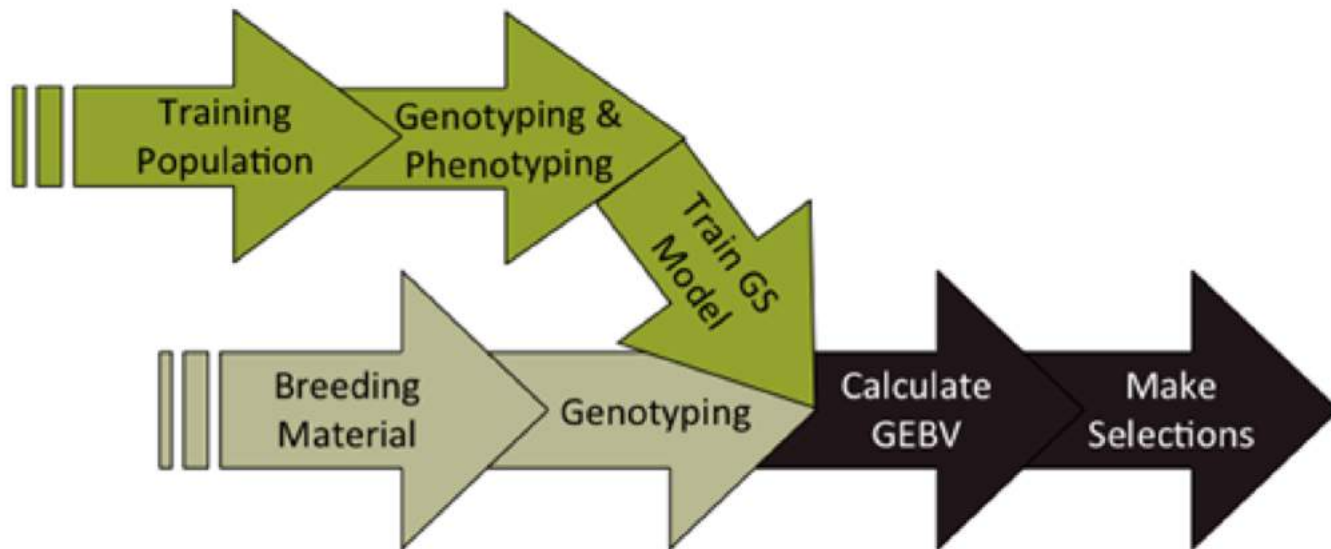# AGR 6322 Advanced Plant Breeding
Fall 2018

Genomic selection

Goals for today

Use of genomic selection in Plant Breeding

UF/IFAS Annual Ryegrass Breeding Program

History

Started 1980s

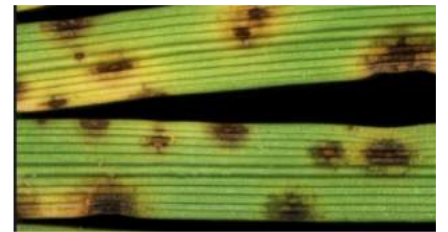Dr. Gordon Prine

Objectives

Forage yield, disease resistance; maturity
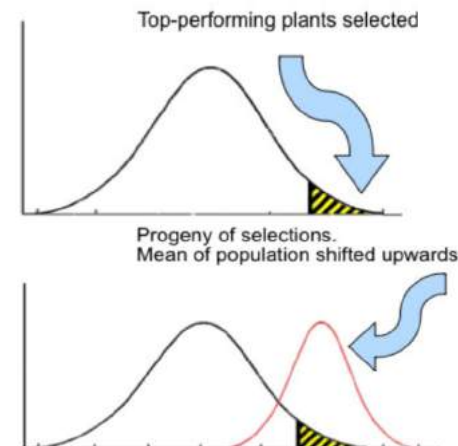

Crown rust (*Puccinia coronata*)

Grey leaf spot (*Pyricularia grisea*)

Leaf spot (*Bipolaris*)

Methodology

Recurrent phenotypic selection (8-10 years)

Top-performing plants selected

Progeny of selections.
Mean of population shifted upwards

More than 40 cultivars!

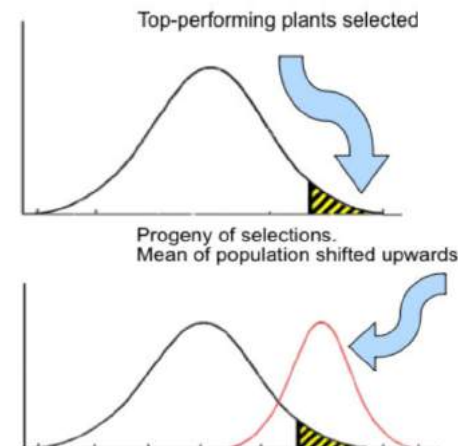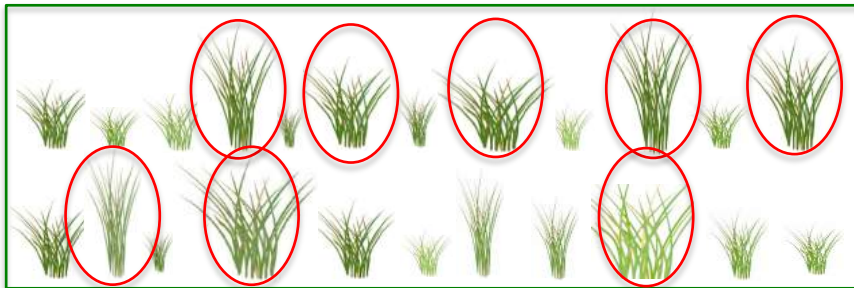Most important forage species in royalty generation!

UF/IFAS Annual Ryegrass Breeding Program

$$\Delta G_a = h^2 S = i h^2 \sigma_p$$

Genetic gain depends of the selection intensity (i), heritability (h^2) and the phenotypic standard deviation.

Methodology

Recurrent phenotypic selection (8-10 years)

Top-performing plants selected

Progeny of selections.
Mean of population shifted upwards

More than 40 cultivars!

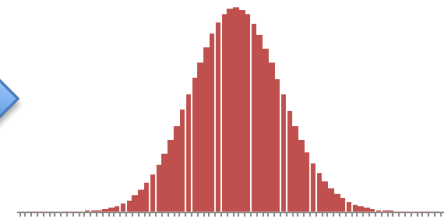Most important forage species in royalty generation!

$$\Delta G_a = h^2 S = i h^2 \sigma_p$$

Phenotypes vs. Genotypes

Field

Phenotype = Genotype + Environment + G*E

BLUP Analysis

$$Phenotype = X\beta + Za + e$$

Controlled crosses
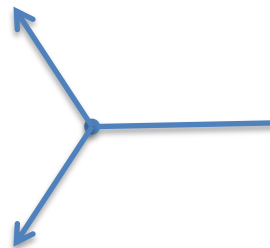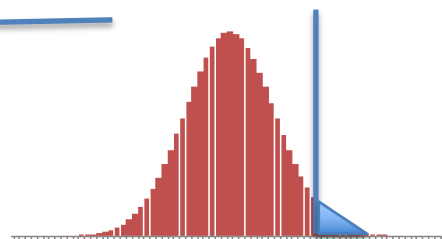
-Ranking Genotypes
-Selection based on BV

Variance Components
P= G + E
G= A + D + I
-Heritability (h^2)
-G*E
-Trait Correlations

