UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA
"LUIZ DE QUEIROZ"
DEPARTAMENTO DE GENÉTICA
LGN5825 Genética e Melhoramento de Espécies Alógamas

Genomic Selection

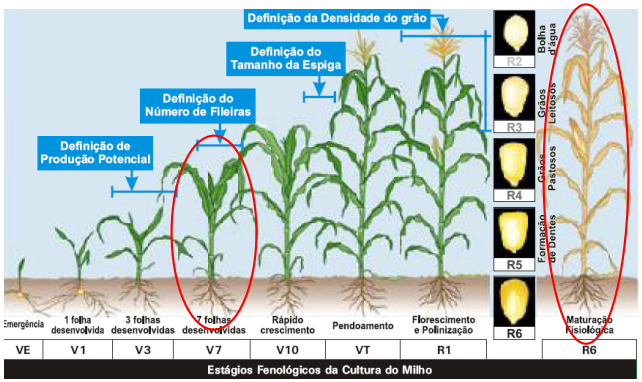
Prof. Roberto Fritsche-Neto

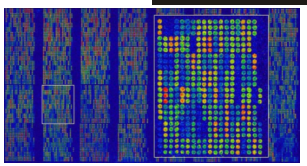
roberto.neto@usp.br

Piracicaba, June 8th, 2018

Early in(direct) selection

- Objectives:
- Reducing the interval between generations
- Using traits more accessible and cheaper evaluation
- Selecting before flowering





- Problems in the traditional selection
- Negative correlations
- Development stage – genes differentially expressed

- What about marker-selection?

$$ISE=\sqrt{\frac{h_m^2}{h_{trait}^2}}r_{(trait,m)} \qquad ISE=\sqrt{\frac{1}{h_{trait}^2}}r_{(trait,m)} \qquad ISE=\frac{r_{(trait,m)}}{\sqrt{h_{trait}^2}}$$

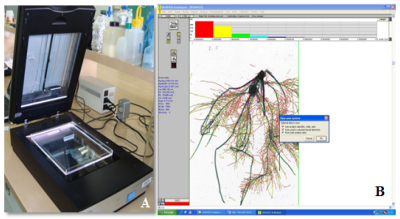
Genomic selection

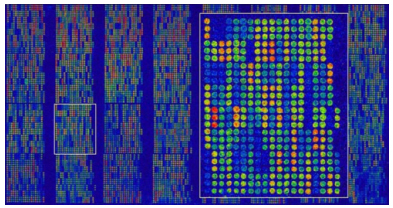
- Simultaneous prediction (without tests) of the genetic effects of large numbers of markers
- Dispersed on wide genome
- Capture the effects of all loci (small and large)
- Explain much of the genetic variation of a character
- Keeps the "black box" about the genetic control
- Minor “aversion” by breeders

- Limitations
- Loss of genetic variance
- Fast inbreeding

General procedures of GS

Training population
sample from population of selection
Phenotyping and genotyping
> 500 individuals





Application in Selection Schemes
Only genotyping
Breeding populations

GEBV = Zh

Y = Xb + Zh + e
Prediction of markers effects
Statistical equations and methods

Validation of markers
New and independent population or
Statistical methods of cross-validation

Cross-validation methods

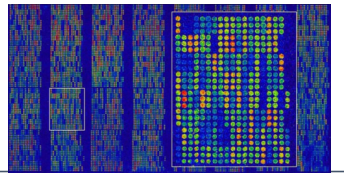
Ex. RR-BLUP
Assumes equal variances among markers (VG / Nm)
The characters differ in the number and the markers that maximize the accuracy of prediction

Genomic Recurrent Selection

Training / Validation Set

Genotyping the lines **OTS**
Phenotyping the hybrids **HTP**
Adjust the models and identify the marker effects

Many cycles of
Genomic selection
(by the markers set that maximize the accuracy)



C_n
C_1
C_0

Improved lines
(hybrids or the newest TS)

Doubled haploids

Recombination

Obtaining progenies

Synthetics or OPV

Evaluation / Selection

RR-BLUP/GS

- It's a multiple regression
- Each marker is a factor/parameter

5 individuals
7 markers

MM	2
Mm	1
mm	0

Individuo	Diâmetro	Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6	Marca 7
1	9.87	2	0	0	0	2	0	0
2	14.48	1	1	0	0	1	1	0
3	8.91	0	2	0	0	0	0	2
4	14.64	1	0	1	0	1	0	0
5	9.55	1	0	0	1	1	1	0

$$X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\mathbf{Y = Xb + Zh + e}$$

$$\begin{bmatrix} \hat{b} \\ \hat{h} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \dfrac{\sigma_e^2}{\sigma_{Am}^2 / n} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{h}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

