Research Article

Ekin
Journal of Crop Breeding and Genetics

2(2):41-46, 2016

www.ekinjournal.com

Ekin International biannual peer-reviewed journal

Genome-Wide Association Mapping Using a Bayesian Mixture Model for Plant Height in *Oryza sativa*

Burak Karacaören*

Faculty of Agriculture, Akdeniz University, 07059, Antalya, TÜRKİYE

* Corresponding author e-mail: burakkaracaoeren@yahoo.com

Citation:
Karacaören B., 2016. Genome-wide association mapping using a bayesian mixture model for plant height in *Oryza sativa*. Ekin J. 2(2):41-46.

Received: 05.01.2016 ***Accepted: 19.03.2016*** ***Published Online: 28.07.2016*** ***Printed: 30.07.2016***

ABSTRACT

Genotypic and phenotypic data could be used to predict inheritance of complex traits for plant breeding in genome wide association mapping studies (GWAS). In GWAS using a single marker model may leads to suboptimal use of genotypic datasets. Alternatively, using whole genome, a Bayesian mixture model may cluster markers into predefined classes. We used 413 diverse accessions of *Oryza sativa* with 36900 Single Nucleotide Polymorphisms (SNPs) markers for plant height. We assumed different genetic architectures for the phenotype. We estimated genotypic heritability as 0.61. Bayesian mixture model detected 144, 446, 54 SNPs with explanatory levels of 0.0001, 0.001 and 0.01 respectively. Chromosome 1 (n=109), and 3 (n=85) had the highest explanatory genetic variances as 23% and 19%, respectively. Correlation between genomic predicted observations and actual observations was found to be 0.94. Since GWAS are mostly based on only one replication as was also the case in this study; results need to be confirmed by independent validation experiments.

Keywords: Genome wide association mapping studies, Bayesian mixture model, Single Nucleotide Polymorphisms Markers, *Oryza sativa*.

Introduction

Plant height is an important complex yield related trait in *Oryza sativa*. However, height itself is an important model phenotype in various organisms. Plant height is easy to measure, highly heritable but underlying biology is found to be complex.

Genotypic and phenotypic data could be used to predict complex traits for plant breeding in genome wide association mapping studies (GWAS) . Single Nucleotide Polymorphisms (SNPs) could be used as markers to detect genomic signals over chromosomes in GWAS. Although GWAS found genomic signals for various phenotypes in various organisms, explanatory proportion of the SNPs remained very low. This unexpected result termed as missing heritability problem (Turkheimer, 2011). For example only 5% of the height phenotypic variance explained by about 50 variants using human GWAS (Yang *et al*., 2010).

Moser *et al*., (2015) suggested to use a bayesian mixture model to overcome associated problems in GWAS including multiple hypothesis testing, linkage disequilibrium and for increasing the power of the experiment. Employing whole SNPs altogether (Meuwissen *et al*., 2001) could be beneficial compared with marker assisted selection (MAS) that employs a few of the markers at the selection stage. Main aim of this study was genome wide association mapping of plant height in *Oryza sativa* using both single SNPs and a bayesian mixture models (Moser *et al*., 2015).

Materials and Methods

Data

The GWAS included 413 Asian rice cultivars from 82 countries. The genome consisted 36900 SNPs distributed over 12 chromosomes. Plant height was measured as distance between from tip of main panicle of the plant to soil surface. More details about the dataset could be found at Zhao *et al*., (2011).

Genome wide Rapid Association Using Mixed Model and Regression

One of the important GWAS assumption is that the

individuals in the samples should be unrelated. Genetic and geographic relationship among the individuals may lead to deviation from this assumption. Aulchenko *et al*., (2007) defined a mixed model to correct for genetic (or when the pedigree file is not available, genomic) effects: Genome wide Rapid Association Using Mixed Model and Regression (GRAMMAR). We used a genomic relationship matrix in the GRAMMAR to correct for relation between plant samples. In addition we obtained first four principal components from the samples to correct for geographical relationship among the samples. In GRAMMAR $y=Xb+Zu+e$ used, where X and Z are design matrices to link plant height (y: response variable) with fixed *b* (top four principal components) and random *u* effects (additive gene effects:), *e* and *u* (weighted with genomic or pedigree information) was assumed to be sampled from normal distribution and their respective variance components was predicted using maximum likelihood procedure as was implemented in GenABEL (Aulchenko *et al*., 2007). We used a single SNP regression model using the GRAMMAR approach by GenABEL to detect associated genes with plant height.

