G2F Case Study Report

Zixiang Wen
Department of Plant, Soil and Microbial Sciences
MICHIGAN STATE UNIVERSITY

Data types and possible strategies

Phenotypic data:

- Inbreed phenotypic data

  2014_inbred_phenotypic_data.csv, 2015_inbred_phenotypic_data.csv

- Hybrid phenotypic data

  g2f_2017_hybrid_data_clean.csv , g2f_2016_hybrid_data_clean.csv

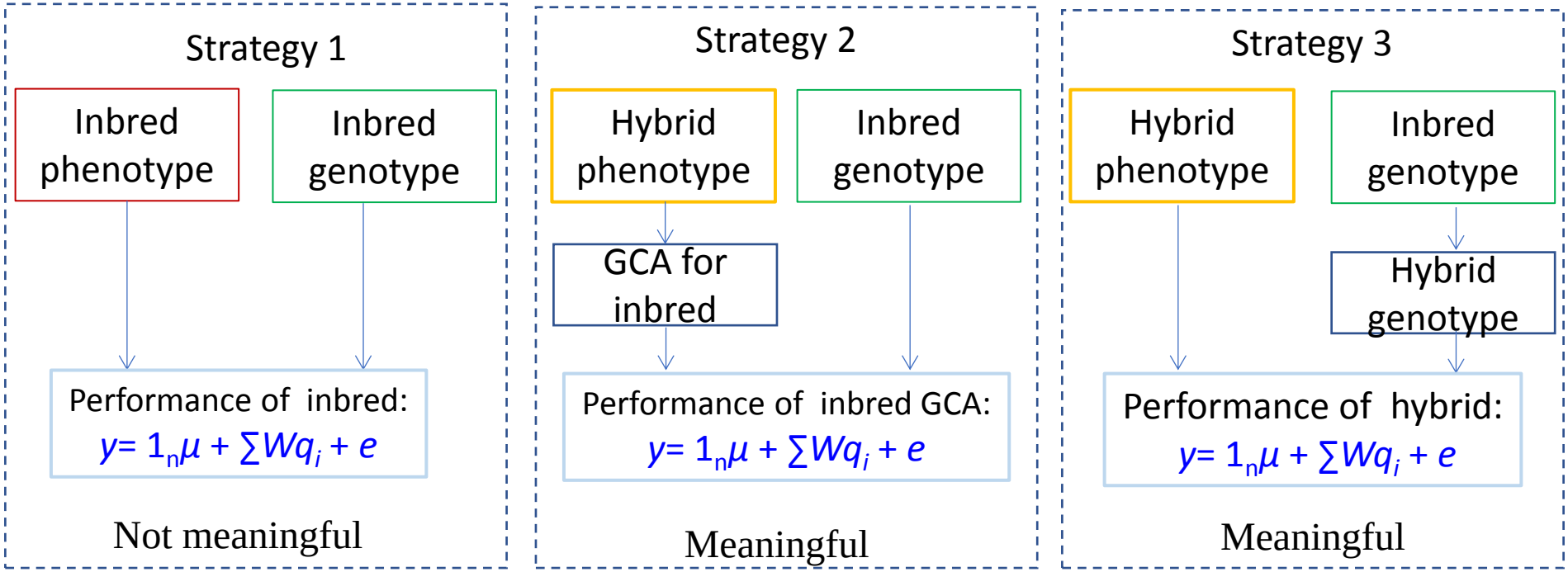
  g2f_2015_hybrid_data_clean.csv , g2f_2014_hybrid_data_clean.csv

Genotypic data:

- Inbred genotypic data:

  g2f_2017_ZeaGBSv27_Imputed_AGPv4.h5

Three strategies



Part 1:

Data cleaning

Datasets have been processed

Phenotypic data:

- g2f_2017_hybrid_data_clean.csv
- g2f_2016_hybrid_data_clean.csv
- g2f_2015_hybrid_data_clean.csv
- g2f_2014_hybrid_data_clean.csv

Genotypic data:

- g2f_2017_ZeaGBSv27_Imputed_AGPv4.h5

Phenotypic data cleaning

Step 1: Extract “Plant height”, “Silk DAP” and “Grain Yield” data

↓

Step 2: Calculate GCA for female parents grouped by location and replicate

↓

Step 3: Combine all GCA data collected across 4 years

↓

Step 4: Use >1.5 IQR as standard to find out outliers

↓

Step 5: Calculate BLUP of GCA for each female parental line (n=1129)

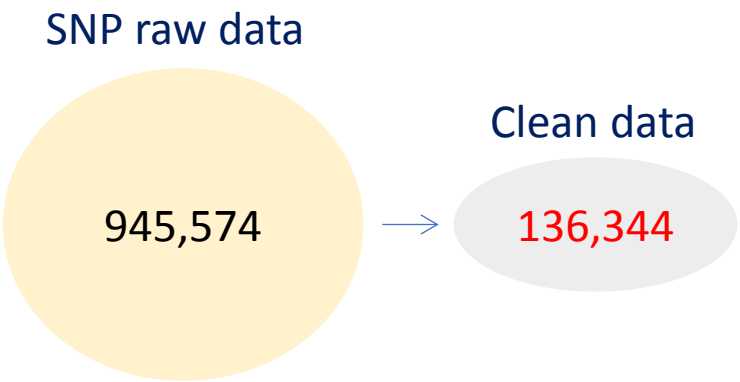
Genotypic data cleaning

Step 1: Convert h5 file to hapmap format (Tassel pipeline)

Step 2: Clean and unify genotype ID

Step 3: SNP pruning (MAF >0.05, heterozygous < 0.1, missing <0.35, LD<0.70)

Step 4: Impute and transform SNP data to numerical data (0,1,2)



Part 2:

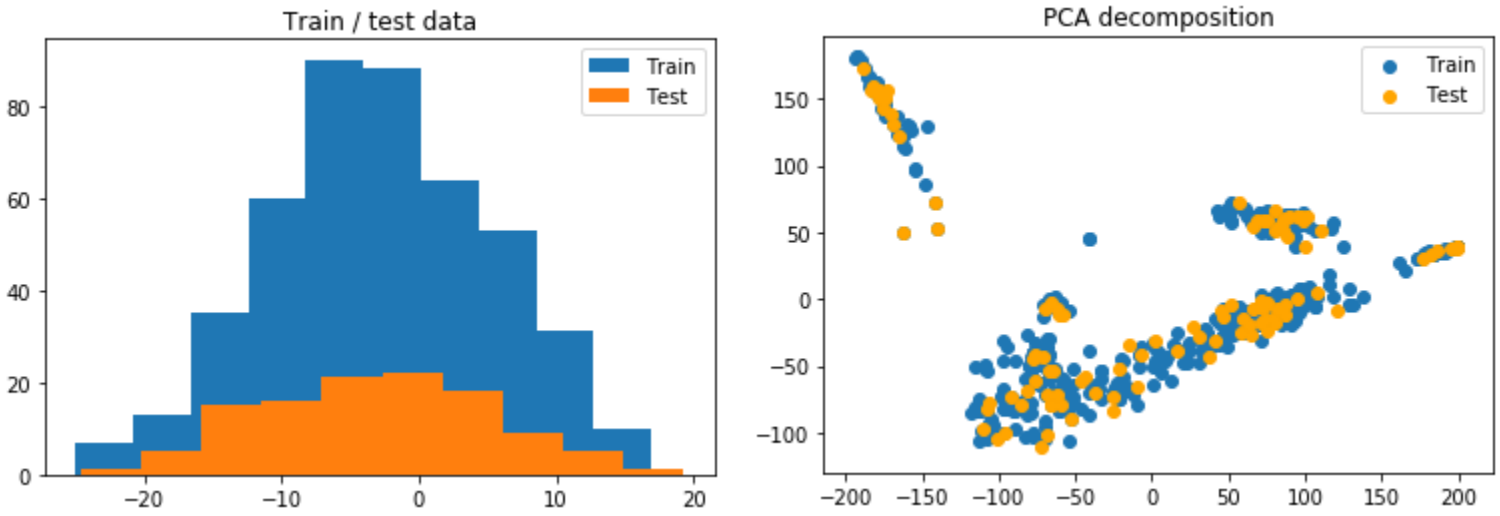
Genomic selection

Procedures for genomic prediction

Step 1: Combine phenotypic and genotypic data

Step 2: Split data into training and test datasets (5-fold)

