UFV, interns, scientific initiation fellows and graduate students, for their support and valuable friendship.

To Laís for the companionship, affection and support.

To my friends during this journey, for the valuable friendship and support of this stage of my life. Thank you.

To Ames friends, for their help, friendship and making me feel at home.

To the Federal University of Viçosa, in particular the Department of Fitotecnia, for the opportunity to take the course.

This study was financed in part by the Coordenação de aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Also, to Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig) and Fundação Arthur Bernardes (Funarbe), for financing the projects.

ABSTRACT

ZUFFO, Leandro, Universidade Federal de Viçosa, October, 2019. **Combining Genotyping-By-Sequencing Data Sets for GWAS and GWS study for Maize Root System Architecture Traits.** Adviser: Rodrigo Oliveira de Lima.

Maize has a sophisticated and complex root architecture that is important for plant anchorage and uptake of nutrients and water. Tools such as genome-wide association (GWAS) and genome-wide selection (GWS) can help the breeders to make decisions about which genotypes should advance in their breeding programs. In this work, we combined two data sets with different genetic backgrounds to perform a GWAS and GWS analysis in different scenarios. Our hypothesis was that combining the genomic data from different backgrounds may increase the power of detecting significant polymorphisms associated with maize seedling root traits and increase prediction accuracy. We used 679 inbred lines, 377 from Ames Panel and 302 from BGEM Panel. The root seedlings phenotype was obtained via image software from plants 14 days old grown in paper rolls in a growth chamber. We evaluated five root traits: depth, lateral root length, primary root length, total number of roots and total root length. Our study is singular by combining two SNP (Single Nucleotide Polymorphism) data sets with different genetic backgrounds to access the prediction accuracy within, across and combining the populations or subpopulations on root traits in maize. After quality control, 232,460 SNPs were used in further analysis. Population structure analysis revealed four groups that were used to build the scenarios for GWS and to control false positives in GWAS analysis. GWAS showed a total of 13 SNPs above the significance threshold. Those SNPs led to 10 candidate genes. At GWS, the combined scenario had the highest accuracy (0.66) across all traits, followed by non-stiff-stalk combined (0.63) and stiff-stalk combined (0.56). The scenarios that we calculated the prediction accuracy across panels showed the low accuracies (all were lower than 0.25). As seen in this study, combine results across studies is useful, even with different backgrounds, to improve the GWS accuracy and detect significant polymorphisms associated with maize seedling root traits and allocate the individuals from the combined data set in groups by using population structure analysis is advantageous. The genes found can be further studied to help understand the genetic basis of root development and improve the root architecture.

Keywords: *Zea mays*. Genomic selection. Association Mapping. Population structure

RESUMO

ZUFFO, Leandro, Universidade Federal de Viçosa, outubro de 2019. **Combinando conjunto de dados GBS para o estudo de GWAS e GWS para características de arquitetura de raiz de milho.** Orientador: Rodrigo Oliveira de Lima.

O milho possui uma arquitetura de raiz sofisticada e complexa que é importante para a ancoragem das plantas e a absorção de nutrientes e água. Ferramentas como a associação genômica ampla (GWAS) e a seleção genômica ampla (GWS) podem ajudar os melhoristas a tomar decisões sobre quais genótipos devem avançar em seus programas de melhoramento. Neste trabalho, combinamos dois conjuntos de dados de duas populações com diferentes ascendências para executar análise de GWAS e GWS em diferentes cenários. Nossa hipótese foi que a combinação de dados genômicos, mesmo com diferentes ascendências, pode aumentar o poder de detectar polimorfismos significativos associados às características radiculares de milho e aumentar a acurácia da predição. Nós utilizamos 679 linhagens de milho, 377 da população Ames e 302 da população BGEM. Os fenótipos de plântulas radiculares foram obtidos via software de imagem com 14 dias de idade cultivadas em rolos de papel em uma câmara de crescimento. Foram avaliados cinco caracteres de raiz: profundidade, comprimento de raízes laterais, comprimento da raiz principal, número total de raízes e comprimento total de raízes. Nosso estudo é singular por combinar dois conjuntos de dados SNP (Single Nucleotide Polymorphism) com diferentes ascendências para acessar a acurácia de predição entre, através e combinando as populações e subpopulações de caracteres de raiz de milho. Após controle de qualidade dos dados genotípicos, 232.460 SNPs foram utilizados nas análises. Análise de estrutura de população revelou total de quatro subpopulações, das quais foram formados os cenários para GWS e controle de falsos positivos na análise de GWAS. Na GWAS, um total de 13 SNPs acima do limite significativo foi detectado. Esses SNPs levaram a 10 genes candidatos. Na GWS, o cenário combinado teve a maior precisão (0,66) em todas as características, seguido pelo NSSS combinado (0,63) e SSS combinado (0,56). Os cenários em que calculamos a acurácia de predição entre populações apresentaram as menores precisões (todas com menos de 0,25). Como visto neste trabalho, combinar resultados entre estudos é útil para melhorar a precisão da GWS e detectar polimorfismos significativos associados às características radiculares de milho, mesmo com diferentes ascendências e separar o conjunto de dados combinado em grupos através de análise de estrutura de população é eficaz. Os genes encontrados podem ser mais estudados para ajudar a entender a base genética do desenvolvimento radicular e aprimorar a arquitetura.

Palavras-chave: *Zea mays*. Seleção genômica. Mapa de associação. Estrutura de população

SUMMARY

1. Introduction

Maize has an important role in humanity as major crop species with wide global geographic distribution (Hake and Ross-Ibarra, 2015). Also, it is the cereal with the highest production in the world, USDA (2019) estimates 1,121 million tons for 2018/19. Maize has a sophisticated and complex root architecture that is important for plant anchorage and uptake of nutrients and water (Lynch, 2013). That unique root architecture allows an efficient uptake from soil and allocation of nutrients to aboveground for high yields. (Rogers and Benfrey, 2015; Hochholdinger et al., 2018). It has been found that shoot growth and demand for nutrients are correlated with root growth (Peng et al., 2010) and closely associated with grain yield (Cai et al., 2012). Root traits are quantitative and controlled by several genes and very difficult to measure (Hochholdinger et al., 2018). Tools such as genome-wide association (GWAS) and genome-wide selection (GWS) can help the breeders to make decisions about which genotypes should advance in their breeding programs.

Maize has a large genetic diversity and rapid linkage disequilibrium (LD) decay, which makes a model crop for association analysis (Flint-Garcia et al., 2005). After the release of B73 as reference genome the number of studies that found markers associated with important agronomic traits on maize increased (Xiao et al., 2017). Pace et al. (2015a) and Sanchez et al. (2018) found 268 and 39 SNPs (single nucleotide polymorphisms), respectively, associated with maize seedling root traits located or linked to gene models related to root development. Wang et al. (2016) evaluated 367 diverse maize inbred lines for drought tolerance at the seedling stage and identified 83 genetic variants that directed to 42 candidate genes.

The success of crop breeding is the selection of the best offspring. In the past, breeders choose a good individual based on their experience in phenotypic selection. With the advances in molecular genetic techniques, GWS become a promising tool to shorten the generation interval, reduce the cost per cycle and save labor costs (Wang et al., 2018). Pace et al. (2015b) applied GWS at root traits on a subset of the Ames panel and found average accuracies between 0.38 and 0.55. Cantelmo et al. (2017) evaluated 838 and 797 single-cross hybrids in winter and summer crop seasons, respectively, in different locations to estimate the predictive accuracy using different training and validation populations. They found ranges of correlations of 0.82 to 0.89 in the winter season, 0.56 to 0.76 in the summer season, and between different seasons and locations were 0.53.

With the advance of genotyping technology, the cost to sequence a genome has decreased. The genotyping-by-sequencing (GBS) method is cost-effective to discover hundreds of thousands of SNPs (Poland and Rife, 2012). Due to this fact, large data sets are publicly available. Torkamaneh and Belzile (2015) using 301 soybean accessions combined two SNP data sets derived from GBS and a SNP array (SoySNP50K) and performed a GWAS on seed oil content. They founded that the number of significant marker-trait associations and the peak significance levels were improved considerably when used the combined data set.

In this work, we combined two data sets from previous studies with a different genetic background to perform a GWAS and GWS analysis in different scenarios. Our hypothesis was that combining the genomic data from different backgrounds may increase the power of detecting significant polymorphisms associated with maize seedling root traits and increase prediction accuracy. The objectives of this study were i) combine genomic data from previous studies, ii) inspect the effect of combining data sets in GWAS, and iii) evaluate GWS accuracy in different sets of training populations.

2. Material and Methods

2.1. Genotypic data

The genotypic data comes from two sources. The first source is a subset of 377 inbred lines from the "Ames Panel" (Romay et al., 2013), a collection of 2,815 public samples conserved at the USDA-ARS North Central Regional Plant Introduction Station in Ames, Iowa, downloaded from the data repository at www.panzea.org (Zhao et al., 2006). The second is a subset of 302 doubled haploid lines derived from crosses between maize landraces from the Germplasm Enhancement of Maize (GEM) project and expired-PVP lines PHZ51 and PHB47 (Sanchez et al. 2018). For further reference, the subsets will be named Ames and BGEM, respectively. Both data sets were genotyped on build version ZeaGBSv27 (Glaubitz et al., 2014) with 955,690 genotyping-by-sequencing markers at the Cornell Institute for Genomic Diversity laboratory and aligned with B73 AGPv2 coordinates as the reference genome.