Cultivar Release
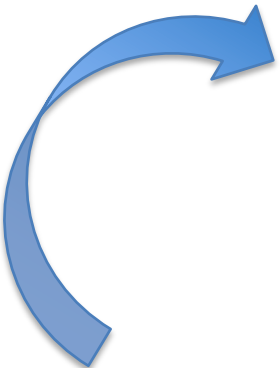
Novel Breeding Tools
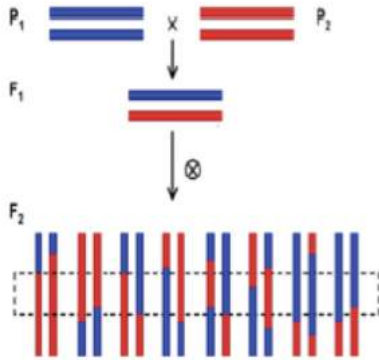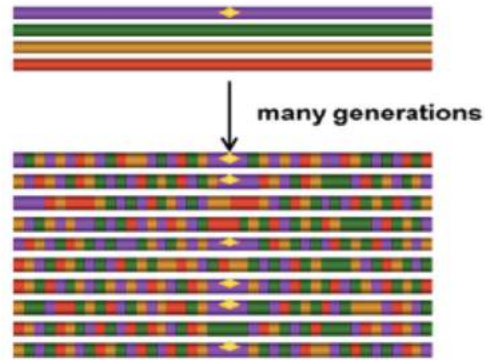
Quantitative Trait Loci (QTL)   **Genome-wide Association**   **Genome-wide Prediction**
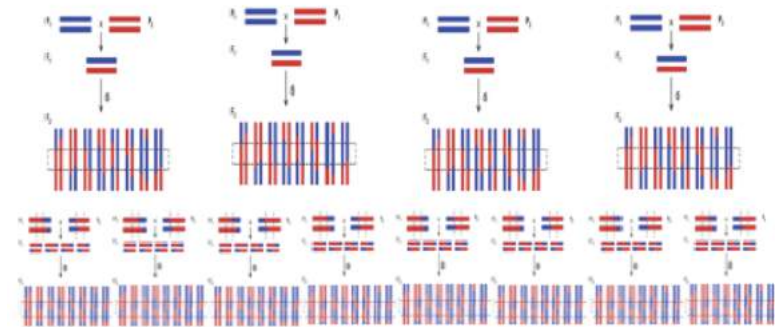
REVIEW & INTERPRETATION

Marker-Assisted Selection in Plant Breeding: From Publications to Practice

Yunbi Xu* and Jonathan H. Crouch

Figure 1. The numbers of articles with the terms *quantitative trait locus* or *quantitative trait loci* (QTL) and *marker-assisted selection* (MAS) by years (1984–2005) from Google Scholar (4 Aug. 2007).

Novel Breeding Tools

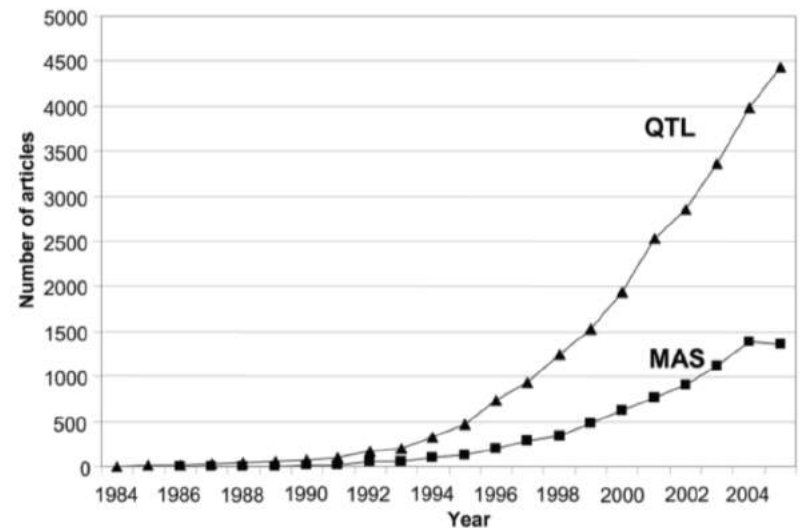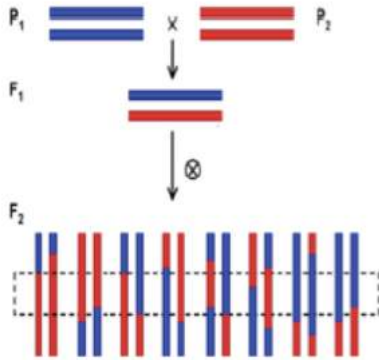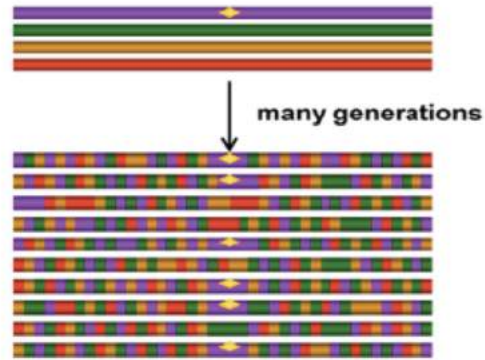Quantitative Trait Loci (QTL)   Genome-wide Association   Genome-wide Prediction

Meuwissen et al., 2001
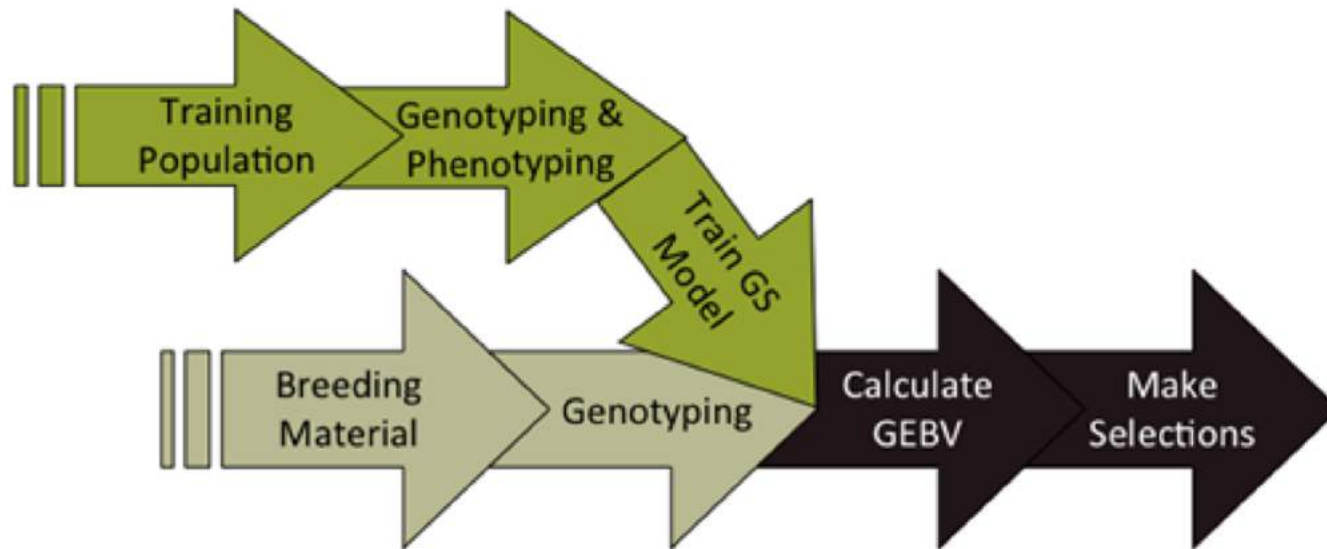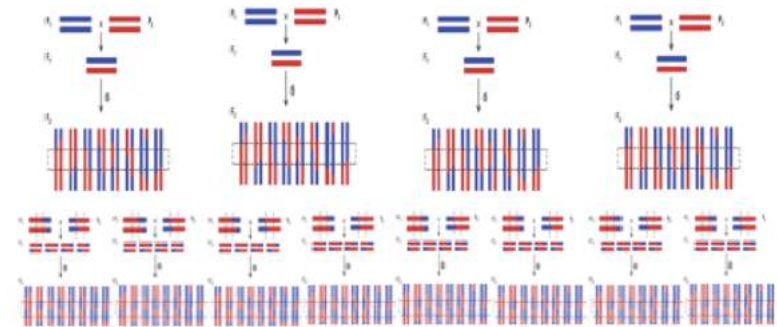
Genomic selection

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes[†] and M. E. Goddard[†,‡]

*Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, [†]Victorian Institute of Animal Science,
Attwood 3049, Victoria, Australia and [‡]Institute of Land and Food Resources,
University of Melbourne, Parkville 3052, Victoria, Australia

"Recent advances in molecular genetic techniques will make dense marker maps available and genotyping many individuals for these markers feasible"

However when this happens number of parameters (p) to be estimated (markers) will much larger than the number of individuals (n) to estimate those parameters

Under this scenario (p>>n) the normal estimators are not efficient, are bias and have problems of overfitting

Genomic selection

Benefits of genomic selection

- Decrease the generation cycle of breeding (e.g. perennials, cattle).

