Incorporating molecular data-driven decisions towards soybean improvement

Zixiang Wen

Department of Plant, Soil and Microbial Sciences

MICHIGAN STATE UNIVERSITY

Outline

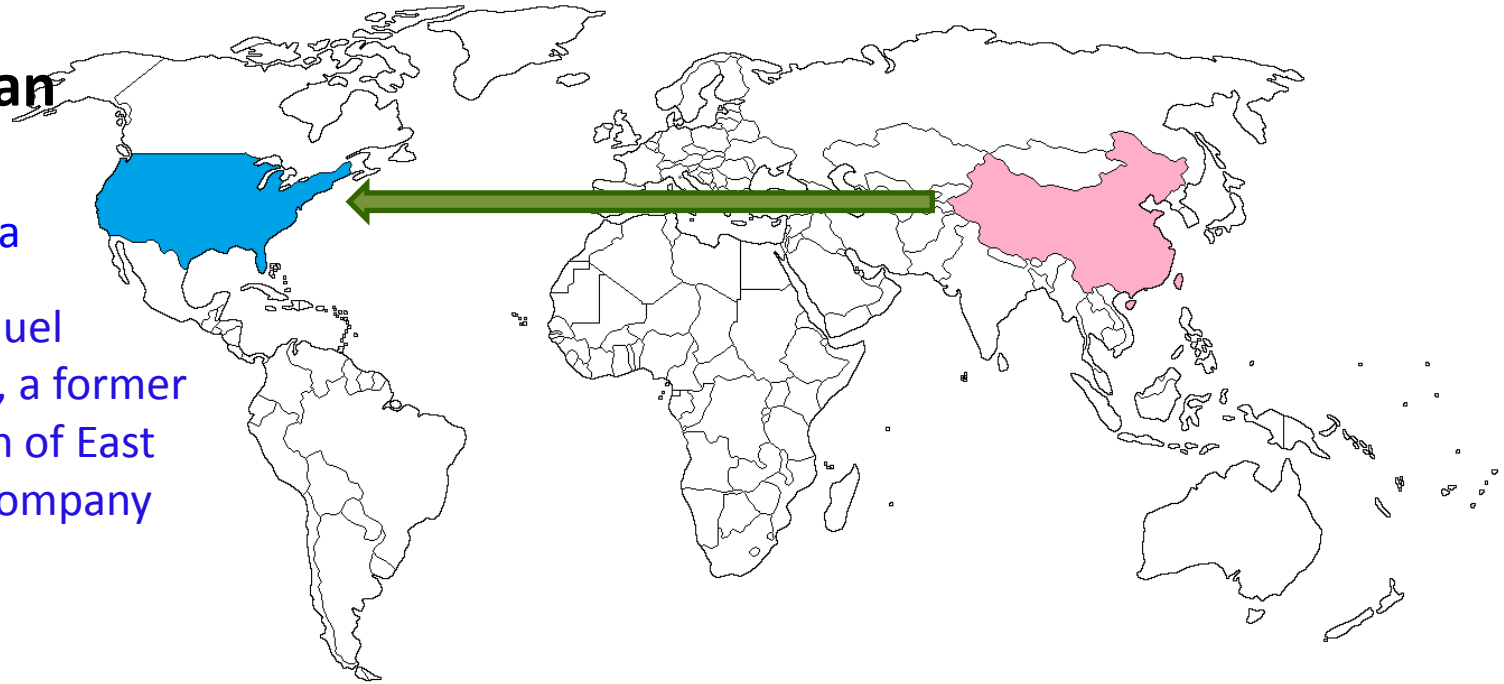
- **1. Introduction**
- **2. Data Exploration**
- **3. Machine learning for prediction of the three traits**
- **4. Summary**

