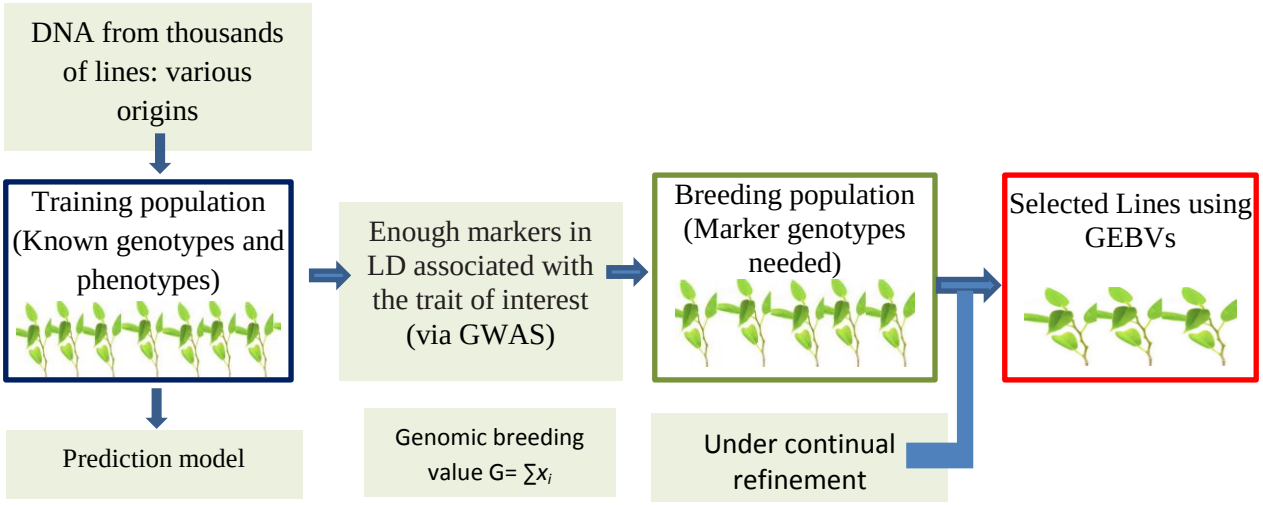
Incorporating molecular data-driven decisions towards soybean improvement

—Milestone report

1. Introduction

Many agriculturally important traits such as yield, quality and some forms of disease resistance are controlled by many genes and are known as quantitative traits. It's labor intensive and time consuming to evaluate those traits. The strategy of genomic selection (GS) estimates the genetic merit of an individual based on molecular genetic information by simultaneously accounting for all DNA markers. GS is a form of marker-assisted selection that selects progeny lines or parents based on the estimated genomic estimated breeding values (GEBVs), which leads to shorter breeding cycle duration as it is no longer necessary to wait for late filial generations' performance trial. The problem I want solve herein is to build a genomic selection model for soybean complex traits (yield, protein and oil content).

The essential steps of GS are also shown as following:



1.1 Data sets

GS usually needs a large amount of molecular markers covering the entire genome in term of all relevant genomic regions are represented by molecular markers and phenotypic data. The complete data set I will use is for 20,087 G. max and G. soja accessions genotyped with 42,509 SNPs is available for Wm82.a1 and Wm82.a2 in HapMap format. It's public available through the following link: https://soybase.org/snps/index.php. The phenotypic data is from a published paper https://www.g3journal.org/content/6/8/2329. Moreover, I also incorporated our own genotype and phenotype data (~1000 accessions) with the above big datasets.

1.2 Data Cleaning

For genotypic data, I adopted the following steps to perform data cleaning. Step 1: Checked the dimension of the two genotypic data sets and found that there were mismatches. Step 2: Merge two genotypic dataset with 'inner' option based on SNP id. Step 3: Found out common lines that have both genotypic and phenotypic data. Step 4: Created a subset of genotypic data that match phenotypic data. Step 5: Replace the missing values with the major value for genotypic data. Step 6: Unify the heterozygous genotype code among two genotypic data sets. Step 7: Drop monomorphic SNP data. Step 8: Drop SNP with minor allele frequency (MAF) less than 5%.Step 9: Drop SNP with high heterozygous ration (large than 10%).Step 10: Wrote out a cleaned genotypic data set.

