Validation of cross progeny variance genomic prediction using simulations and experimental data in winter elite bread wheat

C. Oget-Ebrad, E. Heumez, L. Duchalais, E. Goudemand-Dugué, F.-X. Oury, **J.-M. Elsen, S. Bouchet**

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

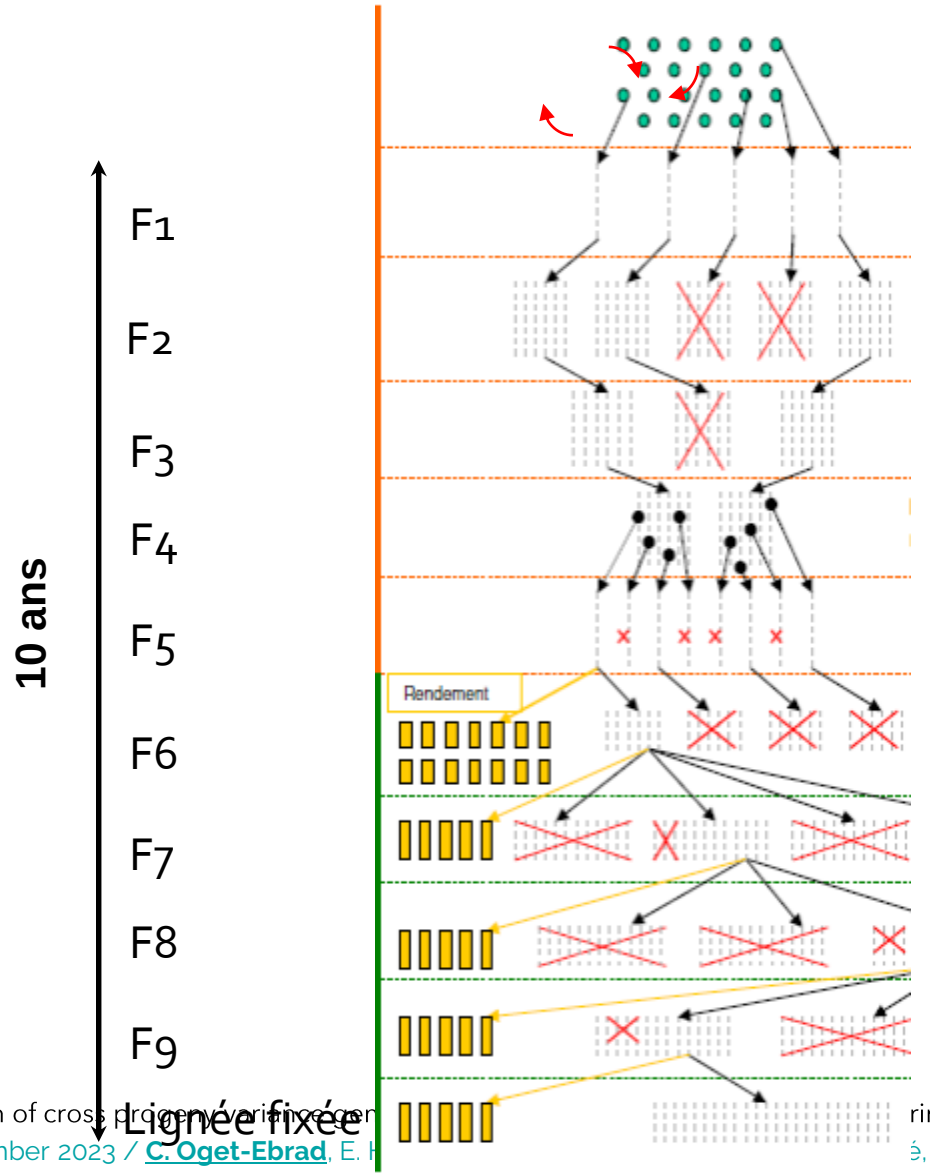
Fsov

FLORIMOND DESPREZ

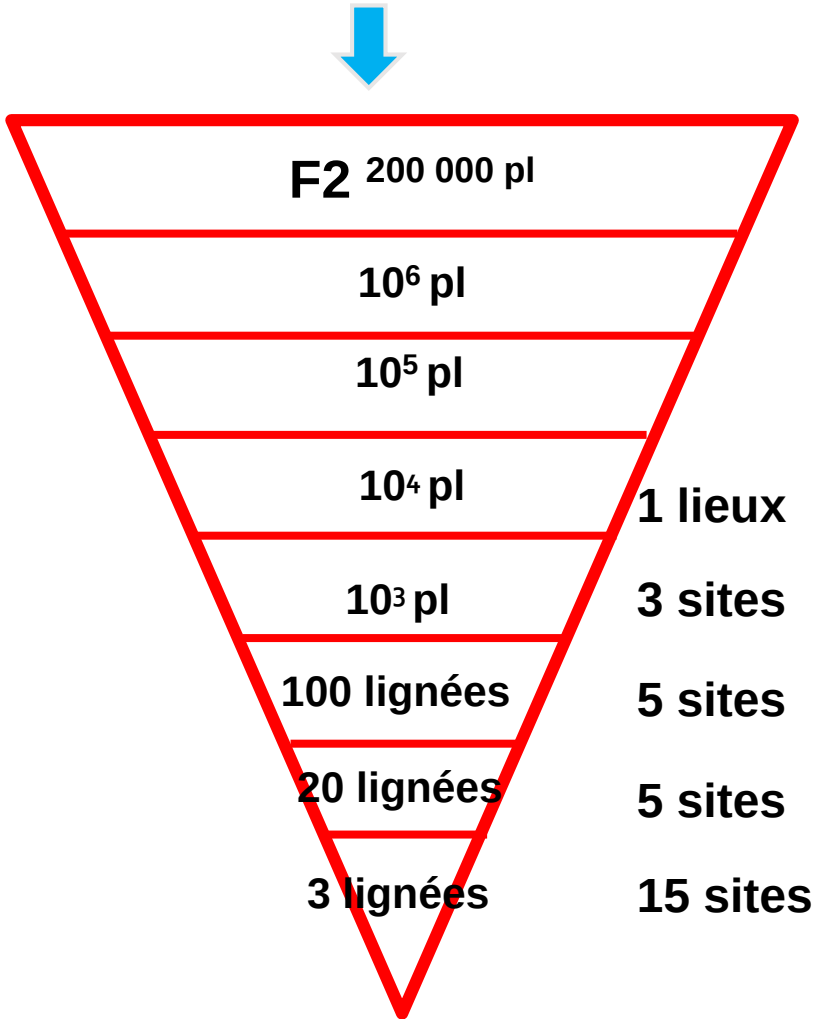
GDEC
Génétique Diversité Ecophysiologie des Céréales

GenPhySE
Génétique, Physiologie et Systèmes d'Elevage

Schéma de sélection classique blé tendre



400 croisements par an



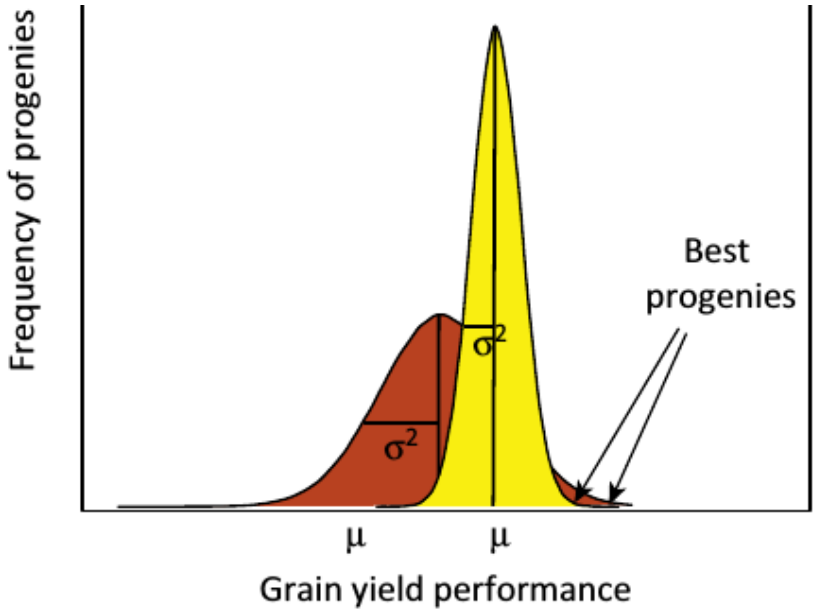
Inscription CTPS (10 000 euros)

Questions

Comment optimiser le programme de sélection grâce aux prédictions génomiques?

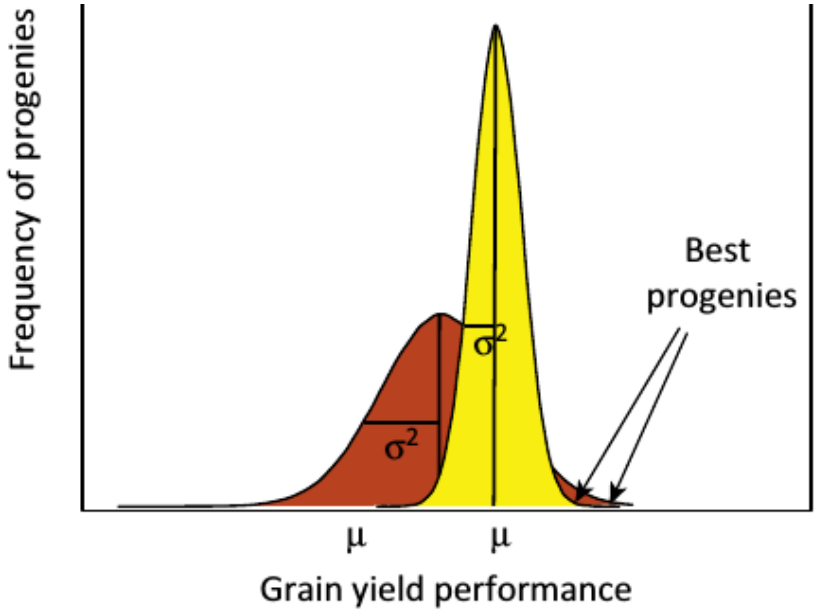
- À quelles étapes du programme du programme de sélection?
- Avec quelle précision?
- Comment améliorer la précision?
- Avec le niveau de précision qu'on a, quel est le gain génétique?

Prédire les meilleurs croisements



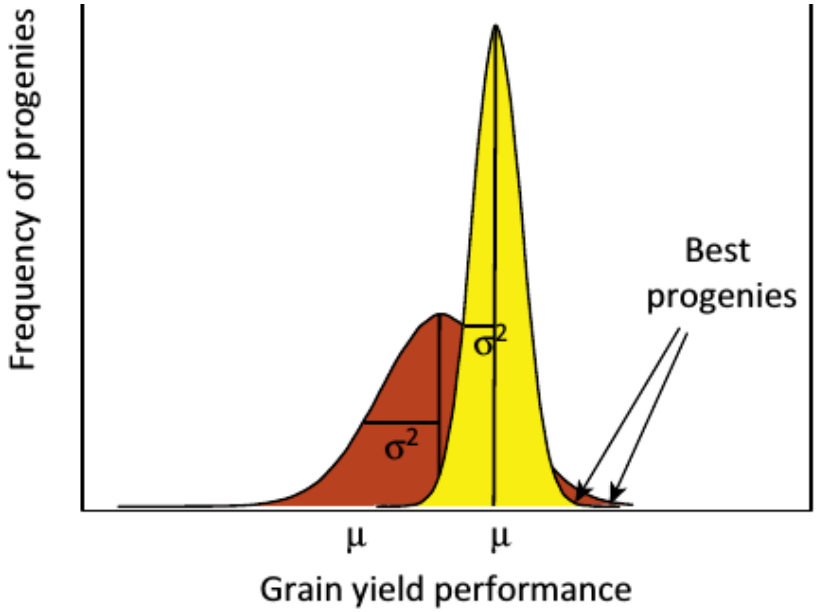
Question aujourd'hui

Précision de notre capacité de prédiction de la valeur d'un croisement



Question aujourd'hui

Précision de notre capacité de prédiction de la valeur d'un croisement

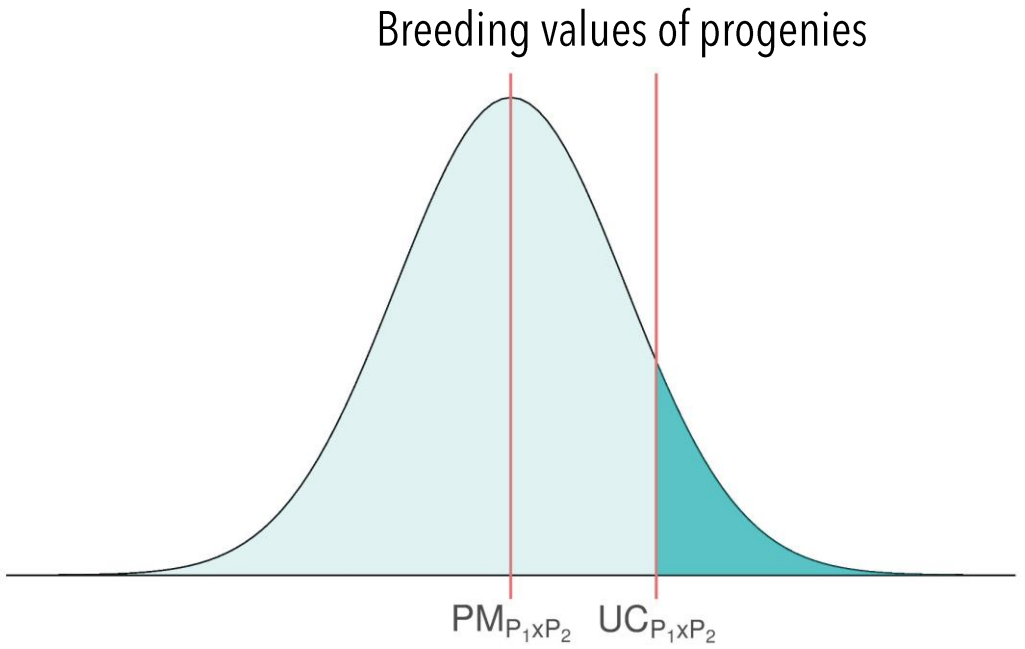


Sur données simulées (quand on connaît les vrais effets des marqueurs): on compare nos prédictions aux vraies valeurs