2.2. Phenotypic data

The phenotypic data for the Ames comes from Pace et al. (2015a), and the BGEM from Sanchez et al. (2018). Both experiments followed the same protocols for phenotyping and

image analysis, described by Pace et al. (2014). The root traits were obtained from 14 days old seedlings grown in a paper roll inside a growth chamber. The seedling phenotypes were measured with the software ARIA (Automatic Root Image Analyzer) (Pace et al., 2014). Five traits were selected to perform this study, depth (DEP, cm): the maximum vertical distance reached by the root system; lateral root length (LRL, cm): cumulative length of all lateral roots; primary root length (PRL, cm): length of the primary root; total number of roots (TNR); total root length (TRL, cm): cumulative length of all the roots. Details about phenotypic data can be found in Table 1.

Table 1. Traits statistics from root seedling

Trait[/1]	BGEM					Ames				
	Mean	SD	Max.	Min.	$\hat{h}^2$	Mean	SD	Max.	Min.	$\hat{h}^2$
DEP	27.50	1.54	30.29	20.91	0.19	23.62	2.34	29.73	15.82	0.26
LRL	231.11	51.92	438.52	107.63	0.38	134.32	47.56	283.64	36.23	0.42
PRL	31.66	4.36	37.38	0.01	0.24	27.42	3.27	38.26	17.07	0.28
TNR	13.77	2.97	23.87	6.98	0.45	10.50	2.86	17.56	3.84	0.49
TRL	263.39	53.14	475.35	128.66	0.37	174.26	49.70	323.24	69.26	0.42

[/1] DEP: depth (cm); LRL: lateral root length (cm); PRL: primary root length (cm); TNR: total number of roots; TRL: total root length (cm); SD: standard deviation; $\hat{h}^2$: estimates of broad-sense heritability.

2.3. Data analysis

The two genotypic data sets were combined, and the quality control was employed. Genotypic markers with more than 20% of missing data, minor allele frequency of 5% and monomorphic markers were removed. The imputation method used was Linkage Disequilibrium K-number neighbor imputation. All the before-mentioned steps were conducted in Tassel 5.0 (Bradbury et al., 2007). The final number of markers was 232,460 evenly distributed across all chromosomes with 0.01% of missing data and average heterozygosity of 0.02%.

Population structure was estimated using all markers via fastStructure (Raj et al., 2014), a fast algorithm based on a variational Bayesian framework. We executed in the combined data set with allowed model complexity $K \in \{1, \ldots, 9\}$, where K denotes the number of populations, convergence criterium of $10e^{-6}$, logistic prior and five cross-validations. As recommended by Raj et al. (2014), the value that maximizes the log-marginal likelihood lower bound of the data was chosen to explain the model complexity. The LD between pairs of SNPs was calculated using Tassel 5.0, then the r^2 decay with distance was fitted using Hill and Weir expectation among adjacent sites (Hill and Weir, 1988).

2.4. Genome-wide association and genome-wide selection analysis

GWAS analysis was conducted on the three data sets, Ames, BGEM, and Combined. For each data set, GWAS was performed via the R package GAPIT (Lipka et al., 2012) using the model FarmCPU (Fixed and random model Circulating Probability Unification) that controls false positives and prevents over-fitting by applying algorithms that resolve confounding problems among testing markers and covariates (Liu et al. 2016). We used in the model the Q matrix from fastSturcture and calculated the kinship matrix via VanRaden method. The utilization of a Q + K matrix can improve statistical power and reduce spurious associations (Yu et al., 2006). To determine the significance threshold for multiple testing and reduce the Type I error, Bonferroni correction (Bonferroni, 1935) was applied. It divides the significance level at each locus by the number of tests, in this study we set alpha = 0.05, then adjusted alpha was 2.15 x 10^{-7}. The LD calculated in each chromosome was used as interval, upstream and downstream of the SNP, to search for candidate genes in all markers with significant p-values. The maize GDB database (http://www.maizegdb.org) was used to find linked candidate genes for each SNP based on the maize B73 RefGen_V2 genome, also to find matches with B73 RefGen_V4. Then, with the gene ID, Ensembl Biomart tool (Kinsella et al., 2011) was used to obtain information of the candidate genes.

The genomic estimated breeding value (GEBV) was estimated in 11 scenarios (Table 2). The accuracy was defined as the correlation between the predicted and observed phenotypic values. First scenarios were estimated in each population separately, then we combined both populations, those were randomly portioned into 60 – 40% as training and validation sets, respectively, and repeated five times. Another scenario we used all data set from one of the populations as training set and the other as validation set, e.g. Ames → BGEM.

We also used the groups from fastStructure (Table 3), stiff-stalk synthetic (SSS) and non-SSS (NSSS) from Ames and BGEM, to use as training and validation set. We used all data set from the group of one population as training to validate at the same group from the other population, e.g. NSSS Ames → NSSS BGEM. Then we combined the groups from both populations to use this large Combined group to estimate the GEBV by randomly portioning into 60 – 40% as training and validation sets, respectively, and repeated five times.

Table 2. Summary of the 11 scenarios used for training and validation set for GWS

Scenario	SSS BGEM	NSS BGEM	SSS Ames	NSS Ames	Training	Validation	Total
Initial	169	131	65	314			
Ames	-	-	39	188	227	152	379
BGEM	101	79	-	-	180	120	300
Ames → BGEM	-	-	65	314	379	300	679
BGEM → Ames	169	131	-	-	300	379	679
NSS Ames → NSS BGEM	-	-	-	314	314	131	445
NSS BGEM → NSS Ames	-	131	-	-	131	314	445
SSS Ames → SSSBGEM	-	-	65	-	65	169	234
SSS BGEM → SSS Ames	169	-	-	-	169	65	234
NSS Combined	-	79	-	188	267	178	445
SSS Combined	101	-	39	-	140	94	234
Combined	101	79	39	188	407	272	679

Moreover, we used four methods to estimate breeding values in GWS analysis. Ridge regression-best linear unbiased prediction (rrB) on R package rrBLUP (Endelman, 2011), BayesA, BayesB, and BayesC on R package BGLR (Perez and de los Campos, 2014) using 30,000 iterations, including a burn-in period of 5,000 and thinning rate of 5.

The models have a similar basic model: $y = X\beta + Z\alpha + \varepsilon$, where y is a vector of phenotypes; β is a vector of non-genetic fixed effects; X is an incidence matrix for the fixed effects; α is a vector of random regression coefficients of all marker effects; Z is an n x k genotypic matrix of markers; and ε is a vector of residuals. The models differ mainly in their priors at α.

The method rrB has a marginal multivariate t distribution with mean zero, and all markers with the same variance. BayesA has higher mass at zero that induce shrinkage of estimates that is size-of-effect dependent (Gianola, 2013), and the prior used is a scaled-t distribution. BayesB and BayesC are a mixture of priors that are similar in point of mass at zero and differ in a scaled-t and Gaussian distribution, respectively. More details for the models can be found in Wang et al. (2018).

3. Results

3.1. Population structure

The point at the log-marginal likelihood slightly changes suggest that the model components identified explain most of the population structure. In our study indicate the presence of four groups. Table 3 present the assignment of the inbred lines in each group,

