Genomic prediction in a multi-generation *Eucalyptus globulus* breeding population

Geoffrey Haristoy[1,2], Laurent Bouffier[2], Luis Fontes[3], Luis Leal[3], Jorge Paiva[4,5], João-Pedro Pina[3], Jean-Marc Gion[1,2]

[1]AGAP, CIRAD, Montpellier, France
[2]BIOGECO, INRAE, Cestas, France
[3]ALTRI, Quinta do Furaduro, Portugal
[4]IBET, Oeiras, Portugal
[5]CECOLAB, Oliveira do Hospital, Portugal

B4EST Final Meeting
June 20th, 2022

Economic importance of *Eucalyptus globulus*

- Estimated planted area (Tomé et al. 2021):
  - 2.3 million ha in the world
  - 0.8 million ha in Portugal

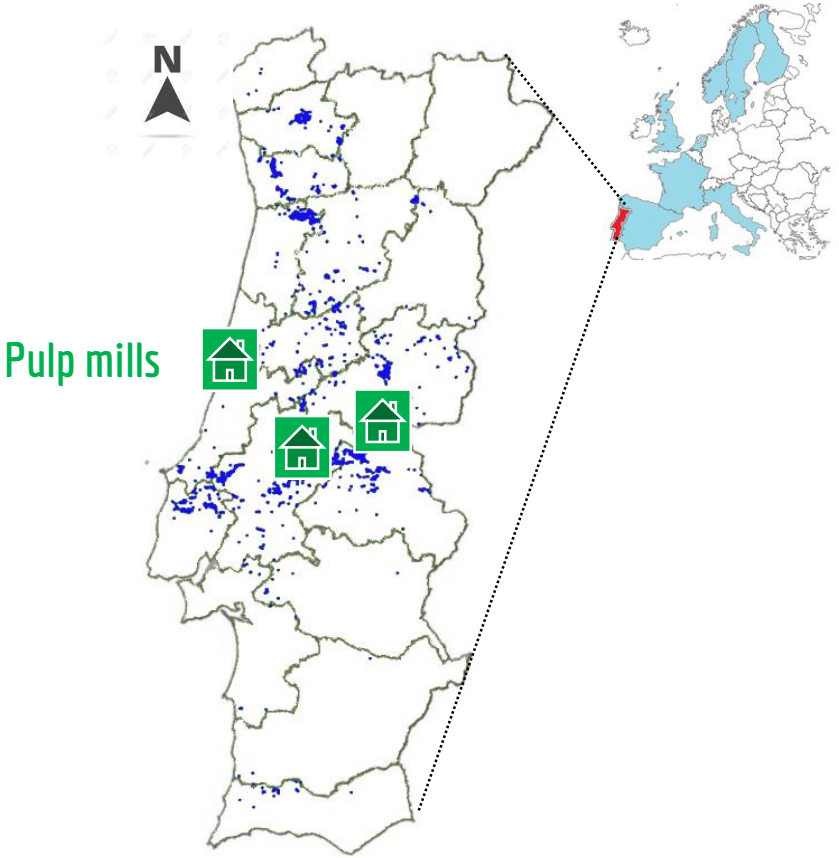
- Various end-uses: paper pulp, firewood, essential oils, ...

- Interesting properties for forestry
  - Fast and continuous growth
  - Can recover after clear-cuttting (up to 3 times)
  - Cloning
  - Good wood properties for pulp industry

Eucalyptus globulus and ALTRI in Portugal

ALTRI's activities : from seed to pulp

- ⟶ Breeding programme (starting in 1966)
- ⟶ Nurseries
- ⟶ Plantations (80 000 ha)
- ⟶ Paper mills



ALTRI's *Eucalyptus globulus* breeding programme

- Recurrent selection for growth and survival

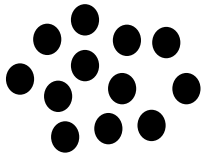
- More than 81 520 genotypes in 31 genetic trials accross Portugal
  → 3 generations of improvment (G0, G1, G2)
  → Around 700 clones

- New challenges for the breeding programme :
  → Ensure documented pedigree
  → Test new approaches : genomic selection
  → *select for new criteria : biotic/abiotic constraints*

- Principle of genomic selection (GS) proposed by Meuwissen et al. (2001)

Cross-validation (CV) : "Genotyped population" split into "**training**" and "**validation**" populations

Genotyped Population



Training Population (TP)

Validation Population (VP)

Using dEBV to train the model

GS model

Genomic estimated breeding value (GEBV)

Accuracy = correlation between dEBV and GEBV

- More than 20 GS proof-of-concept in *Eucalyptus* (Lebedev et *al.*, 2021)
  - -› Pedigree errors rarely cleansed
  - -› few genetic trials
  - -› few generations

1/ Analyse the genetic relatedness within the genotyped population

-› highlight pedigree errors

-› investigate the consequences for breeding value estimation

2/ Assess the accuracy of GS for tree major traits (GBLUP) :

-› growth (height, diameter)

-› survival

3/ Investigate GS accuracy over generations (GBLUP)