Introduction

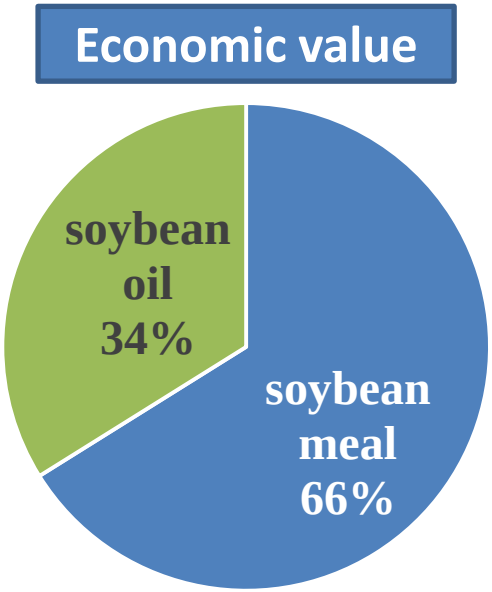
Soybean

1765,
Georgia

By Samuel Bowen, a former seaman of East India Company

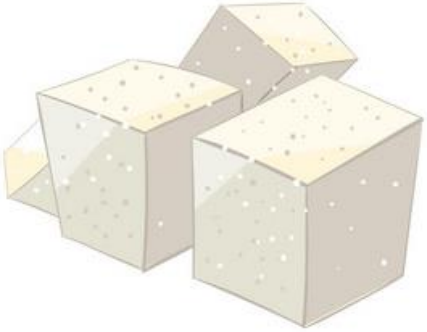


Soybean uses



1 bushel soybean= 11lbs oil +44lbs meal

Soy sauce

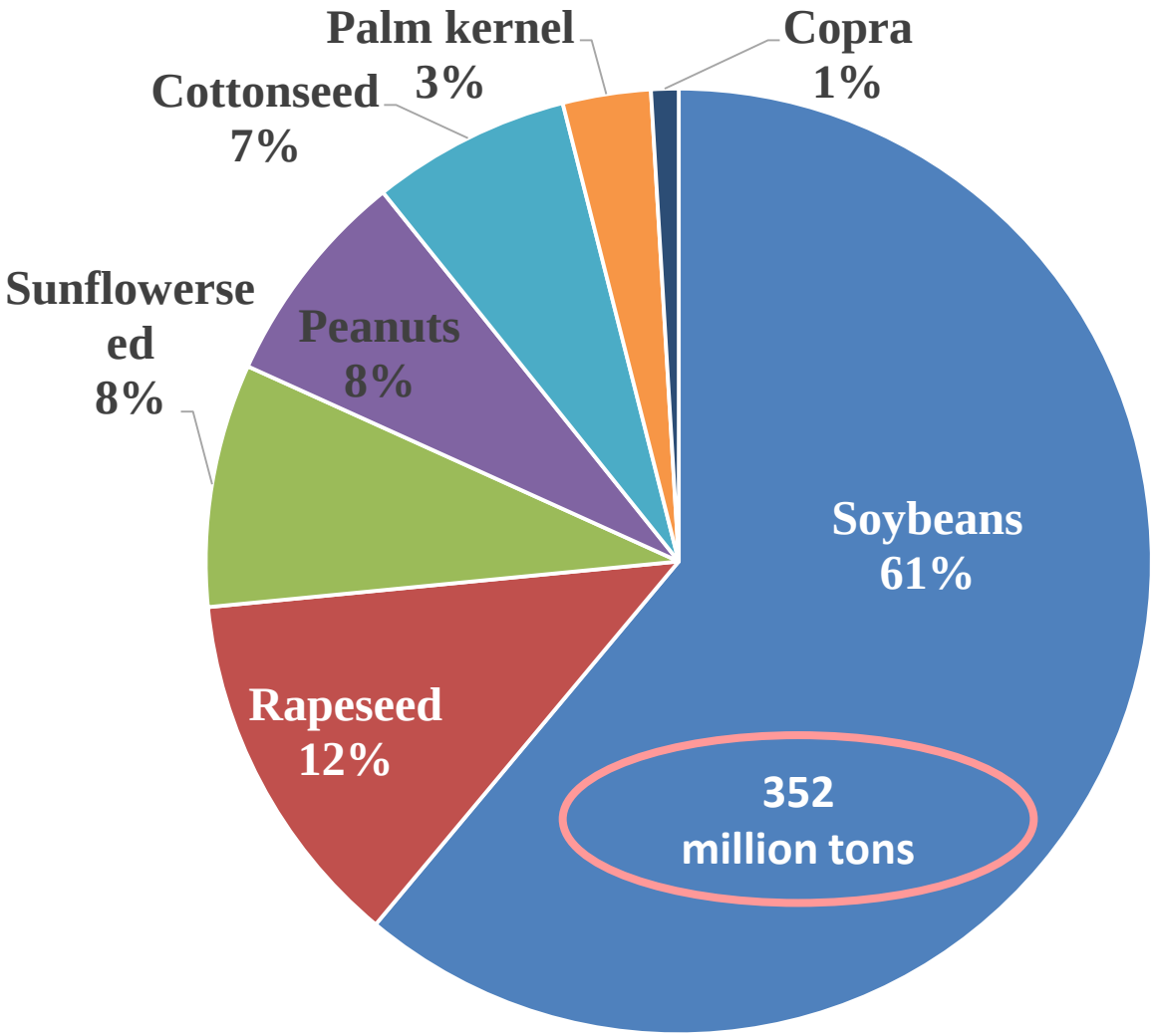


Tofu

Soy milk

Soy sprouts

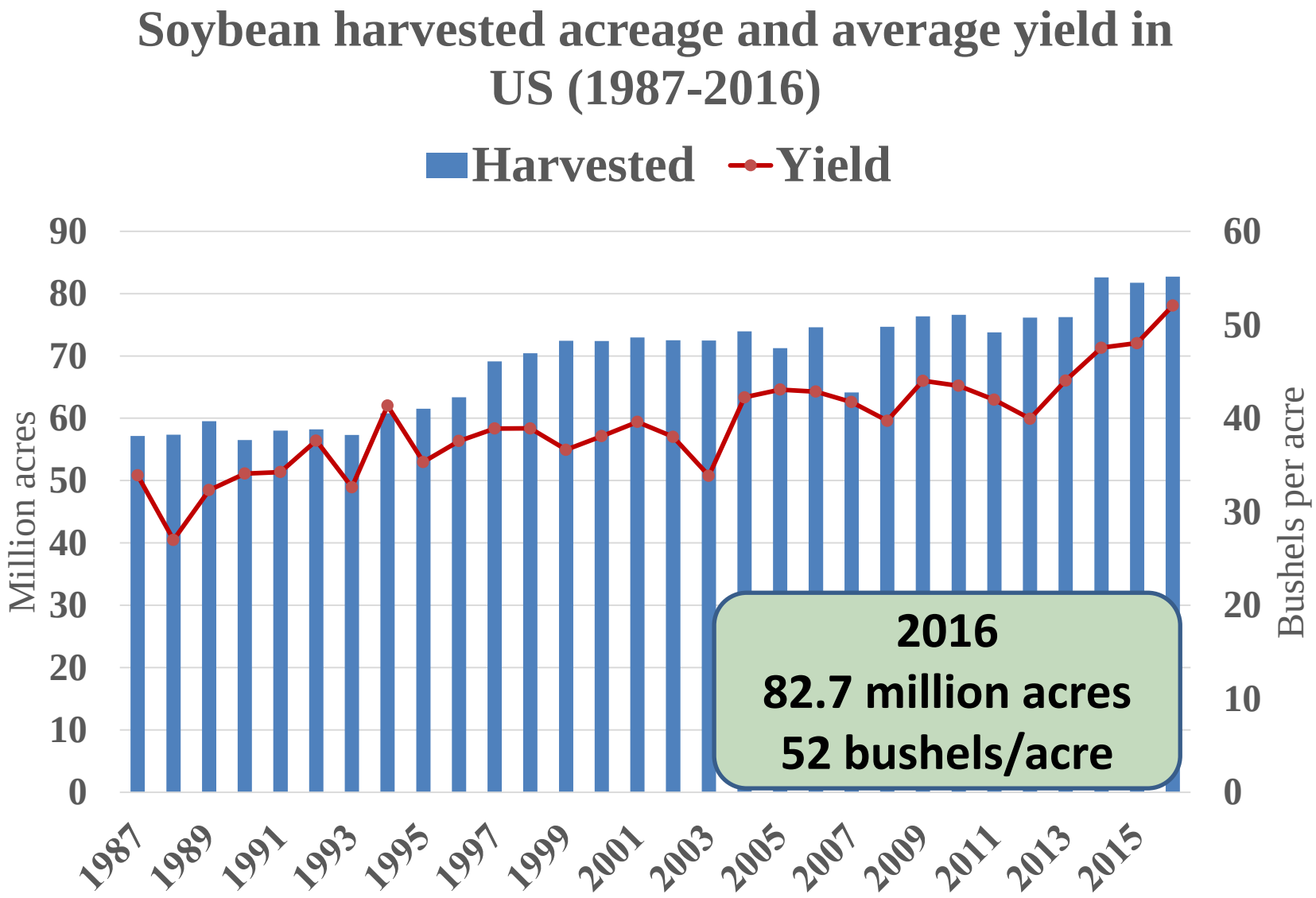
Soybean production



Worldwide oil seed production in 2016/2017

Source: https://www.statista.com/statistics/267271/worldwide-oilseed-production-since-2008/

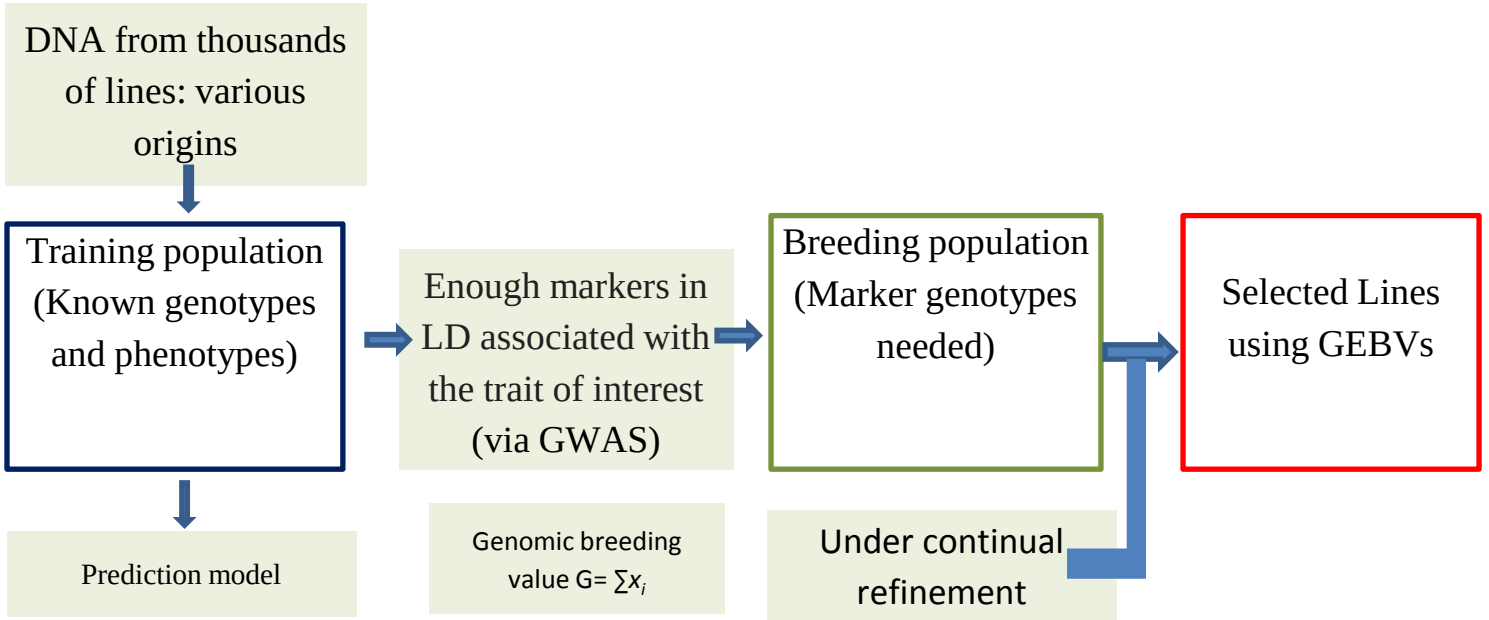
Soybean production



Breeding goals of soybean

- High yield
- High protein
- High oil
- Disease resistance (SDS, aphid, SCN,white mold,phytophthora)

The essential steps of GS are also shown as following:



Data sets

- Data Set 1: 20,087 *G. max* and *G. soja* accessions

Genotypic data: https://soybase.org/snps/index.php

Phenotypic data: ttps://www.g3journal.org/content/6/8/2329

- Data Set 2: 1100 elite soybean accessions from our own breeding program

Genotypic data
: https://drive.google.com/open?id=1E9N505WiXjVG0a6CjOKKTiVY3tkPWu-8

Phenotypic data:
https://drive.google.com/open?id=1dBnTHCeoW2o2yn0l46TbOAaM0jHnDazb

2.1 Data exploration for phenotypic data

Table 1 Describe statistics analysis for three traits

	Yield (Mg/ha)	Protein	Oil
Count	7093	9642	9613
Mean	2.2	44.3	18.6
S.t.d	1.0	2.5	2.0
Min	0.0	37.3	13.0
Max	5.0	51.2	24.1