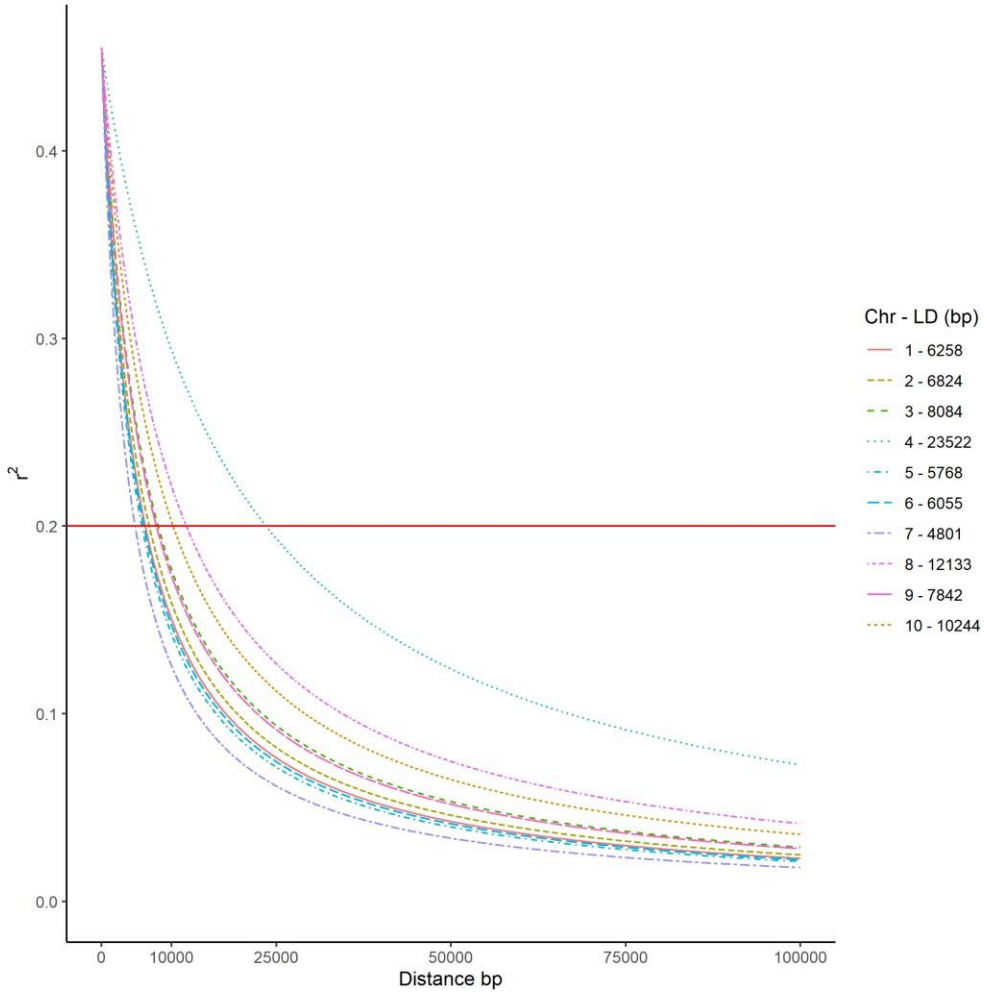
corresponding to the two panels and two main groups SSS and NSSS being separated into SSS BGEM with PHB47, NSSS BGEM with PHZ51, SSS Ames with B73, and NSSS Ames with Mo17.

Table 3. Assignment of the inbred lines in each group determined by population structure analysis

Group	NSSS BGEM	NSSS Ames	SSS BGEM	SSS Ames	Tropical	Popcorn	Sweet	Unclassified	**Total**
1	10	-	155	4	-	-	-	-	169
2	117	1	13	-	-	-	-	-	131
3	2	45	3	14	4	4	6	236	314
4	2	1	-	26	-	-	-	36	65
									679

Figure 2 shows the decay of LD across genetic distance. All chromosomes followed similar decay, except chromosome 4. The LD distance between sites when r^2 was 0.2 ranged from 4,801 to 23,522 bp.

Figure 2. LD decay for all chromosomes and the LD distance in bp when $r^2 = 0.2$



3.2. Genome-wide association

A total of 13 SNPs above the threshold significance were detected across traits and data sets (Table 4). Depth was the unique trait that did not detected significant SNPs (Figure 3). At lateral root length was detected two SNPs on chromosomes 3 and 6, both at Combined data set (Figure 4). At total number of roots and total root length was detected the same SNP on chromosome 3 at Combined data set (Figure 5 and 6, respectively). At primary root length was found six significant SNPs at BGEM and three at Combined data set (Figure 7). From all 13 SNPs above the significant threshold we found 10 candidate genes (Table 5). We considered candidate genes only the SNPs identified within gene regions.

Figure 3. Manhattan plots showing significant SNPs across the genome associated with depth at A) Ames, B) BGEM and C) Combined data sets

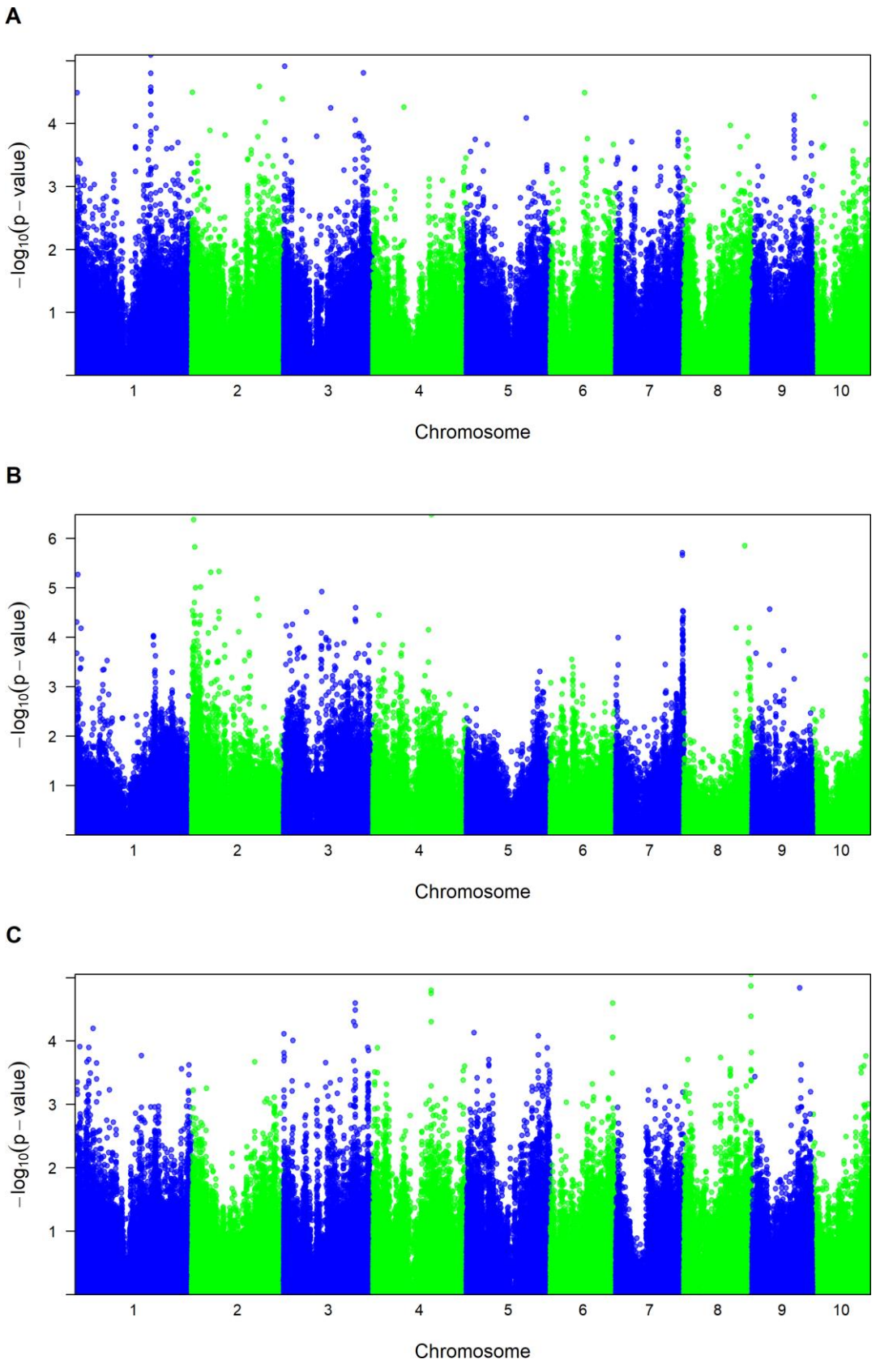


Figure 4. Manhattan plots showing significant SNPs across the genome associated with lateral root length at A) Ames, B) BGEM and C) Combined data sets