As for phenotypic data, I adopted the following steps to perform data cleaning. Step 1: Checked the dimension of phenotypic data and found the mismatch with genotypic data. Step 2: Found out common lines that have both phenotypic and genotypic data. Step 3: Found duplicates in lines' Id and dropped those duplicates. Step 4: Using 1.5 IQR (<-1.5 IQR and >1.5 IQR) as thresholds to find out outliers. Step 5: Wrote out a subset of phenotypic data that match genotypic data and without duplicates and outliers.

There exists missing values and outliers in both genetic and phenotypic data sets. In genotypic dataset, I used 'major' genotypic code replaced missing values.. SNP data with monomorphic SNP, missing data more than 20% and high heterozygous(large than 10%) ratio were considered as outliers. I discarded those markers.. Relative high proportions of missing values were found in yield data. All missing values were discarded. In phenotypic data set, I found low proportion missing value for protein and oil (<0.001) trait Missing values were kept. There were outliers for all three traits. I used 1.5 IQR, namely <-1.5 IQR and >1.5 IQR, as thresholds to find out outliers.

2 Data Exploration

2.1 Data exploration for phenotypic data

First, describe statistics analysis were conducted for the three phenotypic traits (Table 1): Describe statics showed that there are large variations existing in three phenotypic traits. For instance, protein content varies from 37% to 51% with a mean of 44.3%. Oil content varies from 13% to 24% with a mean of 18.6%.

Table 1 Describe statistics analysis for three traits

	Yield ()	Protein	Oil
Count	7093	9642	9613
Mean	2.213421	44.284157	18.581638
S.t.d	0.991092	2.481386	2.008543
Min	0.030000	37.300000	13.000000
Max	4.958000	51.200000	24.100000

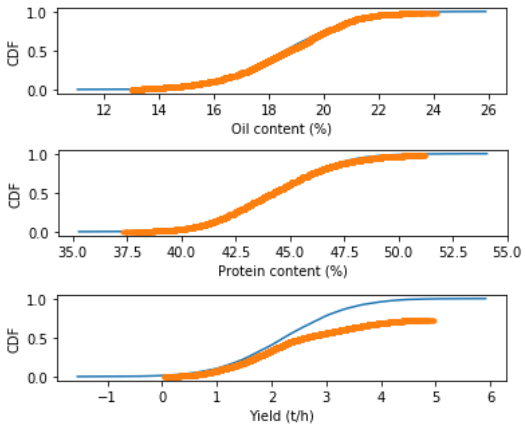


Figure 1 Empirical cumulative distribution functions plot for the three traits data.
Note: Blue lines stand for the ECDF of theoretically normal distribution, oranges line stand for ECDF of data distribution

Second, empirical cumulative distribution functions plot were drew to visualize the feature of the trait data in order from least to greatest and see the whole feature as if is distributed across the data set. For both protein and oil content, the ECDF of theoretically normal distribution overlapped the plot of ECDF for actual data, whereas the plot of ECDF for yield deviated from the theoretical distribution. Further statistics analysis of normal distribution was conducted to verify the ECDF plot results.

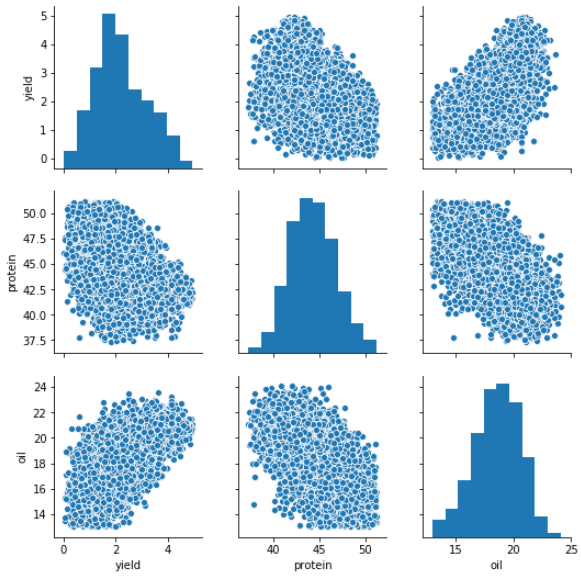


Figure 2 Histogram and scatter plot showing the relationships among the three traits.

Third, relationships among the traits were analyzed using scatter plot and correlation analysis (Figure 2). The results showed that protein and yield are negatively correlated. Protein and oil are negatively correlated, whereas oil and yield are positively correlated.