Step 3: Check the distribution and coverage of the two dataset



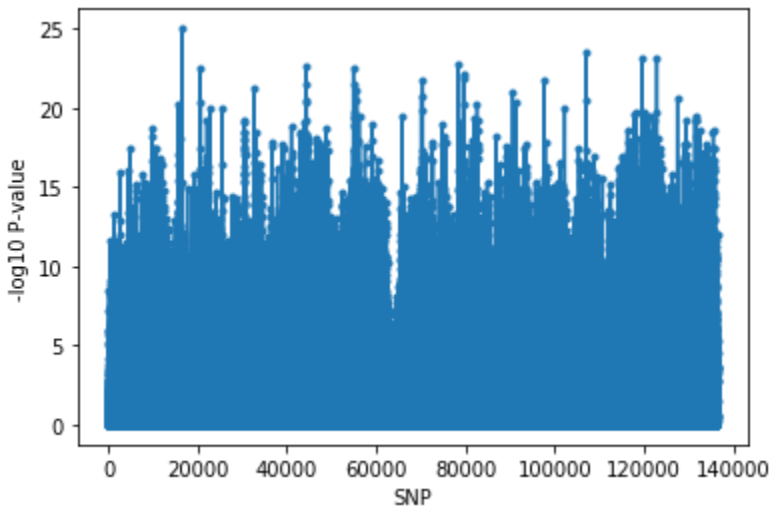
The distribution and coverage of the two dataset for yield