-› Different scenarios of progeny cross-validation

-› Test the ability of GS to predict within-family variability

Data from the *Eucalyptus globulus* breeding programme of ALTRI-florestal

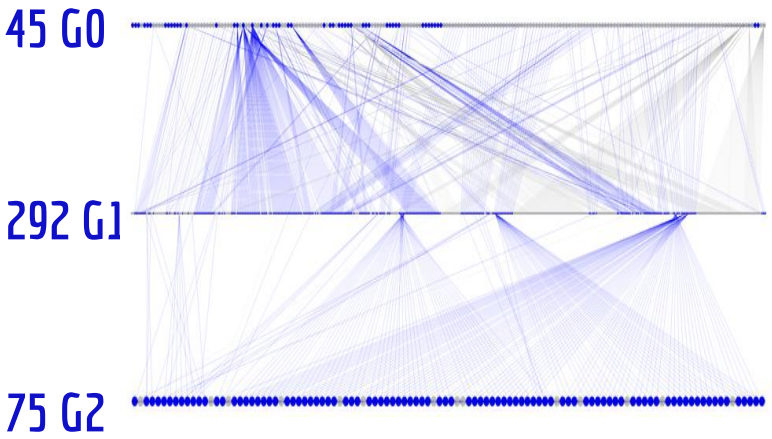
Phenotyped population

81 520 genotypes
31 genetic trials
3 traits: height, diameter, survival

Genotyped population

412 clones

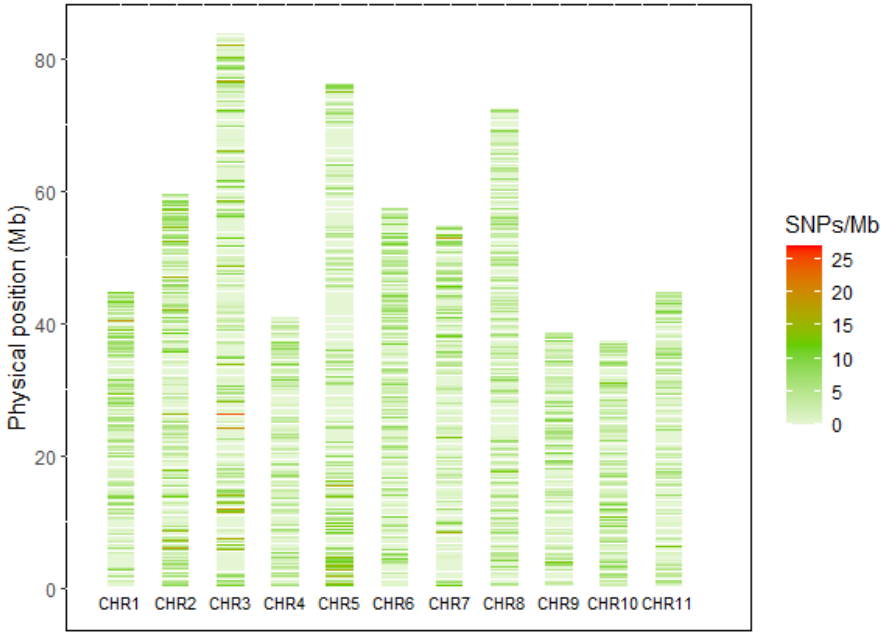
Genotyped population
spreading over 3 generations



= Three large full-sibs families in G2

Eucalyptus multi-species AXIOM array :

- 412 genotypes were genotyped for 68,055 SNPs
- Final set : **14,696** polymorphic SNPs (MAF › 0.05)
  ⟶ uniformly distributed along the 11 chromosomes

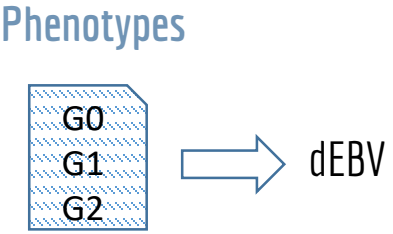


Distribution of the 14,696 SNPs along the genome

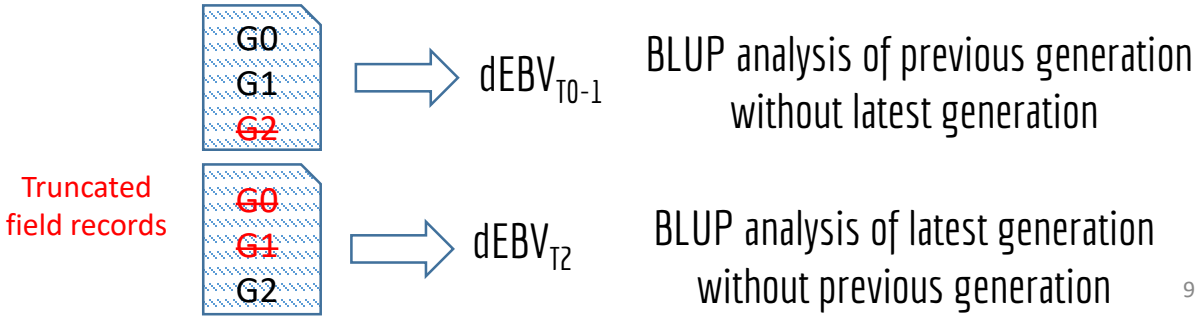
Pseudo-phenotypes used in GS

= Deregressed breeding value (dEBV) derived from pedigree-based BLUP analysis

1. Using all records available
   -› Not ideal for GS across generations



2. Using truncated records (according to generations)



BLUP analysis from :