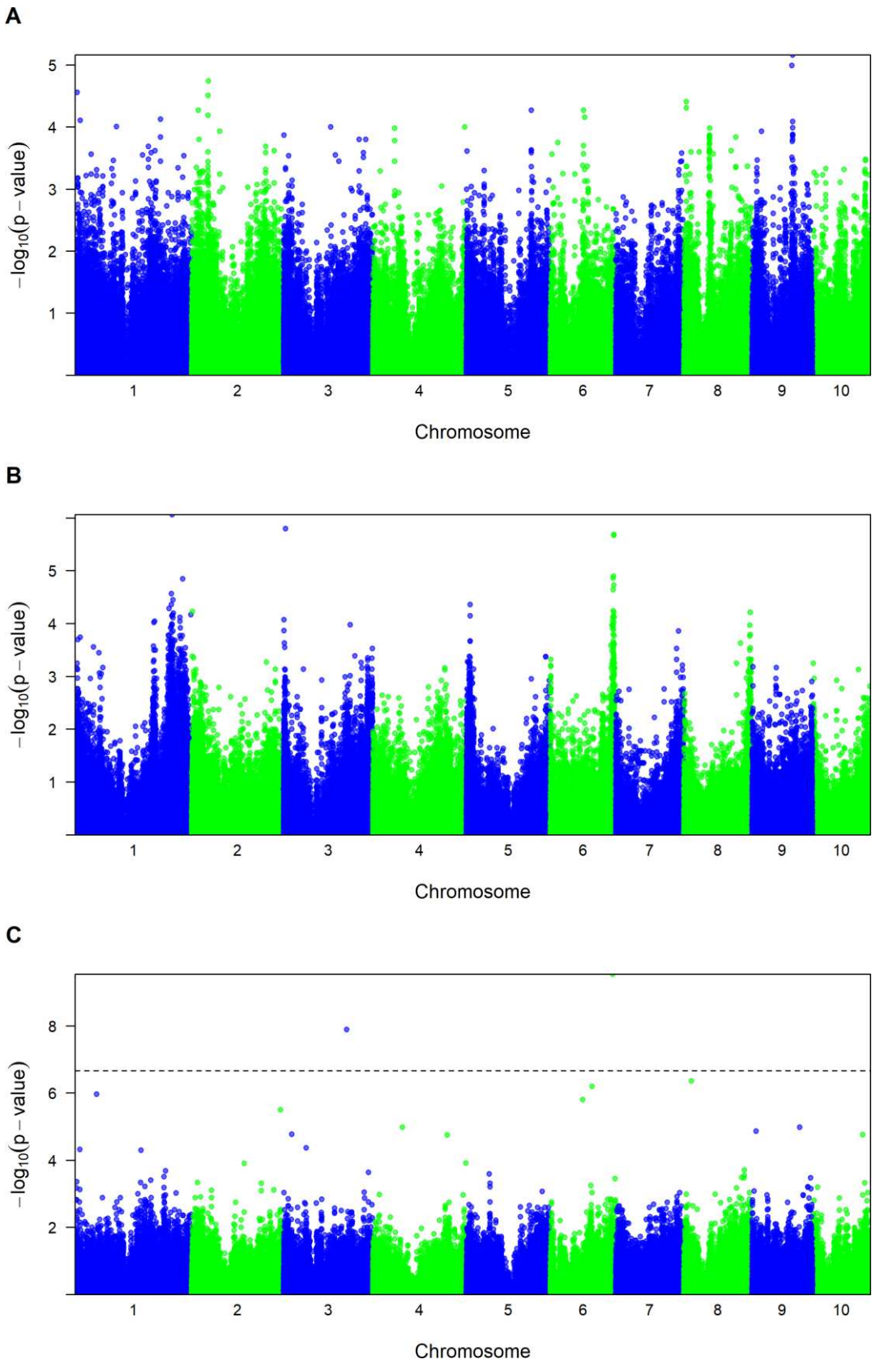


Figure 5. Manhattan plots showing significant SNPs across the genome associated with primary root length at A) Ames, B) BGEM and C) Combined data sets

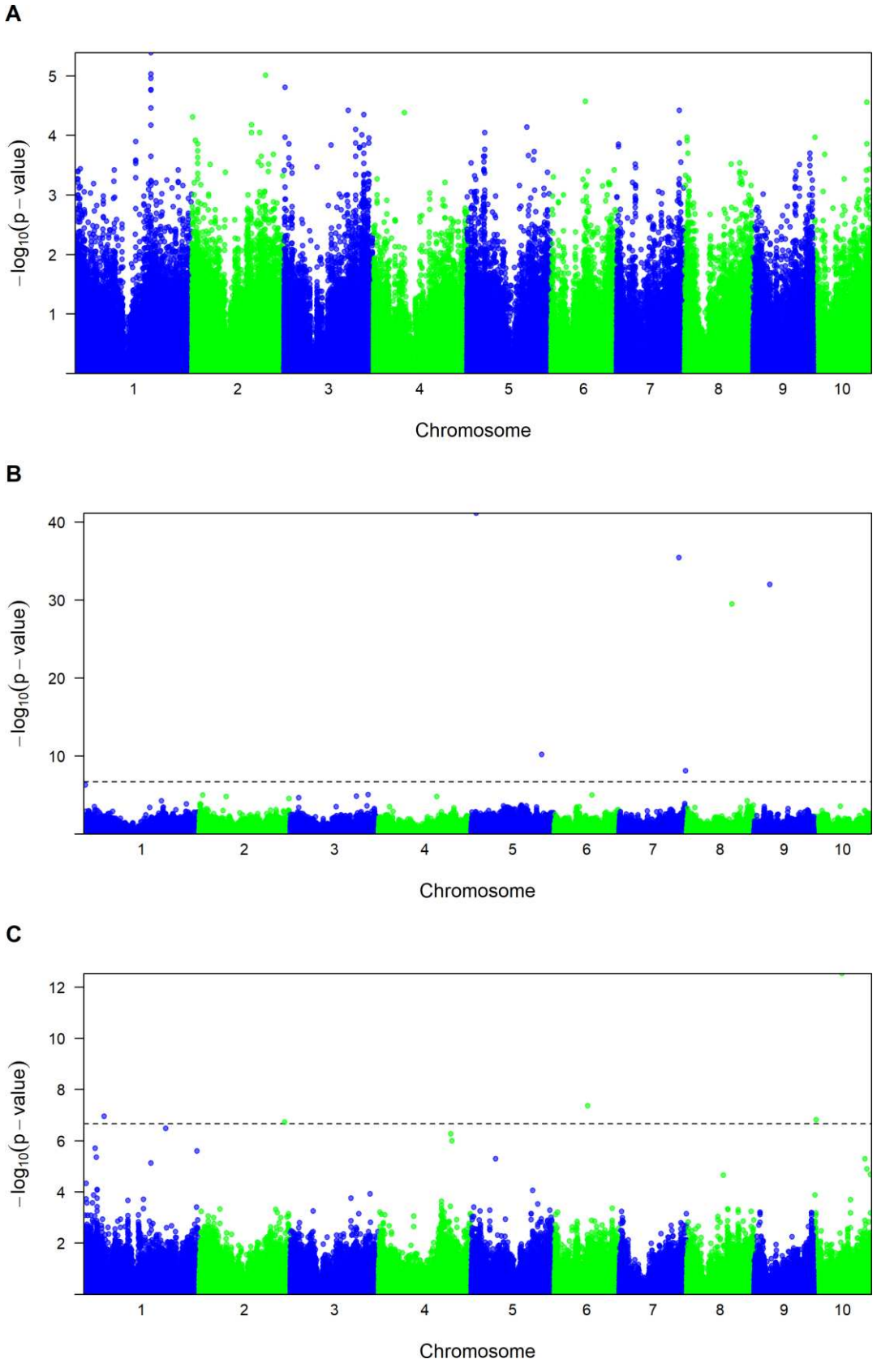


Figure 6. Manhattan plots showing significant SNPs across the genome associated with total number of roots at A) Ames, B) BGEM and C) Combined data sets