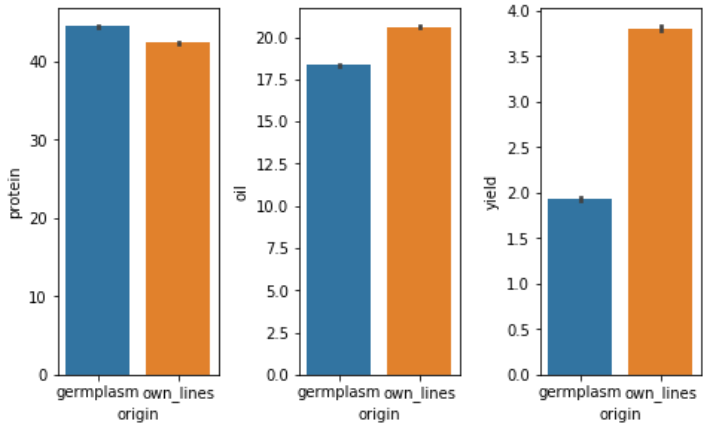


Figure 3 Bar plots showing the difference between our own lines and germplasms

Forth, the difference between my own data and data obtained from database (germplasm) were analyzed (Fig.3). Both two-sample t-test and permutation replicates tests showed that there existed significant difference in the three traits between the two origins.

2.2 Data exploration for genotypic data

After data cleaning for genotypic data, totally 35218 SNP were kept and distributed across 20 soybean chromosomes (Fig.4). Chromosome 18 has largest number of SNP markers (3229) , and Chromosome 20 has smallest number of SNP markers(1651).

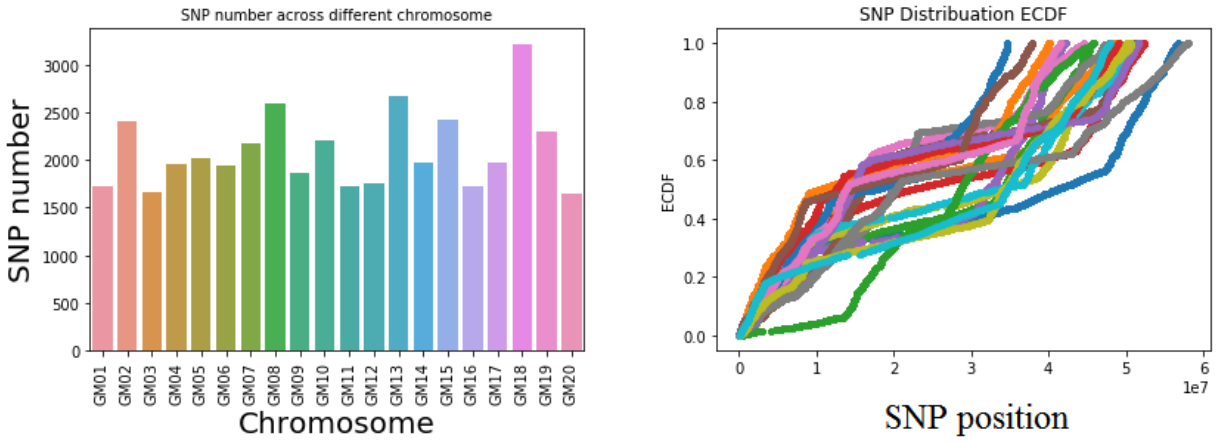


Figure 4 SNP number and position ECDF across 20 soybean chromosomes

Based on the ECDF plot of the position of those SNP markers, I found that SNP markers are not evenly distributed across genome. There exists different pattern of distribution for the SNP across different chromosomes. Based on the analysis of distance among SNP markers (Table 2), SNP density is also different among chromosomes. Chromosome 1 has the sparsest markers distribution with a mean of 33405 bp per maker. Chromsome 13 has the densest markers distribution with a mean of 17076 bp per maker.