- Decrease the cost of testing (e.g. cattle, maize).

- Screening a larger number of genotypes without field testing, thus increasing the selection pressure (e.g. maize, other cereals).

- Predict performance for difficult and/or expensive traits (e.g. cattle, salmon).

- Predict performance for diseases avoiding challenging and losing the germplasm (all species).

- Can be used regardless the genetic architecture of the trait.
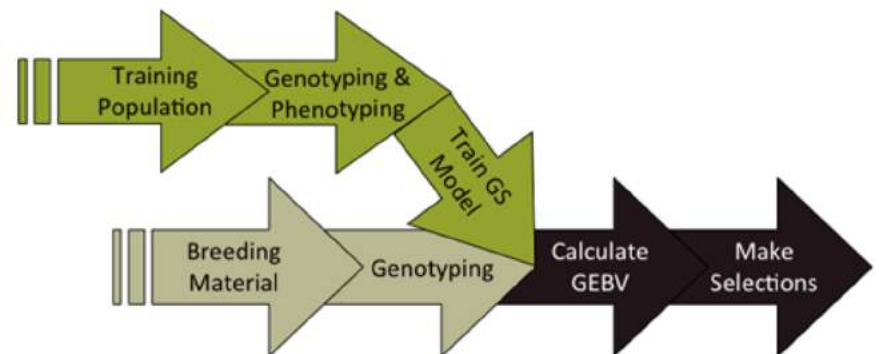
Genomic selection

Basic concept of genomic selection

1. Identify genetic markers covering most of the genome in the breeding population
2. Develop prediction models that capture most of the quantitative variation of interest.
3. Use prediction models to select superior genotypes in next cycle.

Gene/QTL involved in the target traits   Marker
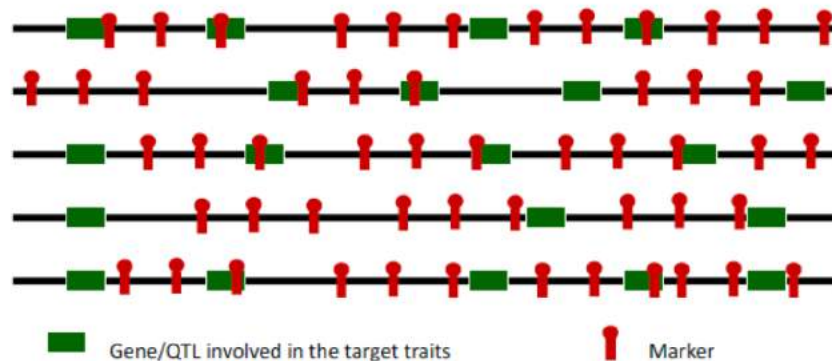
Training Population

- Individuals have to be genotyped and phenotyped
- Population used to fit the model

Testing (Validation) Population

- Model is validated
- Evaluates: cor(GEBV, BV); MSE(GEBV, BV)

Selection Population

- Only genotypes; No phenotypic data
- Population where the models are applied

Genomic selection

Training population

Phenotype

Breeding values (BV)

Molecular markers (m)

Genotype

Prediction model

$$BV_i = 1\boldsymbol{\mu} + \sum_{j=1}^{p} W_{ji}\, m_j + e$$

Selection population

Genotype

Molecular markers (m)

$$GEBV_i = 1\boldsymbol{\mu} + \sum_{j=1}^{p} W_{ji}\, m_j + e$$

Predictive ability ➔ r(Phenotype, GEBV)

• Cross-validation

Factors affecting predictive ability: marker type and density, LD between markers and QTLs, h^2, size and structure of training and validation population (relatedness)
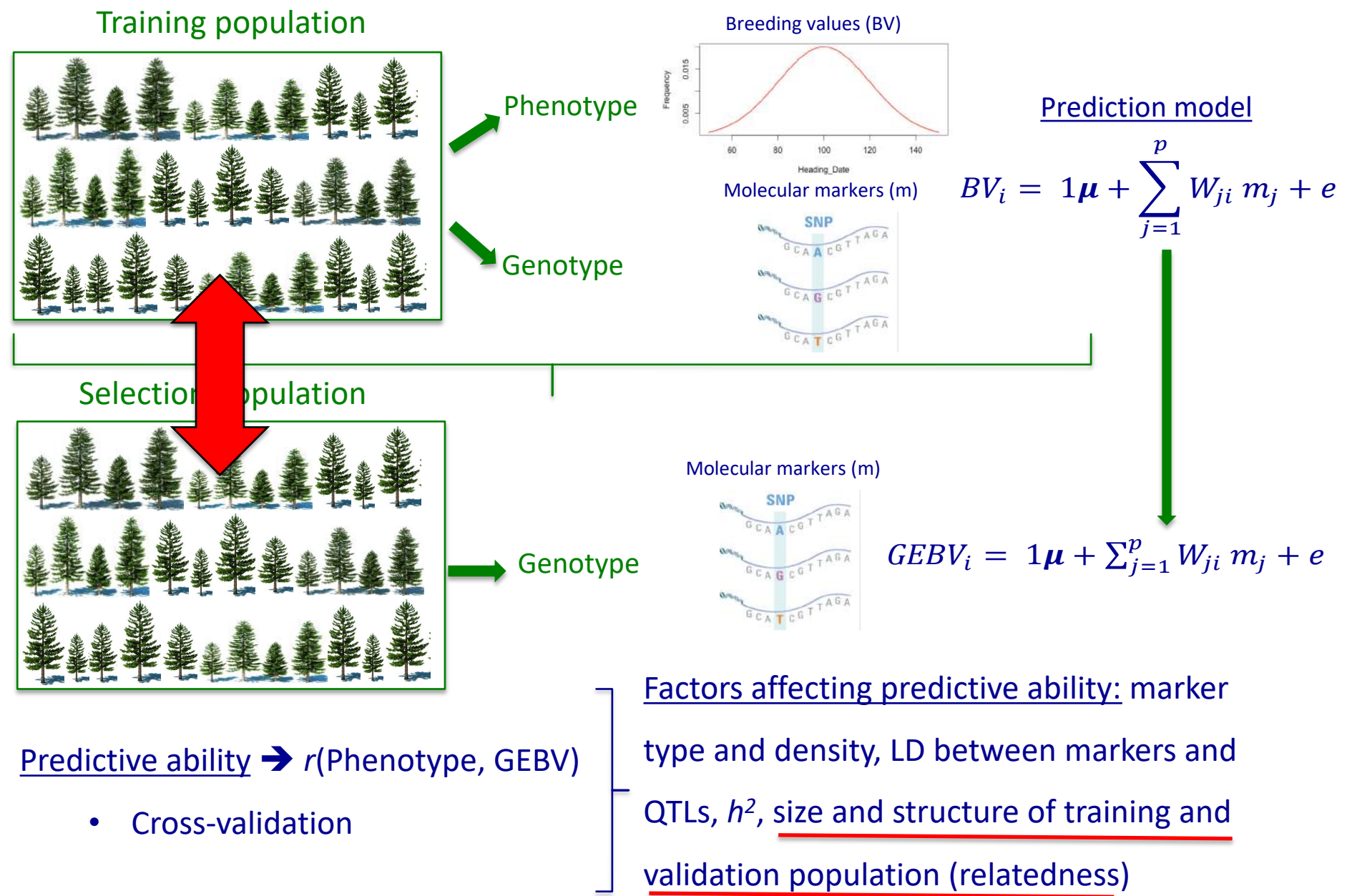
Genomic selection

Predictive ability (accuracy) depends on

- The level of linkage disequilibrium (LD) between the markers and the QTL **(effective population size and genotyping density)**.

- The number of individuals with phenotypes and genotypes in the reference population (training set) from which the marker effects are estimated.

- The heritability of the trait in question.

- The distribution of QTL effects, i.e. number of loci involved.

- Quality of the phenotyping used to construct the prediction model.