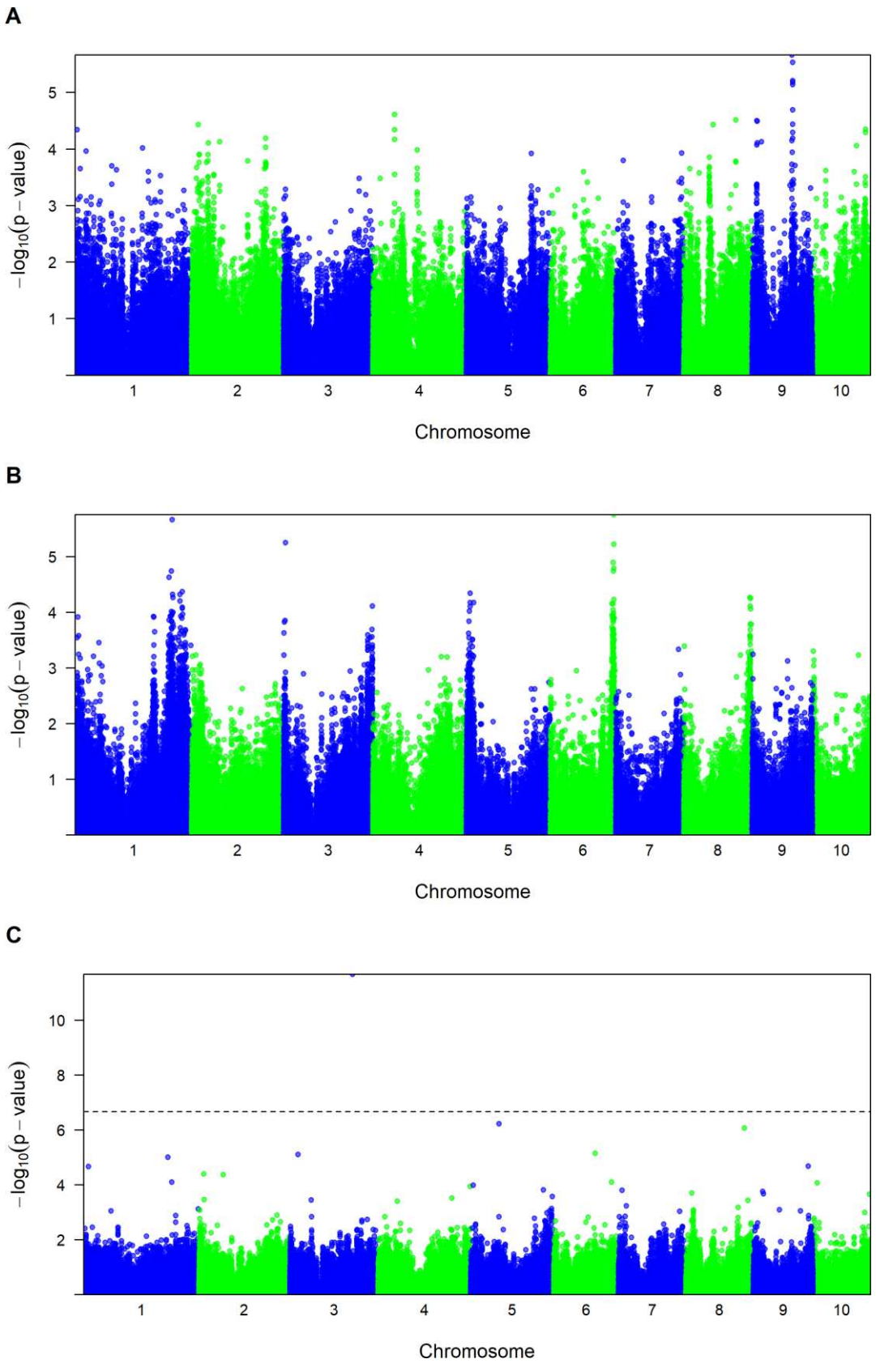


Figure 7. Manhattan plots showing significant SNPs across the genome associated with total root length at A) Ames, B) BGEM and C) Combined data sets

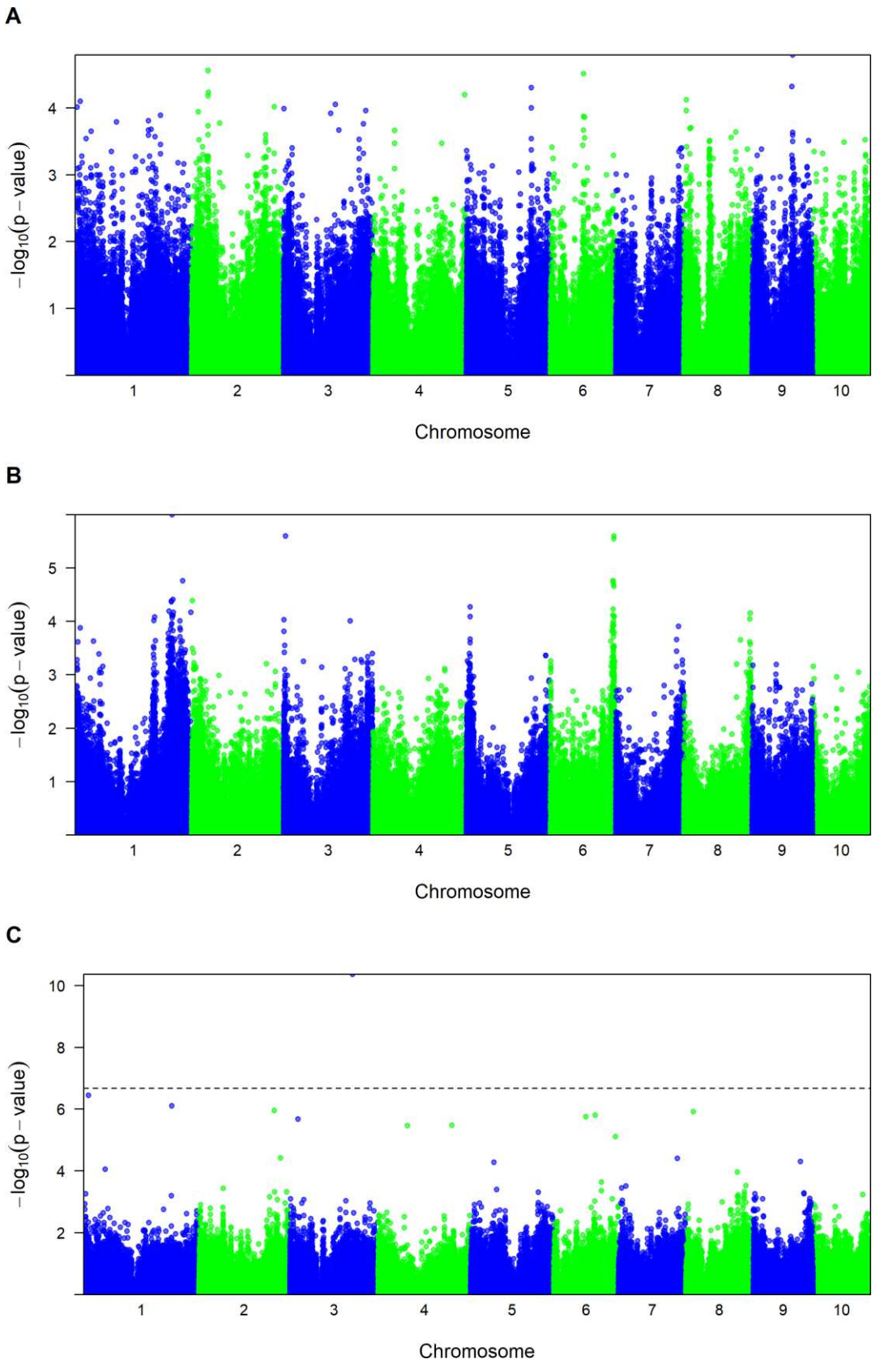


Table 4. Significant markers associated with lateral root length, primary root length, total number of roots and total root length.

Trait	Data Set	Chromosome	SNP	*P* value
Lateral root length	Combined	3	S3_164391691	2.88E-10
Lateral root length	Combined	6	S6_160702812	1.26E-08
Primary root length	BGEM	5	S5_13283601	7.27E-42
Primary root length	BGEM	5	S5_183869147	6.33E-11
Primary root length	BGEM	7	S7_155593969	3.41E-36
Primary root length	BGEM	7	S7_172936588	7.89E-09
Primary root length	BGEM	8	S8_118078481	3.21E-30
Primary root length	BGEM	9	S9_40916795	9.85E-33
Primary root length	Combined	1	S1_52765942	2.99E-13
Primary root length	Combined	6	S6_86338227	4.30E-08
Primary root length	Combined	10	S10_73160806	1.12E-07
Total number roots	Combined	3	S3_164391691	2.13E-12
Total root length	Combined	3	S3_164391691	4.32E-11

Table 5. Candidate genes for significant markers found in GWAS.

Chr	SNP	B73 Gene ID[\1]	Zm Gene ID[\2]	Function[\3]
1	S1_52765942	GRMZM2G145008	Zm00001d028967	Probable ADP-ribosylation factor GTPase-activating protein AGD14
3	S3_164391691	GRMZM2G357926	Zm00001d042434	Late embryogenesis abundant (LEA) Hydroxyproline-rich glycoprotein family
5	S5_13283601	GRMZM2G076676	Zm00001d013528	-
6	S6_86338227	GRMZM2G082874	Zm00001d036454	Tetratricopeptide repeat (TPR)-like Superfamily protein
6	S6_160702812	GRMZM2G474546	Zm00001d038809	Protein kinase superfamily protein
7	S7_155593969	GRMZM2G000361	Zm00001d021708	Pentatricopeptide repeat-containing protein chloroplastic
7	S7_172936588	GRMZM2G166738	Zm00001d022449	Potassium transporter 2
8	S8_118078481	AC197705.4_FG011	Zm00001d010585	UTP--glucose-1-phosphate Uridylyltransferase 3 chloroplastic
9	S9_40916795	GRMZM5G815338	Zm00001d045835	-
10	S10_73160806	GRMZM2G343519	Zm00001d024469	Glutaredoxin family protein

[\1]Based on B73 RefGen_v2; [\2]Based on B73 RefGen_v4; [\3]Obtained from MaizeGDB and Gramene.