Step 4: “Best features” selection

Standard: 10k top SNP with lowest *P*-values in GWAS

Reasons: 1) Reduce overfitting caused by Multicollinearity

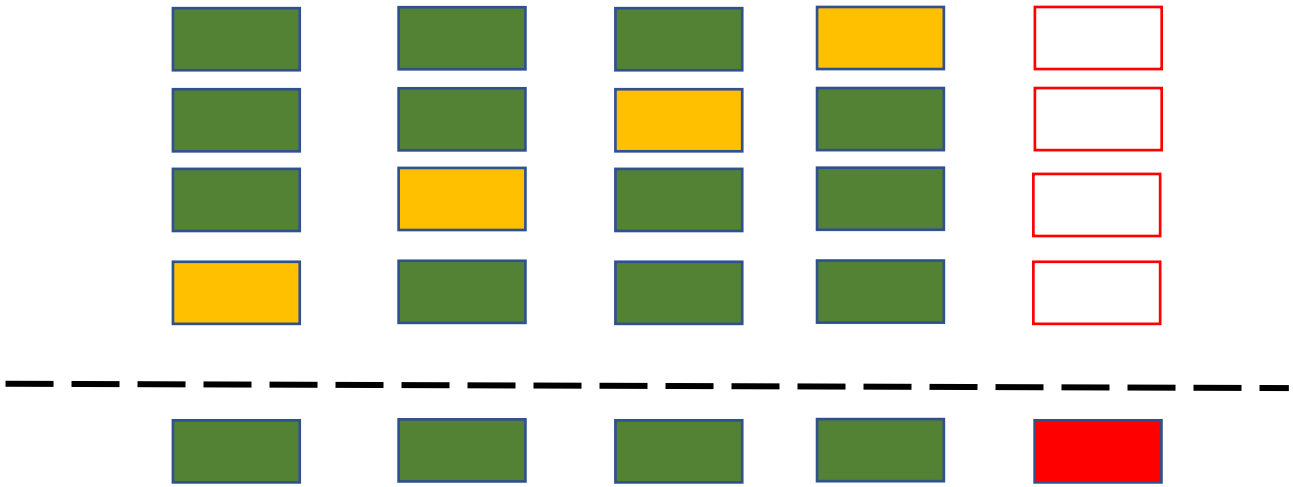
2) Save time



The distribution and coverage of the two dataset for yield

Step 5: Model training and fit