All record available

Truncated records

All record available
Truncated records

All record available
Truncated records

	Pseudo-phenotypes	Selected genotypes	Pseudo-phenotypes	Selected genotypes	Iteration	Accuracy
S0	$dEBV_C$	321 randomly selected	$dEBV_C$	80 randomly selected	100	$r(GEBV_C, dEBV_C)^{***}$
$S1_a$	$dEBV_C$	44 G0	$dEBV_C$	284 G1	-	$r(GEBV_C, dEBV_C)$
$S1_b$	$dEBV_C$	44 G0, 57 G1*	$dEBV_C$	227 G1	100	$r(GEBV_C, dEBV_C)$
$S2_a$	$dEBV_C$	44 G0, 284 G1	$dEBV_C$	73 G2	-	$r(GEBV_C, dEBV_C)$
$S2_b$	$dEBV_C$	44 G0, 284 G1	$dEBV_{C\text{-}T2}$	73 G2	-	$r(GEBV_C, dEBV_{C\text{-}T2})$
$S2_c$	$dEBV_{C\text{-}T01}$	44 G0, 284 G1	$dEBV_C$	73 G2	-	$r(GEBV_{C\text{-}T01}, dEBV_C)$
$S2_d$	$dEBV_{C\text{-}T01}$	44 G0, 284 G1	$dEBV_{C\text{-}T2}$	73 G2	-	$r(GEBV_{C\text{-}T01}, dEBV_{C\text{-}T2})$
$S3_a$	$dEBV_C$	44 G0, 284 G1, 14 G2**	$dEBV_C$	59 G2	100	$r(GEBV_C, dEBV_C)^{***}$
$S3_b$	$dEBV_C$	44 G0, 284 G1, 14 G2**	$dEBV_{C\text{-}T2}$	59 G2	100	$r(GEBV_C, dEBV_{C\text{-}T2})^{***}$
$S4_a$	$dEBV_C$	44 G0, 284 G1, 14 FS2	$dEBV_C$	13 FS2	100	$r(GEBV_C, dEBV_C)^{***}$
$S4_b$	$dEBV_C$	44 G0, 284 G1, 14 FS2	$dEBV_{C\text{-}T2}$	13 FS2	100	$r(GEBV_C, dEBV_{C\text{-}T2})^{***}$

* 20% of the G1 from POP_{GS} ; ** 20% of the three main G2 FS, and 20% of all the other G2 families ; *** mean scenario accuracy over the 100 iterations

(From Haristoy et al., Submitted)