Table 2 Distance distribution among different chromosome

Chr.	Mean	Std	Min	Max
GM01	33045	69569	886	1961970
GM02	20115	40427	427	1023340
GM03	27542	48431	747	603452
GM04	26698	39114	726	379485
GM05	20882	56288	174	1226291
GM06	26355	50252	141	1233091
GM07	20453	35173	526	610279
GM08	18387	40547	83	1309188
GM09	26663	48065	421	934179
GM10	23382	42166	115	886598
GM11	20202	38944	28	644414
GM12	22773	45465	424	926978
GM13	17076	36766	312	1196416
GM14	24852	46173	467	645588
GM15	21277	40807	872	681991
GM16	21976	39528	724	719942
GM17	20992	34526	284	411433
GM18	17970	29316	1186	513926
GM19	21975	34268	425	431415
GM20	29011	52088	660	1541797

2.3. Genome wide association study (GWAS) for the three traits

Using the GWAS to dissect genetic architecture of the three traits this association panels, I successfully identified both known associations (candidate genes or QTLs previously reported in soybean), as well as new candidate loci in the soybean genome. The results of significant SNPs discovered are summarized and plotted in and Figure 6. As shown in the quantile-quantile (QQ) plots (Figure 5), the distribution of observed -log10 P-values from the simple model, which included population structure (top 10 PC), departed from the expected distribution under a model of no association with significant inflation of nominal P-values for all three traits.

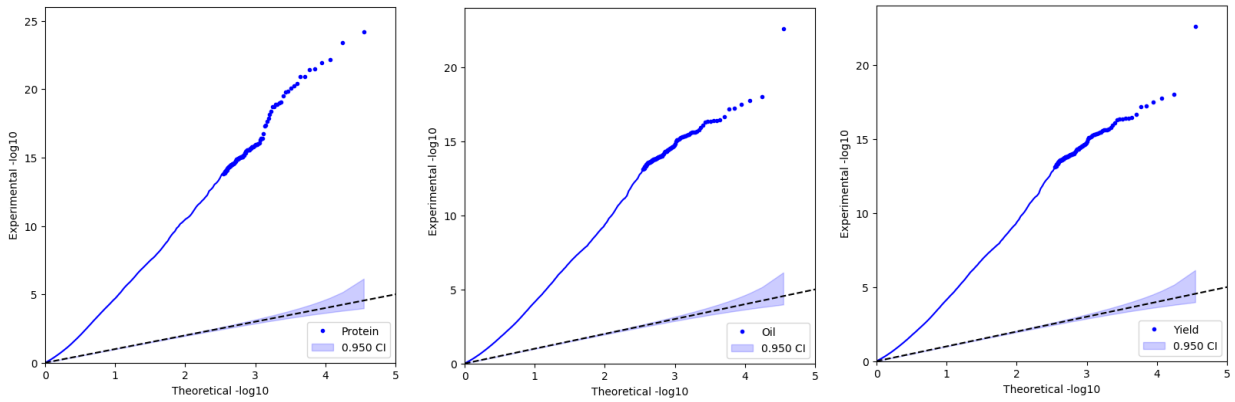


Figure 5 Quantile-quantile (QQ) plot of GLM model for the three traits.