Training data: Green; Validation data: yellow ; Test data: Red



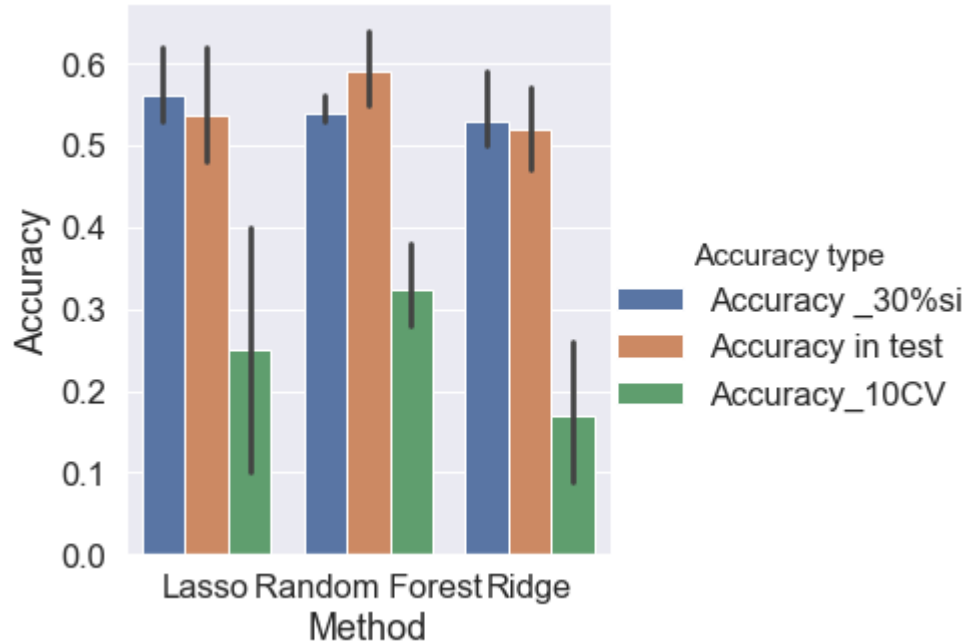
1. Take best parameters (Grid Search)

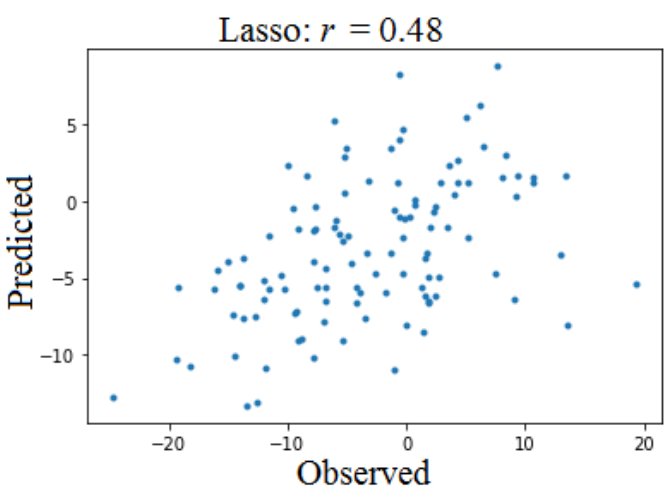
2. Train on training and validation data together (model.fit)

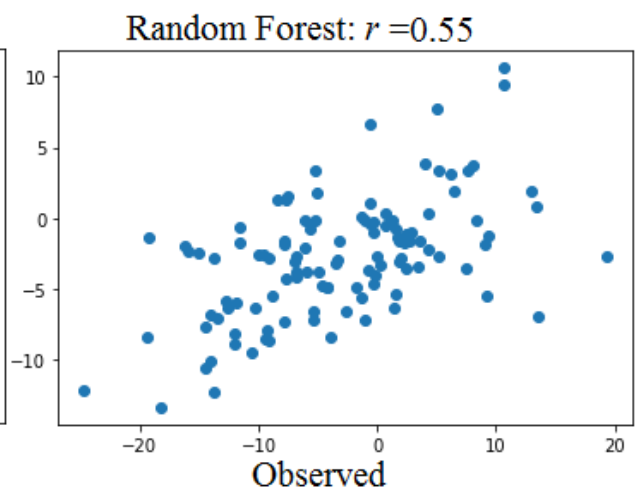
3. Test performance on test data (model.predict)

