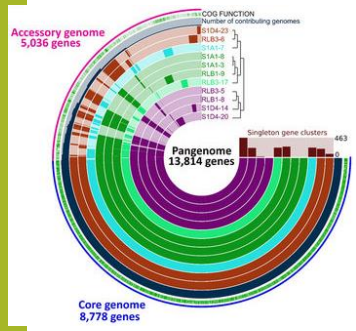
Estimation of cross value

Sophie Bouchet

INRAE, DIGEN

DIGEN



Methodology to evaluate genomic diversity

bioinformatics

bio-analyse

Evaluation of genomic diversity

Molecular biology and development

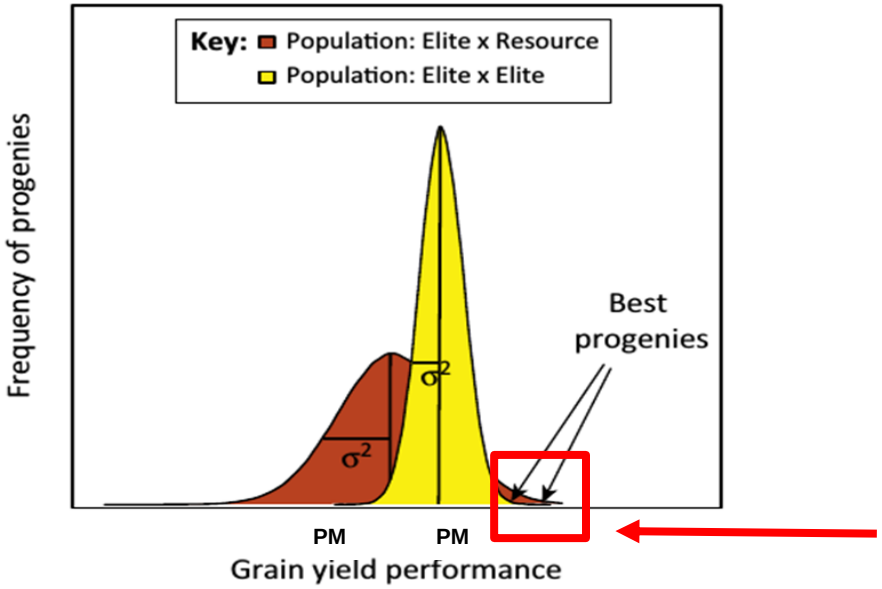
Breeding and pre-breeding methodology

Quantitative genetics

Breeding and pre-breeding

Breeding / pre-breeding

How to rank best crosses?



Rank crosses using right-hand tails

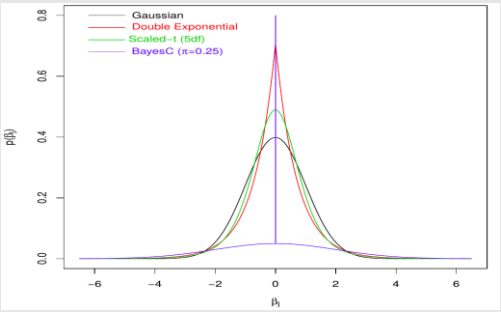
Longin et al. 2014

Data base

phenotypes / genotypes

2000 lines (French registered cultivars: GEVES + INRAE-AO lines)

Estimation of marker effects (genomic prediction model)



Meuwissen, 2001

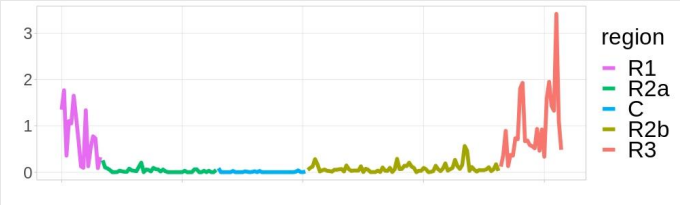
Genotyped diversity panel

632 CRB landraces

Estimation of recombination rate between markers

(~genetic distance c for a bi-parental population:

estimated on all bread wheat polymorphic markers)



Danguy et al, 2021

Vector of 35K SNP effects

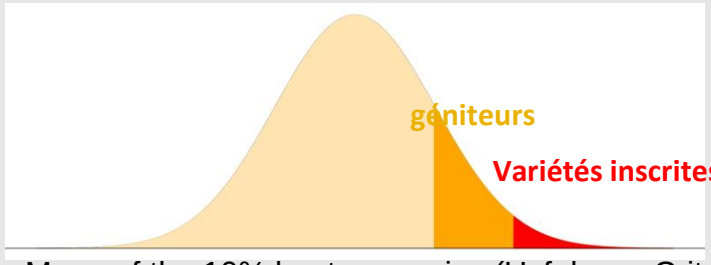
+

Vector of recombination rate

+

Matrix of genitors' genotypes

Estimation of cross value (Cross Selection Criteria ; CSC)