Genomic selection

Predictive ability (accuracy) depends on

Total Height of the Plant

RR-BLUP

Bayes Cpi

More than 1000 Markers needed to reach maximum accuracy

Predictive ability

Number of Molecular Markers (In tenths)
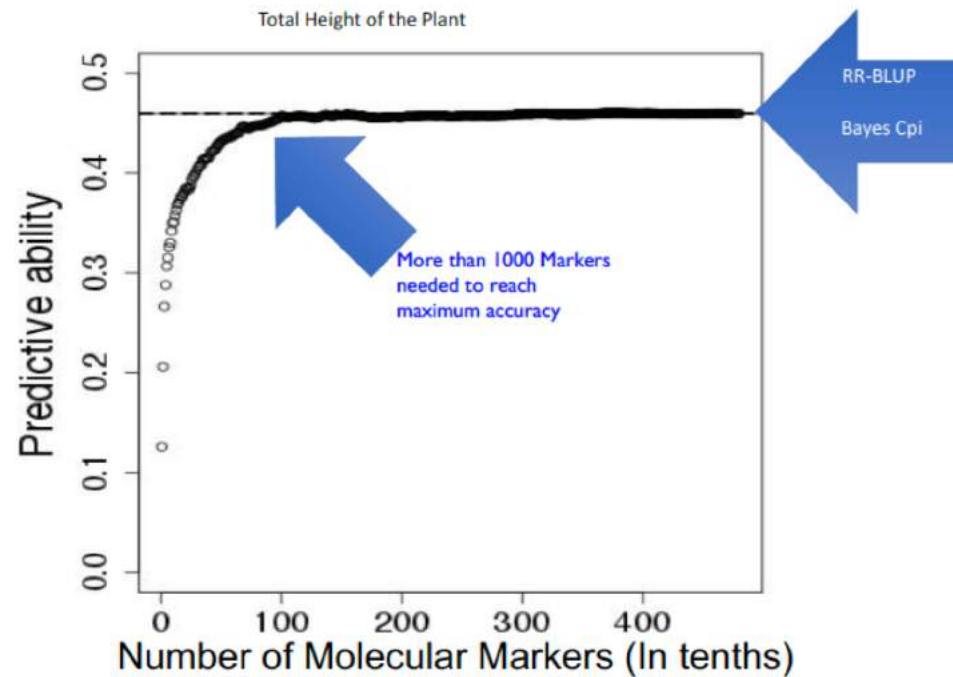
Resende et al. 2012. Genetics, Vol. 190, 1503–1510 April 2012

Population structure in GS

- Genomic Selection
  - ✓ Half-sib progeny ➜ most common strategy in forage and turfgrass species
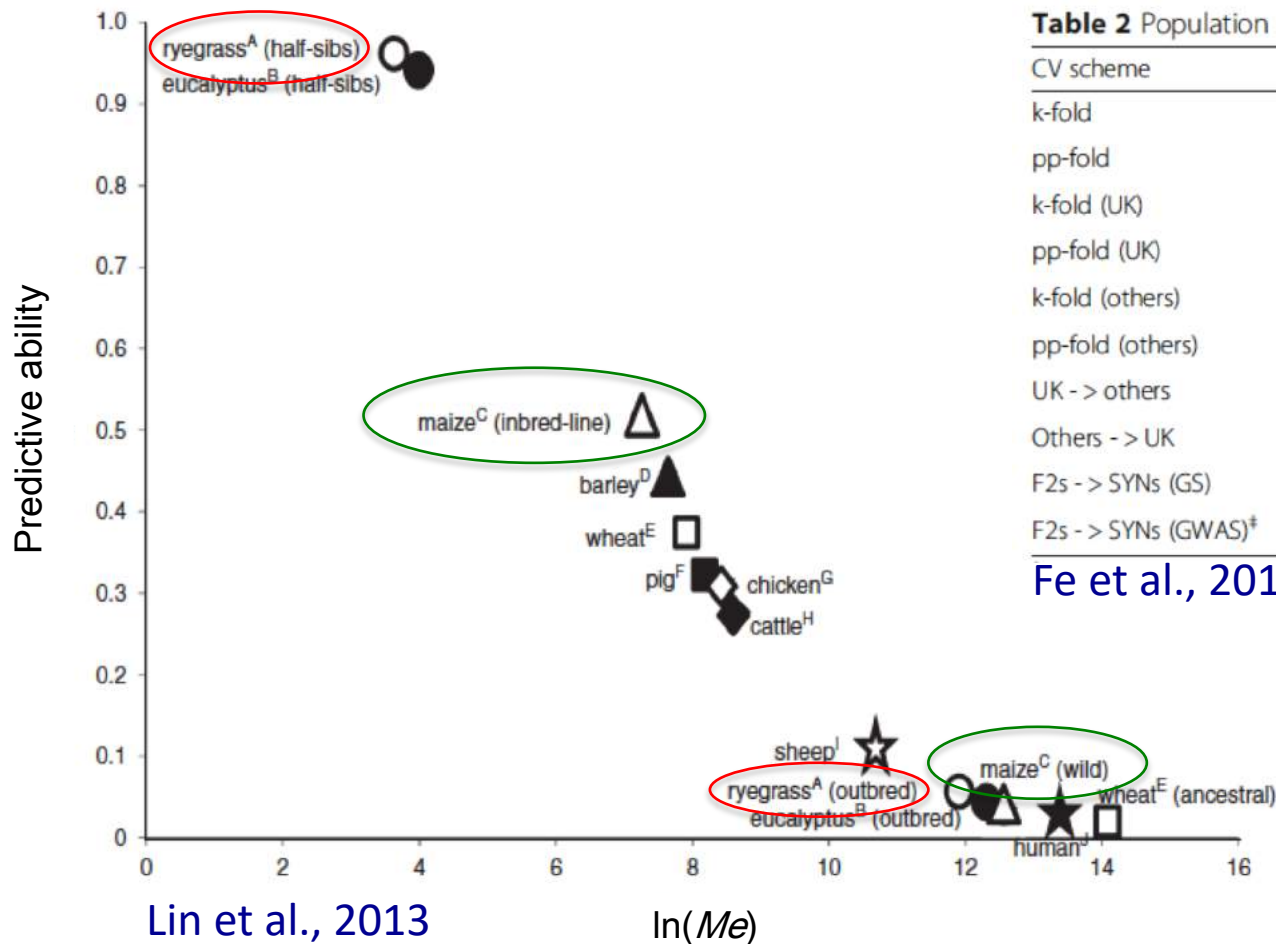 - ✓ Full-sib progeny

Lin et al., 2013

Table 2 Population size and results (with SE) for all CV schemes

CV scheme	Pop.size	$\rho_{\hat{y}f\hat{g}}$[†]	$\rho_{g\hat{g}}$	bias (b)
k-fold	1757	$0.90_{0.01}$[a]	$1.04_{0.07}$	$1.02_{0.01}$
pp-fold	1757	$0.84_{0.01}$[b]	$0.98_{0.06}$	$1.10_{0.02}$
k-fold (UK)	466	$0.78_{0.03}$[a]	$0.86_{0.09}$	$1.06_{0.04}$
pp-fold (UK)	466	$0.52_{0.04}$[b]	$0.57_{0.07}$	$1.30_{0.10}$
k-fold (others)	1291	$0.90_{0.01}$[a]	$1.04_{0.07}$	$1.02_{0.01}$
pp-fold (others)	1291	$0.86_{0.01}$[b]	$0.99_{0.07}$	$1.17_{0.02}$
UK - > others	466	$0.78_{0.02}$[N]	$0.90_{0.07}$	$1.46_{0.03}$
Others - > UK	1291	$0.71_{0.03}$[N]	$0.78_{0.08}$	$0.92_{0.04}$
F2s - > SYNs (GS)	1757	$0.88_{0.05}$[a]	$0.93_{0.24}$	$1.02_{0.06}$
F2s - > SYNs (GWAS)[‡]	1757	$0.74_{0.07}$[b]	$0.78_{0.21}$	$1.33_{0.13}$

Fe et al., 2015

Genomic selection

Predictive ability (accuracy)