Sur données expérimentales (quand on estime les vrais effets des marqueurs): on compare nos prédictions à d'autres prédictions (phénotypes)

Background

- In breeding programs, commercial lines are selected among the **best progenies** of a set of **promising crosses**



Background

- In breeding programs, commercial lines are selected among the **best progenies** of a set of **promising crosses**

- A strategy to select best crosses is based on the **UC** (**Usefulness Criterion**):

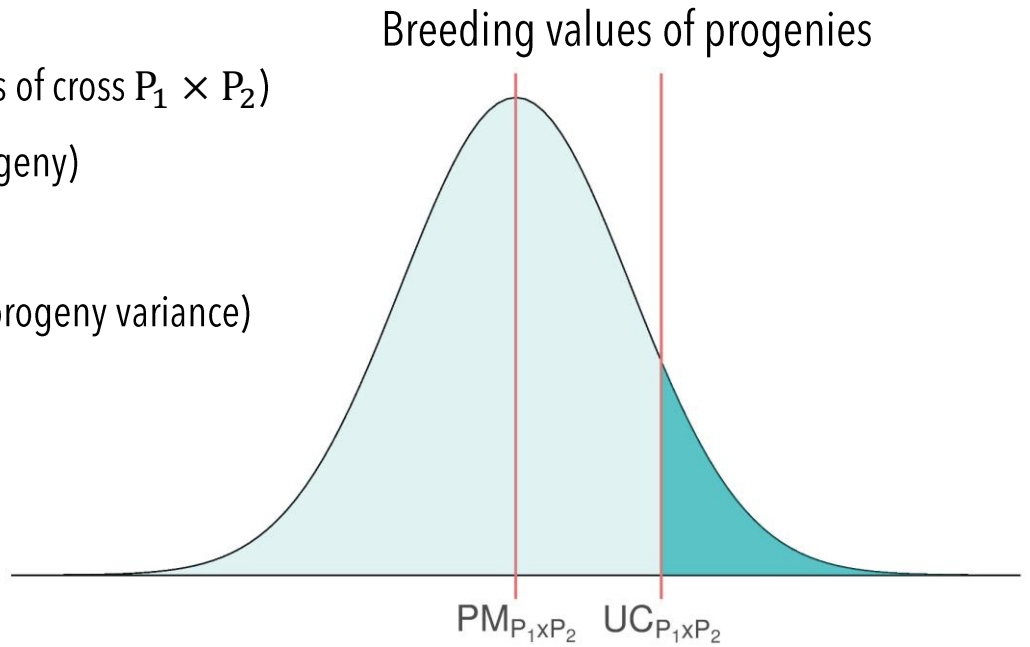
$$UC_{P_1 \times P_2} = PM_{P_1 \times P_2} + i \times SD_{P_1 \times P_2}$$

$UC_{P_1 \times P_2}$: **Usefulness Criterion** (expected mean of best progenies of cross $P_1 \times P_2$)

$PM_{P_1 \times P_2}$: **Parental Mean** (expected mean of a cross $P_1 \times P_2$ progeny)

i: selection intensity (~1.91 corresponding to 7% in our study)

$SD_{P_1 \times P_2}$: **progeny Standard Deviation** (square root of $P_1 \times P_2$ progeny variance)



Objectives

- Estimate **UC**, **PM** and **SD** using a prediction model (Training Population (TP) phenotyped for the same traits and genotyped)

- Compare predictions with true values in simulations
  - (i) Identify the **factors influencing** the prediction ability (trait architecture, heritability, number of QTLs, progeny size…) of PM, SD and UC based on **simulations**
  - (ii) Estimate the **ability of genomic prediction** of PM, SD and UC in **real experimental data**

- Compare predictions with **phenotypes of progeny observed in the field**

Objectives

- UC, PM and SD = cross value components that can be **observed** with **phenotyped progeny in the field** or **predicted** using **prediction models** trained on a Training Population (TP)

- **Objectives:**

  (i) Identify the **factors influencing** the prediction ability (trait architecture, heritability, number of QTLs, progeny size…) of PM, SD and UC based on **simulations**

  (ii) Estimate the **ability of genomic prediction** of PM, SD and UC in **real experimental data**

Mat and Meth: Training Population (TP)

- 2 datasets * 2 geographical areas:

AO = Agri-Obtentions

Mat and Meth : Training Population (TP)

- 2 datasets * 2 geographical areas:
  - **INRAE-AO**: F8-F9 winter bread wheat lines developed by INRAE-AO (2000-2022)
    157-169 (North) and 26-42 (South) lines evaluated each year

AO = Agri-Obtentions
GEVES = *Groupe d'Etude et de contrôle des Variétés Et des Semences* (Variety and Seed Study and Control Group)
VATE = *Valeur Agronomique Technologique et Environnementale* (Agronomic, technological and environmental value)

Mat and Meth : Training Population (TP)

- 2 datasets * 2 geographical areas:
  - **INRAE-AO**: F8-F9 winter bread wheat lines developed by INRAE-AO (2000-2022)
    157-169 (North) and 26-42 (South) lines evaluated each year
  - **GEVES**: VATE winter bread wheat data from the evaluation of varieties for national registration (2000-2022)
    44-46 (North) and 26-27 (South) lines evaluated each year

AO = Agri-Obtentions
GEVES = *Groupe d'Etude et de contrôle des Variétés Et des Semences* (Variety and Seed Study and Control Group)
VATE = *Valeur Agronomique Technologique et Environnementale* (Agronomic, technological and environmental value)

Mat and Meth : Training Population (TP)

- 2 datasets * 2 geographical areas:
  - **INRAE-AO**: F8-F9 winter bread wheat lines developed by INRAE-AO (2000-2022)
    157-169 (North) and 26-42 (South) lines evaluated each year
  - **GEVES**: VATE winter bread wheat data from the evaluation of varieties for national registration (2000-2022)
    44-46 (North) and 26-27 (South) lines evaluated each year
- Crop management methods = **high yield objectives** (optimized pesticide, fungicide and nitrogen amount)

AO = Agri-Obtentions
GEVES = *Groupe d'Etude et de contrôle des Variétés Et des Semences* (Variety and Seed Study and Control Group)
VATE = *Valeur Agronomique Technologique et Environnementale* (Agronomic, technological and environmental value)

Mat and Meth : Training Population (TP)

- 2 datasets * 2 geographical areas:
  - **INRAE-AO**: F8-F9 winter bread wheat lines developed by INRAE-AO (2000-2022)
    157-169 (North) and 26-42 (South) lines evaluated each year
  - **GEVES**: VATE winter bread wheat data from the evaluation of varieties for national registration (2000-2022)
    44-46 (North) and 26-27 (South) lines evaluated each year
- Crop management methods = **high yield objectives** (optimized pesticide, fungicide and nitrogen amount)

Trait	Number of lines with phenotypes + genotypes (23K markers)
Yield	**2,146**
Grain protein content	2,062
Plant height	2,126
Heading date	2,145

Mat and Meth : real experimental data (1/2)

FD = Florimond-Desprez
CF = Clermont-Ferrand (INRAE PHACC)
EM = Estrées-Mons (INRAE GCIE)
AO = Agri-Obtentions

Number of crosses	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***

From 8 to 122 progenies per cross (**mean: 54.8**)

Mat and Meth : real experimental data (1/2)

PrediCropt

FD = Florimond-Desprez
CF = Clermont-Ferrand (INRAE PHACC)
EM = Estrées-Mons (INRAE GCIE)
AO = Agri-Obtentions

CAP = Cappelle (FD)
HOU = Houville (FD)
LU = Lusignan (INRAE FERLUS)
AUZ = Auzeville (INRAE GCA)

Number of crosses	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***

From 8 to 122 progenies per cross (**mean: 54.8**)

Number of plots	2020-2021	2021-2022	2022-2023
CAP	457	327	0
HOU	459	364	301
CF	0	800	804
EM	660	802	1,008
LU	420	834	1,003
AUZ	419	0	820

9,478 plots in total

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (1/2)

- $\widehat{PM}_{P_1 \times P_2} = \frac{\hat{\beta}'(X_{P_1} + X_{P_2})}{2}$

$\hat{\beta}$ = vector of estimated marker effects
X_{P_1}, X_{P_2} = vectors of genotypes for P_1 and P_2

[1] Lehermeier *et al.* (2017)
RIL = Recombinant Inbred Line

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (1/2)

- $\widehat{PM}_{P_1 \times P_2} = \frac{\hat{\beta}'(X_{P_1} + X_{P_2})}{2}$

$\hat{\beta}$ = vector of estimated marker effects
X_{P_1}, X_{P_2} = vectors of genotypes for P_1 and P_2

- $\widehat{SD}^2_{P_1 \times P_2} = \hat{\beta}' V_{P_1 \times P_2} \hat{\beta}$

$V_{P_1 \times P_2}$ = genotypic variance-covariance matrix[1] for biparental RIL progeny

How to compute progeny variance?