A Bayesian mixture model

We used a hierarchical bayesian mixture model (Moser *et al*., 2015) (BayesR) to obtain SNP solutions by using whole genome simultaneously. BayesR assumed a mixture of four normal distributions for the SNP effects to be predicted (assumed to be 0.00001, 0.0001, 0.001, 0.01). We sampled 50000 markov chains and discarded first 20000 as burn in period and recorded every 10th sample for thinning the chain. We compared results of different runs of markov chains to assess convergence.

Results and Discussion

We excluded 8214 SNPs based on minor allele frequency of <5%, and call rate < 90% leaving 28686 SNPs in the dataset. We excluded 259 individuals due to too high identity by state (IBS) (0.95>) leaving 154 individuals in the analyses. Mean IBS estimated as 0.62 (0.15) and mean autosomal heterozygosity estimated as 5.91e-05 (4.49e-05). Genomic heritability was found to be 0.61.

GRAMMAR approach identified strong signals from various locations of the genome (Table 1) however after multiple hypothesis testing only from chromosome 6 (id6002498) had a suggestive genomic signal that could be detectable (Figure 1).

Quantile-quantile plot (Figure 2) showed that still there might be indication of population stratification as inflation factor was found to be 1.34 (0.0001).

We used both full data set ($_n$=413, data1) and quality control filtered data set ($_n$=154, data2) to investigate impact of population sub structure (mainly due to geographical and genetic relationships of the individuals) to the Bayesian mixture analyses. Since we used different proportion of explanatory variances for each sub classes we could investigate SNP effects from chromosomes and/ or from certain loci. We presented the results of BayesR in Table 2 using both Data1 and Data2.

When we filtered out the highly correlated individuals (Data2) BayesR detected 3376 SNPs compared with the full data set (Data1) as 644 SNPs. .

Table 2 suggested that the highest explanatory proportions obtained from chromosome 1 as 0.23 and 0.19 for Data1 and Data2 respectively. Bayesian mixture model detected 144, 446, 54 SNPs with explanatory levels of 0.0001, 0.001 and 0.01 using Data1. Bayesian mixture model detected 2957, 356, 155 SNPs with explanatory levels of 0.0001, 0.001 and 0.01 using Data2.

We detected mostly small SNPs effects from various part of the genome using both Data1 (Table 3) and Data 2 (Table 4). Correlation between genomic predicted observations and actual observations found to be 0.94 and 0.99 for Data1 and Data2 respectively. As was expected using homogenized samples (Data2) lead to higher accuracy for predicting phenotypes using genotypic information. Since the GWAS are mostly based on only one replication (as was also the case in this study); results needs to be confirmed by independent validation experiments.

Conclusion

Employing homogenized sample in GWAS using IBS information to overcome population stratification leads to better genomic predictions of plant height for *Oryza sativa* by the Bayesian mixture model.

Acknowledgment

This study was supported by the Akdeniz University Project Number FDK-2004-106.

Table 1. Summary of genome wide rapid association using mixed model and regression (P: raw p values, Pc corrected p values using 1000 permutations)

SNP	Chromosome	Chi square	P	Pc
id6002498	6	37.76	7.97E-10	0.077
id11006324	11	27.22	1.81E-07	0.549
id12000343	12	25.98	3.44E-07	0.66
id2001384	2	25.36	4.74E-07	0.706
id1020583	1	24.94	5.91E-07	0.739
id1020512	1	24.53	7.31E-07	0.773
id1020569	1	24.53	7.31E-07	0.773
id1020642	1	24.53	7.31E-07	0.773
id3013805	3	23.73	1.11E-06	0.827
id6010525	6	23.62	1.17E-06	0.831

Figure 1. Manhattan plot of GWAS result using GRAMMAR approach. The x-axis of the Manhattan plot shows the genomic position, the y-axis represents the log10 base transformed p-values.