Review this paper to learn more about the factors affecting the accuracy of GWS

ORIGINAL PAPER

Genomic selection in forest tree breeding

Dario Grattapaglia · Marcos D. V. Resende

Review

Genomic selection in crops, trees and forages: a review

Z. Lin[A,B,C,D], B. J. Hayes[A,B,C], and H. D. Daetwyler[A,B,C]

[A]School of Applied Systems Biology, La Trobe University, Bundoora, Vic. 3083, Australia.
[B]Department of Environment and Primary Industries, Biosciences Research Division, AgriBio, 5 Ring Road, Bundoora, Vic. 3083, Australia.
[C]Dairy Futures Cooperative Research Centre, AgriBio, 5 Ring Road, Bundoora, Vic. 3083, Australia.
[D]Corresponding author. Email: zibei.lin@depi.vic.gov.au

Genomic selection

Analytic methods for GS

- **BLUP-Based:** **G-BLUP, RR-BLUP, RR-BLUP_B**
- **Bayes-Based:** **BayesA, BayesB, BayesCπ**, BayesR
- **LASSO-Based:** **Bayesian Lasso Regression, Improved Lasso**
- **Semi-Parametric Regression: RKHS**
- **Non-Parametrics:** Suport Vector Machine, Neural-Networks
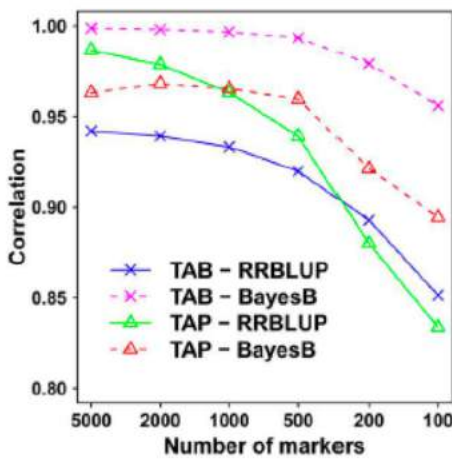- **Others...**

Table 1 Predictive ability of genomic selection models using four different methods

Trait category	Trait	h^2	Methods			
			RR–BLUP	BLASSO	Bayes A	Bayes Cπ
Growth	HT	0.31	0.39	0.38	0.38	0.38
	HTLC	0.22	0.45	0.44	0.44	0.44
	BHLC	0.35	0.49	0.49	0.49	0.49
	DBH	0.31	0.46	0.46	0.46	0.46
Development	CWAL	0.27	0.38	0.36	0.36	0.36
	CWAC	0.45	0.48	0.46	0.47	0.47
	BD	0.15	0.27	0.25	0.27	0.27
	BA	0.33	0.51	0.51	0.51	0.51
	Rootnum_bin	0.10	0.28	0.28	0.27	0.28
	Rootnum	0.07	0.24	0.26	0.25	0.24
Disease resistance	Rust_bin	0.21	0.29	0.28	0.34	0.34
	Rust_gall_vol	0.12	0.23	0.24	0.28	0.29
Wood quality	Stiffness	0.37	0.43	0.39	0.42	0.42
	Lignin	0.11	0.17	0.17	0.17	0.17
	LateWood	0.17	0.24	0.24	0.23	0.24
	Density	0.09	0.20	0.22	0.23	0.22
	C5C6	0.14	0.26	0.25	0.25	0.25

h^2 is the narrow-sense heritability of the trait.

Resende et al., 2012

Genomic selection

Analytic methods for GS

• Genomic BLUP (GBLUP) is a Genomic Selection method that uses the same framework than BLUP analysis, but replaces:

– The **numerator relationship matrix (A)** derived from the pedigree by,

– The **realized relationship matrix (G_A)** derived from molecular markers.

• G_A is also known as **observed relationship matrix** or **genomic matrix.**

• The **A** matrix is derived based on the infinitesimal model and represents and average relationship.

• The relationship matrix derived from the markers is more informative because the relationships estimates include the *Mendelian sampling.*

• There are several different algorithms to compute the **G_A** matrix from SNP data:

Hayes and Goddard (2008)

Van Raden (2008) – 2 methods

Yang *et al.* (2010) – Human genetics … and others!

• Relationship matrices work well to model the variance-covariance of additive effects assuming a large number of markers is used.

Genomic selection

Analytic methods for GS

Genomic Relationship Matrix for Correcting Pedigree Errors in Breeding Populations: Impact on Genetic Parameters and Genomic Selection Accuracy

Patricio R. Munoz, Marcio F. R. Resende Jr., Dudley A. Huber, Tania Quesada, Marcos D. V. Resende, David B. Neale, Jill L. Wegrzyn, Matias Kirst, and Gary F. Peter★

c) Full Sib: Original Pedigree   d) Full Sib: Corrected Pedigree
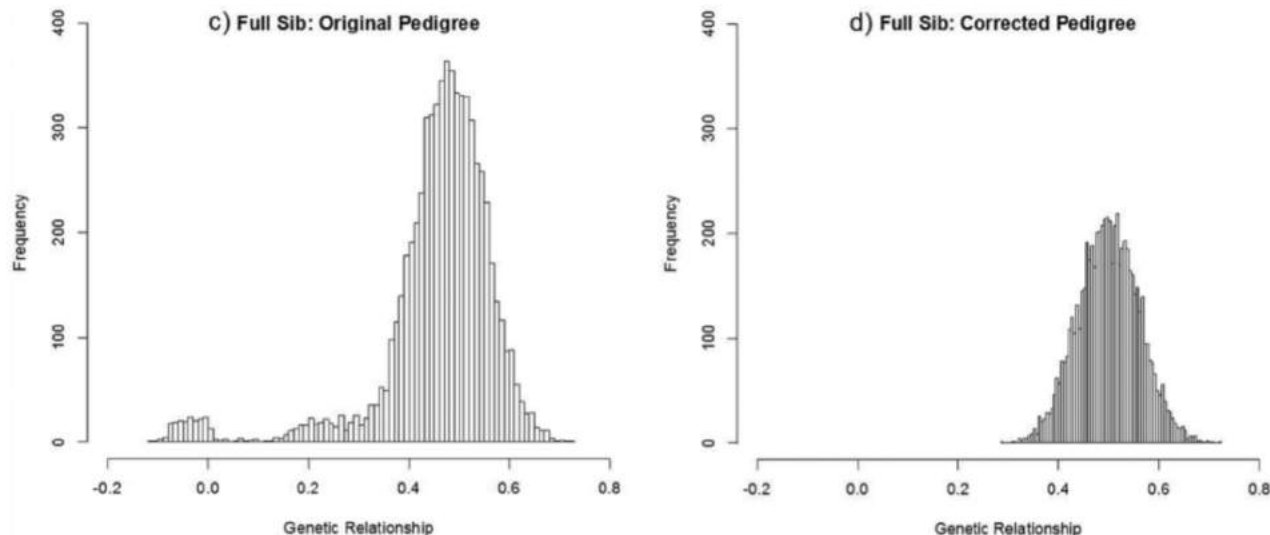
Genome-wide Family Prediction

Training population

Phenotype

$$GEBV_i = \sum_{j=1}^{p} W_{ji}\, m_j$$

Genotype

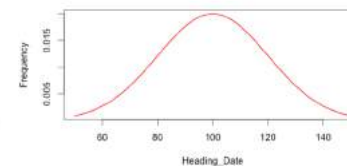Cross-pollinated species

Annual ryegrass

Alfalfa

Training population

Family pooled phenotypic and genotypic data
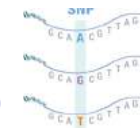
Family phenotypic values (FV)

Allele frequency (m)

Family prediction model

$$GWFP = \sum_{j=1}^{p} W_{jF}\, m_j$$
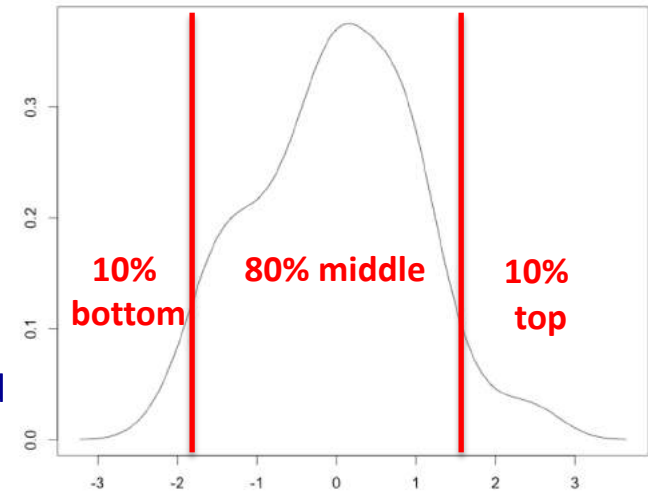
"Genomic Wide Family Prediction (GWFP)"

Objectives

1) Identify the optimum number of individuals per family to accurately calculate allele frequency and phenotypic mean values per family.