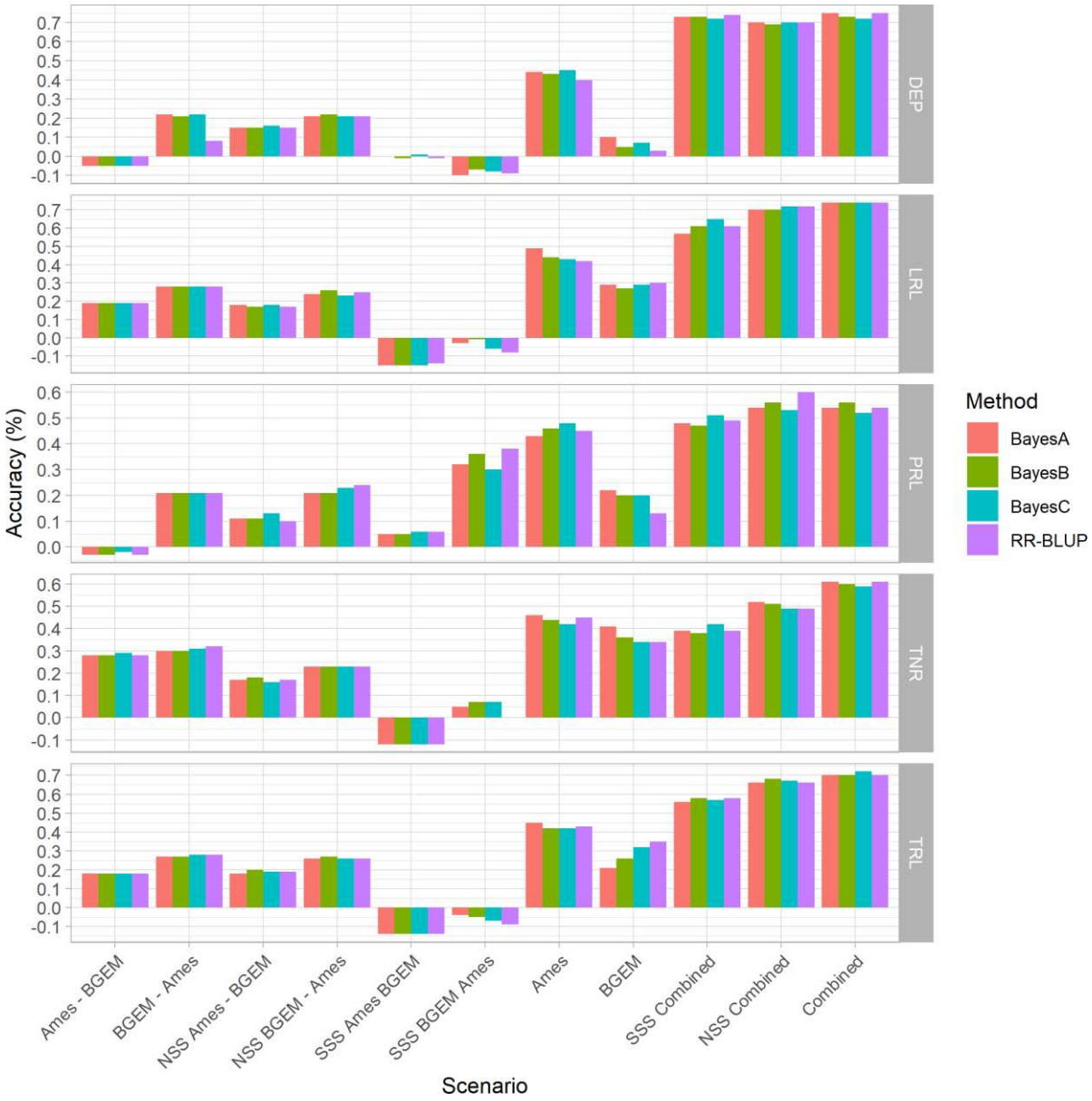
3.3. Genome-wide selection

The accuracy of the five traits from all four methods and 11 scenarios are shown in Figure 4. Comparing the prediction accuracies among the traits and scenarios, we detected the highest accuracy of 0.75, 0.74, 0.72, 0.61, and 0.60 for DEP, LRL, TRL, TNR, and PRL respectively. The lowest accuracy was found for LRL (-0.15).

The trait LRL had the highest accuracy (0.74) across all methods, followed by DEP (0.73), TRL (0.71), TNR (0.60) at Combined, and PRL (0.56) at NSSS Combined. The accuracy of the methods in each scenario per trait usually had a small difference, with the largest range of 0.14 for DEP and TRL at BGEM – Ames, and BGEM, respectively. Relating the difference among the accuracies of the methods, the frequentist approach had the biggest difference against the Bayesians, 0.14 with BayesA and BayesC, and 0.13 with BayesB, while the Bayesian approaches showed narrower ranges, BayesB with BayesA, and BayesB with BayesC showed no difference above 0.06. Ranking by the method that had the highest accuracy in each scenario and traits, the method BayesC had the best performance on 22, followed by rrB on 13, while BayesB and BayesA were similar on 10 of 55 combinations. However, the overall average of each method showed similar accuracies: 0.30 to BayesA, BayesB and BayesC and 0.29 to rrB.

Regarding the scenarios, combining data sets improved accuracy. The Combined scenario had the highest accuracy (0.66) across all traits, followed by NSSS Combined (0.63) and SSS Combined (0.56). The scenarios that we calculated the prediction accuracy across sets showed the lowest accuracies, all had accuracies lower than 0.25, with SSS Ames – BGEM being the less accurate (-0.07).

Figure 4. GWS accuracies for all traits, methods, and scenarios. [2\] DEP: depth; LRL: lateral root length; PRL: primary root length; TRL: total root length; TNR: total number of roots.



4. Discussion

4.1. Population structure

The analysis of population structure is important to reduce spurious associations and allocate the individuals in the right group for breeding programs and genetic studies, especially if does not have information available. Torkamaneh et al. (2019) using whole-genome SNP data for 1,007 soybean accessions revealed 12 subpopulations via fastSTRUCTURE. Adu et

al. (2019) studied 94 tropical maize inbred lines using cluster and population structure analysis via STRUCTURE and concluded that population structure analysis was able to allocate the inbred lines in three heterotic groups and it is more reliable than cluster analysis. In our data sets, we have the classification and pedigree of almost all inbred lines, nevertheless, the Ames Panel has 272 (72%) of inbred lines unclassified. FastSTRUCTURE revealed four groups at Combined data set (Table 4). As reported by Sanchez (2018), at BGEM panel some lines were miss-grouped into the opposite heterotic group because had a high donor (exotic) parent proportion, an average of 60%. In the first group, we have mainly SSS BGEM inbred lines with the recurrent parent PHB47 a known SSS inbred line. The four lines from Ames Panel that are allocated in the first group are PHB47, Mo401, NC250, and H84, all have in common B37 as parent. At the second group we have largely NSSS BGEM inbred lines with the recurrent parent PHZ51 a known NSSS inbred line, as well the PHZ51 from Ames Panel. Those results are similar with Sanchez (2018), that found two groups, consistent with BC1-derived introgression line in two ex-PVPs. The third group clustered the NSSS Ames, mostly of the unclassified and Mo17, a known NSSS inbred line. The fourth group was formed with SSS Ames and B73, a known SSS inbred line. These results are consistent with Pace et al. (2015a) that used STRUCTURE to stratify 384 lines from the Ames Panel and found two subpopulations. One larger with 319 lines composed of mostly non-stiff-stalk inbred lines with some tropical, popcorn and mixed lines, and the other subpopulation includes mostly individuals from stiff-stalk with B73.

In the present study, the average LD decay across all chromosomes was close to 9kb, this find is in agreement with results found by Dinesh et al. (2016). They conducted a genetic diversity and population structure analysis among 64 CIMMYT maize inbred lines using SNP markers from GBS and found at $r^2 = 0.2$ the LD decay range of 4.31 – 15.88 kb across chromosomes. The chromosomes four and eight show higher LD, a similar report was found by Thirunavukkarasu et al. (2013) that genotyped 240 subtropical elite maize inbred lines and conducted a genetic diversity, populations structure, and linkage disequilibrium analysis on the SNP data.

4.2. Genome-wise association

The root system is important to extract the water, nutrients and anchor the plant in the soil. However, it is hard to phenotype root traits. Different methods were developed for a rapid, accurate and high-throughput screening of root architecture at the seedling stage in the

laboratory as paper roll and hydroponic system (Wang et al., 2019). To improve breeding success, it is important to integrate genomic tools to identify desirable alleles and use them to access germplasm (Andorf et al., 2019). Wang et al. (2019) evaluated 297 maize inbred lines in paper rolls to get accurate phenotypes at seedling stage, then for GWAS a set of 131,271 SNPs were analyzed with GLM and MLM (Mixed Linear Model). A total of 355 and 28 marker-trait associations were identified by GLM and MLM, respectively, and 96 candidate genes were located, five candidate genes were identified as promising. Zaidi et al. (2016) phenotyped a subset of 396 diverse tropical maize lines from the CIMMYT Asia association mapping panel for root traits under drought and well-watered conditions. GWAS analysis using MLM on 331,390 SNPs detected 167 SNPs associated with root functional, structural, shoot biomass and grain yield. They found among these SNPs favorable associations in chromosome 3 for drought stress tolerance in maize. Identify and understand SNPs associated with root traits is fundamental for breeding programs through marker-assisted introgression into elite germplasm, especially now with recurrent climate changes.