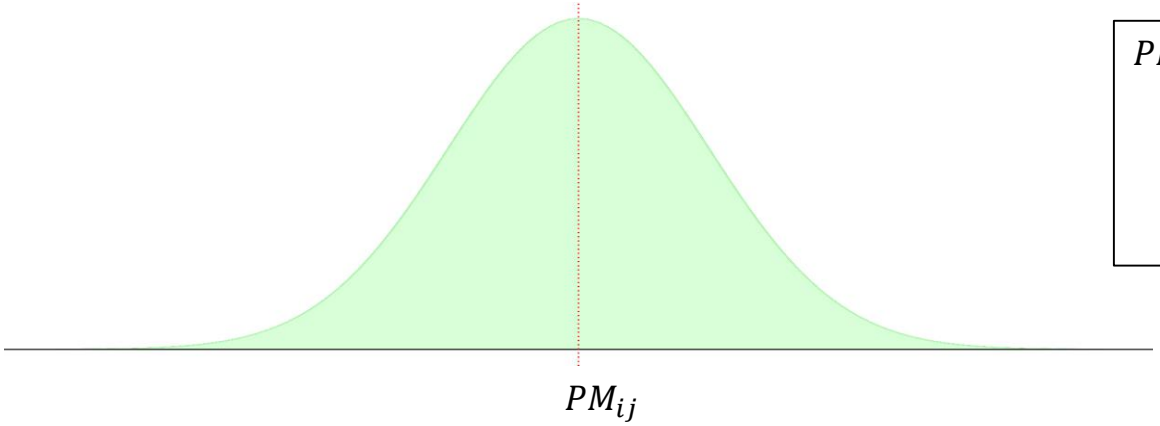
Mean of the 10% best progenies (Usfulness Criterion ; UC), probability to get one progeny > threshold)

Danguy et al, submitted

How to optimize mating plan ?

#1: Expected mean of progeny (PM)

Progeny of P_i * P_j ~ N(PM_{ij}, σ_{ij}^2)



PM_{ij} = Expected mean of progeny

= Mean of parents

= ($\widehat{g_i}$ + $\widehat{g_j}$)/2

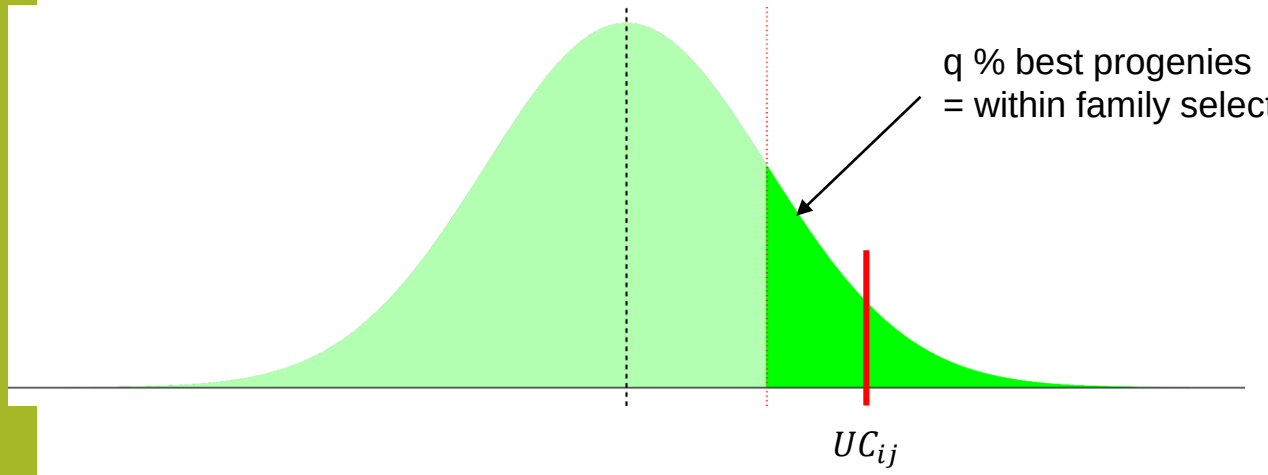
PM = The most classical criteria to choose crosses
+ no genotyping

How to optimize mating plan ?

Schnell and Utz 1975
Zhong and Jannink 2007

#2: Expected mean of best progenies (UC)

Progeny of P_i * P_j ~ N(PM_{ij}, σ_{ij}^2)



$$UC_{ij}^{q} = PM_{ij} + i_q \sigma_{ij}$$

Which q% to choose ?

q = 7% → UC1 = $PM_{ij} + 2 * \sigma_{ij}$
q = 0.01% → UC2 = $PM_{ij} + 4 * \sigma_{ij}$

How to compute progeny variance?

Formula of Lehermeier et al. 2017 for Doubled Haploids

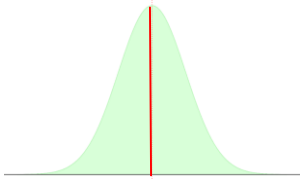
$$\sigma^2_{P_i \times P_j} = V1_{ij} + V2_{ij}$$

Genetic diversity of parents

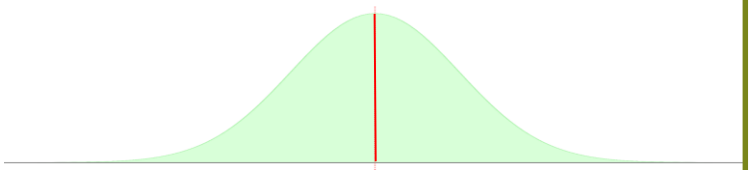
$$V1_{ij} = \sum_{m=1}^{M} \beta^2{}_m * 4p_{m\,ij}(1 - p_{m\,ij})$$