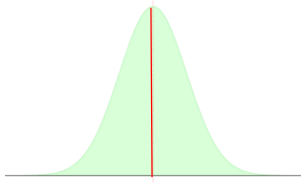
$$\sigma^2_{P_i \times P_j} = V1_{ij} + V2_{ij}$$

Genetic diversity of parents

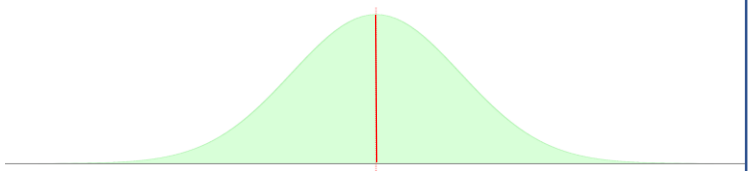
$$V1_{ij} = \sum_{m=1}^{M} \beta^2{}_m * 4p_{m\,ij}(1 - p_{m\,ij})$$

β = effects of allele
p = allelic frequency

Similar parents



Very different parents

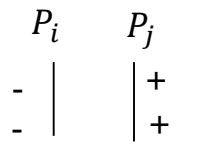


Recombination of desirable alleles

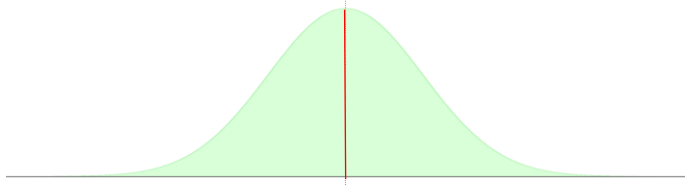
$$V2_{ij} = \sum_{l<m} \beta_l \beta_m * 4D_{lm\,ij}(1 - 2r_{lm})$$

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

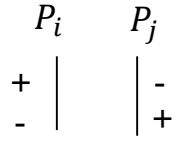
r = Frequency of recombinants



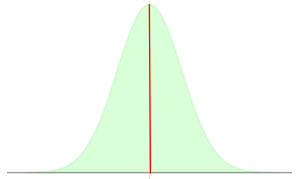
Coupling (D = 0.25)



Recombination is not desirable



Repulsion (D = - 0.25)



Recombination is desirable

Formula of Lehermeier et al. 2017 for Doubled Haploids

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (1/2)

- $\widehat{PM}_{P_1 \times P_2} = \frac{\hat{\beta}'(X_{P_1} + X_{P_2})}{2}$

$\hat{\beta}$ = vector of estimated marker effects
X_{P_1}, X_{P_2} = vectors of genotypes for P_1 and P_2

- $\widehat{SD}^2_{P_1 \times P_2} = \hat{\beta}' \boxed{V_{P_1 \times P_2}} \hat{\beta}$

$$(V_{P_1 \times P_2})^{RIL(k)}_{jl} = 4D^*_{jl}\left(\sum_{gen=1}^{k}\left(0.5(1-2c_{jl})\right)^{gen}\right)$$

$V_{P_1 \times P_2}$ = genotypic variance-covariance matrix[1] for biparental RIL progeny

D^*_{jl} = LD between alleles at loci j and l among parental lines
k = number of generations
c_{jl} = recombination rate between loci j and l

[1] Lehermeier *et al.* (2017)
RIL = Recombinant Inbred Line

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (1/2)

- $\widehat{PM}_{P_1 \times P_2} = \frac{\hat{\beta}'(X_{P_1} + X_{P_2})}{2}$

$\hat{\beta}$ = vector of estimated marker effects
X_{P_1}, X_{P_2} = vectors of genotypes for P_1 and P_2

- $\widehat{SD}^2_{P_1 \times P_2} = \hat{\beta}' \boxed{V_{P_1 \times P_2}} \hat{\beta}$

$$(V_{P_1 \times P_2})^{RIL(k)}_{jl} = 4D^*_{jl}\left(\sum_{gen=1}^{k}\left(0.5(1-2c_{jl})\right)^{gen}\right)$$

$V_{P_1 \times P_2}$ = genotypic variance-covariance matrix[1] for biparental RIL progeny

D^*_{jl} = LD between alleles at loci j and l among parental lines
k = number of generations
c_{jl} = recombination rate between loci j and l

- $\widehat{UC}_{P_1 \times P_2} = \widehat{PM}_{P_1 \times P_2} + i \times \widehat{SD}_{P_1 \times P_2}$

i = selection intensity (~1.91 corresponding to 7% in our study)

[1] Lehermeier *et al.* (2017)
RIL = Recombinant Inbred Line

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (1/2)

- $\widehat{PM}_{P_1 \times P_2} = \frac{\hat{\beta}'(X_{P_1} + X_{P_2})}{2}$

$\hat{\beta}$ = vector of estimated marker effects
X_{P_1}, X_{P_2} = vectors of genotypes for P_1 and P_2

- $\widehat{SD}^2_{P_1 \times P_2} = \hat{\beta}' \boxed{V_{P_1 \times P_2}} \hat{\beta}$

$$\left(V_{P_1 \times P_2}\right)_{jl}^{RIL(k)} = 4D^*_{jl}\left(\sum_{gen=1}^{k}\left(0.5\left(1 - 2c_{jl}\right)\right)^{gen}\right)$$

$V_{P_1 \times P_2}$ = genotypic variance-covariance matrix[1] for biparental RIL progeny

D^*_{jl} = LD between alleles at loci j and l among parental lines
k = number of generations
c_{jl} = recombination rate between loci j and l

- $\widehat{UC}_{P_1 \times P_2} = \widehat{PM}_{P_1 \times P_2} + i \times \widehat{SD}_{P_1 \times P_2}$

i = selection intensity (~1.91 corresponding to 7% in our study)

- **Predictions of these 3 cross value components** performed with the `PopVar` R package

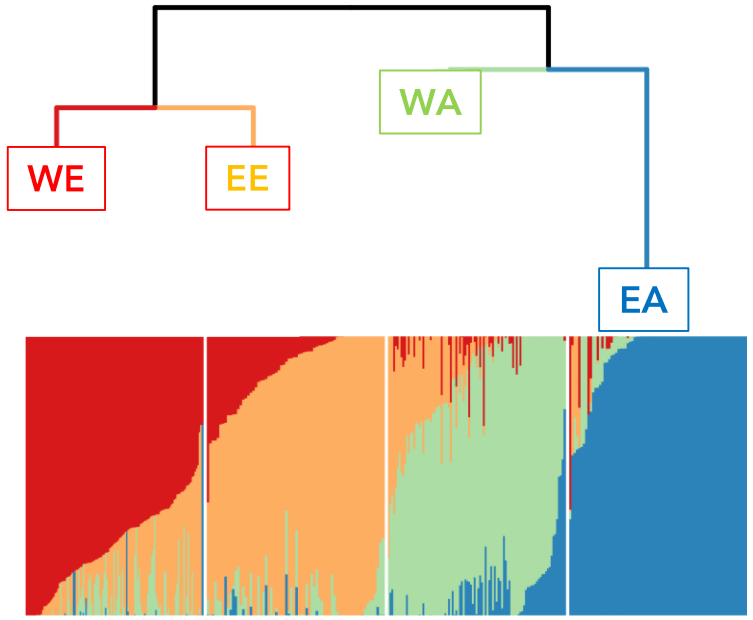
[1] Lehermeier *et al.* (2017)
RIL = Recombinant Inbred Line

Vector of recombination c (phD Alice Danguy Des Déserts)

371 bread wheat landraces (without introgression) sampled worldwide (Balfourier et al. 2019, Science Advances)
130k SNP of TABW410k (Kitt et al. 2021, Zenodo)

4 differentiated bread wheat populations

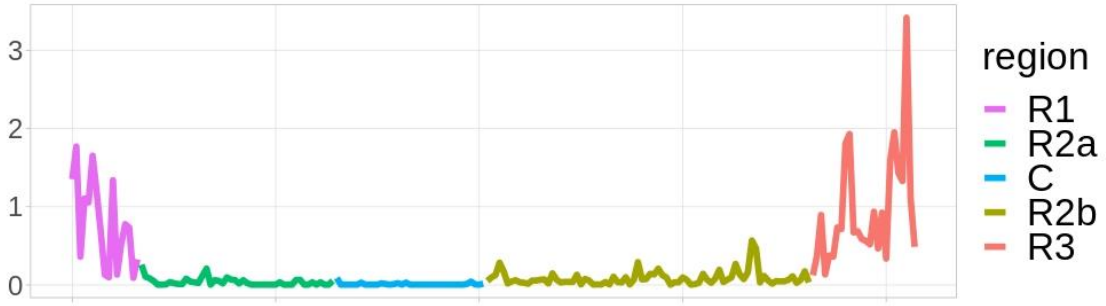
Europe ⟷ Asia (Fst)



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

Vector of recombination c (phD 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



JOURNAL ARTICLE

Evolution of Recombination Landscapes in Diverging Populations of Bread Wheat

Alice Danguy des Déserts, Sophie Bouchet, Pierre Sourdille ✉, Bertrand Servin ✉

Genome Biology and Evolution, Volume 13, Issue 8, August 2021, evab152, https://doi.org/10.1093/gbe/evab152

Published: 29 June 2021 **Article history** ▾

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (2/2)

- Different **prediction models** were tested to **estimate marker effects**: BayesA, BayesB, BayesC, Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Ridge Regression BLUP (Vg1)

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (2/2)

- Different **prediction models** were tested to **estimate marker effects**: BayesA, BayesB, BayesC, Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Ridge Regression BLUP (Vg1)

- We developed **2 alternative approaches** to compute **gametic variance** (tested only with Ridge Regression BLUP):

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (2/2)

- Different **prediction models** were tested to **estimate marker effects**: BayesA, BayesB, BayesC, Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Ridge Regression BLUP ($Vg1$)

- We developed **2 alternative approaches** to compute **gametic variance** (tested only with Ridge Regression BLUP):

  - Taking into account the **marker effect estimation error** (algebraic version of the PMV of Lehermeier *et al.* 2017):

$$\left(\widehat{\mathrm{SD}}^2_{\mathrm{P_1 \times P_2}}\right)_{\mathrm{Vg_2}} = \mathrm{Vg_1} + \textbf{term for marker effect estimation error}$$

Mat and Meth : cross value components prediction ability

Analytic formulae (genomic predictions) (2/2)

- Different **prediction models** were tested to **estimate marker effects**: BayesA, BayesB, BayesC, Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Ridge Regression BLUP (Vg1)

- We developed **2 alternative approaches** to compute **gametic variance** (tested only with Ridge Regression BLUP):

  - Taking into account the **marker effect estimation error** (algebraic version of the PMV of Lehermeier *et al.* 2017):

    $$\left(\widehat{\mathrm{SD}}^2_{\mathrm{P}_1\times\mathrm{P}_2}\right)_{\mathrm{Vg}_2} = \mathrm{Vg}_1 + \textbf{term taking into account marker effect estimation error}$$

  - Considering that the uncertainty of the estimation of marker effects is **modulated by the genomic constitution of each parent**:

    $$\left(\widehat{\mathrm{SD}}^2_{\mathrm{P}_1\times\mathrm{P}_2}\right)_{\mathrm{Vg}_3} = \mathrm{Vg}_2 + \textbf{term taking into account genomic constitution of each parent}$$

Materials: cross value components prediction ability

Analytic formulae (genomic predictions) (2/2)

- Different **prediction models** were tested to **estimate marker effects**: BayesA, BayesB, BayesC, Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Ridge Regression BLUP (Vg1)

- We developed **2 alternative approaches** to compute **gametic variance** (tested only with Ridge Regression BLUP):
  - Taking into account the **marker effect estimation error** (algebraic version of the PMV of Lehermeier *et al.* 2017):

$$\left(\widehat{SD}^2_{P_1\times P_2}\right)_{Vg_2} = \boxed{\hat{\beta}' V_{P_1\times P_2}\hat{\beta}}_{Vg_1} + \text{trace}\left\{V_{P_1\times P_2}\,\text{var}(\beta|X,y)\right\}$$

$$\text{var}(\beta|X,y) = \hat{\sigma}^2_\beta\left(I - X'\left(XX' + I\frac{\hat{\sigma}^2_r}{\hat{\sigma}^2_\beta}\right)^{-1}X\right)$$

$\hat{\sigma}^2_\beta, \hat{\sigma}^2_r$ = markers and residuals estimated variances

X = vector of TP's genotypes

  - Considering that the uncertainty of the estimation of marker effects is **modulated by the genomic constitution of each parent**:

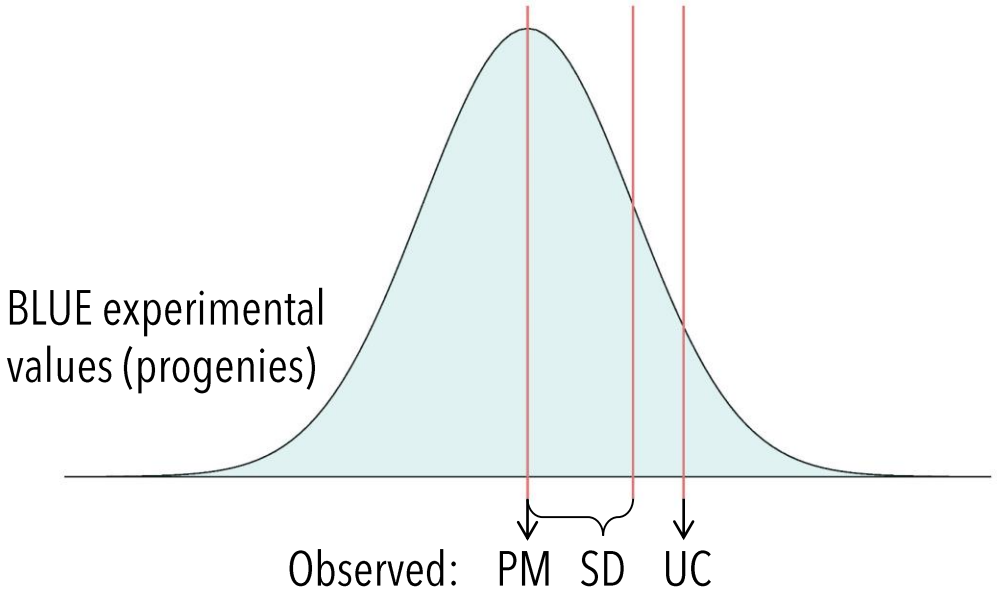
$$\left(\widehat{SD}^2_{P_1\times P_2}\right)_{Vg_3} = \boxed{\hat{\beta}' V_{P_1\times P_2}\hat{\beta} + \text{trace}\left\{V_{P_1\times P_2}\,\text{var}(\beta|X,y)\right\}}_{Vg_2} + 0.25X'_{P_1\times P_2}\,\text{var}(\beta|X,y)X_{P_1\times P_2}$$