2) Explore the implementation of GWFP in two loblolly pine (*Pinus taeda* L.) populations:

   a) Breeding population CCLONES, 63 families (5-20 individuals per family) (Resende et al., 2012)

   - Four traits (stem diameter, rust, tree stiffness and lignin content).

   - Illumina Infinium assay, resulting in 4,740 polymorphic SNPs.

   b) Simulated population with the same pedigree as CCLONES (de Almeida Filho et al., 2016)

   - Two traits with different genetic architectures: oligogenic (30 QTLs) and polygenic (1,000 QTLs)

   - Simulated genome had 5,000 polymorphic loci.

- Phenotypic and genotypic data was pooled at the family level *in silico*.

- Marker effects were estimated with *Bayes-B* and *Bayes-RR* using the package *BGLR* 1.0.4 (Pérez and de los Campos, 2014) in R.
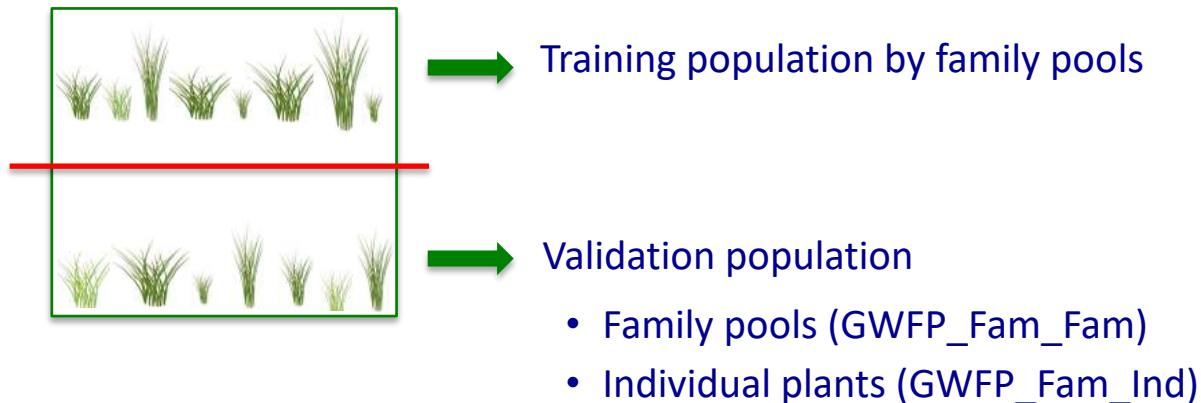
- Cross validation: 10-fold and leave-one-out.

"Genomic Wide Family Prediction (GWFP)"

Validation populations using phenotypic values:

1. <u>Bottom</u>: 10% families with the smallest phenotypic values
2. <u>Top:</u> 10% families having the largest values
3. <u>Bottom-top:</u> 5% from bottom and 5% from trop
4. <u>Middle:</u> 10% families showing phenotypes between extremes
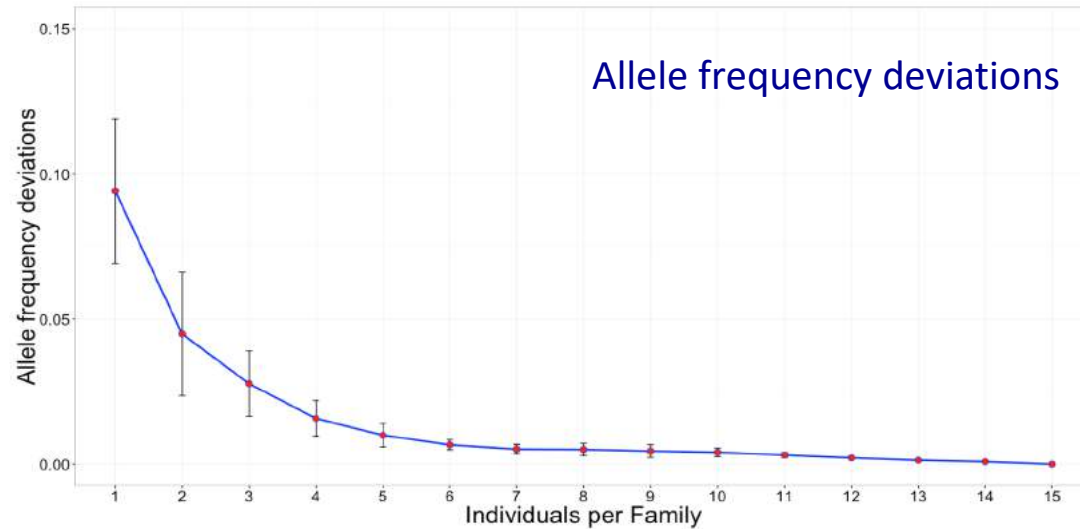5. <u>Combined:</u> 3% of families from bottom, 3% from and top, and 4% from the middle.

10% bottom   **80% middle**   **10% top**

<u>Families with >10 individuals per family</u>

→ Training population by family pools

→ Validation population
   - Family pools (GWFP_Fam_Fam)
   - Individual plants (GWFP_Fam_Ind)

Optimum number of individuals per family

Families with >14 individuals per family

Allele frequency deviations

Families were chosen to represent all possible segregation ratios (1:1 and 1:2:1) in ten SNPs.
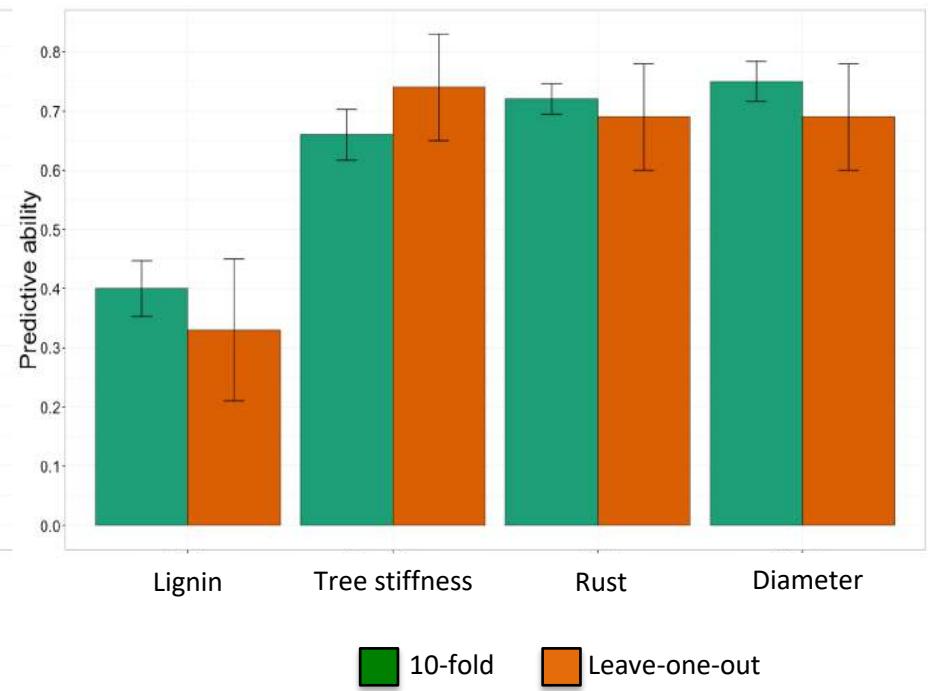