β = effects of allele
p = allelic frequency

Similar parents



Very different parents



Recombination of desirable alleles

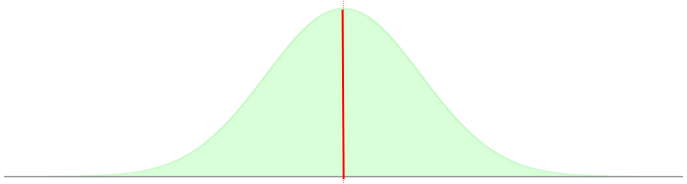
$$V2_{ij} = \sum_{l<m} \beta_l \beta_m * 4D_{lm\,ij}(1 - 2r_{lm})$$

D: covariance between alleles
= linkage desequilibirium
(monomorphism: D = 0)

r = Frequency of recombinants

P_i	P_j
-	+
-	+

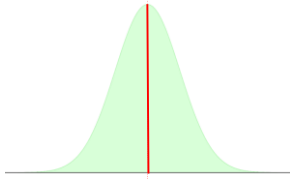
Coupling (D = 0.25)



Recombination is not desirable

P_i	P_j
+	-
-	+

Repulsion (D = - 0.25)



Recombination is desirable

How to compute progeny variance?

$$\sigma^2_{P_i \times P_j} = \sum_{m=1}^{M} \beta^2{}_m * 4p_{m\,ij}(1 - p_{m\,ij}) + \sum_{l<m} \beta_l \beta_m * 4D_{lm\,ij}(1 - 2r_{lm})$$

QTLs

effects

Polymorphism at QTLs

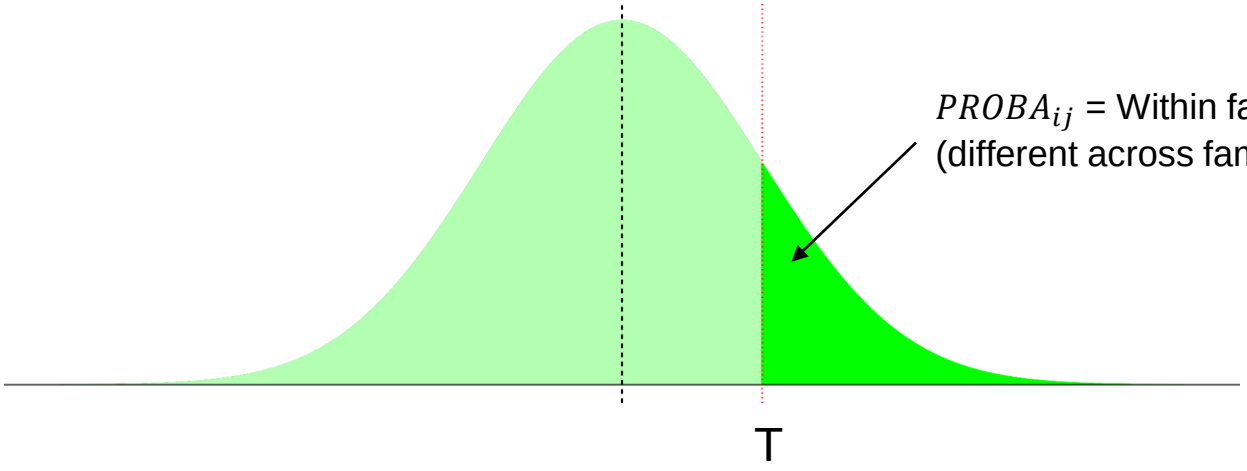
Estimated with Genomic Prediction model

How to optimize mating plan ?

Wellmann 2019
Bijma et al. 2020

#3: Probability to produce a progeny ≥ T (PROBA)

Progeny of P_i * P_j ~ N(PM_{ij}, σ_{ij}^2)



$PROBA_{ij}$ = Within family selection rate (different across families)

T

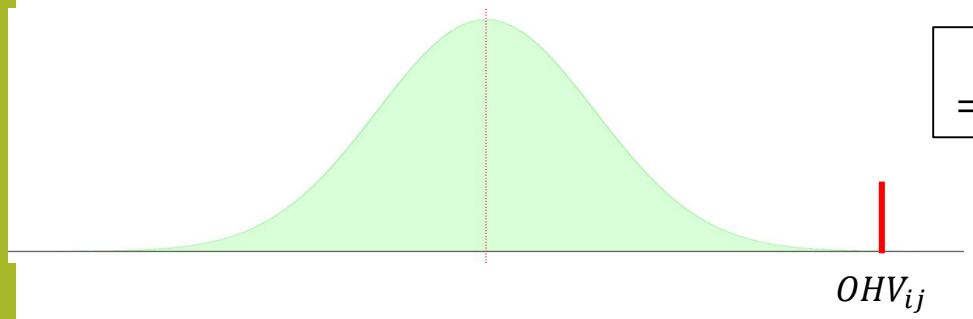
= Breeding value of the best parental line

Daetwyler et al. 2015

How to optimize mating plan ?