One conventional cross-validation scenario

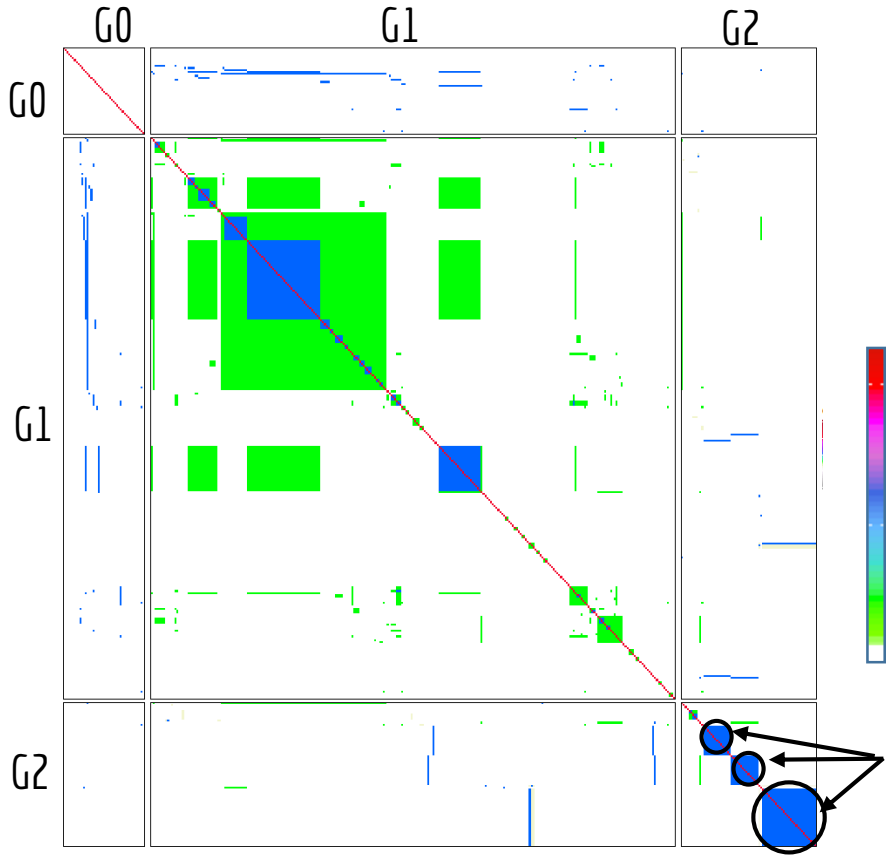
10 progeny cross-validation scenario

	TP		VP			
	Pseudo-phenotypes	Selected genotypes	Pseudo-phenotypes	Selected genotypes	Iteration	Accuracy
S0	$dEBV_C$	321 randomly selected	$dEBV_C$	80 randomly selected	100	$r(GEBV_C, dEBV_C)^{***}$
$S1_a$	$dEBV_C$	44 G0	$dEBV_C$	284 G1	-	$r(GEBV_C, dEBV_C)$
$S1_b$	$dEBV_C$	44 G0, 57 G1*	$dEBV_C$	227 G1	100	$r(GEBV_C, dEBV_C)$
$S2_a$	$dEBV_C$	44 G0, 284 G1	$dEBV_C$	73 G2	-	$r(GEBV_C, dEBV_C)$
$S2_b$	$dEBV_C$	44 G0, 284 G1	$dEBV_{C\text{-}T2}$	73 G2	-	$r(GEBV_C, dEBV_{C\text{-}T2})$
$S2_c$	$dEBV_{C\text{-}T01}$	44 G0, 284 G1	$dEBV_C$	73 G2	-	$r(GEBV_{C\text{-}T01}, dEBV_C)$
$S2_d$	$dEBV_{C\text{-}T01}$	44 G0, 284 G1	$dEBV_{C\text{-}T2}$	73 G2	-	$r(GEBV_{C\text{-}T01}, dEBV_{C\text{-}T2})$
$S3_a$	$dEBV_C$	44 G0, 284 G1, 14 G2**	$dEBV_C$	59 G2	100	$r(GEBV_C, dEBV_C)^{***}$
$S3_b$	$dEBV_C$	44 G0, 284 G1, 14 G2**	$dEBV_{C\text{-}T2}$	59 G2	100	$r(GEBV_C, dEBV_{C\text{-}T2})^{***}$
$S4_a$	$dEBV_C$	44 G0, 284 G1, 14 FS2	$dEBV_C$	13 FS2	100	$r(GEBV_C, dEBV_C)^{***}$
$S4_b$	$dEBV_C$	44 G0, 284 G1, 14 FS2	$dEBV_{C\text{-}T2}$	13 FS2	100	$r(GEBV_C, dEBV_{C\text{-}T2})^{***}$

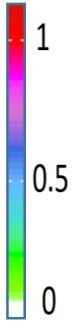
* 20% of the G1 from POP_{GS} ; ** 20% of the three main G2 FS, and 20% of all the other G2 families ; *** mean scenario accuracy over the 100 iterations

(From Haristoy et al., Submitted)

1. Genetic structure of GS population



- Relationship using A coefficients
  - -› few contributors
  - -› highlight familiy structure



-› self-relationship

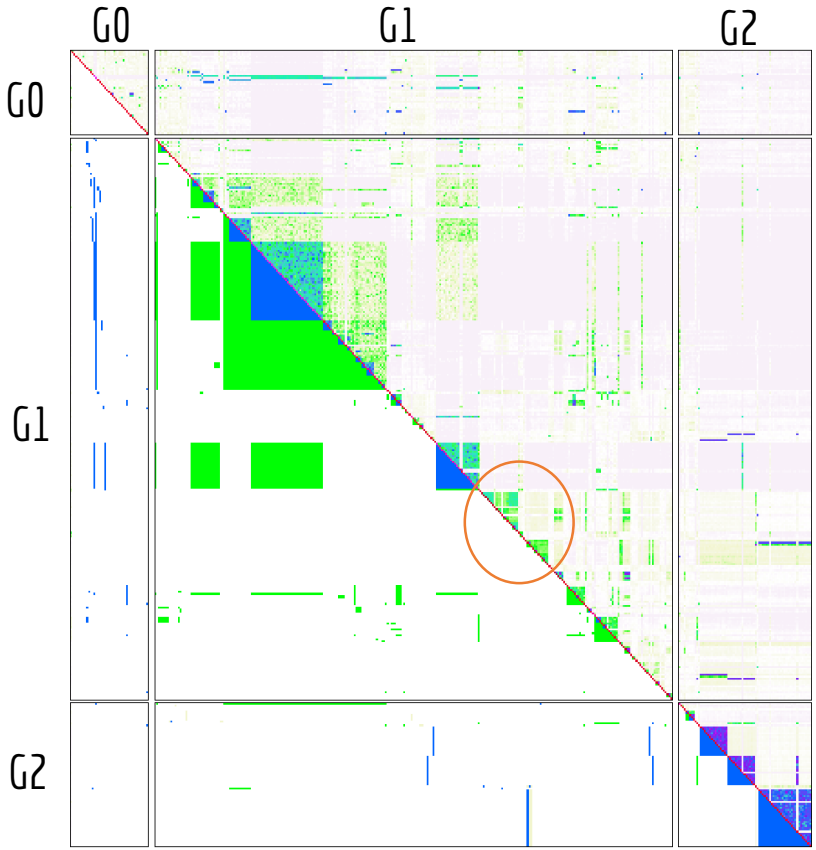
-› Parent/progeny or full-sibs

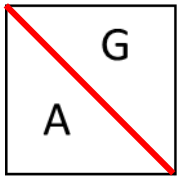
-› Grandparents/progeny and half sibs

-› unrelated

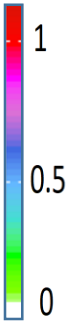
Three full-sibs families

1. Genetic structure of GS population





- Relationship using A coefficients
  -› 6 classes of relationships
- The G matrix reveals:
  unknown relationships