In this study, we used two data sets from different backgrounds of maize seedling traits grown in a paper roll. Our hypothesis is that combining data sets it is possible to increase the analysis power. GWAS at each population found significant SNPs just in one trait (PRL) and population (BGEM), otherwise at Combined data set we were able to find significant SNPs in all traits. Those SNPs led to 10 candidate genes. The gene Zm00001d028967 found in primary root length belongs to the Ras superfamily that regulates different cellular process such as cell growth and cell proliferation (Hall, 1990). Jincheng-Yuan et al (2015) isolated an ADP-ribosylation factor in maize and analysis of real-time quantitative PCR showed higher expression levels in root and embryo than mature leaves, silks and seeds. Zm00001d042434 was detected at lateral root length, total number of roots and total root length at Combined, it is linked to the LEA gene family in maize and associated with tolerance of abiotic stress. Li and Cao (2015) identified LEA genes in maize and found expression across different tissues examined, the LEA_3 group showed high accumulation in root and steam leaf, other LEA proteins showed expression in embryo, those results indicate an important role for abiotic stress. Zm00001d036454 was found in primary root length at Combined and belong to the Tetratricopeptide repeat superfamily protein. Wei and Han (2017) investigated TPR gene family on *Arabidopsis*, rice, poplar and maize and found many genes associated with root and steam growth or functions, meristem growth, post-embryonic cell division and differentiation of distal tissues in the root. Zm00001d038809 was detected on lateral root length at Combined

and belong to the protein kinase superfamily. Fan et al (2018) investigated the expression of protein kinases in maize, rice and *Arabidopsis* and found genes with different functions over the plant, but a large number was high expressed in maize root. Zm00001d021708 was found on primary root length at BGEM and belongs to the pentatricopeptide repeat proteins (PPRs) superfamily. Wei and Han (2016) conducted a genome-scale analysis to identify gene members of PPRs, they found significantly expressed genes in crown root, primary root and seminal root under stress. Zm00001d022449 was detected on primary root length at BGEM and it is a maize gene related to Potassium transporter. Potassium is an essential nutrient for plant in several cellular processes. Zhang et al. (2012) founded a series of cis elements involved in $Ca2^{+}$ responsive, abiotic stress, and seed development. Zm00001d010585 was detected on primary root length at BGEM, any information related with maize or root was found in the literature. Zm00001d024469 was detected on primary root length at Combined and belong to the Glutaredoxin family protein. According to Ding et al. (2019), the genes are involved in plant growth and development, primary metabolism, and abiotic stress. These results provide genes with possible influence on root traits that can be used for breeding after validation in field conditions.

4.3. Genome-wide selection

In maize breeding, the GWS is becoming an important tool to predict un-phenotyped genotypes, increase breeding efficiency by early selection and accelerate genetic gain (Crossa et al., 2010; Edriss et al., 2017). Researchers have used SNP data to apply GWS to predict maize single crosses (Viana et al., 2018), evaluate maize yield under genotype x environment interaction (Millet et al., 2019) and combine with different types of omics data to improve prediction accuracy (Xu et al., 2017; Schrag et al., 2018). Our study is singular by combining two SNP data sets with different genetic backgrounds to access the prediction accuracy within, across and combining the populations or subpopulations on root traits in maize.

Several factors can affect the accuracy of genomic selection, alike training population size, genetic relationship, heritability, marker density, and statistical methods (Technow et al., 2013; Wang et al., 2018; Cerrudo et al., 2018; Edwards et al., 2019; Zhang et al., 2019). In our study, prediction across populations showed low accuracies. In contrast, the prediction within populations presented better results for all traits at Ames and almost all at BGEM, and higher accuracies were found as a result of combining the data sets. In addition, we noticed that the training population size is an important factor when related individuals are present in the

training and validation sets, since we had higher accuracy at SSS Combined with 140 individuals as training than the scenarios across populations, as Ames – BGEM with almost three-fold (379) individuals as training. Similar to our results, Mastrodomenico et al. (2019) evaluating 522 single-cross maize hybrids and the parents across 10 environments, observed that increasing the training population size was more effective when a greater number of parents was available in the training and validation sets. Zhang et al. (2017) evaluated 22 bi-parental tropical maize populations on six trait-environment combinations under different levels of training population size, marker density, and heritability reported that prediction accuracy within-population improved with an increase in training population size.

Different statistical methods have been used to determine breeding values from genomic prediction models (Crossa et al., 2017; Alves et al., 2019). Previous studies have reported better performance of the Bayesian approach in a genomic selection over other methods (Daetwyler et al., 2010; Shikha et al., 2017). The assumptions on the marker effect need to match the genetic architecture of target traits to the method be effective (Clark et al., 2011; Zhao et al., 2013). Bayesian models focus on QTL-marker associations, while rrB rely on kinship more strongly (Habier et al., 2007). Hence, if are few large effects QTLs the Bayesian approach will present better accuracy over rrB, otherwise, if are many small effects QTLs, the accuracy will be similar (Lin et al., 2014). In our study, BayesC outperformed the other methods by a slight difference, where the largest difference was 0.04 across traits. Probably because we have a big marker density with 232,460 SNPs.

The main goal of breeding is to select in early generations better progenies with favorable alleles. GWAS and GWS are powerful tools in genetic mapping and to save labor costs to predict and select un-phenotyped materials. The recent increase development of new DNA marker technology allied with the omics era will produce a large amount of data. Combine large data sets will be routine in breeding programs enabling geneticists to analyze data gathered using different approaches to identify reliable biological causes on phenotypic variation in complex traits. However, innovations in statistical methods need to improve with computational speed to handle the continuously increasing sample size and marker density.

5. Conclusions

As seen in this study, combine results across studies is useful, even with different genetic backgrounds, to improve the GWS accuracy and detect significant polymorphisms

associated with maize seedling root traits. One approach to handle these large data sets, as proven in this study, is to allocate the individuals in groups by population structure analysis. The genes found can be further studied to help understand the genetic basis of root development and improve the root architecture.

6. References

Alves FC, Granato ÍSC, Galli G, et al (2019) Bayesian analysis and prediction of hybrid performance. Plant Methods 15:14. doi: 10.1186/s13007-019-0388-x

Andorf C, Beavis WD, Hufford M, et al (2019) Technological advances in maize breeding: past, present and future. Springer Berlin Heidelberg

Boakyewaa Adu G, Badu-Apraku B, Akromah R, et al (2019) Genetic diversity and population structure of early-maturing tropical maize inbred lines using SNP markers. PLoS One 14:e0214810. doi: 10.1371/journal.pone.0214810

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste, Studi in Onore del Professore Salvatore Ortu Carboni. Rome, 13-60

Bradbury PJ, Zhang Z, Kroon DE, et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi: 10.1093/bioinformatics/btm308

Cai H, Chen F, Mi G, et al (2012) Mapping QTLs for root system architecture of maize (Zea mays L.) in the field at different developmental stages. Theor Appl Genet 125:1313–1324. doi: 10.1007/s00122-012-1915-6

Cantelmo NF, Von Pinho RG, Balestre M (2017) Genome-wide prediction for maize single-cross hybrids using the GBLUP model and validation in different crop seasons. Mol Breed 37:. doi: 10.1007/s11032-017-0651-7

Cao S, Loladze A, Yuan Y, et al (2017) Genome-Wide Analysis of Tar Spot Complex Resistance in Maize Using Genotyping-by-Sequencing SNPs and Whole-Genome Prediction. Plant Genome 10:. doi: 10.3835/plantgenome2016.10.0099

Cerrudo D, Cao S, Yuan Y, et al (2018) Genomic Selection Outperforms Marker Assisted Selection for Grain Yield and Physiological Traits in a Maize Doubled Haploid Population Across Water Treatments. Front Plant Sci 9:. doi: 10.3389/fpls.2018.00366

Clark SA, Hickey JM, van der Werf JH (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18. doi: 10.1186/1297-9686-43-18

Combs E, Bernardo R (2013) Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. Plant Genome 6:. doi: 10.3835/plantgenome2012.11.0030

Crossa J, Campos G de los, Pérez P, et al (2010) Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186:713–724. doi: 10.1534/genetics.110.118521

Crossa J, Pérez-Rodríguez P, Cuevas J, et al (2017) Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci 22:961–975. doi: 10.1016/j.tplants.2017.08.011

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185:1021–1031. doi: 10.1534/genetics.110.116855

De Storme N, Geelen D (2014) Callose homeostasis at plasmodesmata: molecular regulators and developmental relevance. Front Plant Sci 5:. doi: 10.3389/fpls.2014.00138

Dinesh A, Patil A, Zaidi P, et al (2016) Genetic diversity, linkage disequilibrium and population structure among CIMMYT maize inbred lines, selected for heat tolerance study. Maydica 61.3:

Ding, S., He, F., Tang, W., Du, H., & Wang, H. (2019). Identification of Maize CC-Type Glutaredoxins That Are Associated with Response to Drought Stress. Genes , Vol. 10. https://doi.org/10.3390/genes10080610

Edriss V, Gao Y, Zhang X, et al (2017) Genomic Prediction in a Large African Maize Population. Crop Sci 57:2361. doi: 10.2135/cropsci2016.08.0715

Edwards SM, Buntjer JB, Jackson R, et al (2019) The effects of training population design on genomic prediction accuracy in wheat. Theor Appl Genet 132:1943–1952. doi: 10.1007/s00122-019-03327-y

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J 4:250. doi: 10.3835/plantgenome2011.08.0024

Fan, M., Ma, W., Liu, C., Zhang, C., Wu, S., Chen, M., … Lin, F. (2018). Evolution and Expression Characteristics of Receptor-Like Cytoplasmic Protein Kinases in Maize, Rice and Arabidopsis. International Journal of Molecular Sciences , Vol. 19. https://doi.org/10.3390/ijms19113680