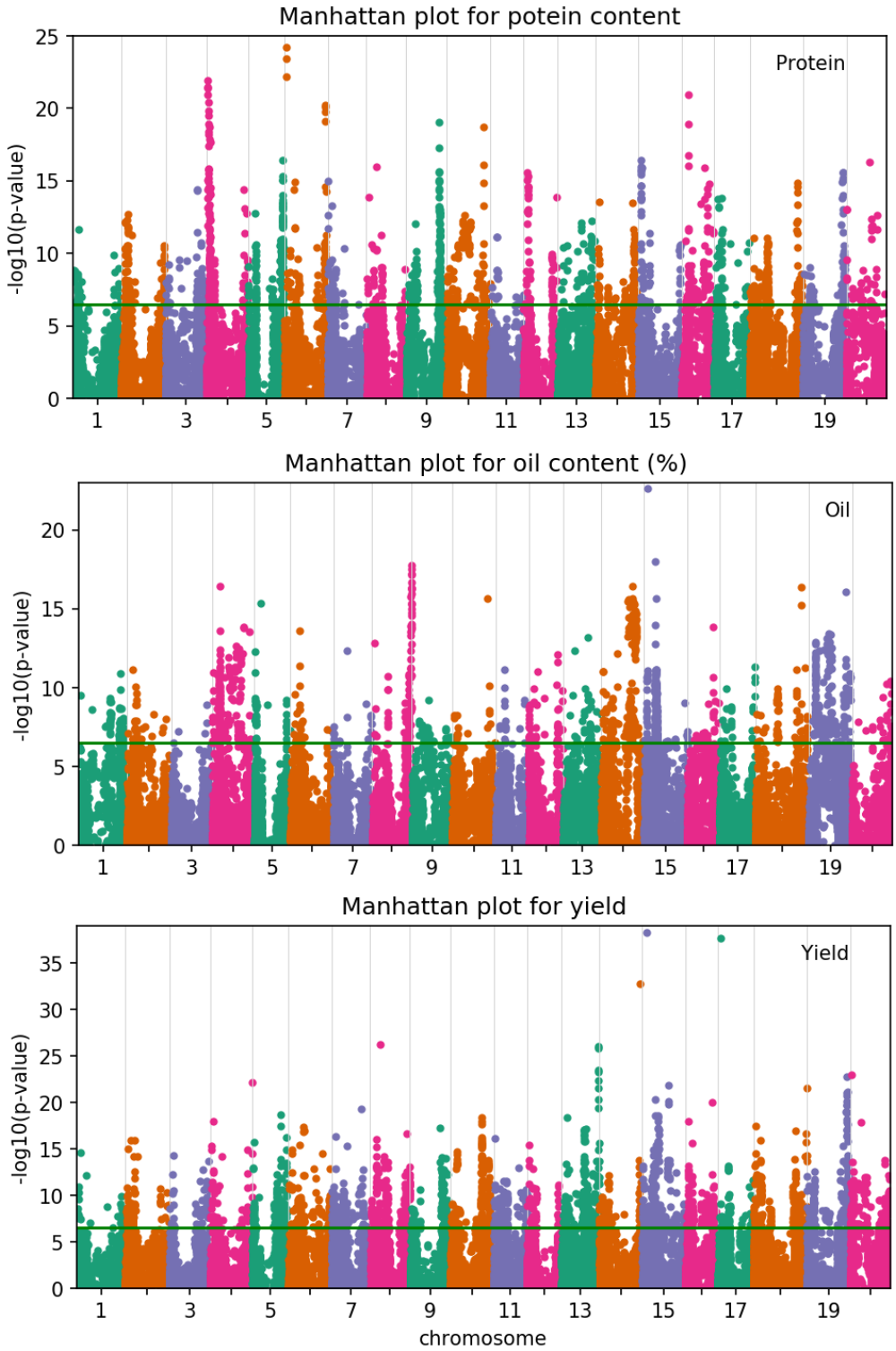


Figure 6 Genome-wide association study of protein, oil and yield based on GLM. (**a**) Manhattan plots of the simple model for in the association panel P1. The −log10 *P*-values from a genome-wide scan are plotted against the position on each of the 20 chromosomes. The horizontal red line indicates the genome-wide significance threshold (FDR<0.05