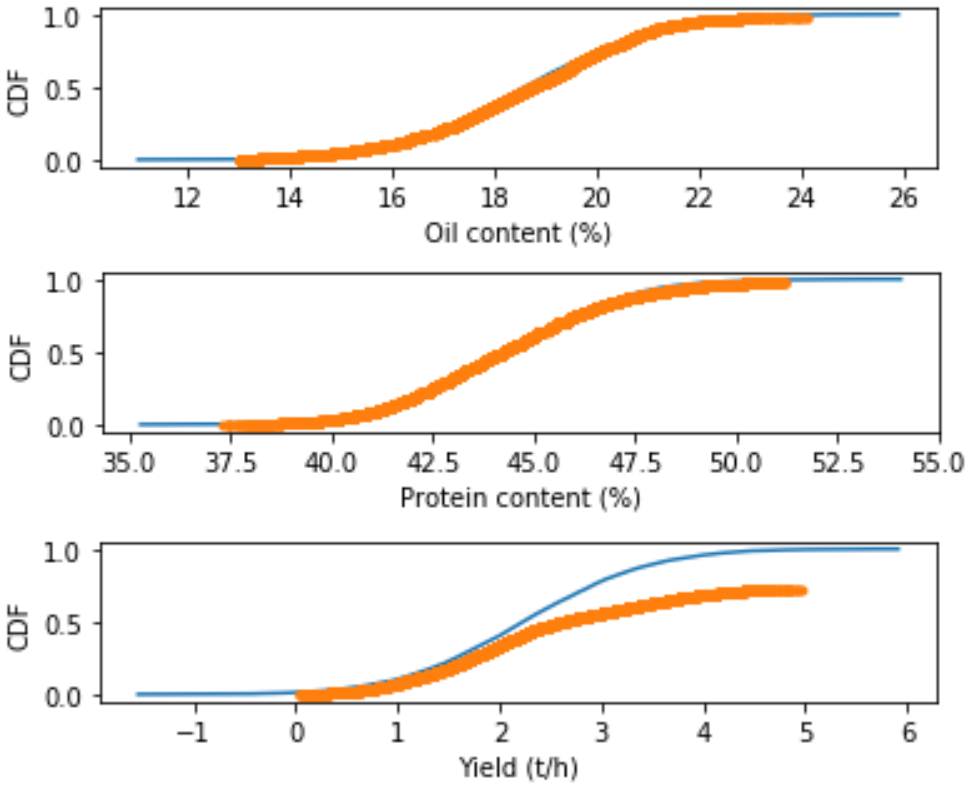


Figure 1 Empirical cumulative distribution functions plot for the three traits data. Note: Blue lines stand for the ECDF of theoretically normal distribution, oranges line stand for ECDF of data distribution

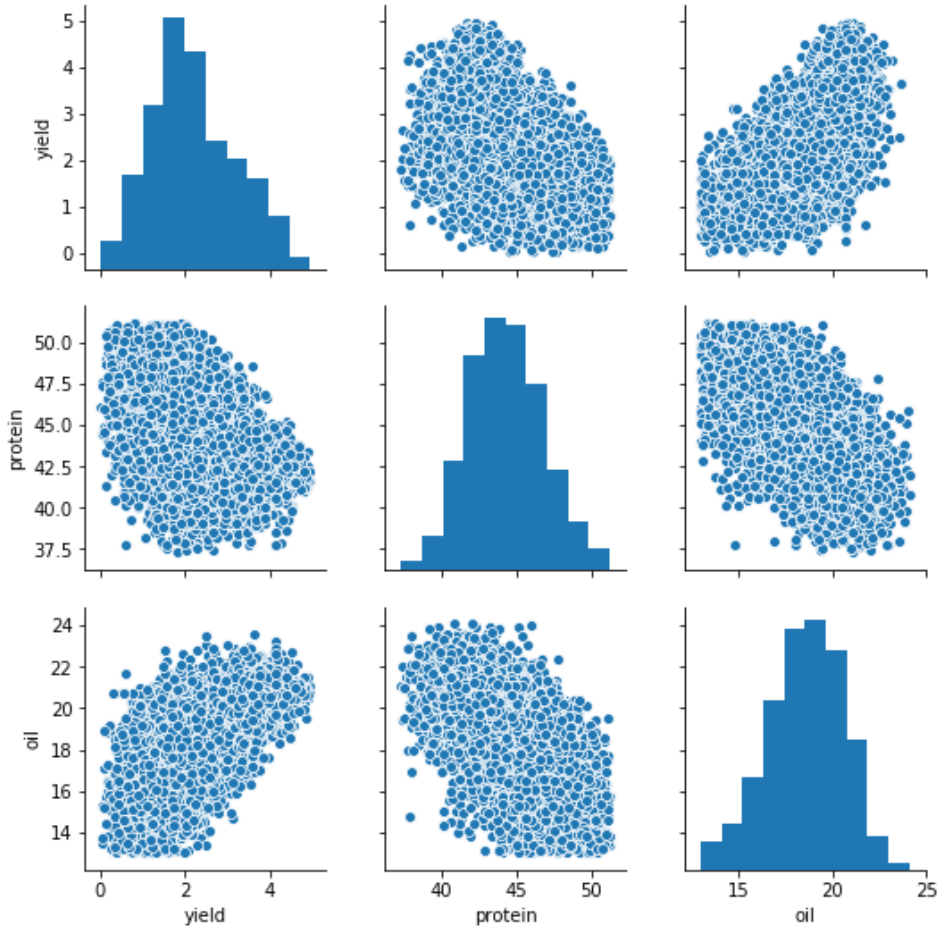


Figure 2 Histogram and scatter plot showing the relationships among the three traits.

The results showed that protein and yield are negatively correlated. Protein and oil are negatively correlated, whereas oil and yield are positively correlated.