Flint-Garcia SA, Thuillet AC, Yu J, et al (2005) Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064. doi: 10.1111/j.1365-313X.2005.02591.x

Gianola D (2013) Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194:573–596. doi: 10.1534/genetics.113.151753

Glaubitz JC, Casstevens TM, Lu F, et al (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS One 9:e90346. doi: 10.1371/journal.pone.0090346

Habier D, Fernando RL, Dekkers JCM (2007) The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177:2389–2397. doi: 10.1534/genetics.107.081190

Hake S, Ross-Ibarra J (2015) Genetic, evolutionary and plant breeding insights from the domestication of maize. Elife 4. doi: 10.7554/eLife.05861

Hall, A. (1990). The cellular functions of small GTP-binding proteins. Science, 249(4969), 635–640. https://doi.org/10.1126/science.2116664

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54–78. doi: 10.1016/0040-5809(88)90004-4

Hochholdinger F, Yu P, Marcon C (2018) Genetic Control of Root System Development in Maize. Trends Plant Sci 23:79–88. doi: 10.1016/j.tplants.2017.10.004

Jincheng-Yuan, Jinhui-Song, Hailian-Ma, Xiaoqing-Song, Huiping-Wei, & Yinghui-Liu. (2015). Ectopic expression a maize ADP-ribosylation factor gene in Arabidopsis, increase plant size and growth rate. Journal of Plant Biochemistry and Biotechnology, 24(2), 161–166. https://doi.org/10.1007/s13562-013-0248-0

Li, X., & Cao, J. (2016). Late Embryogenesis Abundant (LEA) Gene Family in Maize: Identification, Evolution, and Expression Profiles. Plant Molecular Biology Reporter, 34(1), 15–28. https://doi.org/10.1007/s11105-015-0901-y

Lim J, Helariutta Y, Specht CD, et al (2000) Molecular analysis of the SCARECROW gene in maize reveals a common basis for radial patterning in diverse meristems. Plant Cell 12:1307–1318. doi: 10.1105/tpc.12.8.1307

Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: a review. Crop Pasture Sci 65:1177. doi: 10.1071/CP13363

Lipka AE, Tian F, Wang Q, et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399. doi: 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 Genetics, 12(2), e1005767. Retrieved from https://doi.org/10.1371/journal.pgen.1005767

Lynch JP (2013) Steep, cheap and deep: an ideotype to optimize water and N acquisition by maize root systems. Ann Bot 112:347–357. doi: 10.1093/aob/mcs293

Mastrodomenico AT, Bohn MO, Lipka AE, Below FE (2019) Genomic selection using maize ex-plant variety protection germplasm for the prediction of nitrogen-use traits. Crop Sci 59:212–220. doi: 10.2135/cropsci2018.06.0398

Millet EJ, Kruijer W, Coupel-Ledru A, et al (2019) Genomic prediction of maize yield across European environmental conditions. Nat Genet 51:952–956. doi: 10.1038/s41588-019-0414-y

Nuccio ML, Wu J, Mowers R, et al (2015) Expression of trehalose-6-phosphate phosphatase in maize ears improves yield in well-watered and drought conditions. Nat Biotechnol 33:862–869. doi: 10.1038/nbt.3277

Pace J, Lee N, Naik HS, et al (2014) Analysis of maize (Zea maysL.) seedling roots with the high-throughput image analysis tool ARIA(Automatic root image analysis). PLoS One 9:1–10. doi: 10.1371/journal.pone.0108255

Pace J, Gardner C, Romay C, et al (2015a) Genome-wide association analysis of seedling root development in maize (Zea mays L.). BMC Genomics 16. doi: 10.1186/s12864-015-1226-9

Pace J, Yu X, Lübberstedt T (2015b) Genomic prediction of seedling root length in maize (*Zea mays* L.). Plant J 83:903–912. doi: 10.1111/tpj.12937

Peng Y, Niu J, Peng Z, et al (2010) Shoot growth potential drives N uptake in maize plants and correlates with root growth in the soil. F Crop Res 115:85–93. doi: 10.1016/j.fcr.2009.10.006

Pérez P, de los Campos G (2014) Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198:483–495. doi: 10.1534/genetics.114.164442

Poland JA, Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome J 5:92. doi: 10.3835/plantgenome2012.05.0005

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945 LP – 959

Raj A, Stephens M, Pritchard JK (2014) FastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics 197:573–589. doi: 10.1534/genetics.114.164350

Rogers ED, Benfey PN (2015) Regulation of plant root system architecture: implications for crop advancement. Curr Opin Biotechnol 32:93–98. doi: 10.1016/j.copbio.2014.11.015

Romay MC, Millard MJ, Glaubitz JC, et al (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14:R55. doi: 10.1186/gb-2013-14-6-r55

Sanchez DL, Liu S, Ibrahim R, et al (2018) Genome-wide association studies of doubled haploid exotic introgression lines for root system architecture traits in maize (Zea mays L.). Plant Sci 268:30–38. doi: 10.1016/j.plantsci.2017.12.004

Schrag TA, Westhues M, Schipprack W, et al (2018) Beyond genomic prediction: Combining different types of omics data can improve prediction of hybrid performance in maize. Genetics 208:1373–1385. doi: 10.1534/genetics.117.300374

Shikha M, Kanika A, Rao AR, et al (2017) Genomic Selection for Drought Tolerance Using Genome-Wide SNPs in Maize. Front Plant Sci 8:1–12. doi: 10.3389/fpls.2017.00550

Technow F, Bürger A, Melchinger AE (2013) Genomic Prediction of Northern Corn Leaf Blight Resistance in Maize with Combined or Separated Training Sets for Heterotic Groups. G3: Genes|Genomes|Genetics 3:197–203. doi: 10.1534/g3.112.004630

Thirunavukkarasu N, Hossain F, Shiriga K, et al (2013) Unraveling the genetic architecture of subtropical maize (Zea mays L.) lines to assess their utility in breeding programs. BMC Genomics 14:. doi: 10.1186/1471-2164-14-877

Torkamaneh D, Belzile F (2015) Scanning and filling: Ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. PLoS One 10:1–16. doi: 10.1371/journal.pone.0131533

Torkamaneh D, Laroche J, Valliyodan B, et al (2019) Soybean Haplotype Map (GmHapMap): A Universal Resource for Soybean Translational and Functional Genomics. bioRxiv 534578. doi: 10.1101/534578

Viana JMS, Pereira HD, Mundim GB, et al (2018) Efficiency of genomic prediction of non-assessed single crosses. Heredity (Edinb) 120:283–295. doi: 10.1038/s41437-017-0027-0

Wang, Wei, Li, et al (2019) Integrating GWAS and Gene Expression Analysis Identifies Candidate Genes for Root Morphology Traits in Maize at the Seedling Stage. Genes (Basel) 10:773. doi: 10.3390/genes10100773

Wang X, Wang H, Liu S, et al (2016) Genetic variation in ZmVPP1 contributes to drought tolerance in maize seedlings. Nat Genet 48:1233–1241. doi: 10.1038/ng.3636

Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: Current status and prospects. Crop J 6:330–340. doi: 10.1016/j.cj.2018.03.001

Wei, K., & Han, P. (2017). Comparative functional genomics of the TPR gene family in Arabidopsis, rice and maize. Molecular Breeding, 37(12), 152. https://doi.org/10.1007/s11032-017-0751-4

Xiao Y, Liu H, Wu L, et al (2017) Genome-wide Association Studies in Maize: Praise and Stargaze. Mol Plant 10:359–374. doi: 10.1016/j.molp.2016.12.008

Xu Y, Xu C, Xu S (2017) Prediction and association mapping of agronomic traits in maize using multiple omic data. Heredity (Edinb) 119:174–184. doi: 10.1038/hdy.2017.27

Yu J, Pressoir G, Briggs WH, et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. doi: 10.1038/ng1702

Zaidi PH, Seetharam K, Krishna G, et al (2016) Genomic regions associated with root traits under drought stress in tropical maize (Zea mays L.). PLoS One 11:1–18. doi: 10.1371/journal.pone.0164340

Zhang A, Wang H, Beyene Y, et al (2017) Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. Front Plant Sci 8:1–12. doi: 10.3389/fpls.2017.01916

Zhang H, Yin L, Wang M, et al (2019) Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. Front Genet 10:. doi: 10.3389/fgene.2019.00189

Zhang, Z., Zhang, J., Chen, Y., Li, R., Wang, H., & Wei, J. (2012). Genome-wide analysis and identification of HAK potassium transporter gene family in maize (Zeamays L.). Molecular Biology Reports, 39(8), 8465–8473. https://doi.org/10.1007/s11033-012-1700-2

Zhao W, Canaran P, Jurkuta R, et al (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res 34:D752–D757. doi: 10.1093/nar/gkj011

Zhao Y, Zeng J, Fernando R, Reif JC (2013) Genomic Prediction of Hybrid Wheat Performance. Crop Sci 53:802. doi: 10.2135/cropsci2012.08.046