#4: Best « theoretical » progeny (OHV)

Progeny of P_i * P_j ~ N(PM_{ij}, σ_{ij}^2)



OHV_{ij}
= Sum of most desirable alleles of parents at each locus

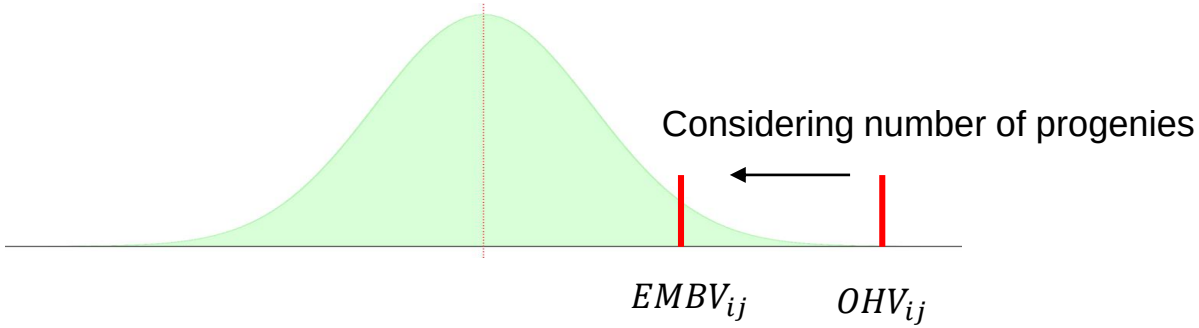
Probability of OHV ?

Müller & al. 2018

How to optimize mating plan ?

#5: Expected value of the best progeny considering the number of progenies allocated to the cross (EMBV)

Progeny of P_i * P_j ~ N(PM_{ij}, σ_{ij}^2)



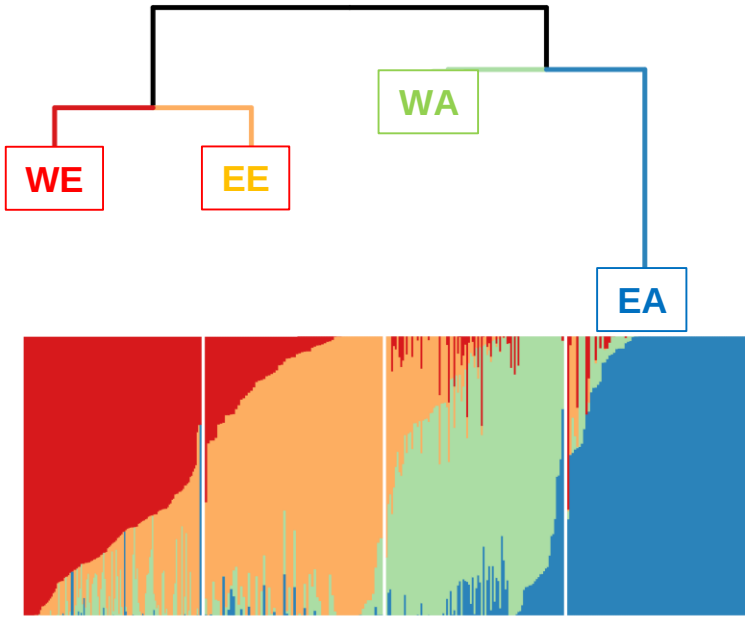
EMBV of a cross depends on number of progenies

Estimation of recombination rate
phD, Alice Danguy Des Déserts

371 bread wheat landraces sampled worldwide (Balfourier et al. 2019, Science Advances)
130k SNP of TABW410k (Kitt et al. 2021, Zenodo)

4 differentiated bread wheat populations

Europe ⟷ Asia (Fst)

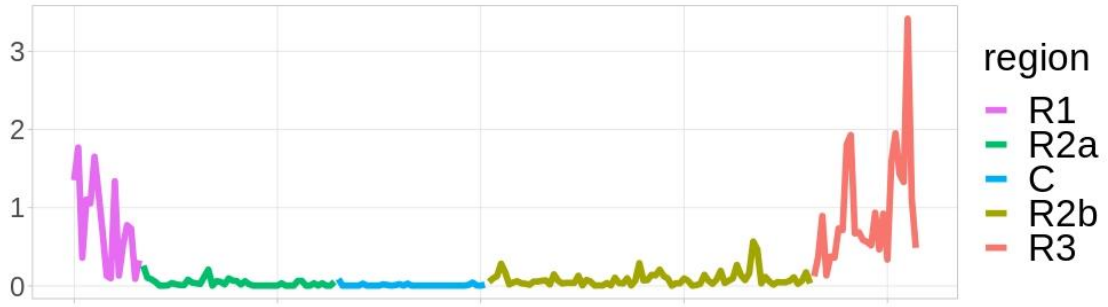


- Do the recombination profiles of these 4 populations vary ?
- Run PHASE (Li et Stephens 2003, Genetics) to estimate a proxy of c (recombination rate)

Estimation of recombination rate

Alice Danguy Des Déserts

- Recombination rates are globally colinear between populations and with bi-parental c estimates
- The more divergent the populations, the more LD patterns differentiate
- We use WE recombination vector when we work on French material



FSOV Predicropt

Claire Oget

Estimate our ability to predict cross value

- Estimation of 6 Cross Selection Criteria (CSC) (Danguy et al, submitted)for Yield, Prot, Heading date, Height
- Evaluation of traits in the field: 100 crosses x 60 progenies

Estimation of marker effects

Optimisation of the training population

Geno: 18501 SNP
Pheno:

Dataset	Zone	# environments	# lines (genotyped)
GEVES	North	494	436 (408)
	South	270	231 (214)
INRAE_AO	North	194	2543 (1581)
	South	94	596 (375)
Both	**Both**	**966**	**3192 (2107)**

Estimation of marker effects

Optimisation of the training population

- 1 Environment = 1 year x 1 location
- Considering INRAE-AO & GEVES separately and together
- Considering North & South separately and together

I. **Rank the environments**

a. Spatially adjusted means for each environment
b. BLUPs for each trait using all the environments except the one to be testted (excluding common lines): TP / VP

c. Marker effects estimation from each TP
d. GEBVs prediction of each VP
e. Prediction accuracy = Pearson correlation between BLUPs & GEBVs → rank environments from the least accurate to the most

II. Remove iteratively the worst environment and estimate GEBV accuracy using cross validation

III. Conclusion:

We do not improve significantly accuracy by removing the worst environments: the data base is clean

The best training population to predict a North or South trial is the total data base (GEVES + INRAE-AO, North + South trials)

GEBV accuracy (cross-validation 60%TP / 40% VP)

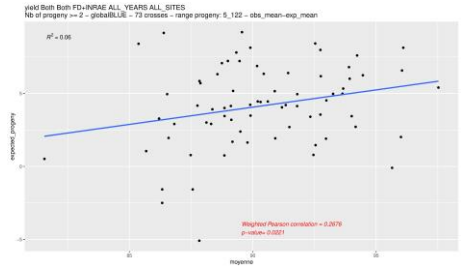
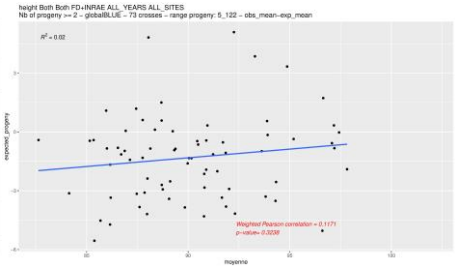
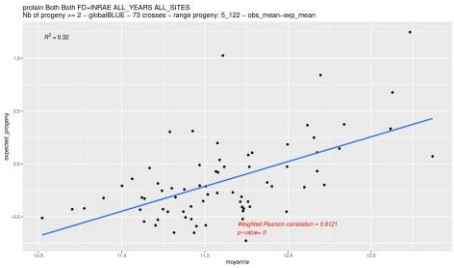
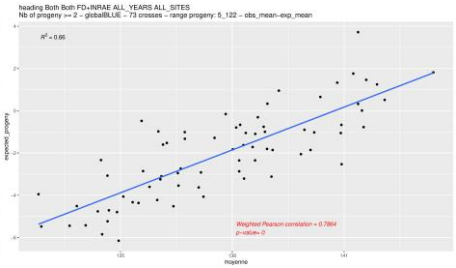
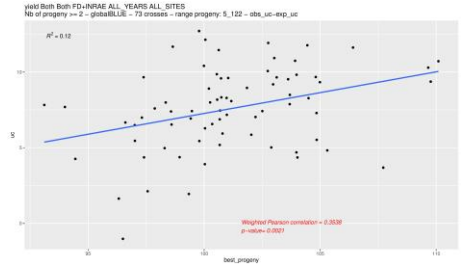
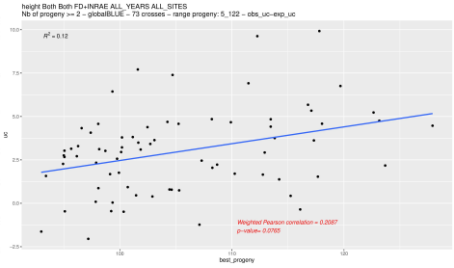
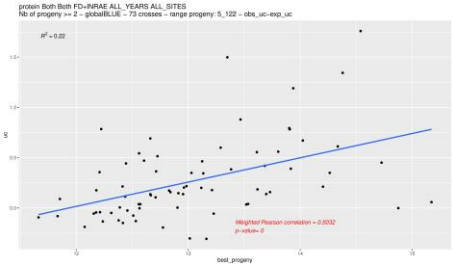
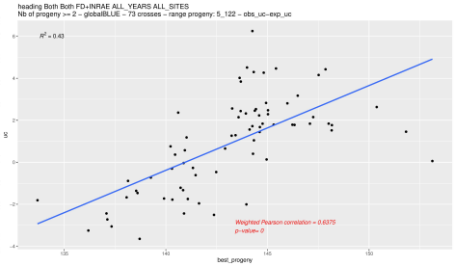
	Yield	Height	Protein	Heading
BayesA	0.62 *(0.018)*	0.60 *(0.025)*	0.60 *(0.019)*	0.81 *(0.013)*
BayesB	0.61 *(0.018)*	0.59 *(0.023)*	0.58 *(0.020)*	0.80 *(0.011)*
BayesC	0.61 *(0.019)*	0.52 *(0.026)*	0.58 *(0.019)*	0.50 *(0.057)*
BL	0.63 *(0.037)*	0.54 *(0.060)*	0.60 *(0.018)*	0.77 *(0.012)*
BRR	0.63 *(0.018)*	0.52 *(0.022)*	0.59 *(0.020)*	0.65 *(0.016)*
rrBLUP	0.62 *(0.017)*	0.52 *(0.017)*	0.60 *(0.018)*	0.65 *(0.015)*