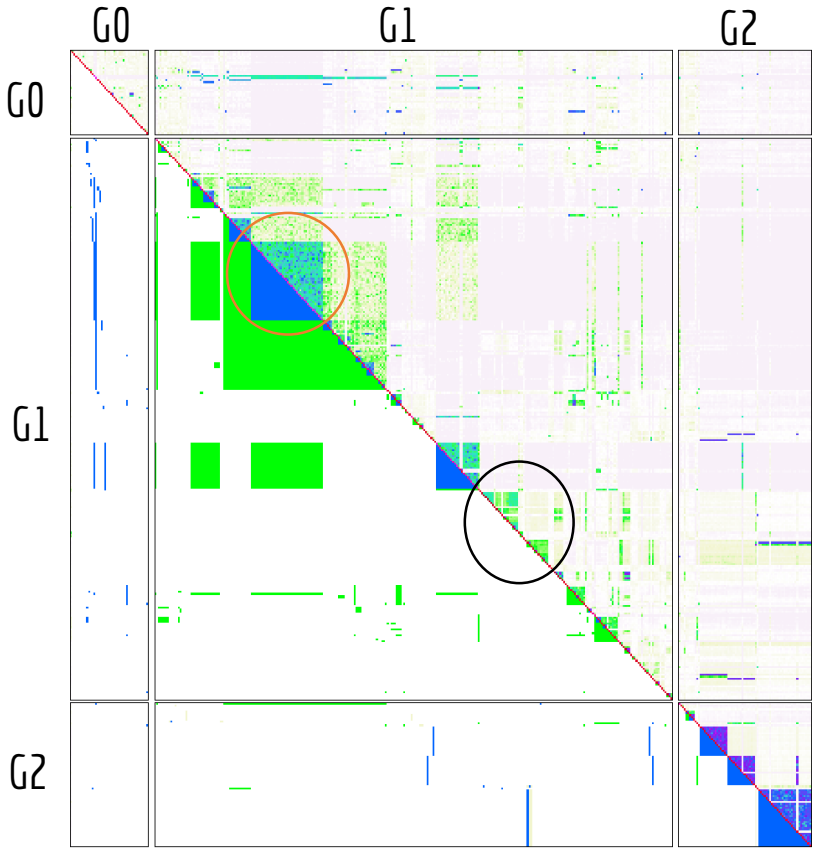
-› self-relationship

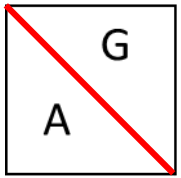
-› Parent/progeny or full-sibs

-› Grandparents/progeny and half sibs

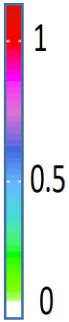
-› unrelated

1. Genetic structure of GS population





- Relationship using A coefficients
  -› 6 classes of relationships
- The G matrix reveals:
  unknown relationships
  mendelian sampling within families



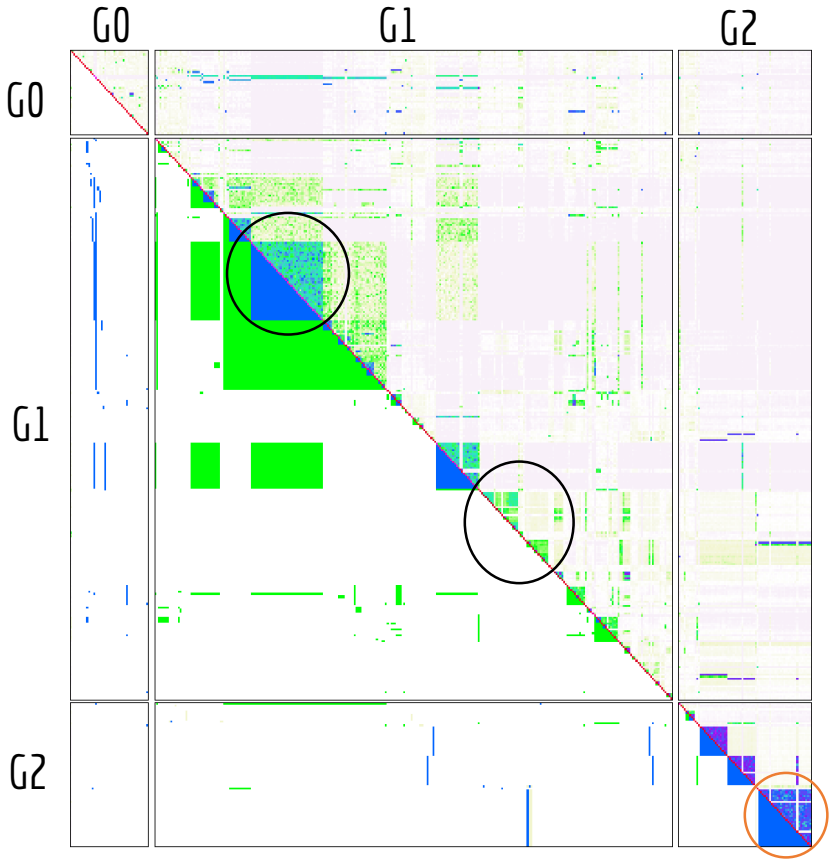
-› self-relationship

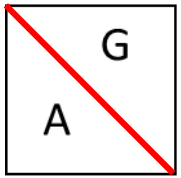
-› Parent/progeny or full-sibs

-› Grandparents/progeny and half sibs

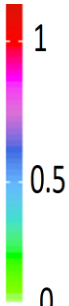
-› unrelated

1. Genetic structure of GS population





- Relationship using A coefficients
  -› 6 classes of relationships
- The G matrix reveals:
  unknown relationships
  mendelian sampling within families
  Pedigree errors