Phenotypic mean deviations
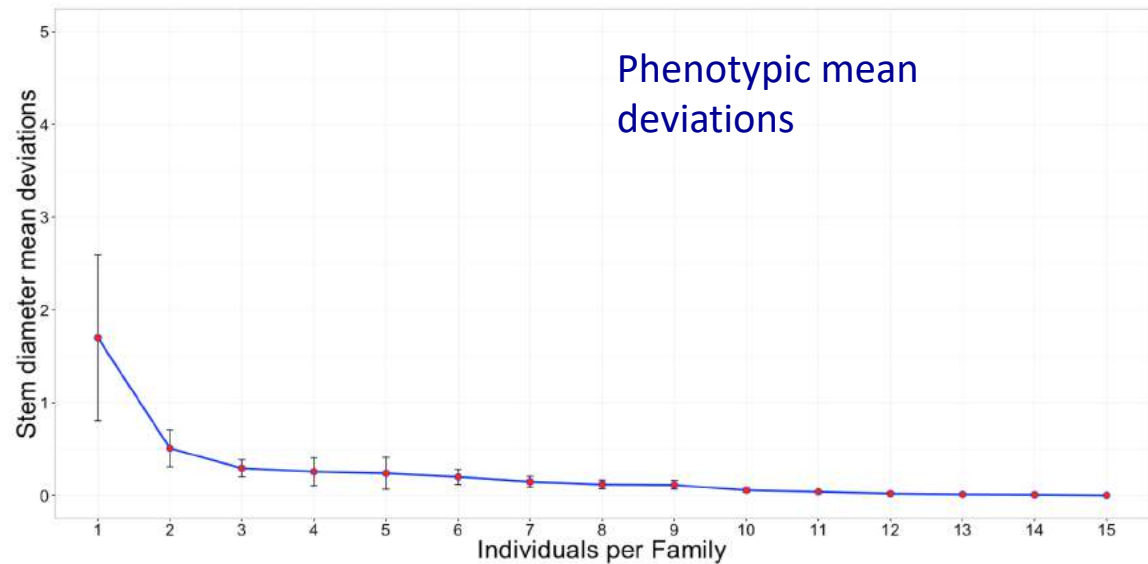
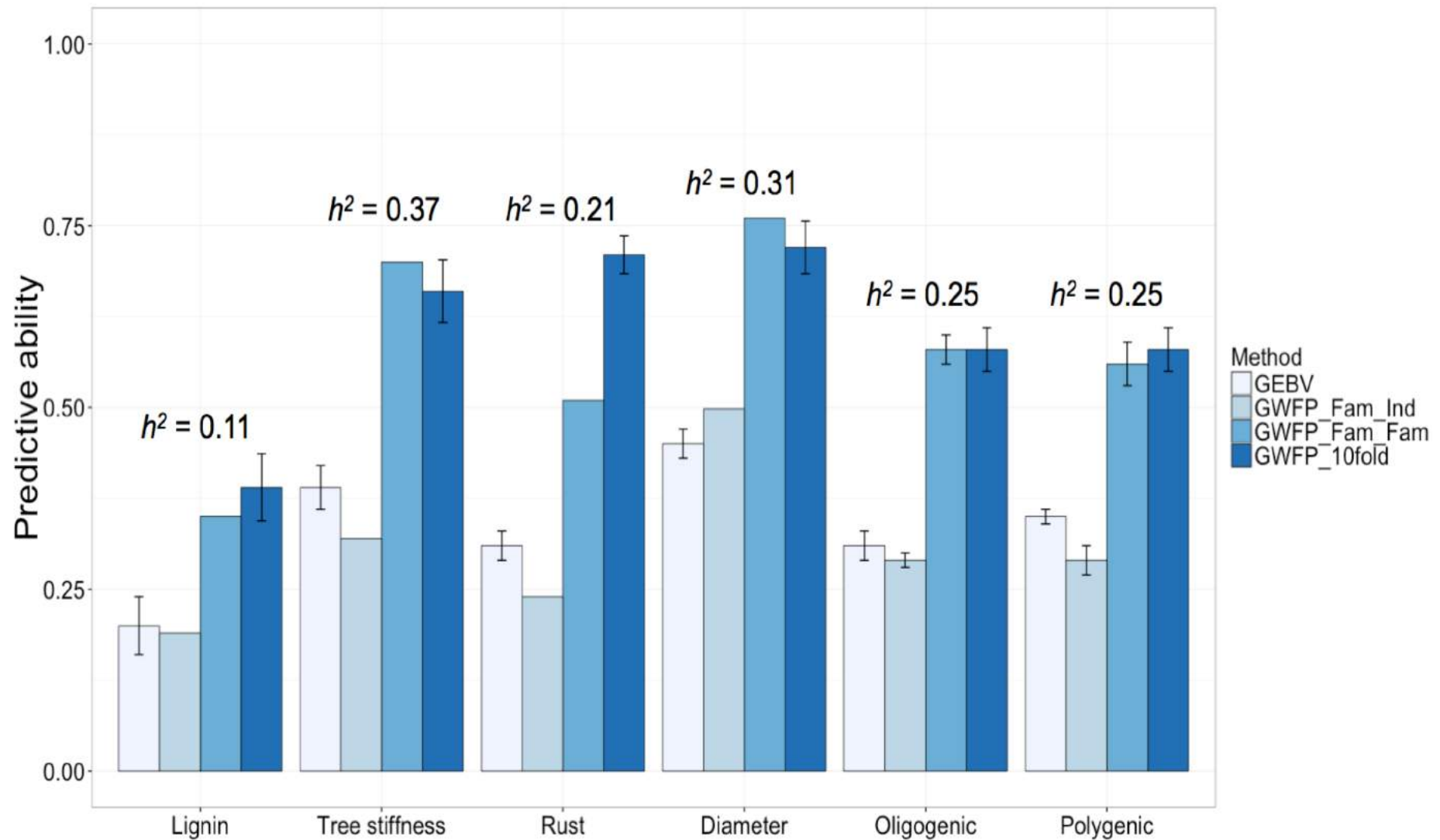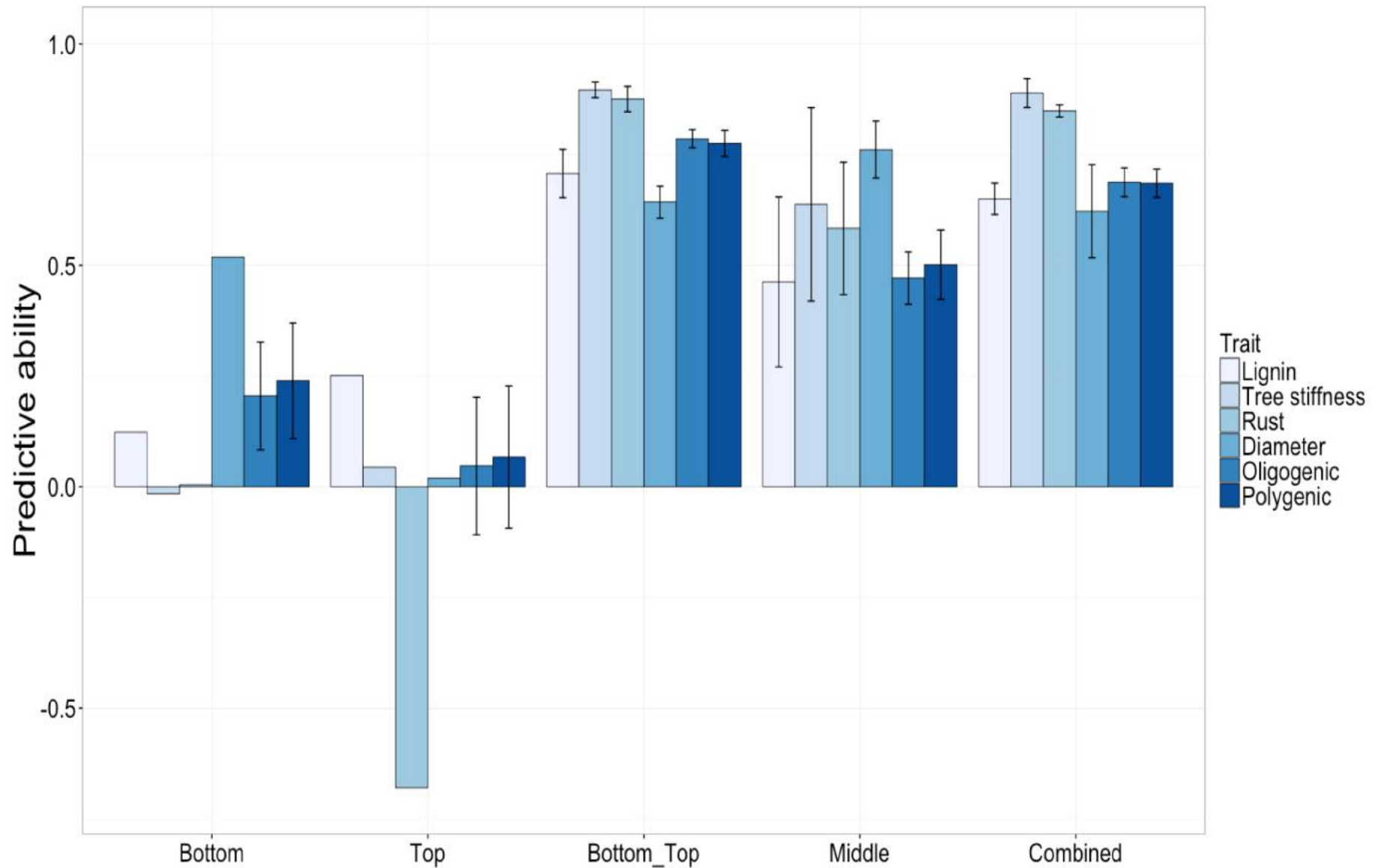# Average predictive ability obtained with *BayesB and Bayes-RR,* and 10-fold and LOO CV
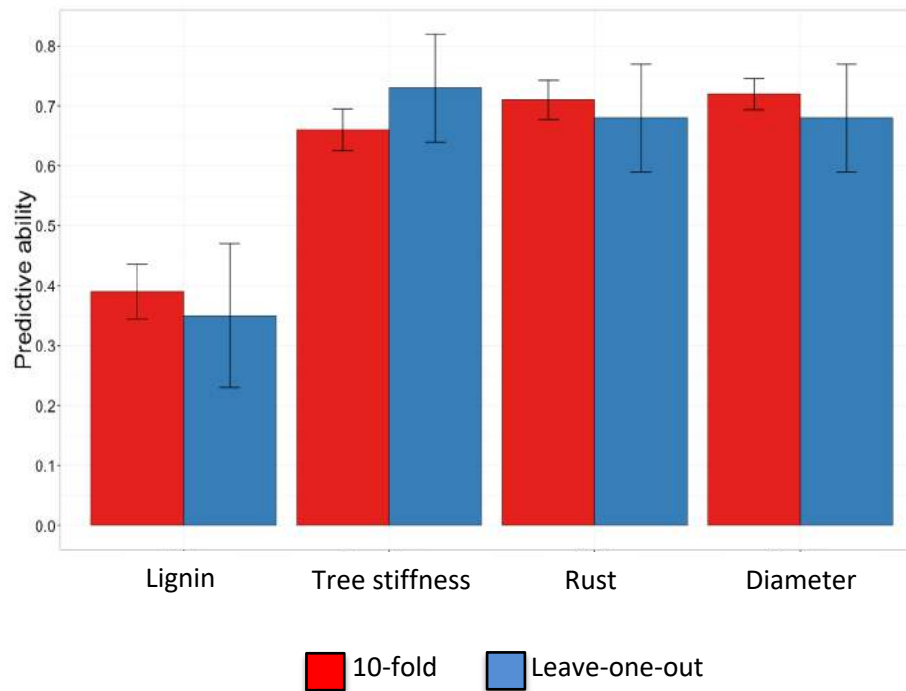
Average predictive ability obtained with *BayesB* and 10-fold CV