$X_{P_1\times P_2}$ = vector of genotypes for the F1 of cross $P_1 \times P_2$

Mat and Meth : cross value components prediction ability

Experimental values

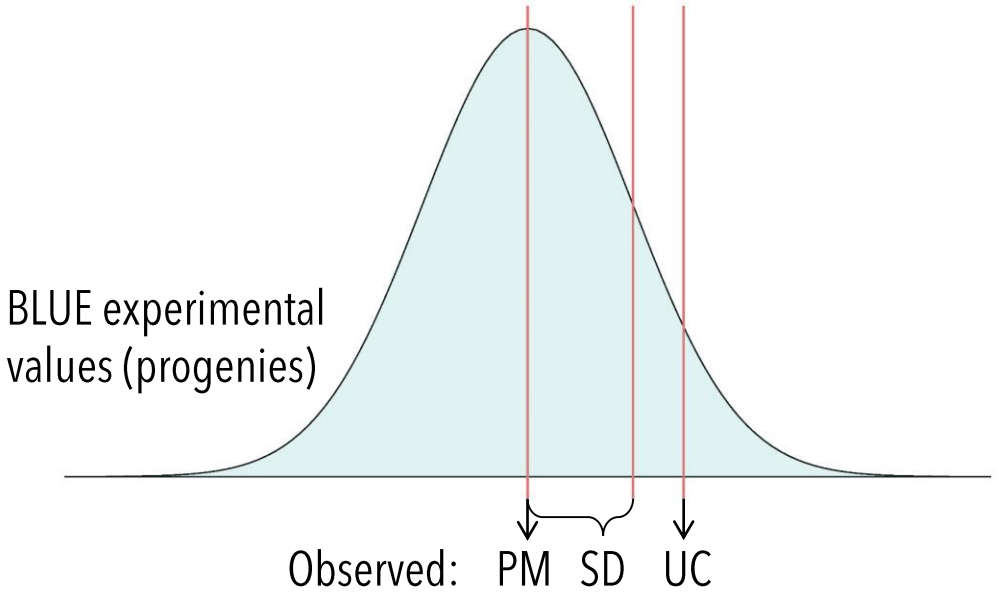
- **BLUE values of progenies,** mean (PM), standard deviation (SD) and mean of the 7% best progeny (observed UC) (or the value of the best progeny when progeny size < 15)



Mat and Meth : cross value components prediction ability

Experimental values

- **BLUE values of progenies,** mean (PM), standard deviation (SD) and mean of the 7% best progeny (observed UC) (or the value of the best progeny when progeny size < 15)



- Prediction ability = **weighted Pearson's correlation** (to adjust for the number of progenies per cross) between genomic predictions and experimental cross value components

Mat and Meth : simulation design

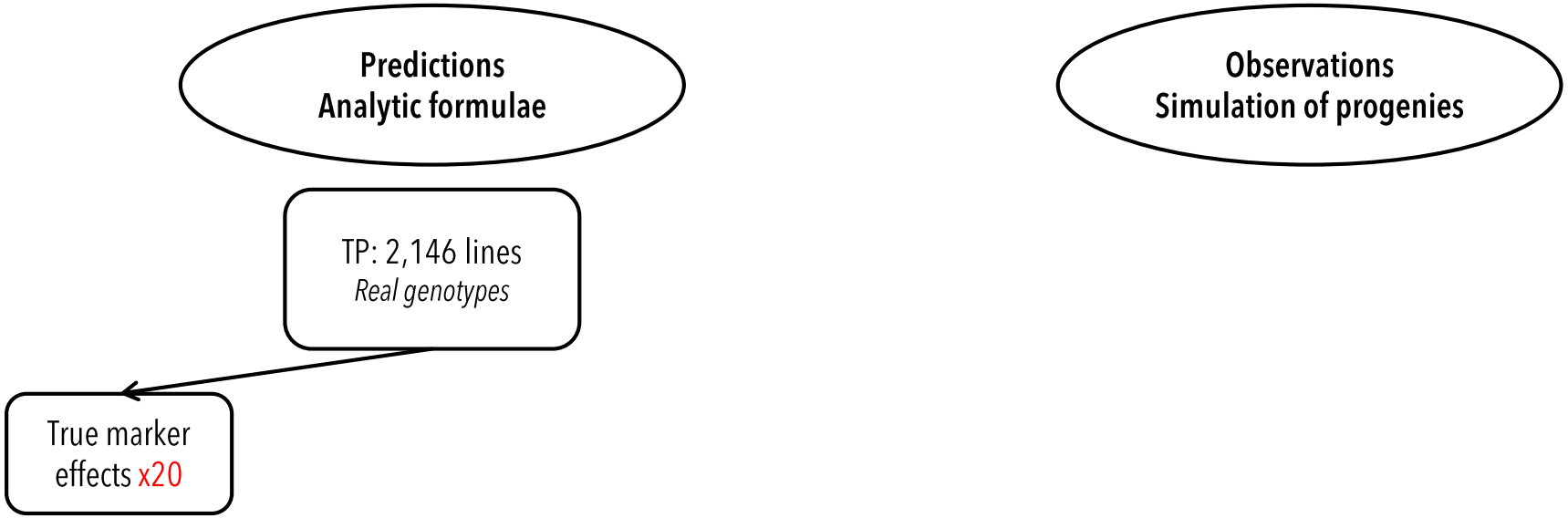
Mat and Meth : simulation design

Predictions
Analytic formulae

TP: 2,146 lines
Real genotypes

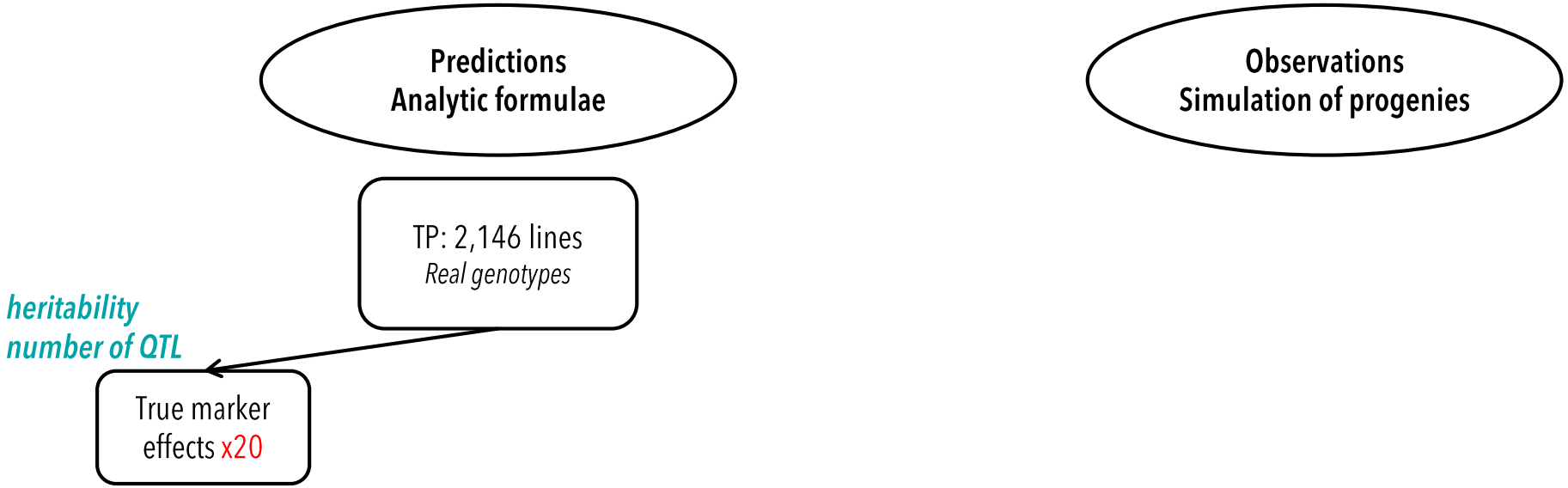
Observations
Simulation of progenies

Mat and Meth : simulation design



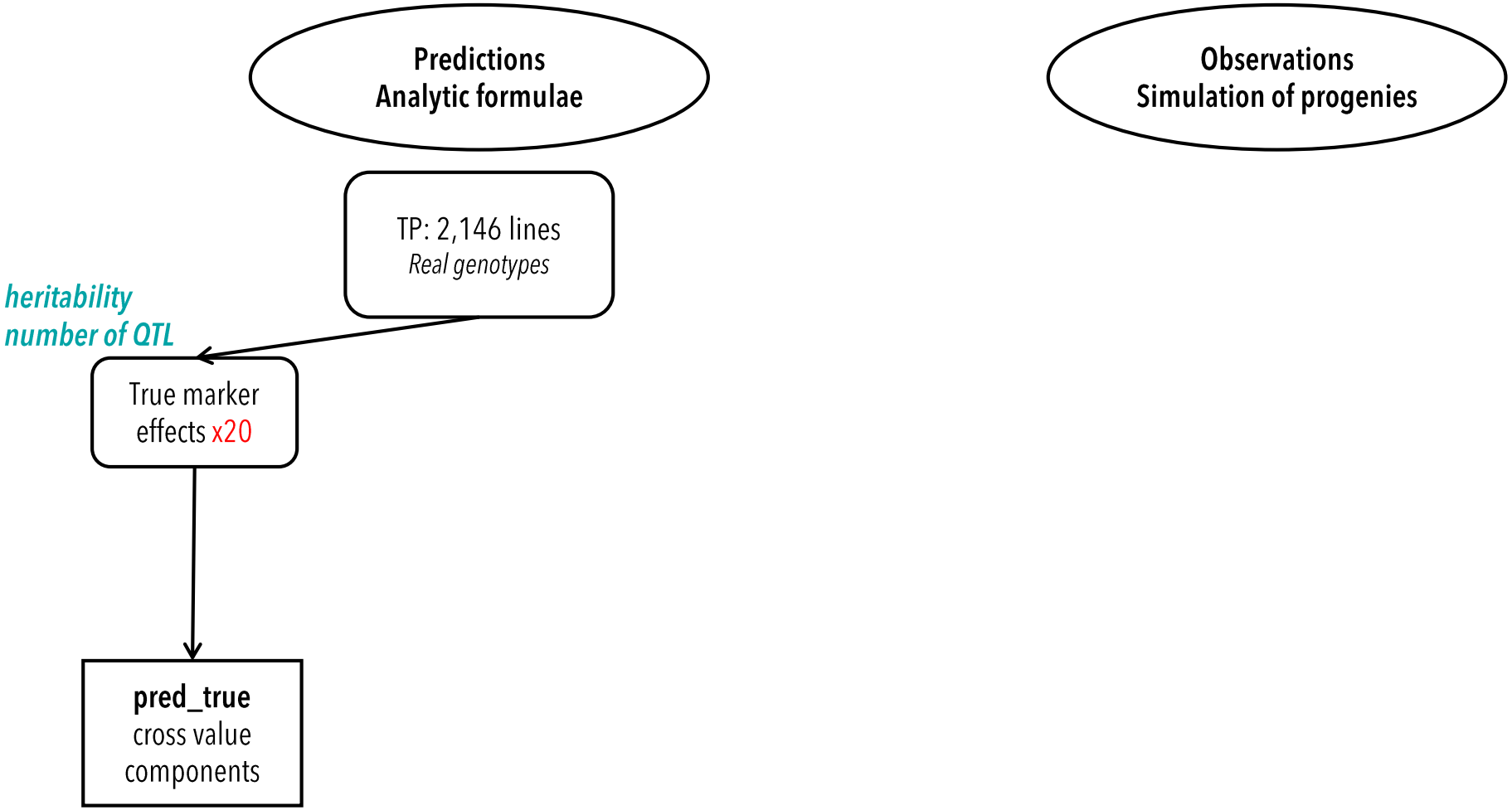
TP = Training Population
QTL = Quantitative Trait Loci