-› self-relationship

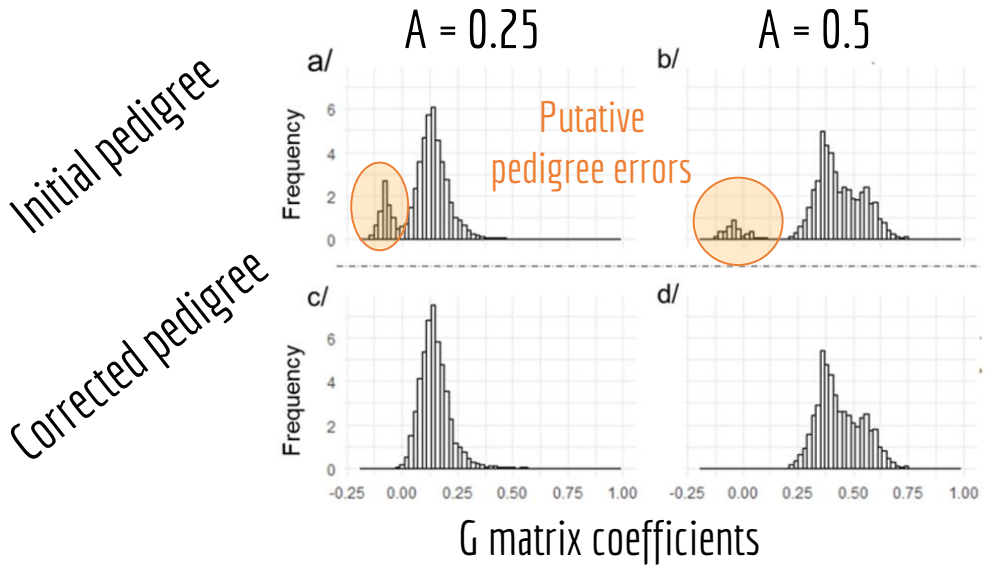
-› Parent/progeny or full-sibs

-› Grandparents/progeny and half sibs

-› unrelated

1. Genetic structure of GS population

Distribution of G coefficients according to A coefficients



Out of our 412 genotypes :

→ 11 pairs of genotypes were genetically identical

→ 45 genotypes had at least one wrong parent

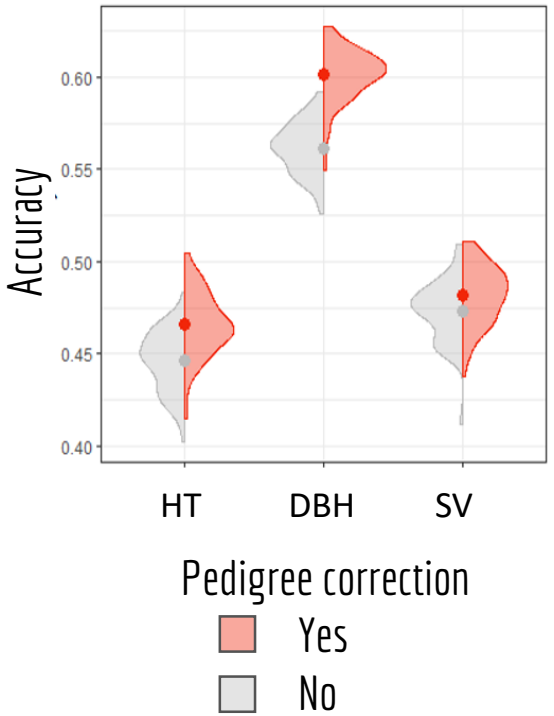
→ Correction of pedigree errors to avoid biais in breeding value estimation

→ After deregression, used as pseudo-phenotype in GS models

2. GS accuracy for tree major traits using conventional cross-validation

	TP	VP
S0	G0 G1 G2 321 dEBV	G0 G1 G2 80 dEBV

Conventional CV
(5 fold, 100 iterations)



Kind of "average" accuracy (mix of different generations)

Consistent with what is generally found in the literature
(r^2 ranging from 0.4 to 0.8)