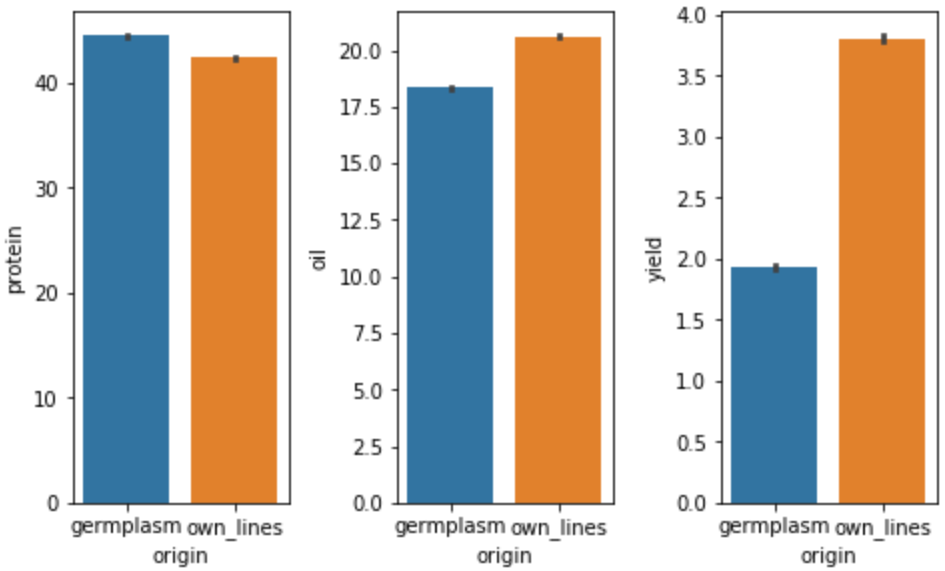


Figure 3 Bar plots showing the difference between our own lines and germplasms

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

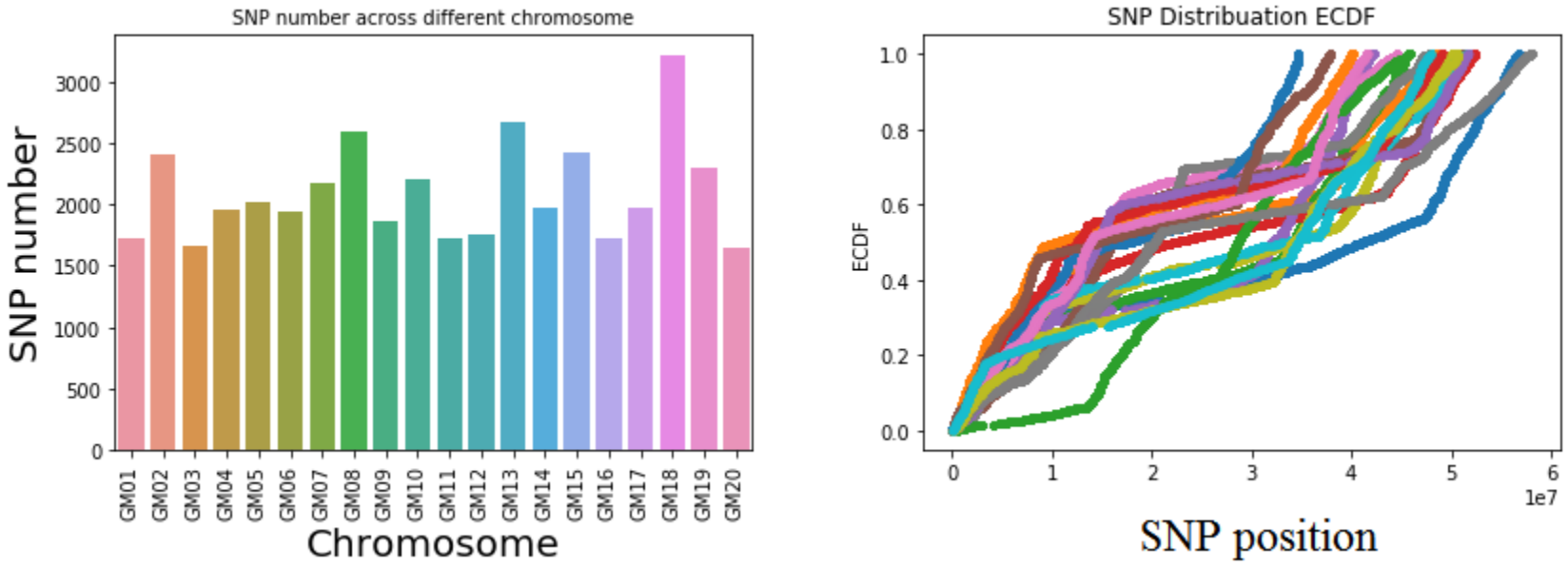


Figure 4 SNP number and position ECDF across 20 soybean chromosomes