Mat and Meth : simulation design



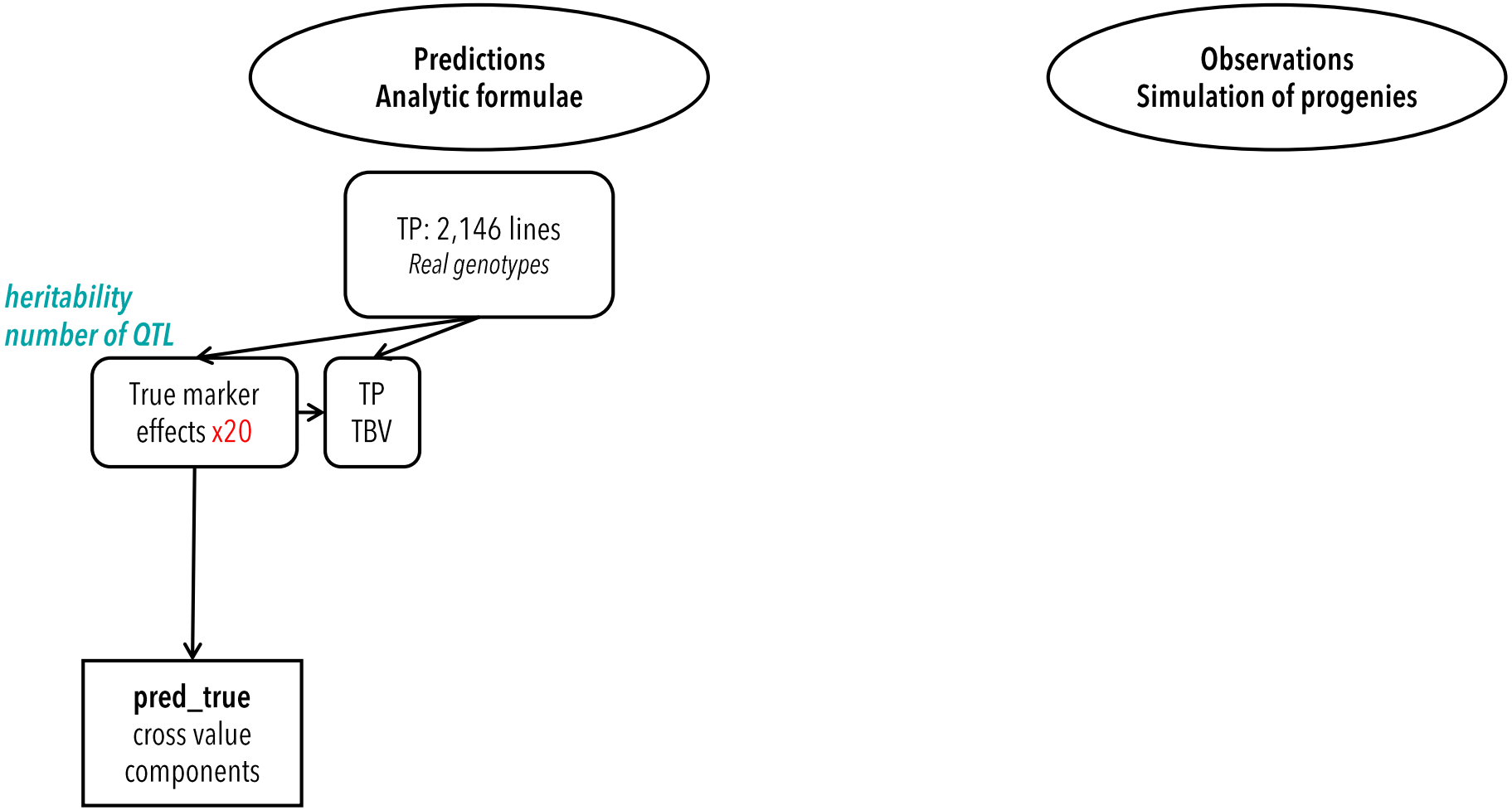
TP = Training Population
QTL = Quantitative Trait Loci

Mat and Meth : simulation design



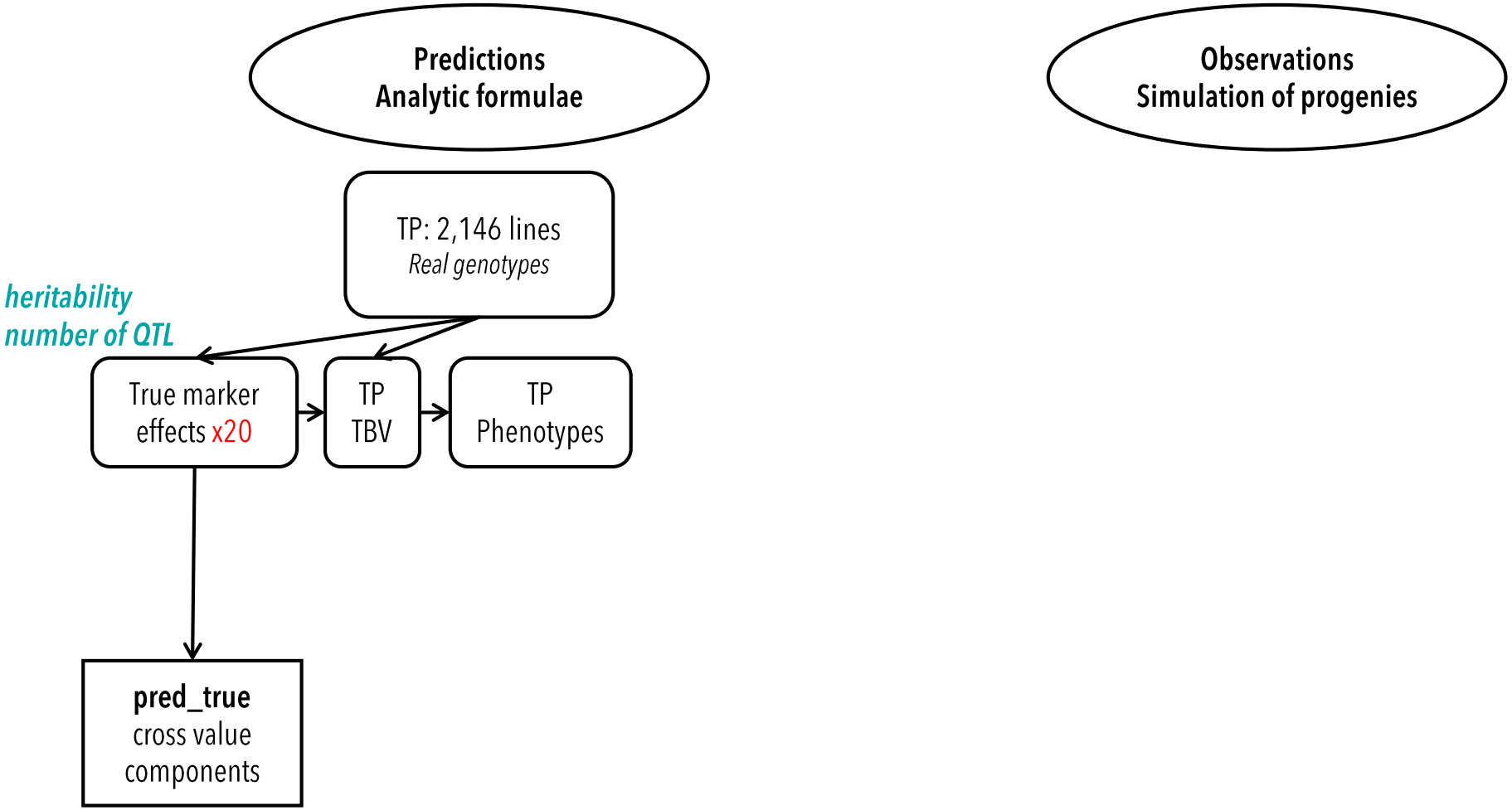
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



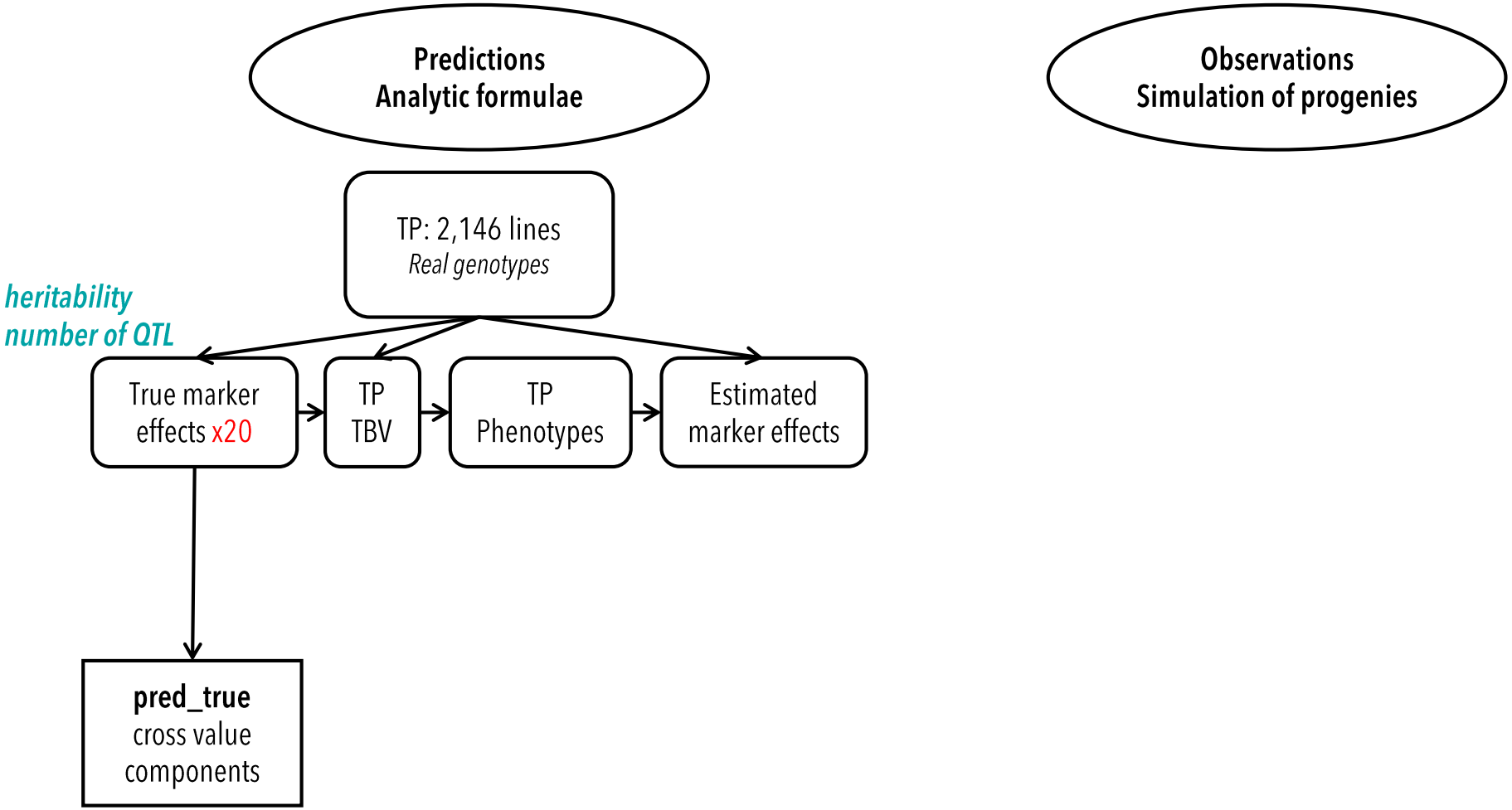
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