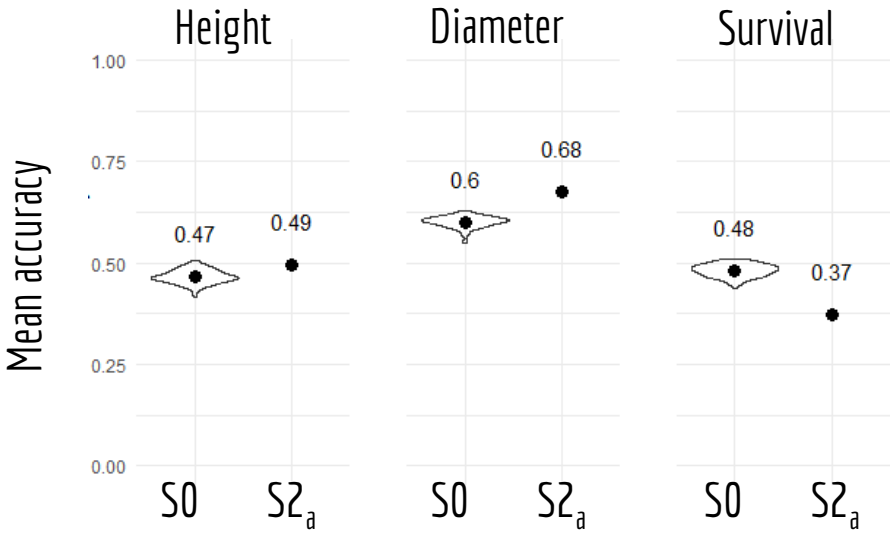
-› Pedigree correction improved prediction accuracies

3. GS accuracy evolution across one generation (progeny validation)

Expectation : LD is degraded through generations (r^2 S1 < r^2 S0)

S0 vs $S2_a$

	TP	VP
S0	G0 G1 G2 321 dEBV	G0 G1 G2 80 dEBV
$S2_a$	G0 G1 328 dEBV	G2 73 dEBV



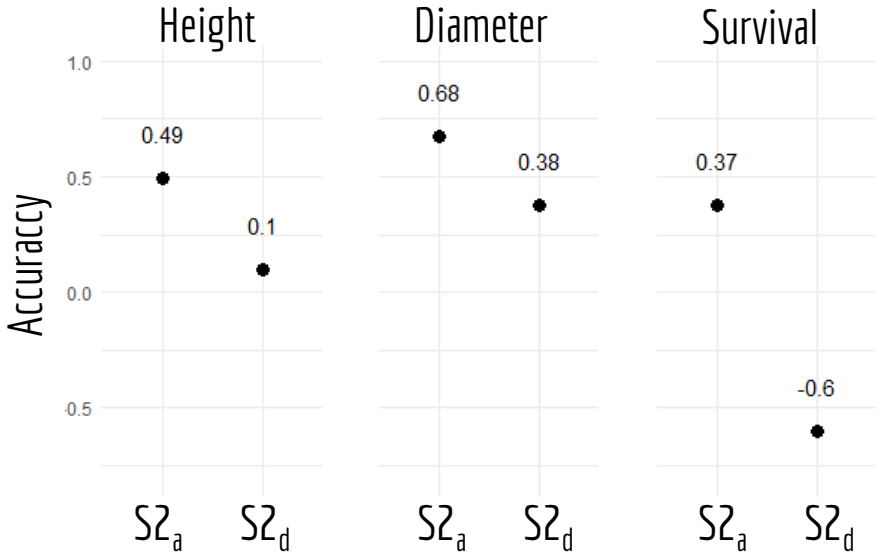
-› Accuracy maintained from one generation to the next for growth traits

3. GS accuracy evolution across one generation (progeny validation)

Expectation : LD is degraded through generations (r^2 S1 < r^2 S0)

$S2_a$ vs $S2_d$

	TP	VP
$S2_a$	G0 G1 328 dEBV	G2 73 dEBV
$S2_d$	G0 G1 328 $dEBV_{T01}$	G2 73 $dEBV_{T2}$



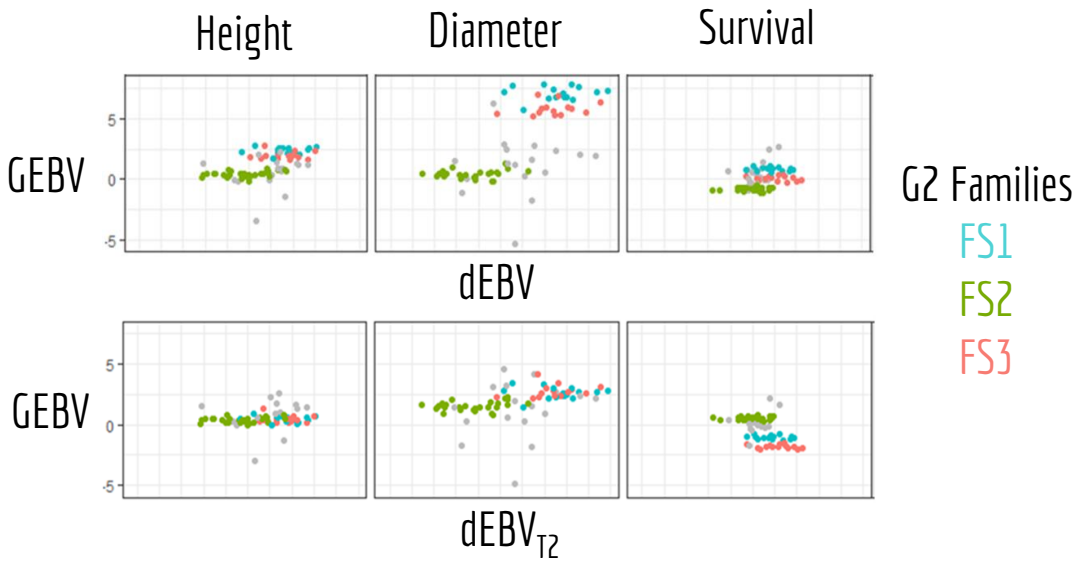
Drop of the accuracy in $S2_d$ compared to $S2_a$ for all traits
Negative accuracy for survival in $S2_d$?