Average predictive ability obtained with *BayesB* using phenotypic classes as the validation population

Potential Impact

- ✓ Results revealed great potential for using GWFP in breeding programs that select families/populations to develop cultivars
  - o More accurate phenotyping at the plot level
  - o Significant drop in DNA extractions and genotyping costs: 923 trees in the CCLONES breeding population from 63 families.

- ✓ The GWFP_Fam_Ind approach could be applied to any other crop that select individual plants, taking advantage of the lower costs to genotype the training population.

- ✓ GWAFFs (Byrne et al., 2013)
  - • Fe et al., 2015

🔓 OPEN ACCESS   📄 PEER-REVIEWED

RESEARCH ARTICLE

Genome Wide Allele Frequency Fingerprints (GWAFFs) of Populations via Genotyping by Sequencing

Stephen Byrne ✉, Adrian Czaban, Bruno Studer, Frank Panitz, Christian Bendixen, Torben Asp

Published: March 4, 2013 • http://dx.doi.org/10.1371/journal.pone.0057438

- ✓ GWFP will be implemented in annual ryegrass using 200 half-sib families as the training population.
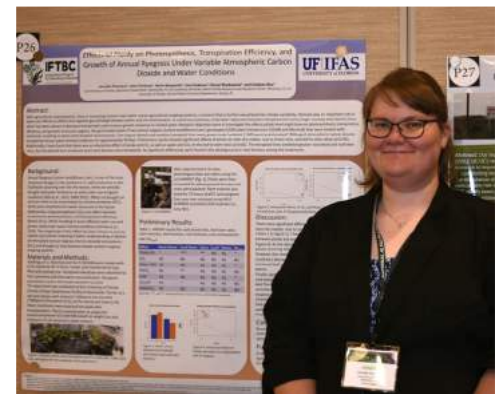
Ryegrass – Genomic Prediction

April-May 2016

February 2017

Jennifer Timmers, PhD student

February 2018

March 2018

April 2018

May 2018

Alfalfa – Genomic Prediction

2rd year after establishment ➔ 33 plants (one from each cultivar)

33 Half-sib

160 Full-sib

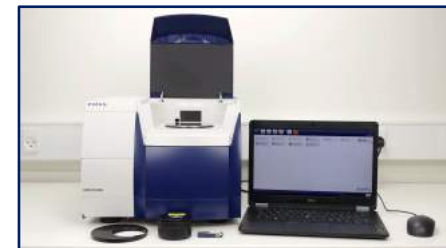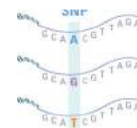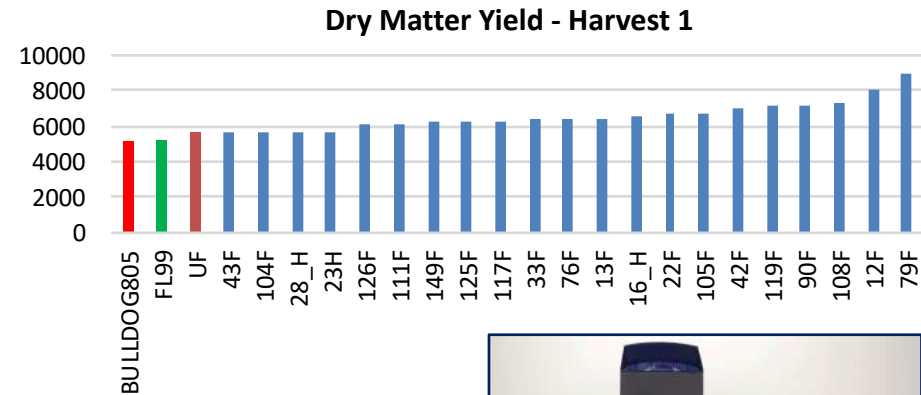
Janam Acharya
MS student

Traits

- Biomass

- Leaf/Stem ratio

- Nutritive value (NIRS)

- Persistence

Phenotype

Allele

frequency (m)

Dry Matter Yield - Harvest 1

Genomic selection

- An important investment will be needed to genotype a large number of individuals for GWS, thus maximizing the use of the markers is a MUST

- Using the genomic information we can:

Construct GWS models

- ~950 individuals
- ~450,000 possible crosses
- Mating allocation under a whole genome evaluation scenario (including dominance) – (Toro and Varona, 2010)
- Estimate of the variance of the future progeny of a given cross

Pedigree corrections

Predicting Crosses