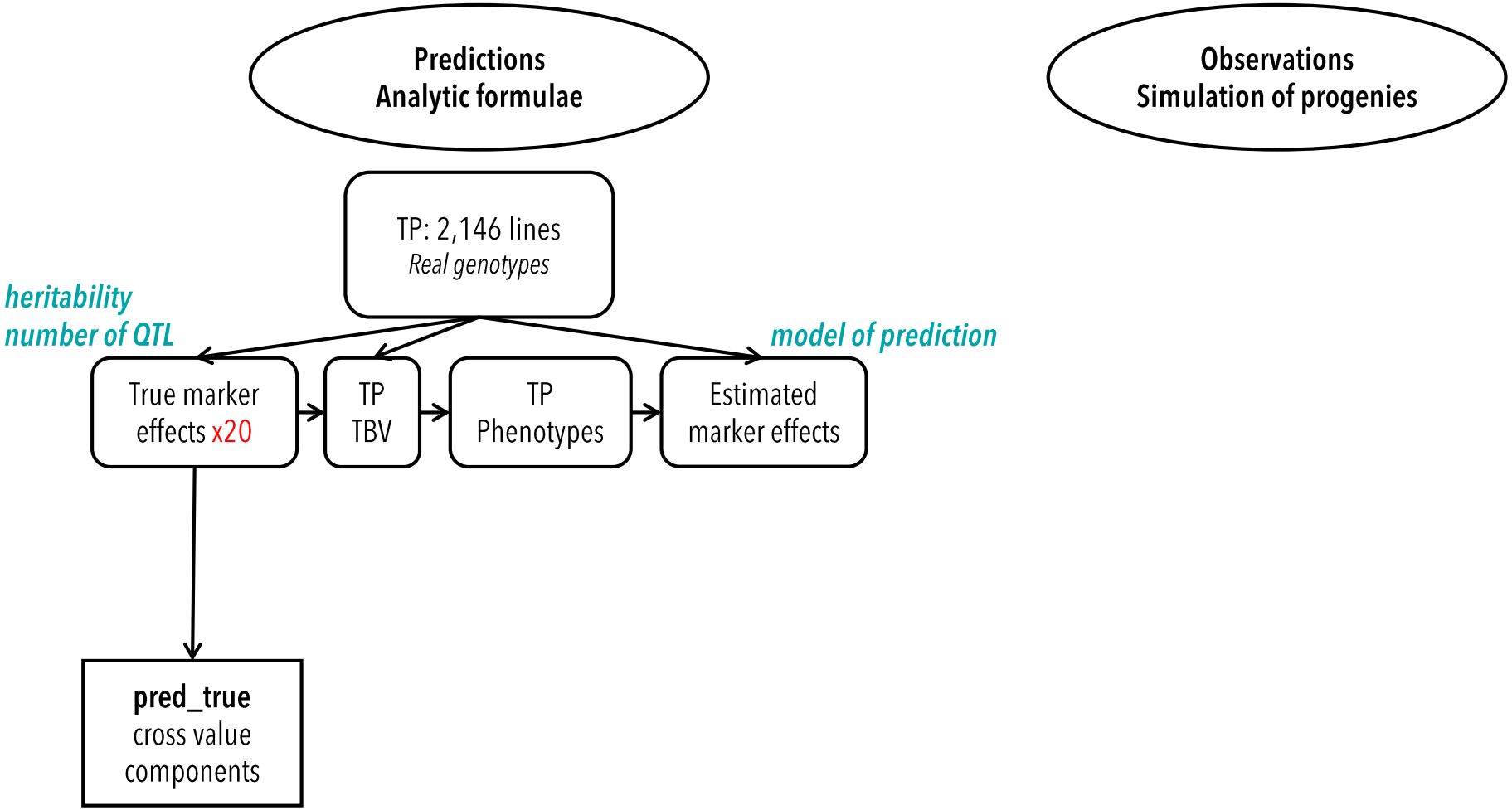
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



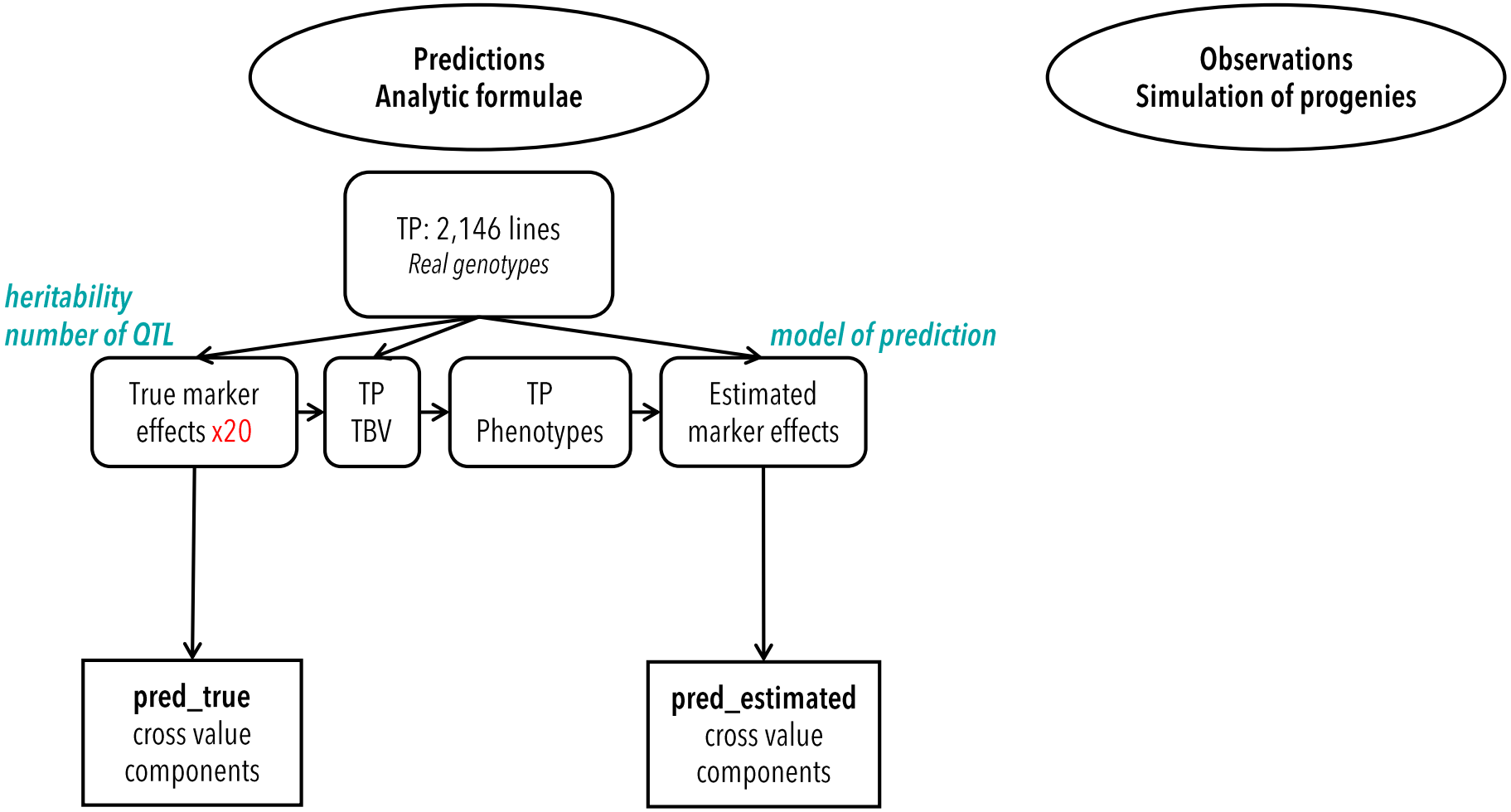
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



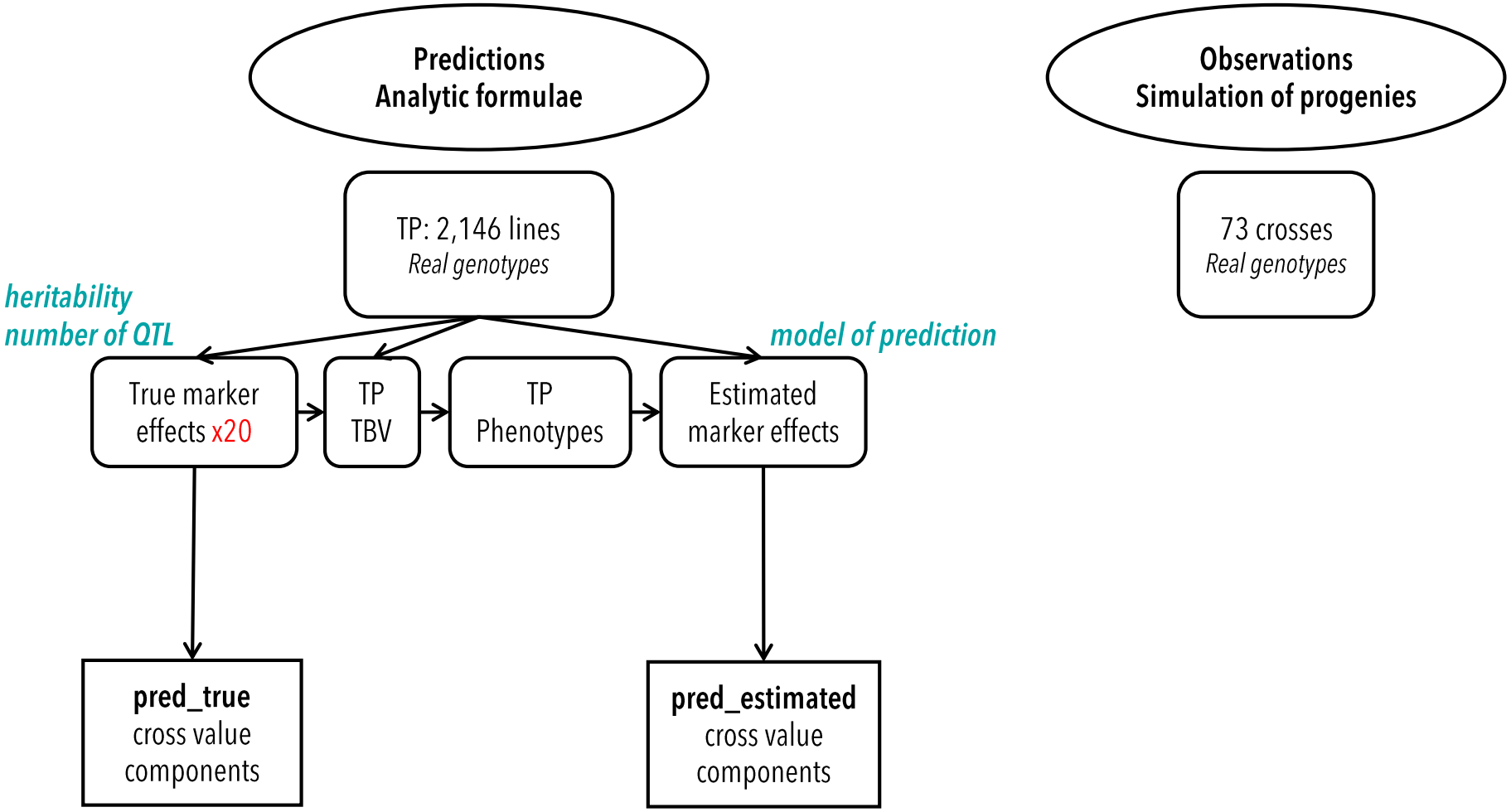
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



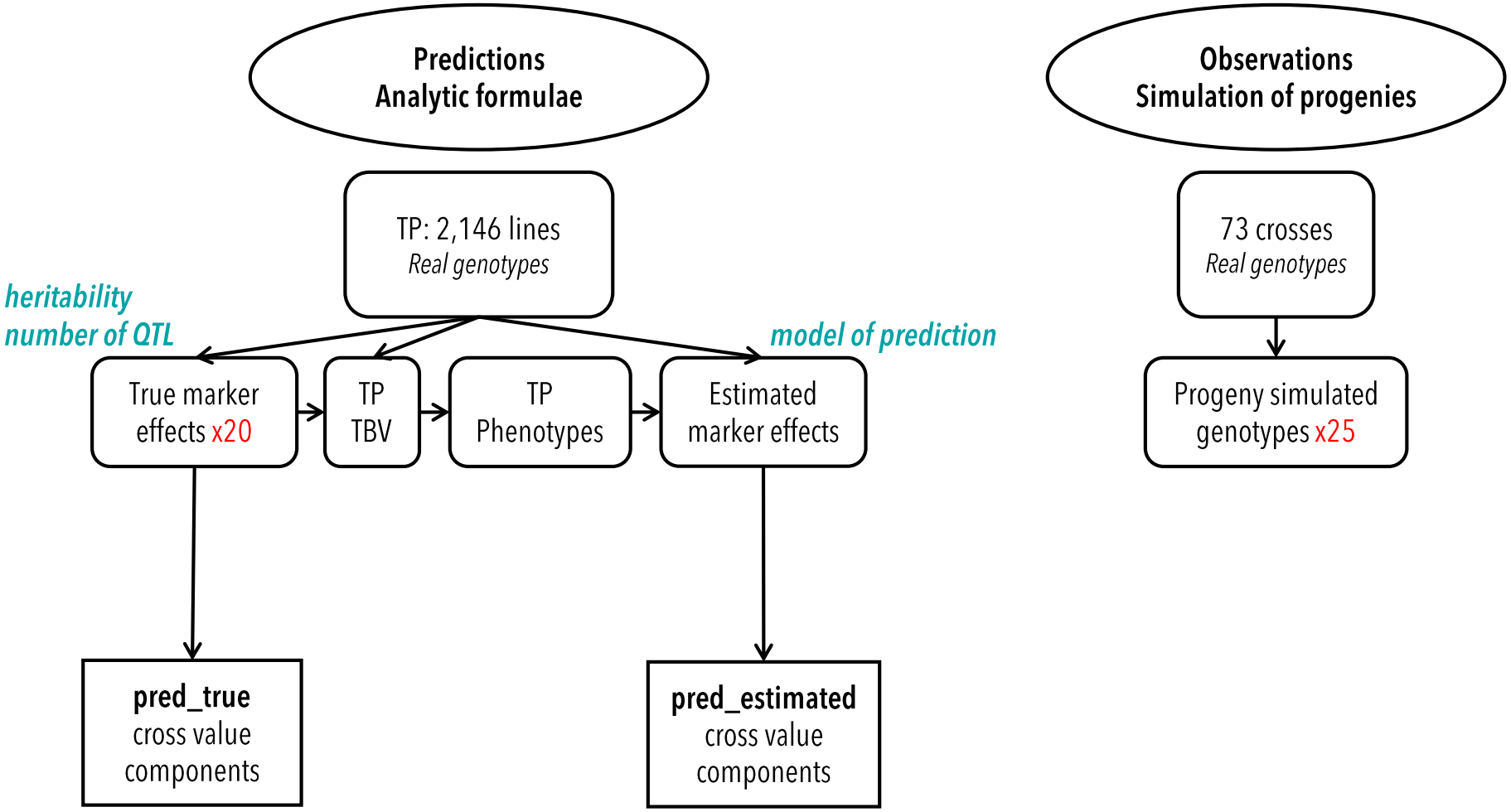
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