3. GS accuracy evolution across one generation (progeny validation)

Expectation : LD is degraded through generations (r^2 S1 < r^2 S0)

$S2_a$ vs $S2_d$

	TP	VP
$S2_a$	G0 G1 328 dEBV	G2 73 dEBV
$S2_d$	G0 G1 328 $dEBV_{T01}$	G2 73 $dEBV_{T2}$



G0/G1 EBV seems strongly dependent on the progeny (particularly for the parents of FS1 and FS3)

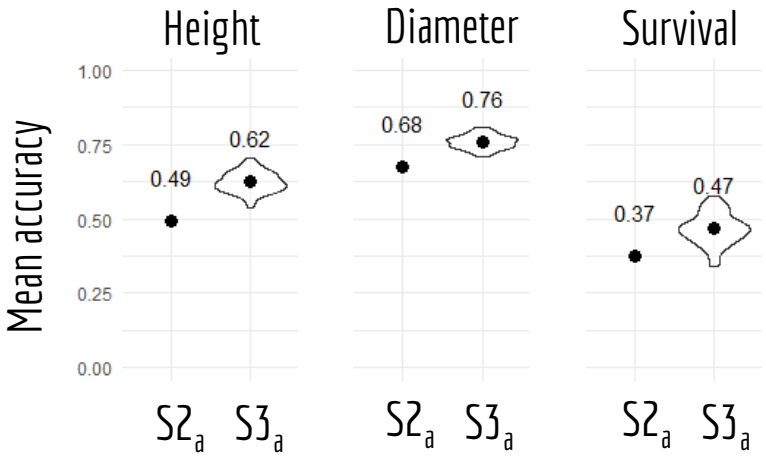
-› Different abiotic and/or biotic constraints affecting previous (G0 and G1) and last generation (G2) ?

3. Effect of adding some G2 in TP on average accuracy

Expectation : Higher accuracy when adding G2 in the training population

$S2_a$ vs $S3_a$

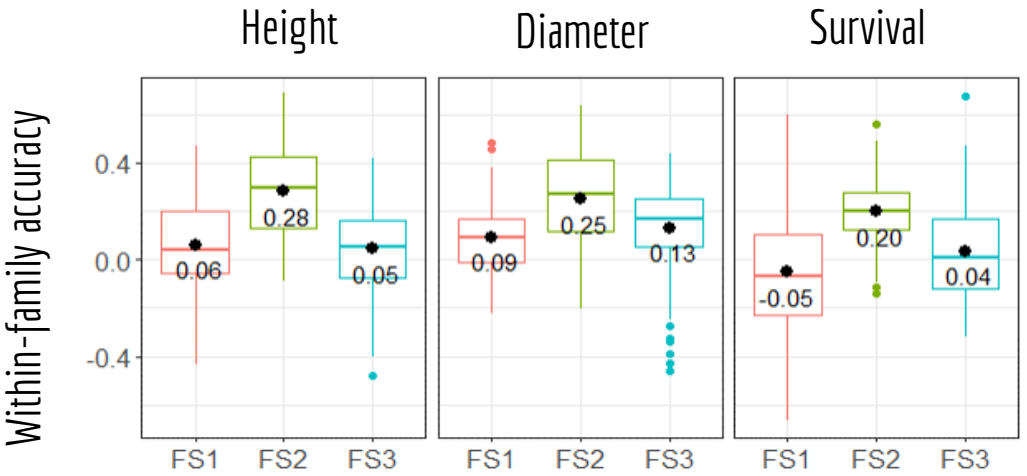
	TP	VP
$S2_a$	G0 G1 328 dEBV	G2 73 dEBV
$S3_a$	G0 G1 G2 328 + 14 dEBV	G2 59 dEBV



-› **Adding only 14 G2 into training set drastically improved predictions**

3. Effect of adding some G2 in TP for predicting within-family variability

$S3_a$	G0 G1 G2 328 + 14 dEBV	G2 59 dEBV



Various results according families and the trait considered (from -0.05 up to 0.28 in average)

-› Poor ability to predict intra-family variability

- G matrix analysis revealed the real structure of the population
- Pedigree correction always resulted in an improvement of the accuracy
- Reasonable accuracy when aggregating data from the first two generations to predict in the latest
  → more challenging when using pseudo-phenotypes from truncated data sets
- Adding only few G2 in the training set drastically improved accuracies of the remaining G2
- Within-family prediction remains challenging

Outlook to further analysis

- Enlarging genotyping of full-sibs family + optimization of TP and VP to improve within-family prediction
- Testing more complex models (multi-traits, ssgBLUP, …)
- Extend the analysis to more contrasted environments (include GxE)

Perspectives for ALTRI 's breeding

- Allows pedigree correction
- Allows pre-screening of best families before trial establishment

-› Strategies for partial and optimized genotyping of the breeding population still required to be investigated

Thank you for your attention

B4EST Final Meeting
June 20th, 2022