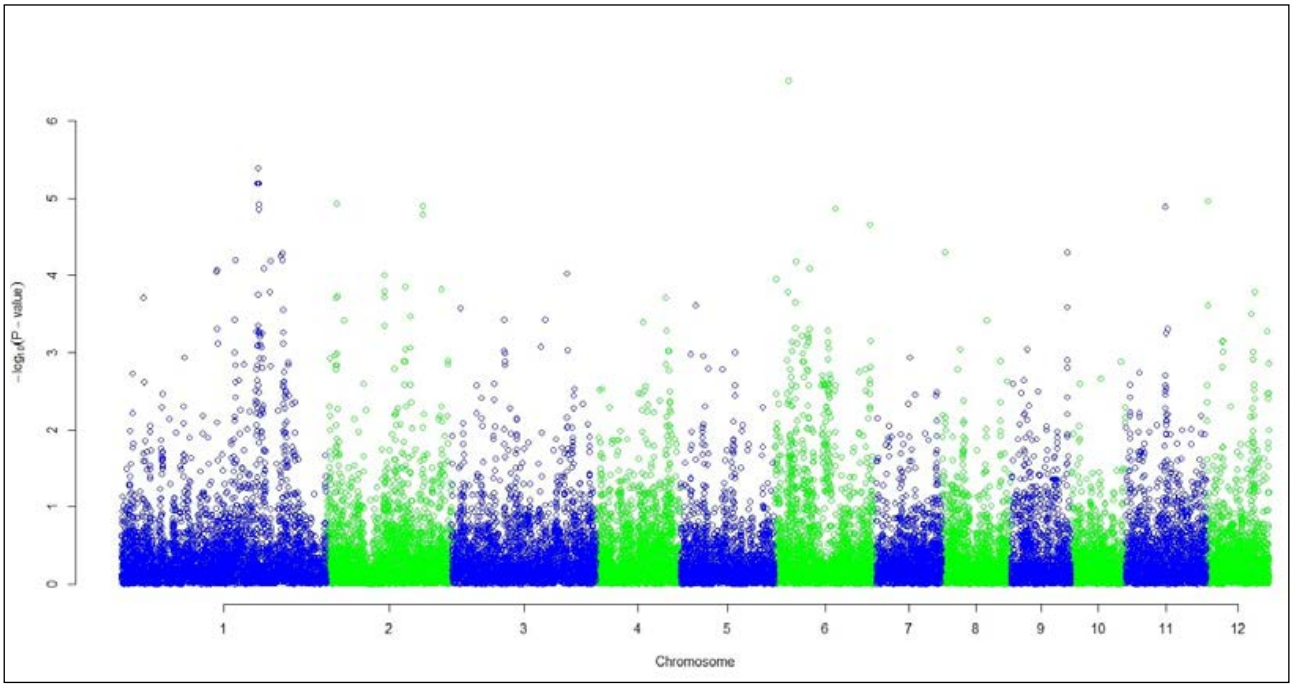

Figure 2. Quantile-Quantile plot of GRAMMAR GWAS result.