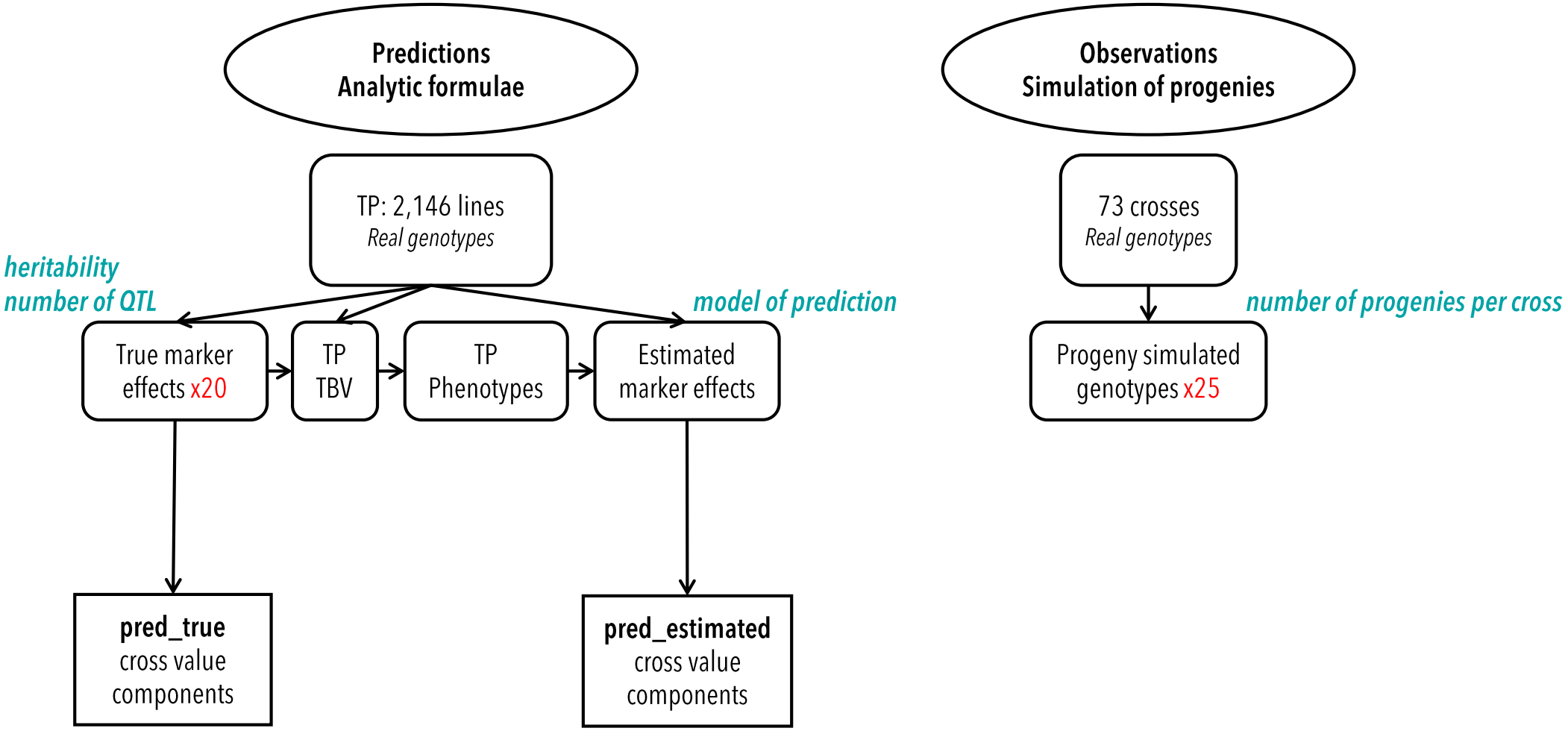
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



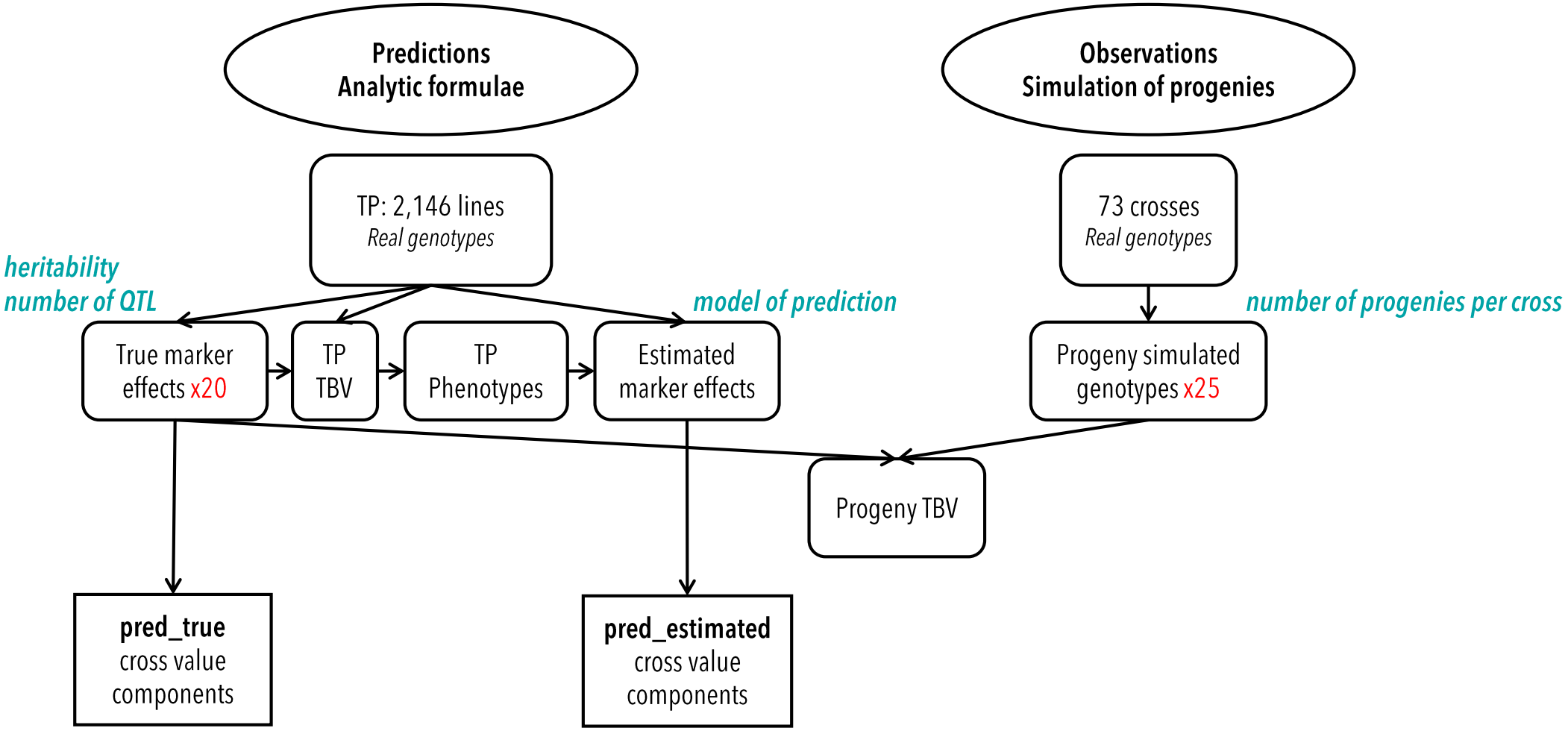
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



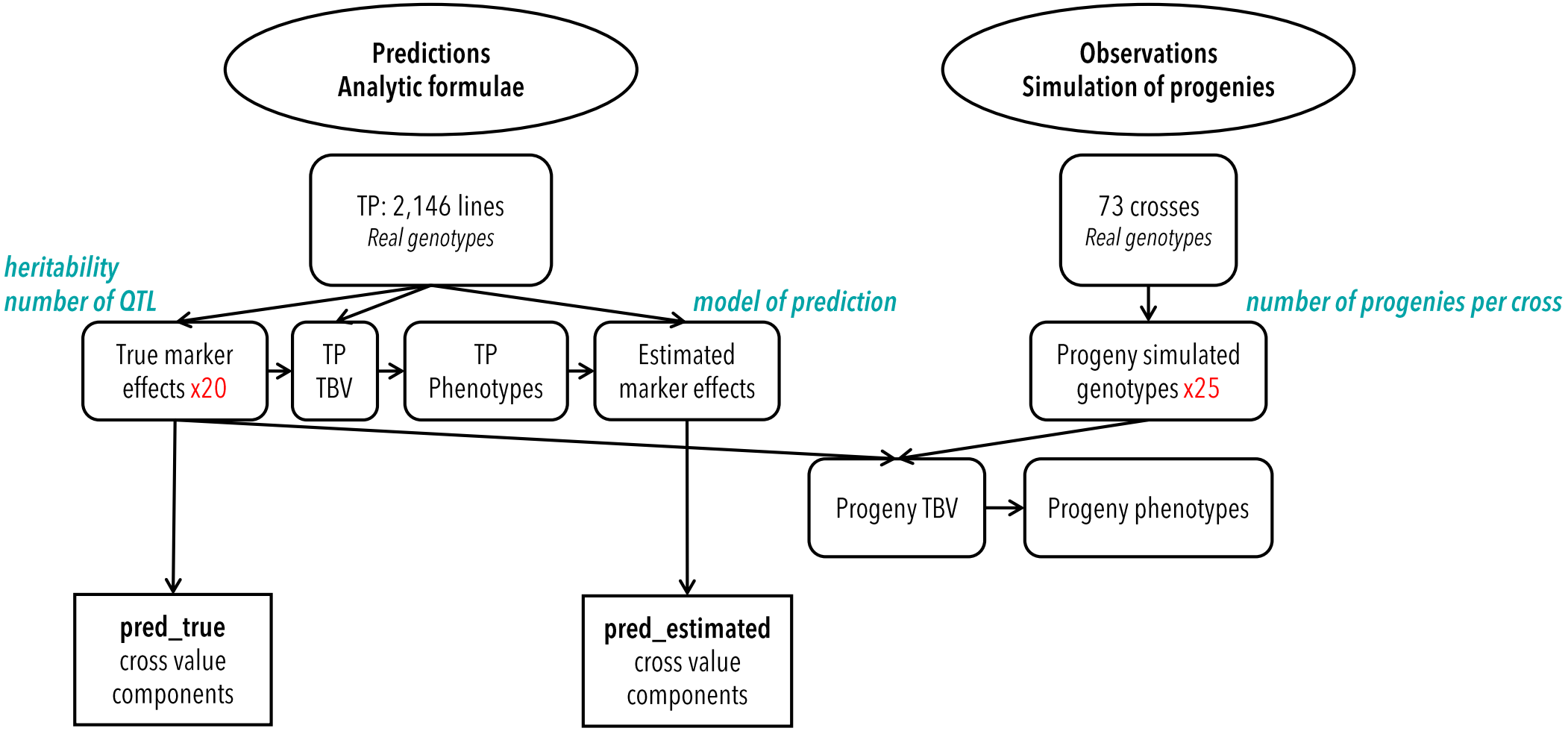
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