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

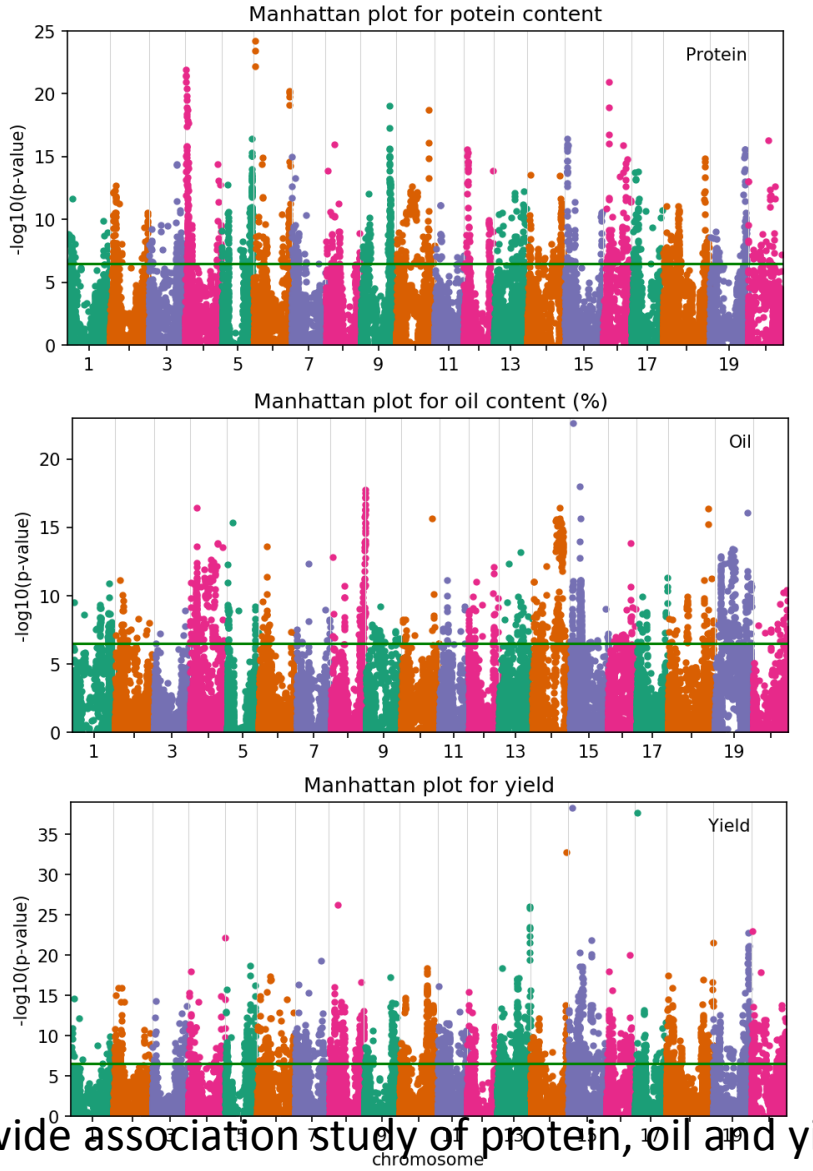


Figure 6 Genome-wide association study of protein, oil and yield based on GLM

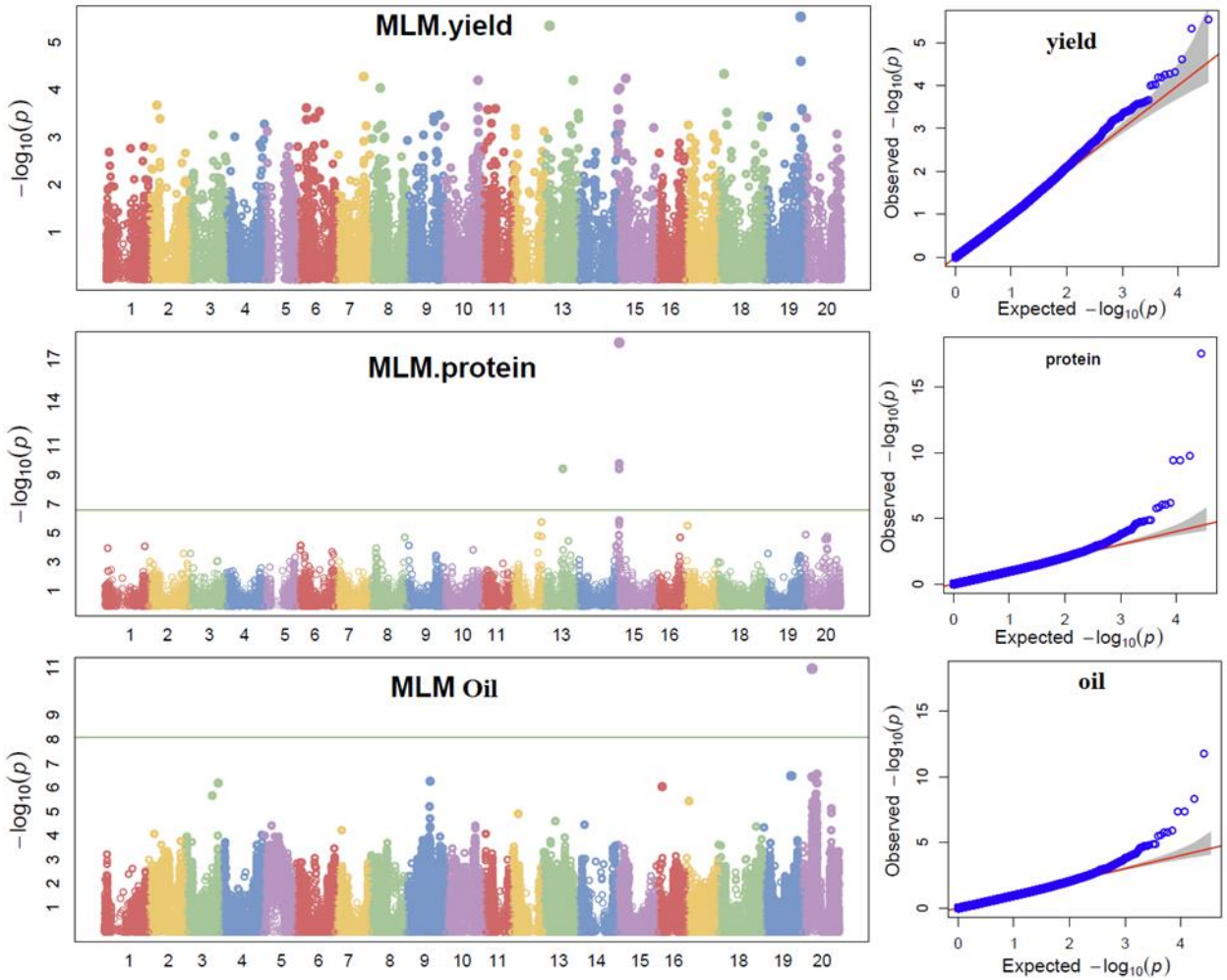


Figure 7 Genome-wide association study of protein, oil and yield based on MLM.