FSOV Predicropt trials

Croisements	2020	2021	2022	*Total*
FD	8	8	0	***16***
CF	0	0	7	***7***
EM	5	21	11	***37***
AO	12	20	10	***42***
Total	***25***	***49***	***28***	***102***

Parcelles	2020-2021	2021-2022	2022-2023	2023-2024	*Total*
FD	916	689	300	0	***1 905***
CF	0	789	800	0	***1 589***
EM	483	790	1 000	0	***2 273***
LU	420	834	1 000	0	***2 254***
AUZ	416	0	840	450	***1 706***
Total	***2 235***	***3 102***	***3 940***	***450***	***9 727***

predicted values *vs.* observed values (73 crosses)

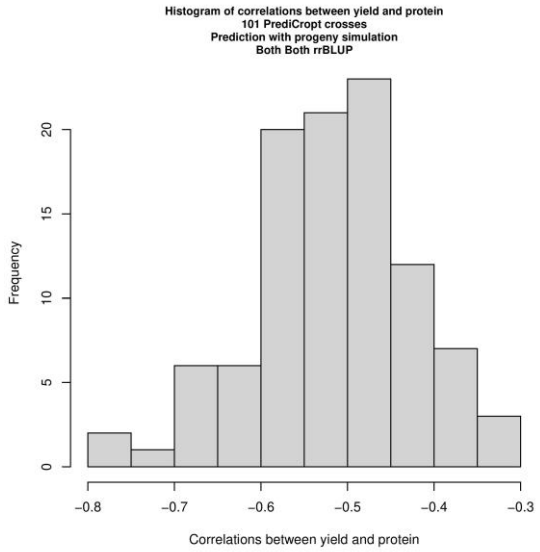
	Yield	Height	Protein	Heading date
PM				
UC				

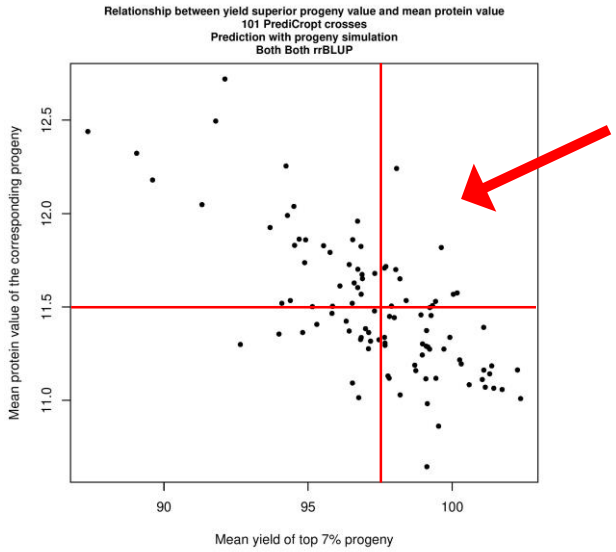
Quelle proportion de croisements faire sur la base de l'UC?

predicted values *vs.* observed values (73 crosses)

Trait	Yield	Height	Protein	Heading
Cor obs./exp. mean	0.27	0.12	0.61	0.79
p-value	*0.02*	*0.32*	*8.71E-09*	*1.68E-16*
Cor obs./exp. sd	-0.15	0.46	-0.04	0.42
p-value	*0.21*	*3.58E-05*	*0.71*	*2.36E-04*
Cor obs./exp. uc	0.35	0.21	0.50	0.64
p-value	*2.14E-03*	*0.08*	*5.71E-06*	*1.32E-09*

Prediction of the correlation yield/prot, GPD+ potential





There is a variation of yield / prot correlation between crosses. Can we predict it? Is it correlated with GPD+ potential? Can we predict the crosses that can product high yield + high prot individuals? For what population size?

Accuracy of UC prediction

Progeny simulation

Trait	Yield		Height		Protein		Heading	
Model	rrBLUP	BayesA	rrBLUP	BayesA	rrBLUP	BayesA	rrBLUP	BayesA
Cor obs./exp. mean	0.29	0.29	0.10	0.11	0.57	0.56	0.79	0.85
p-value	*0.01*	*0.01*	*0.39*	*0.34*	*1.48E-07*	*2.08E-07*	*4.67E-17*	*2.93E-21*
Cor obs./exp. sd	-0.17	-0.19	0.45	0.60	-0.03	0.07	0.36	0.42
p-value	*0.16*	*0.11*	*7.28E-05*	*2.75E-08*	*0.83*	*0.55*	*1.98E-03*	*1.91E-04*
Cor obs./exp. uc	0.43	0.42	0.14	0.44	0.49	0.50	0.63	0.60
p-value	*1.69E-04*	*1.81E-04*	*0.23*	*8.19E-05*	*9.66E-06*	*8.34E-06*	*1.87E-09*	*2.65E-08*