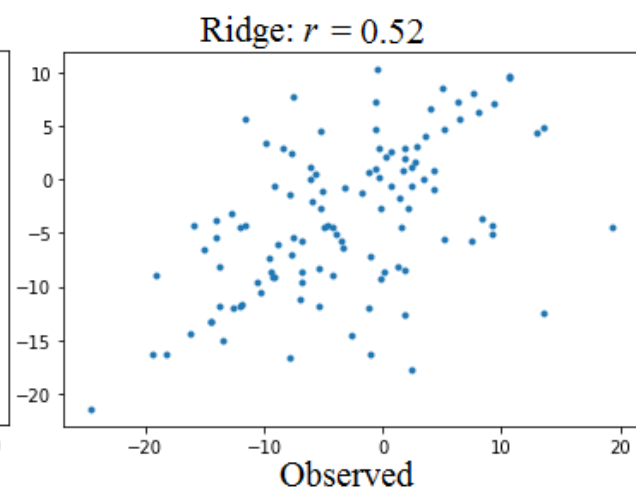
Genomic prediction results for yield

- Training size: 451
- Test size: 113
- Model: Lasso, Random forest, Ridge
- Accuracy: 0.55 ~ 0.64



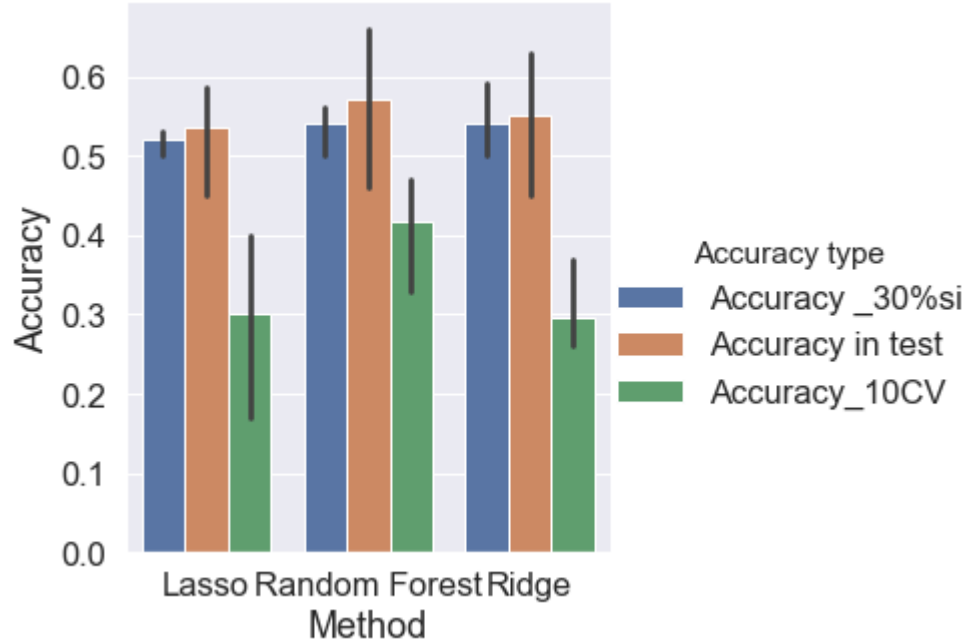


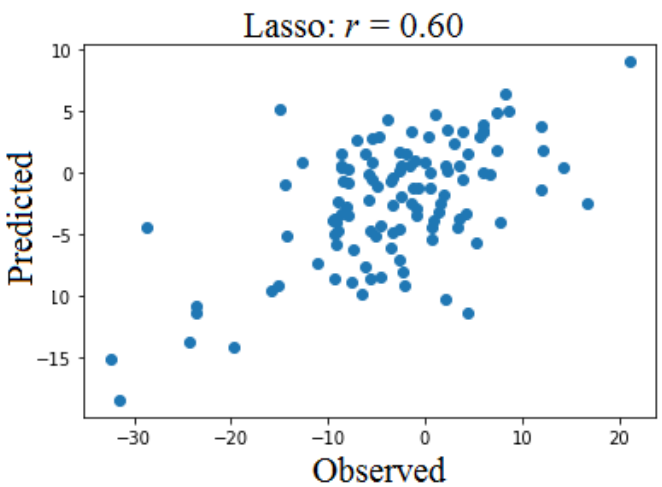


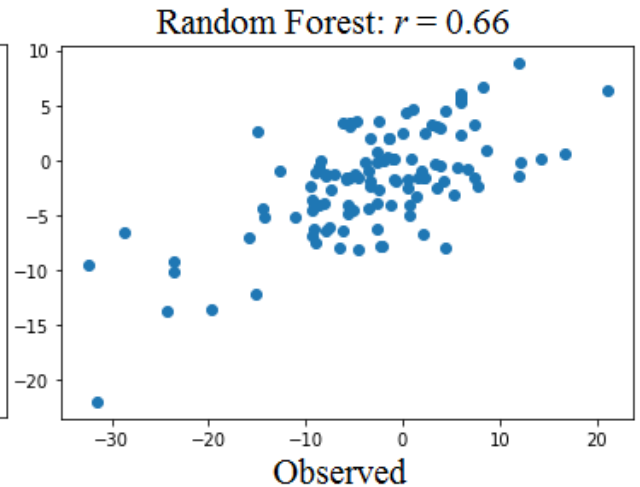


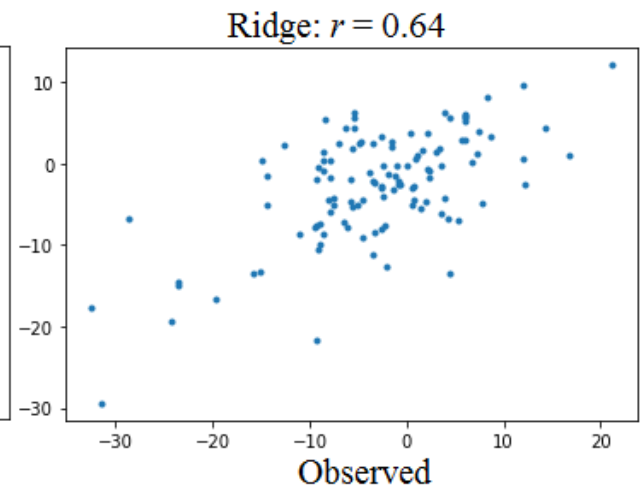
Genomic prediction results for plant height

- Training size: 460
- Test size: 115
- Model: Lasso, Random forest, Ridge
- Accuracy: 0.46~0.66