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

	SNP	Chr.	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

3. Machine learning for prediction of the three traits

To perform machine learning for the prediction of the three traits, tow algorithms were used. One is lasso (least absolute shrinkage and selection operator).

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The second algorithms used in the analysis is Random Forest regression. After training with 100 decision trees, predictions for unseen samples *x′* can be made by averaging the predictions from all the individual regression trees on *x′*:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

3.1 Machine learning for prediction of protein content

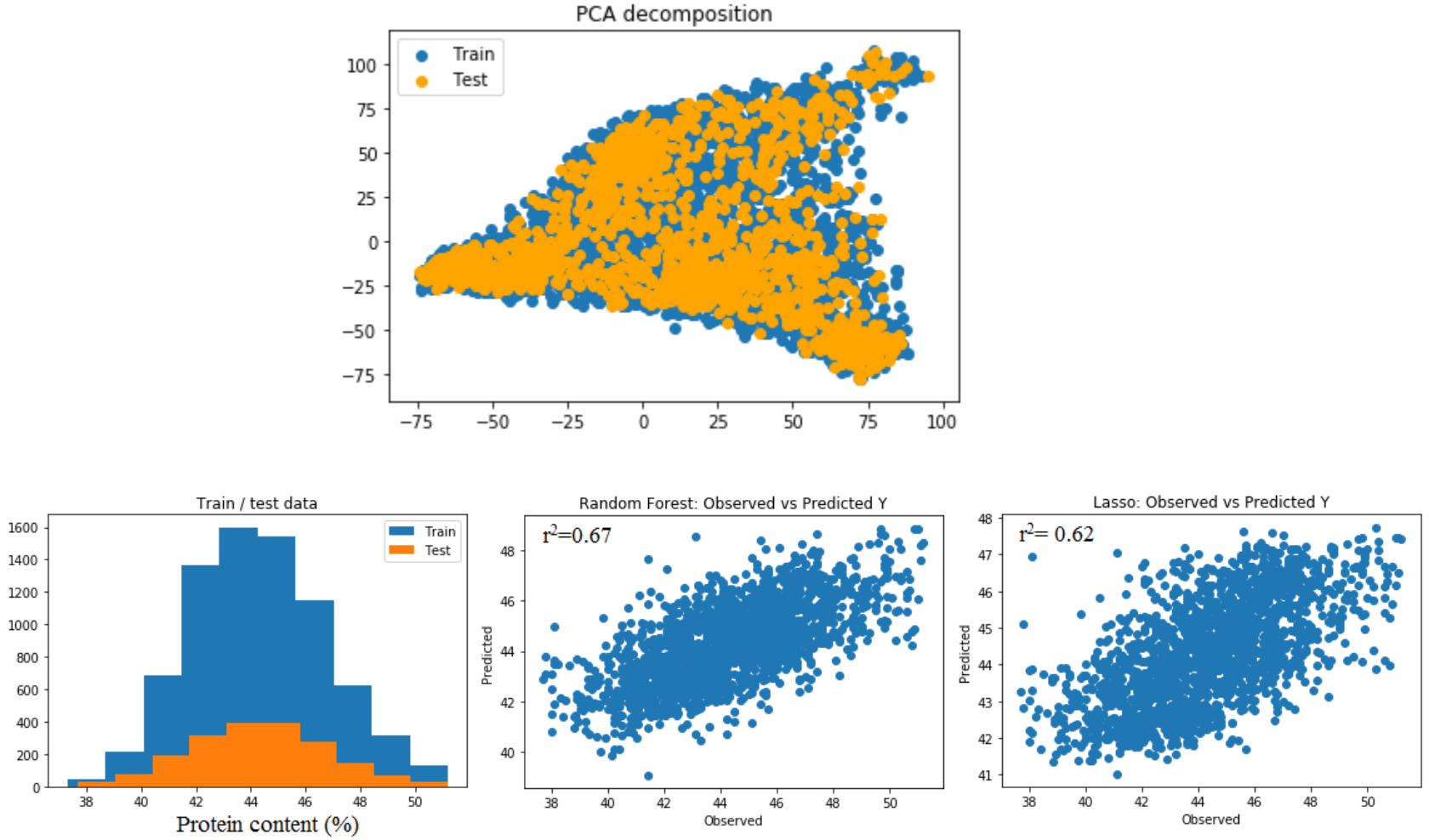


Figure 8 Genomic predictions for protein content with Lasso and Random Forest

3.2 Machine learning for prediction of oil content

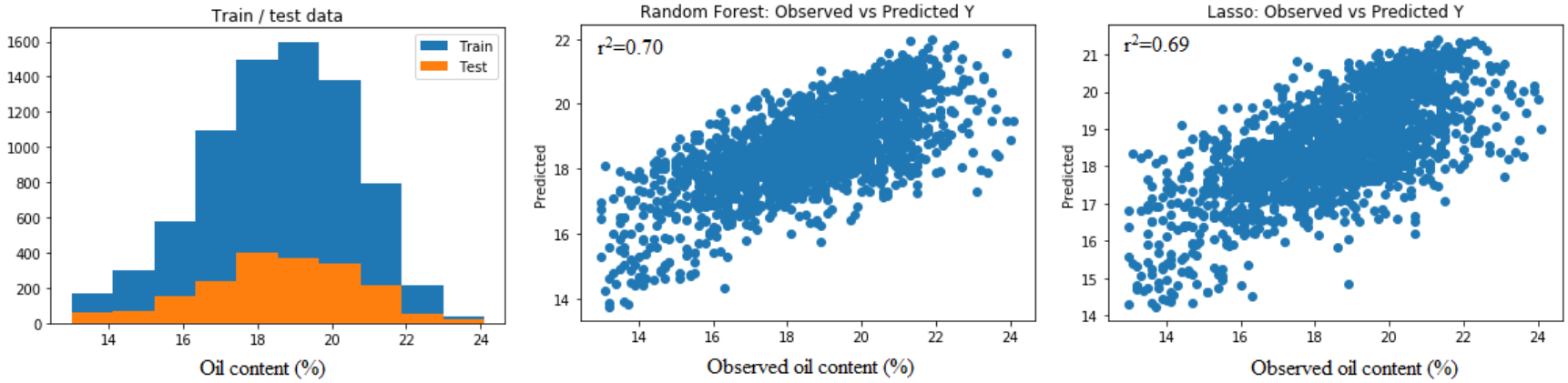