RR-BLUP/GS

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \dfrac{\sigma_e^2}{\sigma_{Am}^2 / n} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{h}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \longrightarrow \begin{bmatrix} \hat{b} \\ \hat{h} \end{bmatrix} = \begin{bmatrix} 12.4519 \\ -0.3526 \\ 0.2761 \\ 1.4467 \\ -1.3701 \\ -0.3526 \\ 0.5436 \\ -1.63765 \end{bmatrix}$$

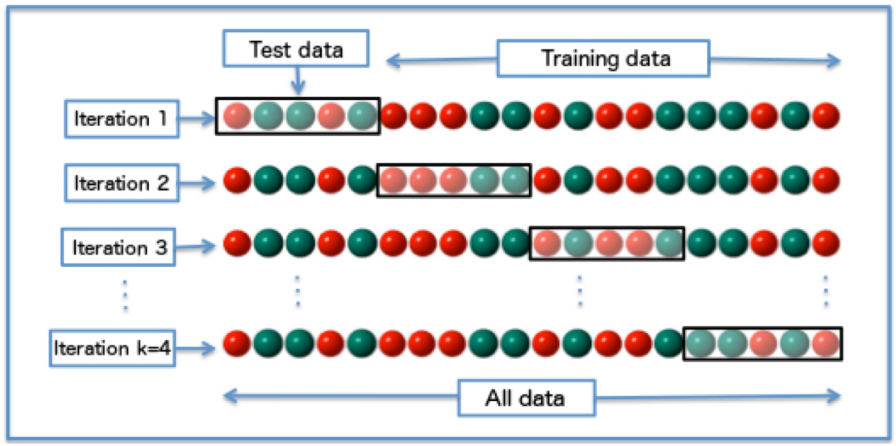
- Genomic estimated breeding values
- h = marker effects

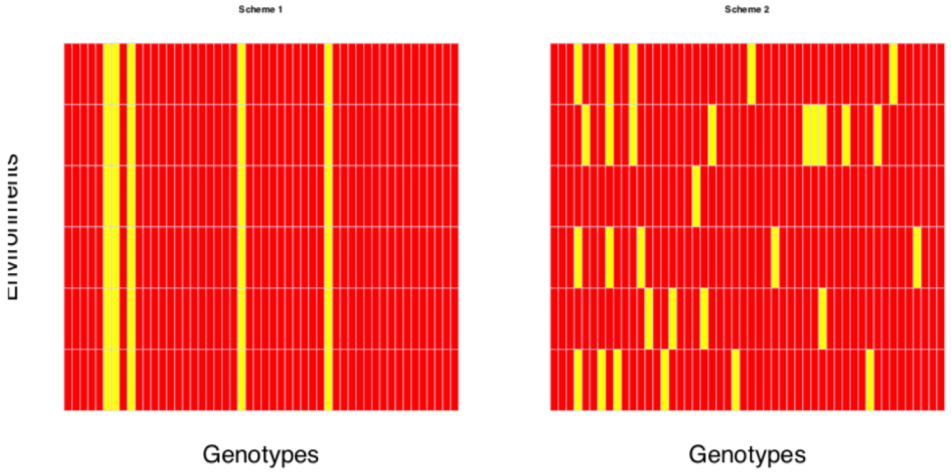
$$\mathbf{GEBV = Zh}$$

$$\begin{bmatrix} -1.4104 \\ 0.1145 \\ -2.7230 \\ 0.7415 \\ -1.5317 \end{bmatrix} = \begin{matrix} 2 & 0 & 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{matrix} \mathbf{x} \begin{bmatrix} -0.3526 \\ 0.2761 \\ 1.4467 \\ -1.3701 \\ -0.3526 \\ 0.5436 \\ -1.63765 \end{bmatrix}$$

Cross-validation

- Estimate how accurately a predictive model will perform in practice
- Avoid overfitting
- Partitioning the original sample into a training set to train the model, and a test set to evaluate it
- E.g., partitioning the data set into two sets of 80% for training and 20% for test
- In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples
- Multi-environment models - Two main schemes (CV1 and CV2)
- ***k*-fold cross-validation vs. Repeated random sub-sampling validation**





G-BLUP method

- Equivalent to the RR-BLUP but less computing consuming
- Easy to extend it to other factors or kind of kernels

Y = Xb + Za + e

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \end{bmatrix} \begin{bmatrix} \hat{u} \\ \tilde{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

$$G_a = \frac{WW'}{\sum_{i=1}^{n}(2p_iq_i)}$$

Efeitos aditivos: W

$$W = \begin{cases} Se\ MM; & 2 & \rightarrow 2-2p = 2q \\ Se\ Mm; & 1 & \rightarrow 1-2p = q-p \\ Se\ mm; & 0 & \rightarrow 0-2p = -2p \end{cases}$$