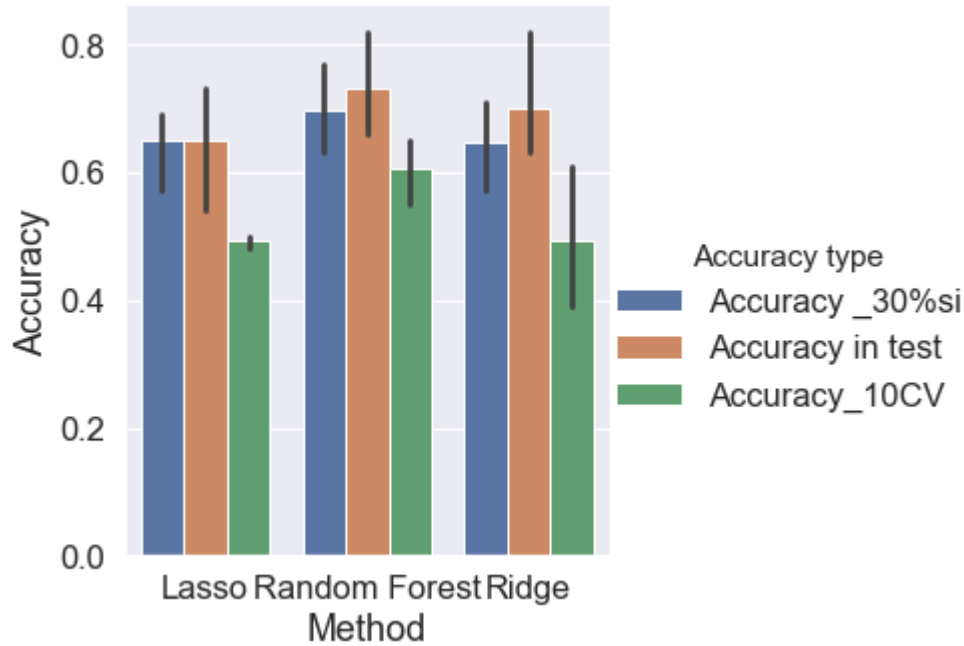


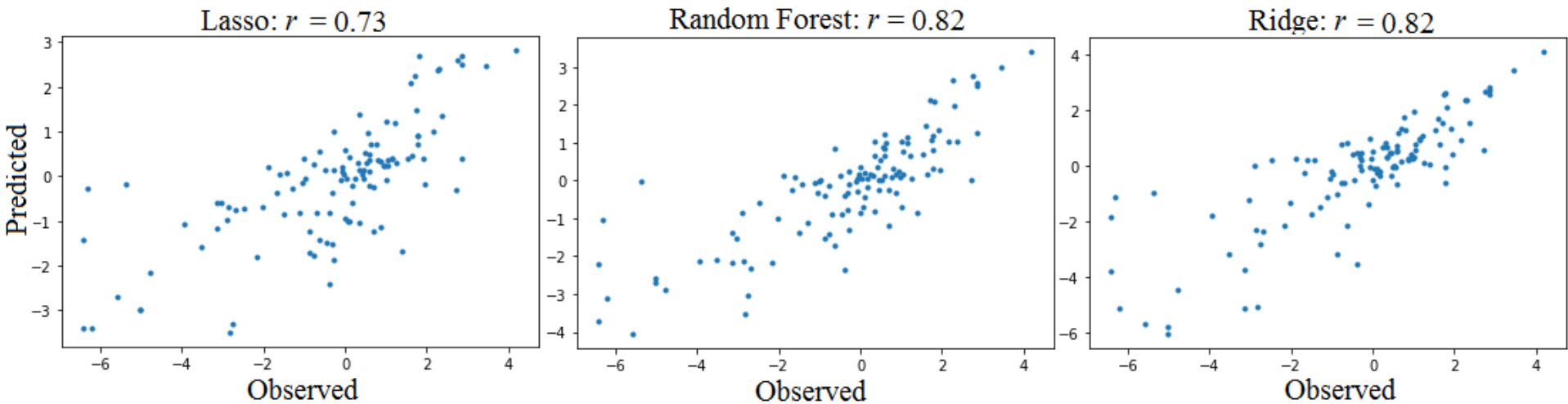




Genomic prediction results for silk DAP

- Training size: 460
- Test size: 115
- Model: Lasso, Random forest, Ridge
- Accuracy: 0.66 ~ 0.82