Analytic Formula

Trait	Yield		Height		Protein		Heading	
Model	rrBLUP	BayesA	rrBLUP	BayesA	rrBLUP	BayesA	rrBLUP	BayesA
Cor obs./exp. mean	0.29	0.29	0.10	0.12	0.57	0.57	0.80	0.84
p-value	*0.01*	*0.01*	*0.39*	*0.31*	*1.50E-07*	*1.90E-07*	*4.13E-17*	*2.33E-20*
Cor obs./exp. sd	-0.17	-0.16	0.45	0.59	-0.02	0.08	0.36	0.36
p-value	*0.15*	*0.17*	*6.46E-05*	*2.88E-08*	*0.87*	*0.51*	*2.05E-03*	*1.52E-03*
Cor obs./exp. uc	0.43	0.43	0.15	0.48	0.49	0.50	0.63	0.53
p-value	*1.77E-04*	*1.64E-04*	*0.20*	*1.92E-05*	*9.42E-06*	*7.07E-06*	*1.76E-09*	*1.53E-06*

Analytic formula (quick) give the same results than progeny simulations
Variable selection models (BAYES A in particular) are better for sd estimation when there are major QTLs (heading date, plant height)

Ability to predict the trait mean of the 7% best yield progenies of a cross

Progeny simulation

Trait	Yield	
Model	rrBLUP	BayesA
Cor obs./exp. cor resp. Height	0.26	0.16
p-value	*0.02*	*0.17*
Cor obs./exp. cor resp. Protein	0.39	0.38
p-value	*7.45E-04*	*8.45E-04*
Cor obs./exp. cor resp. Heading	0.61	0.65
p-value	*1.13E-08*	*4.85E-10*

Analytic Formula

Trait	Yield	
Model	rrBLUP	BayesA
Cor obs./exp. cor resp. Height	0.27	0.16
p-value	*0.02*	*0.18*
Cor obs./exp. cor resp. Protein	0.39	0.38
p-value	*7.51E-04*	*8.29E-04*
Cor obs./exp. cor resp. Heading	0.61	0.65
p-value	*1.09E-08*	*5.73E-10*

Figures to keep in mind

r	Yield	Height	Protein	Heading
accuracy GEBV (data INRAE-AO)	0.56	0.35	0.56	0.38
accuracy GEBV (data GEVES-INRAE-AO)	0.63	0.6	0.6	0.81
accuracy UC (r)	0.43	0.5	0.5	0.64
repetability phenotype	0.55	0.54	0.72	0.89

Predict mating plan design pipeline of optimisation

Vector of cross value (CSC)

+

Constraints : budget, diversity threshold,

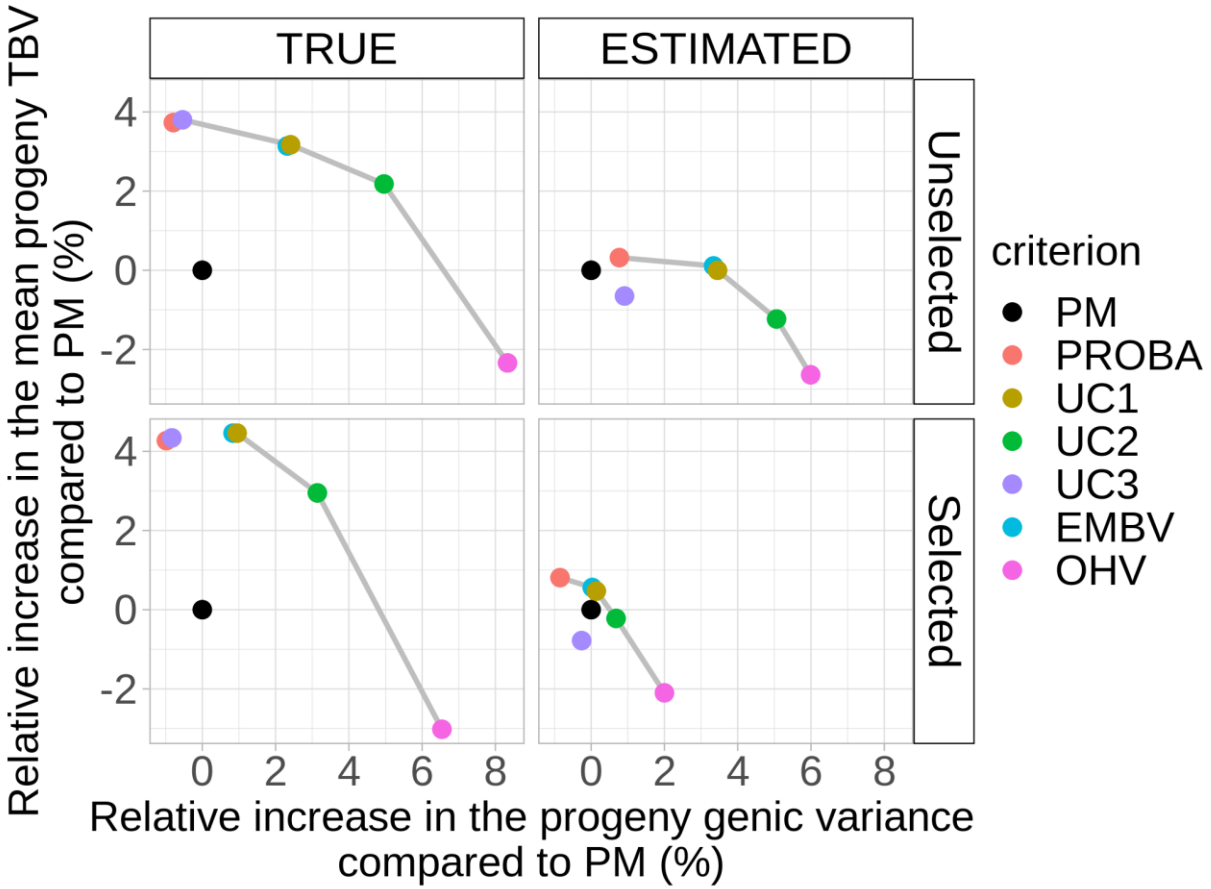
Major alleles: quality, disease…

Genetic algorithm
(Danguy et al., submitted)