Figure 9 Genomic prediction for oil content with Lasso and Random Forest

- Correlations between predicted and observed reaching up to 0.70 with Random Forest model
- Random Forest model slightly outperformed the Lasso model in term of accuracy in test data.

3.3 Machine learning for prediction of yield

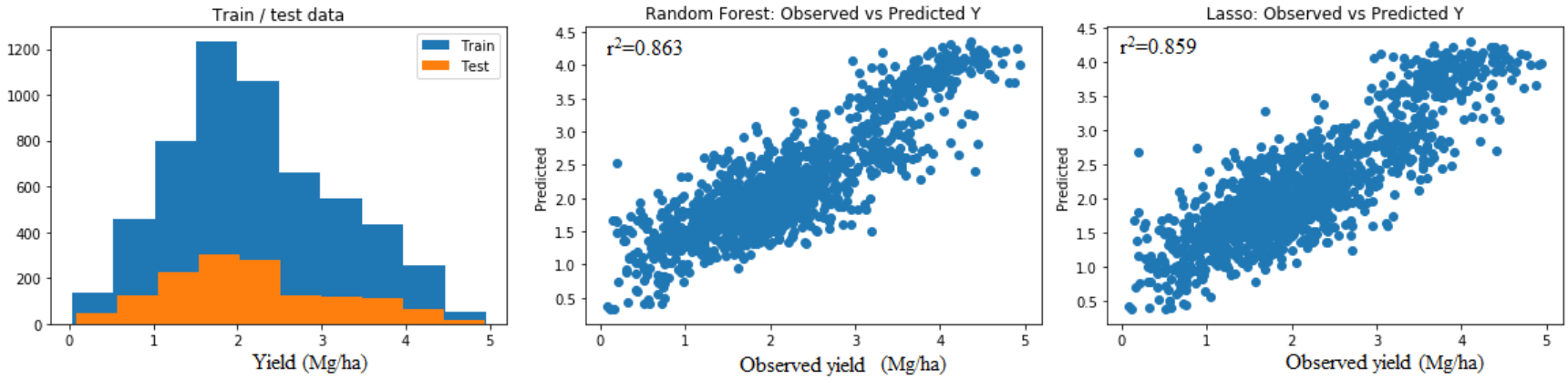


Figure 10 Genomic prediction for yield with Lasso and Random Forest

- Correlations between predicted and observed reaching up to 0.863 with Random Forest model.
- However, Lasso model outperformed Random Forest the model in term of accuracy in cross validation.

4. Summary

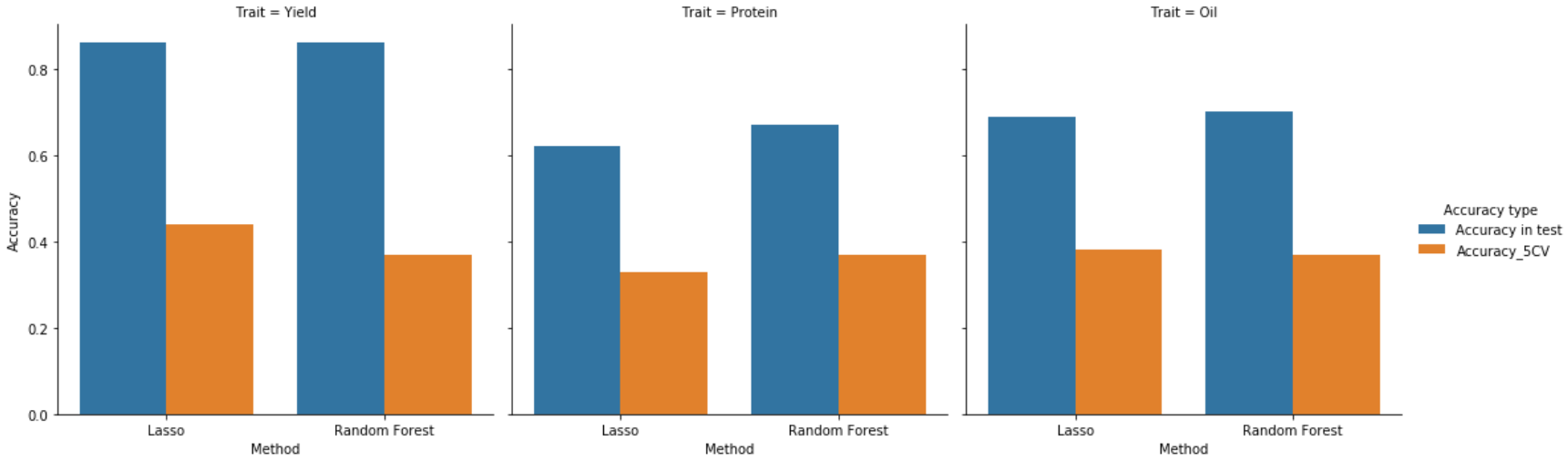


Figure 11 Genomic prediction accuracy for three traits with Lasso and Random Forest models

Resulting genomic prediction models explained an appreciable amount of the variation in accession, with correlations between predicted and observed reaching up to 0.86 for yield, 0.70 and 0.67 for oil and protein respectively