How Should I Select the Individuals of my Training Population to Make Selections in Genomic Selection?

Julio Isidro Sánchez
10th International Oat Conference
11-15 July 2016 Saint Petersburg-Rusia
Tools

Phenotypic Selection

MAS

SNPs, CAPs, SCARs, RAPDs, AFLPs, STSs, RFLPs, SSRs
Without the right tools, 
We are just playing.
MOST GENETIC CHARACTERS ARE QUANTITATIVES

MAS

GENOME

GS

GENOME

Trait of interest
Marker
Tools

Falling fast
In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore’s law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.
Number of Citations of GS per year. Web of science.
WHAT IS GENOMIC SELECTION?

Genomic Selection is a "new" tool in plant and animal breeding that use statistical modeling to predict how a plant/animal will perform (Breeding Value) before it is phenotyped.
Big Picture in Plant Breeding

\[ R = \frac{i r \sigma_A}{t} \]

- \( i \) = Intensity of selection
- \( r \) = Accuracy of selection
- \( \sigma_A \) = Additive genetic variance (standard deviation)
- \( t \) = Time
Increase Genetic Gain by:

- Increase Accuracy of Selection
- Decrease Generation interval

\[ R = \frac{\text{ir} \sigma^2}{t} \]
Genomic Selection scheme

Base Population

Calibration Set
- Phenotypes
- Genotypes

Test Set
- Genotypes

Optimization

Training Set
- Estimate Marker effects: \( \hat{g}_i \)
- Train GS model
  - \( \text{rrBLUP} \)
  - Bayes B
  - LASSO
  - RHKS

\[ \text{GEBUS} = \sum_{j=1}^{p} x_{ij} \hat{g}_i \]

Predictions

Selection
Question

Does population structure have an impact on the optimization of the training population?

1. Random Sampling
2. Stratified Sampling
3. Coefficient of determination (CD)
4. Prediction Error Variance (PEV)
5. Stratified Coefficient of Determination (StratCD)
Mild Population Structure in Wheat

Strong Population Structure in Rice
Improving genetic diversity using kinship matrix

\[
PEV = \text{diag} \left[ \frac{c'(Z'MZ + \lambda G^{-1})^{-1} c}{c' c} \right] \times \sigma^2 \epsilon
\]

\[
CD = \text{diag} \left[ \frac{c'(G - \lambda(Z'MZ + \lambda G^{-1})^{-1} c}{c' G c} \right]
\]

Coefficient of determination use kinship matrix in its calculation
OPTIMIZATION

100 pop size

Method 1- STRATIFIED SAMPLING

Random

Wheat Dataset

627 Calibration Set

50X

500 Test Set

554
50
100
200
300

1127

100 pop size

34 C1
29 C2
21 C3
16 C4

326

30 TRS

Optimization on Training Population Size

100 pop size

C1-188 CLUSTERING

C2-175

C3-153

C4-111
OPTIMIZATION

Method 2 CDmean and PEVmean

Wheat Dataset

Random

627
Calibration Set

527_{RS}

100_{TR}

1127

Test Set

Maximize CDmean

Minimize PEVmean

Optimization Training Population Size

25
50
100
200
300

50X
Improving genetic diversity using kinship matrix

\[ R = \frac{i r \sigma_A}{t} \]
**Yield**

- Plot a)
  - Accuracy vs. Popsize
  - Lines represent different methods: CDmean, PEVmean, StratCDmean, Stratified Sampling, Random

**Test Weight**

- Plot b)
  - Accuracy vs. Popsize
  - Lines represent different methods: CDmean, PEVmean, StratCDmean, Stratified Sampling, Random

**Lodging**

- Plot c)
  - Accuracy vs. Popsize
  - Lines represent different methods: CDmean, PEVmean, StratCDmean, Stratified Sampling, Random

**Heading Date**

- Plot d)
  - Accuracy vs. Popsize
  - Lines represent different methods: CDmean, PEVmean, StratCDmean, Stratified Sampling, Random

**Plant Height**

- Plot e)
  - Accuracy vs. Popsize
  - Lines represent different methods: CDmean, PEVmean, StratCDmean, Stratified Sampling, Random
Population structure has an impact on the optimization of the training population.

- **Mild PS**---CDmean and StratCDmean

- **Strong PS**---Stratified Sampling
Improving genetic diversity using kinship matrix

Wheat CDmean

Wheat PEVmean

Wheat StratCDmean
Large genotypic variance obtained by CDmean doesn´t translate into large phenotypic variance
There isn’t a best selection criterion to optimize the TRS under population structure. PS plays an important role in optimization of TRS in GS.

Before optimization, population structure must be evaluated.

Highest accuracies with methods that capture more phenotypic variance.

CDmean is an optimal criterion for long-term selection.
Optimization of genomic selection training populations with a genetic algorithm

Deniz Akdemir, Julio I Sanchez and Jean-Luc Jannink

Figure 1 Arabidopsis data. The difference between the accuracies of the models trained on optimized populations versus random samples. Positive values indicate the cases for which the optimized population performed better as compared to a random sample.
OAT Nutritional traits

Base Population

Calibration Set
- Phenotypes
- Genotypes

Training Set

Test Set
- Genotypes

GEBUS = \sum_{j=1}^{p} x_{ij} \hat{g}_i

Train GS model:
- rrBLUP
- Bayes B
- LASSO
- RHKS

Predictions

Selection