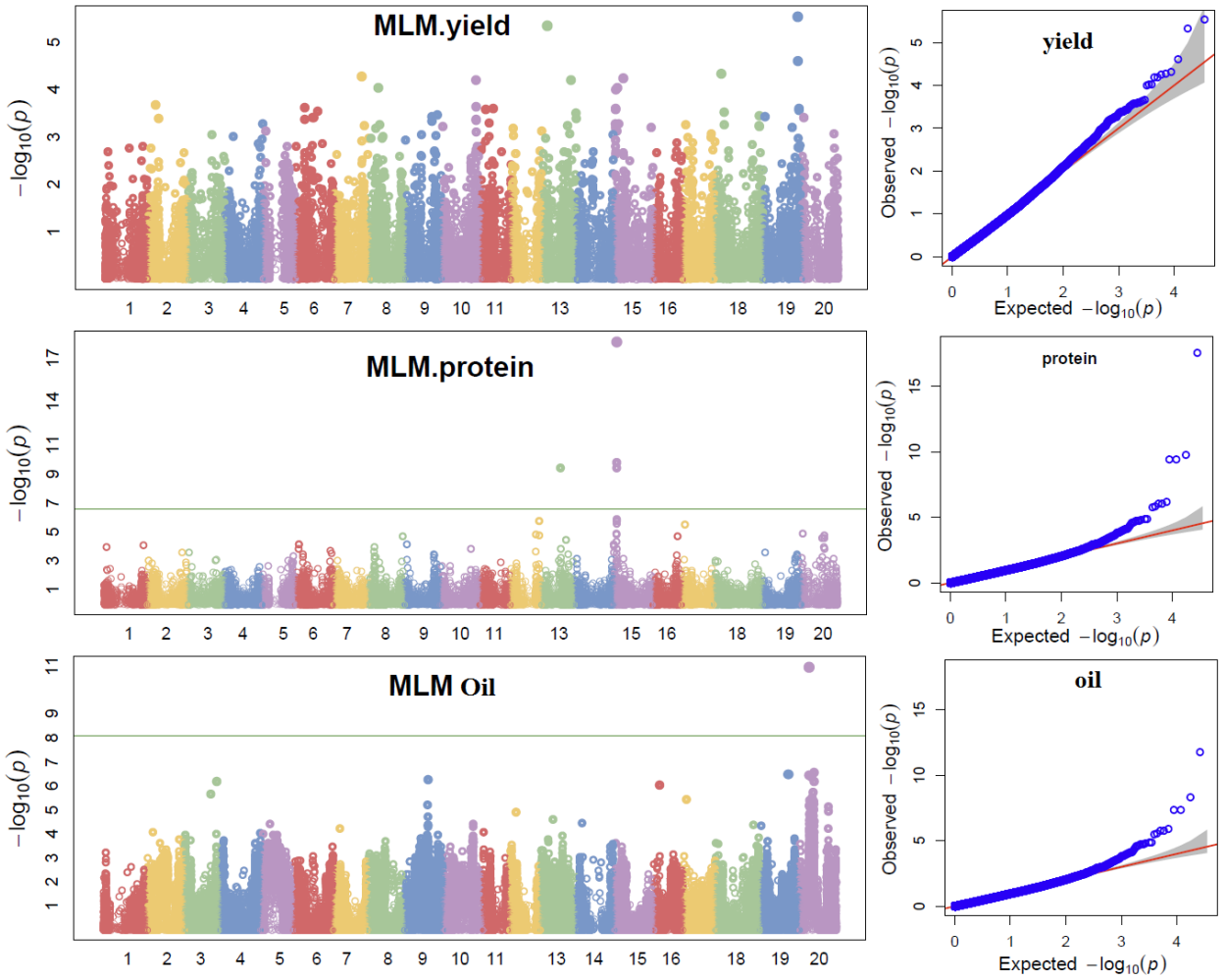


Figure 7 Genome-wide association study of protein, oil and yield based on MLM. (**a**) Manhattan plots of the simple model for in the association panel. The −log10 *P*-values from a genome-wide scan are plotted against the position on each of the 20 chromosomes. The horizontal red line indicates the genome-wide significance threshold (FDR<0.05).

While the MLM model, which includes top 3 PC and kinship matrix, allowed us to reduce the excess low *P*-values for DS, DI and DX (Figure 7). Based on the QQ plots, lower inflation of nominal *P*-values was consistently observed when the MLM model was used than when the simple model was used.

The following table (Table 3) listed the top 5 significant SNP associated with traits. Overall, the genome wide association study not only provide a basis for further efforts to pinpoint causal variants and to clarify how the implicated genes affect the three agronomic traits, but also provided a good candidates for feature selection of prediction models. In other words, only significant SNP associated trait will be selected as features to prediction the performance of the traits.

Table3 A subset of SNPs(top 5) significantly associated with the three traits

	SNP	Chromosome	Position	P-value	MAF	n	R^2	effect
Yield	Gm_19_44642440	19	44642440	2.93E-06	0.17	7041	0.74	0.07
	Gm_13_6704149	13	6704149	4.63E-06	0.32	7041	0.74	-0.05
	Gm_19_44761515	19	44761515	2.50E-05	0.47	7041	0.74	0.04
	Gm_18_7942395	18	7942395	4.76E-05	0.07	7041	0.74	0.10
	Gm_7_35103803	7	35103803	5.35E-05	0.12	7041	0.74	0.07
Protein	Gm_15_3828587	15	3828587	7.95E-19	0.12	9590	0.45	0.41
	Gm_15_3919945	15	3919945	1.66E-10	0.21	9590	0.45	-0.26
	Gm_13_24858209	13	24858209	3.86E-10	0.11	9590	0.45	-0.30
	Gm_15_3918803	15	3918803	4.06E-10	0.20	9590	0.45	0.26
	Gm_15_3702534	15	3702534	1.32E-06	0.03	9590	0.45	0.56
Oil	Gm_17_865220	17	865220	7.76E-07	0.07	9590	0.36	0.45
	Gm_17_4249592	17	4249592	1.14E-06	0.07	9590	0.36	-0.46
	Gm_20_43423685	20	43423685	1.73E-06	0.11	9590	0.35	-0.32
	Gm_20_43428880	20	43428880	2.35E-06	0.11	9590	0.35	-0.32
	Gm_05_3049162	5	3049162	2.47E-06	0.07	9590	0.35	0.41