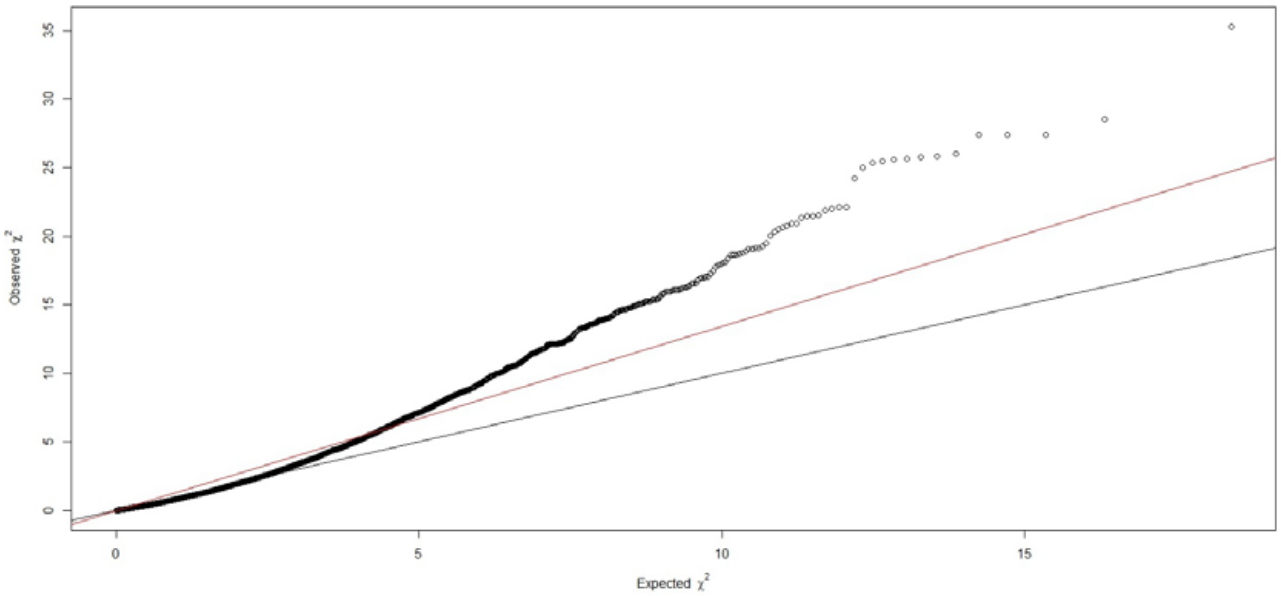

Table 2. Summary of Bayesian mixture model analyses for full data (Data1) and reduced data (Data2).

	Data1 (n=413)		Data2 (n=154)	
Chromosome	**%Genetic Variance**	**# of SNPs**	**%Genetic Variance**	**# of SNPs**
1	0.23	108	0.19	583
2	0.07	69	0.14	377
3	0.19	85	0.15	418
4	0.03	52	0.06	241
5	0.09	45	0.11	296
6	0.14	63	0.12	308
7	0.03	43	0.08	174
8	0.02	36	0.07	187
9	0.05	38	0.09	203
10	0.05	29	0.03	162
11	0.07	39	0.09	240
12	0.05	37	0.06	187

Table 3. Summary of Bayesian mixture genome wide association model for Data1.

Chromosome	SNP	Base Pair	Effect	%Genetic Variance
1	dd1000754	7334172	12.30	0.06
3	id3016879	34446028	9.34	0.05
5	id5013556	27621664	9.07	0.04
3	id3001242	2224752	6.63	0.03
10	id10001390	4682100	6.03	0.03
6	id6002006	2655955	5.65	0.03
1	id1024441	38537795	5.61	0.03
1	id1018291	30408458	5.26	0.03
3	id3016453	33688227	4.92	0.02
6	id6006541	10552135	4.89	0.02

Table 4. Summary of Bayesian mixture genome wide association model for Data2.

Chromosome	SNP	Base Pair	Effect	%Genetic Variance
11	id11007149	18919417	11.48	0.03
7	id7004642	25018437	5.49	0.02
1	id1018227	30312328	5.37	0.02
2	id2003308	6479823	4.86	0.01
2	id2005948	14158673	4.70	0.01
9	id9005931	17645328	3.98	0.01
6	id6009548	16836564	3.92	0.01
5	wd5001329	10805291	3.80	0.01
1	id1028014	42619920	3.57	0.01
2	id2011435	26085576	3.36	0.01

References

Aulchenko YS, Ripke S, Isaacs A and van Dujin CM (2007). GenABEL: An R library for genome-wide association analysis. Bioinformatics. 23: 1294-1296.

Meuwissen THE, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome wide dense marker maps. Genetics. 157:1819-1829..

Moser G, Lee HS, Hayes BJ, ... and Visscher MP (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model, PLoS. Genet. 11: e1004969.

Turkheimer E (2011). Still missing. Res. Hum. Dev. 8: 227-241.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK,... and Montgomery GW (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565-569.

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH,... and McClung AM (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nature. Com. 2: 467.