GEBV = Za

Y = Xb + Za + Zd + e

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'X & Z'Z & Z'Z + G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \tilde{a} \\ \tilde{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix}$$

Efeitos de dominância: S

$$S = \begin{cases} Se\ MM; & 0 & \rightarrow -2q^2 \\ Se\ Mm; & 1 & \rightarrow 2pq \\ Se\ mm; & 0 & \rightarrow -2p^2 \end{cases}$$

$$G_d = \frac{SS'}{\sum_{i=1}^{n}(2p_iq_i)^2}$$

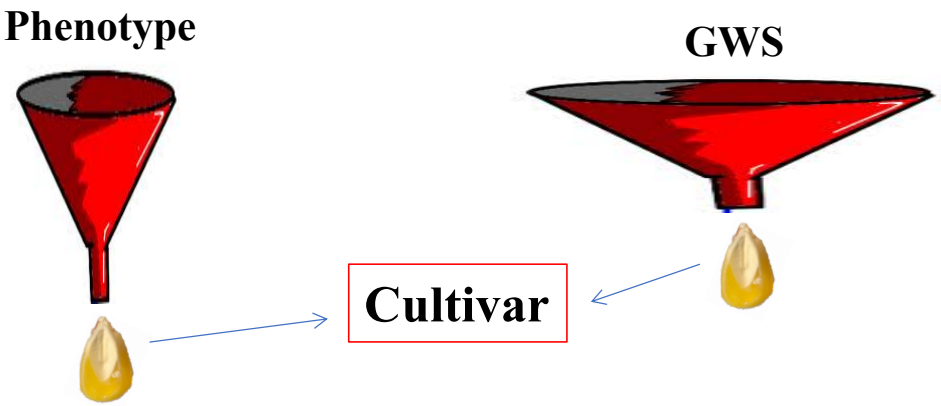
GEBV = Za + Zd

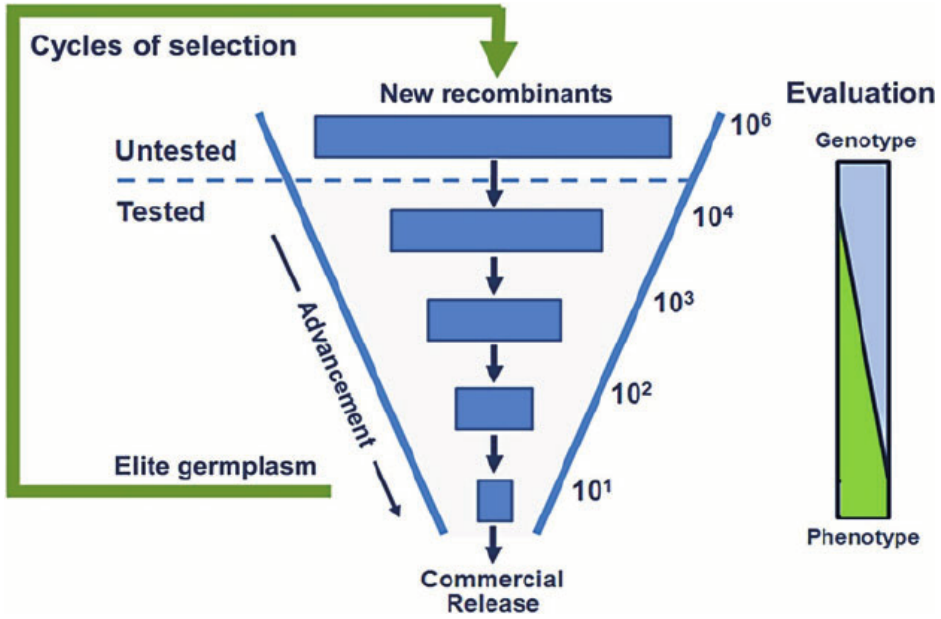
Factors Affecting Prediction Accuracy

- Marker density and LD decay
- Effective population size – diversity
- Training set – populations structure, who phenotyping, stage
- Genetic relationship between training population and selection candidates
- Rare alleles (MAF < 5%)
- Missing data and imputation method (*call rate* < 95%)
- Statistical model
- Correlated traits (multi-trait models)
- Progeny size and ploidy
- Crossover GE
- Number of cycles of GS

GS applied to breeding programs

- GS modifies significantly the way to select





- Reducing the time to develop cultivars
- Increasing the effective size and selection intensity
- Increasing the genetic gain per unit time