Optimisation of mating design

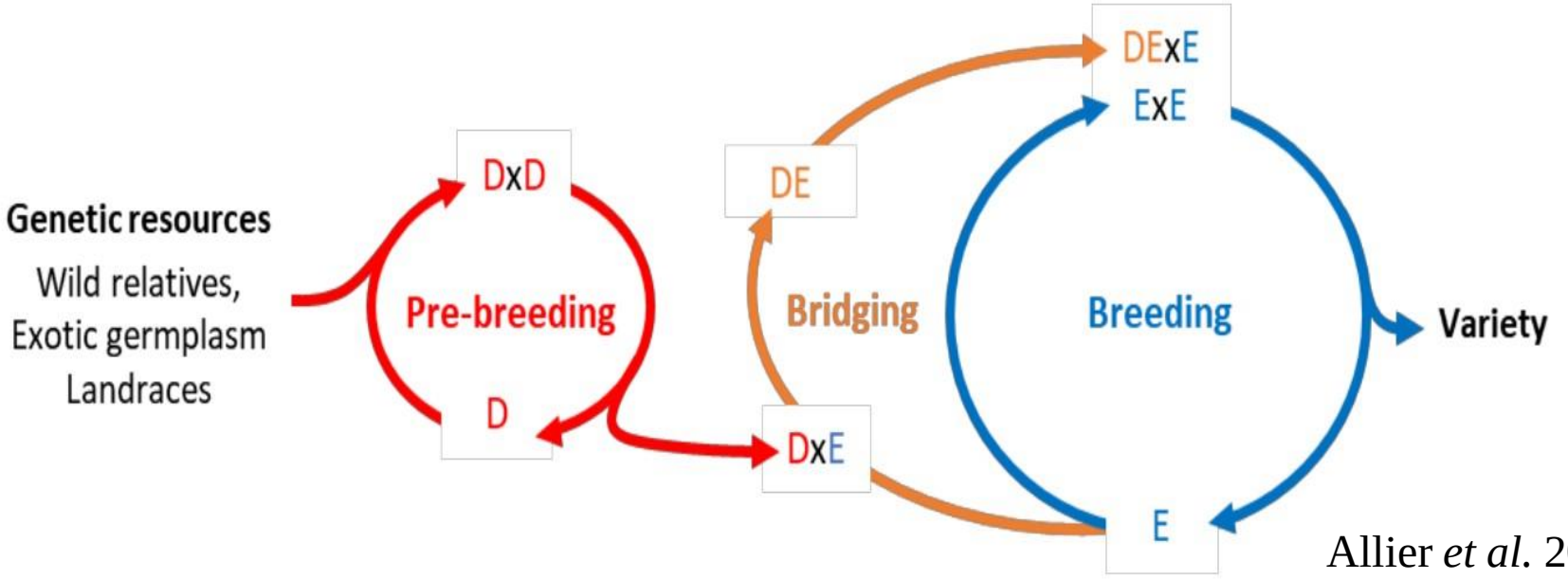
cross	nb of progenies
$P_1 * P_2$	D_{12}
$P_i * P_j$	D_{ij}
….	

Conclusion



- Optimisation of cross design using CSC based on sd estimation increase genetic gain and / or maintain more diversity
- Genetic gain depends on var (sd) / var (PM) in the cross progenies

Objectif: Optimisation des plans de croisements / ré-introduction de diversité



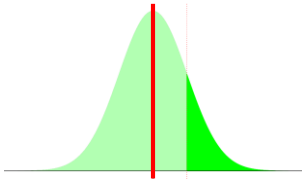
Allier *et al.* 2020

Contraintes maladies
Contraintes qualité
Contraintes diversité

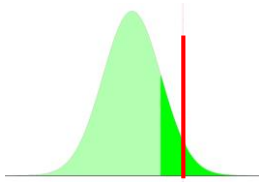
Merci pour votre attention

- Unités expérimentales (essais au champ)
- FD (croisements et essais)

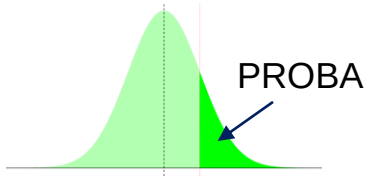
Summary of criteria



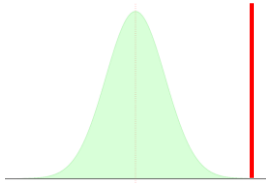
PM



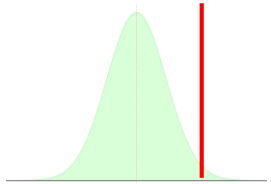
UC



T



OHV



EMBV

Which criteria provides best progenies?