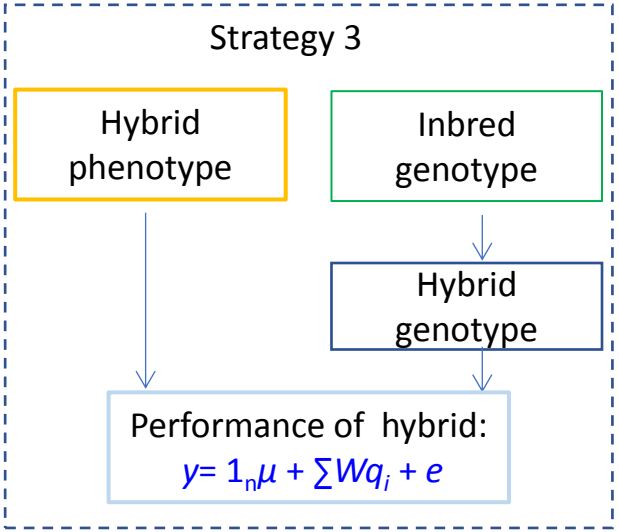




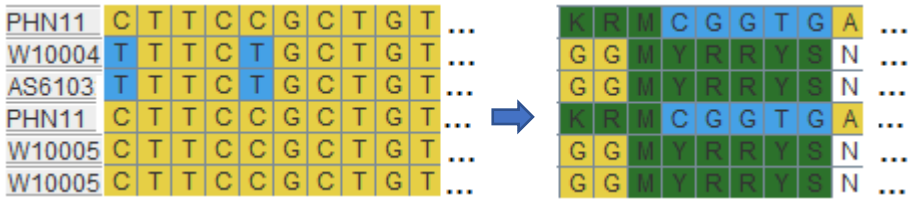
Works needed to be done in the future

- 1. Hyperparameter tuning for Random Forest.
- 2. Try Bayes method.
- 3. Enlarge population size for both genotyping and phenotyping
- 4. Separate Heterotic group for training model.
- 5. Predict hybrid performance directly by generating hybrid genotypes in silico based on genotypic data of parental lines.
- 6. Considering both the additive and dominant effect, as well as integrating environmental factors.

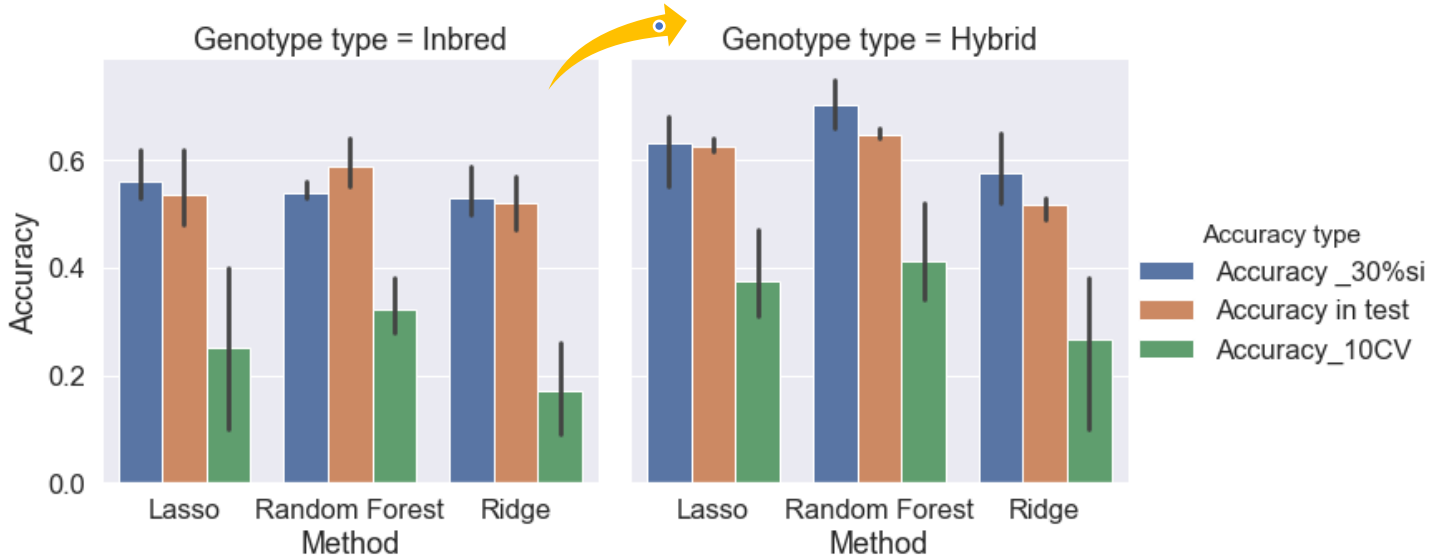
Predict hybrid performance directly by generating hybrid genotypes in silico



Generating hybrid genotypes in silico



- Prediction accuracy increased by 10%
- Accuracy at top 30% selection intensity reach 0.75



Thanks for your attention!

Most analyses were implemented in python

- Data cleaning:
- Pandas
- numpy
- R (lme4)
- Genomic selection:
- Sklearn.model_selection ,
- linear_model, sklearn.metrics,
- GridSearchCV, Lassso, Ridge, RandomForestRegressor