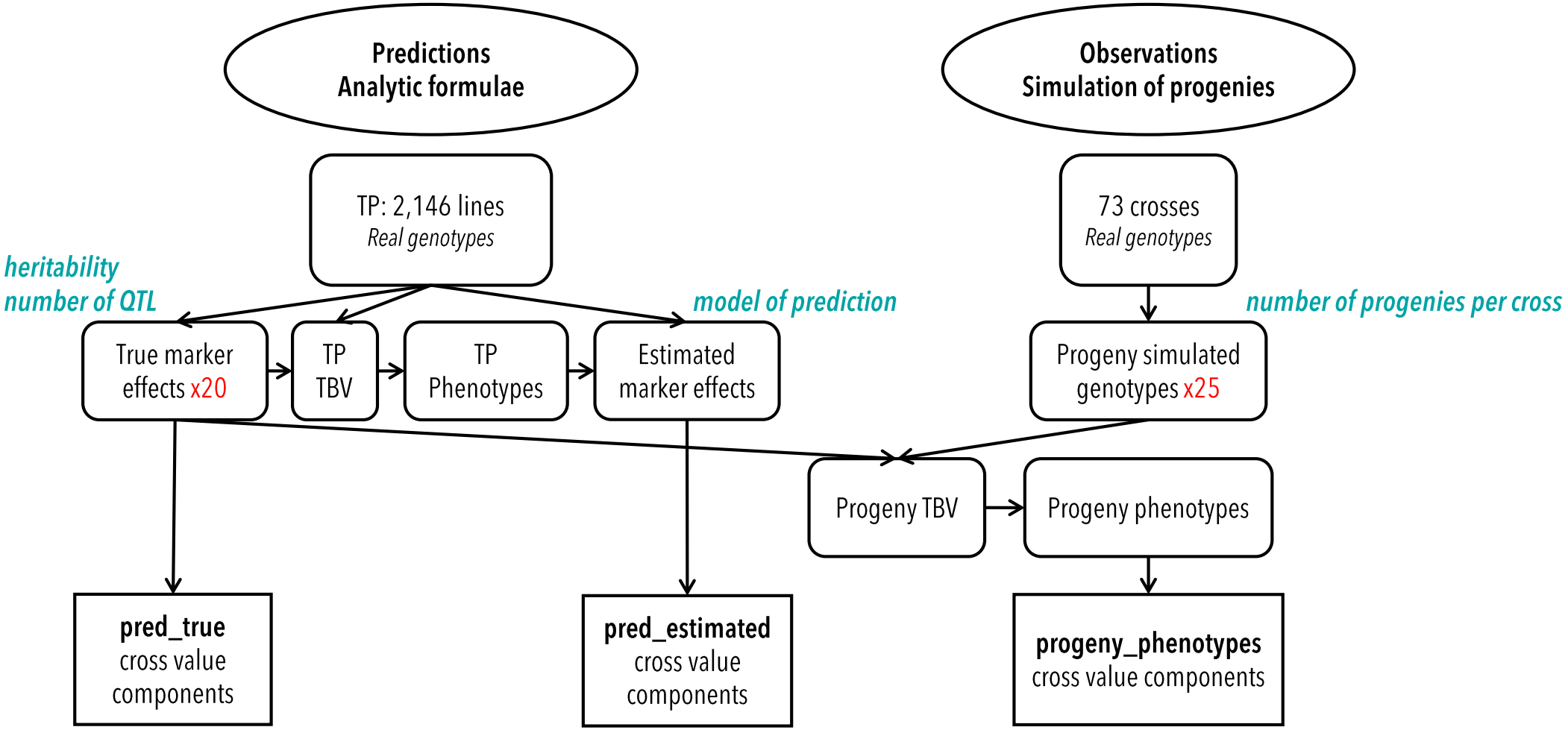
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



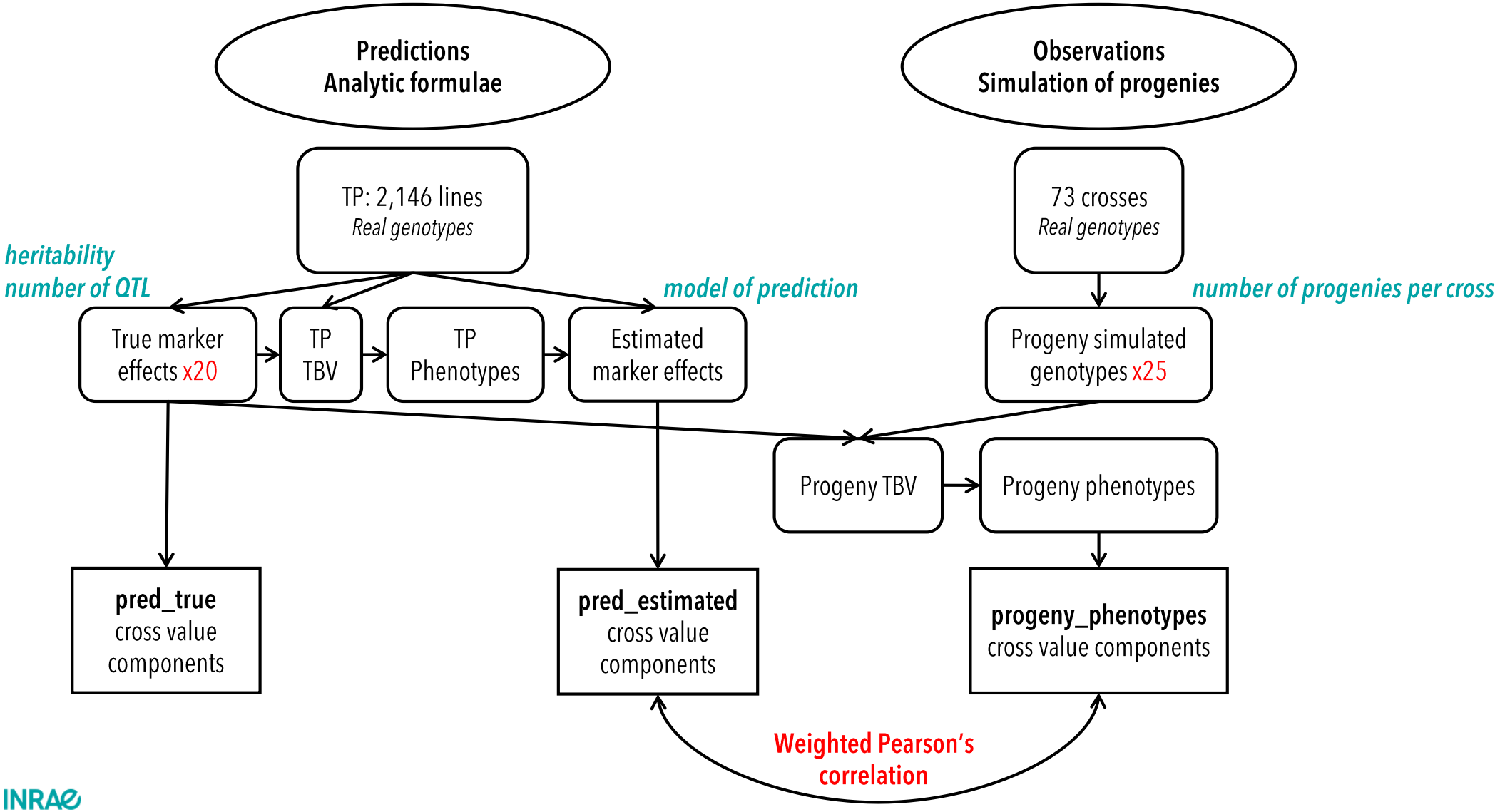
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



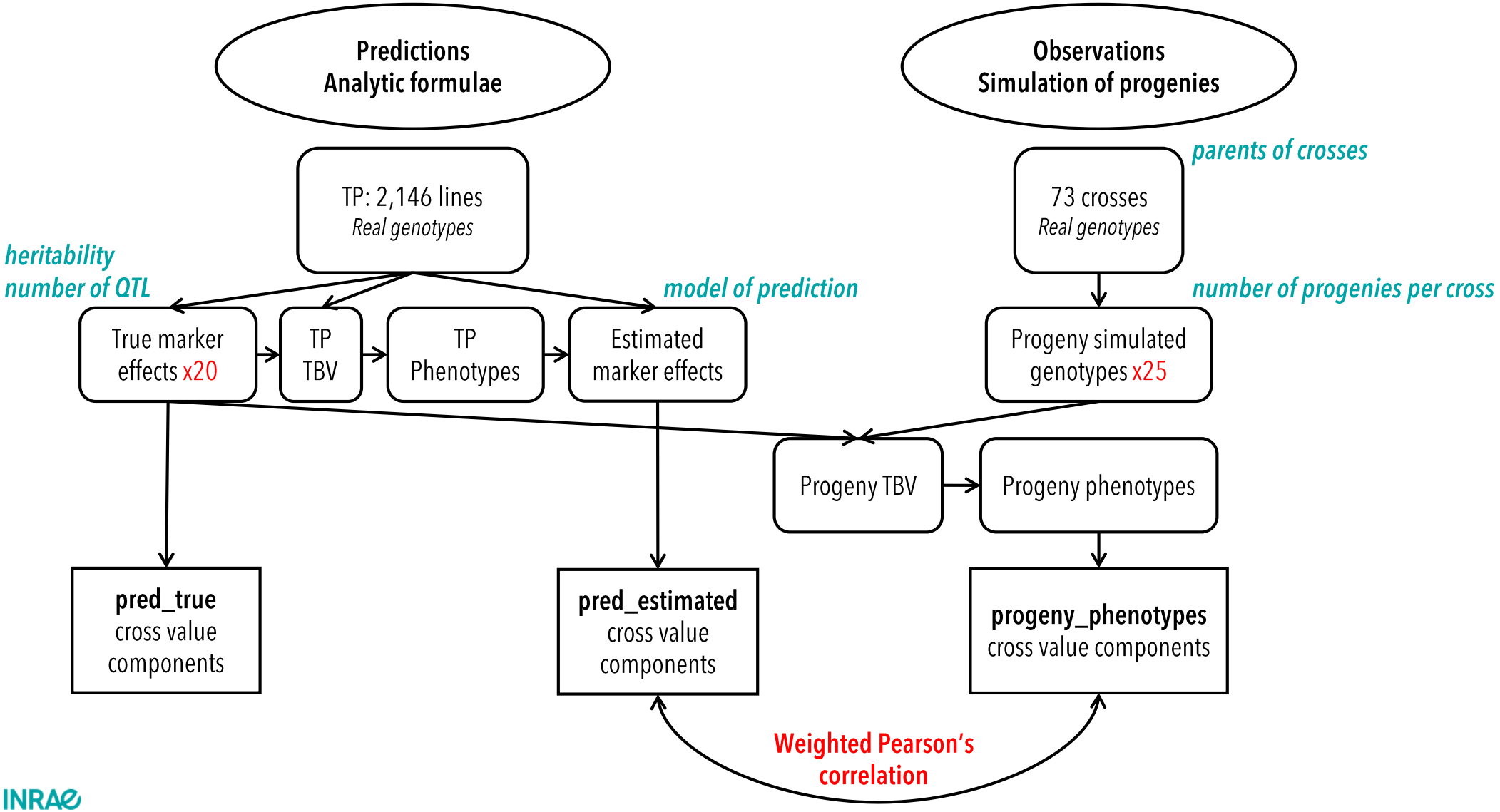
TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



TP = Training Population
QTL = Quantitative Trait Loci
TBV = True Breeding Value

Mat and Meth : simulation design



Results: Training Population (TP) quality

- **Genotypes**:
  - After quality control (MAF & call rate) and imputation of missing genotypes → **2,146** lines and **23,140** markers

Results: Training Population (TP) quality

- **Genotypes**:
  - After quality control (MAF & call rate) and imputation of missing genotypes → **2,146** lines and **23,140** markers

- **Phenotypes** (BLUE values):
  - Means of correlations between years: **0.69** (yield), **0.78** (grain protein content), **0.87** (plant height), **0.91** (heading date)

Results: Training Population (TP) quality

- **Genotypes**:
  - After quality control (MAF & call rate) and imputation of missing genotypes → **2,146** lines and **23,140** markers

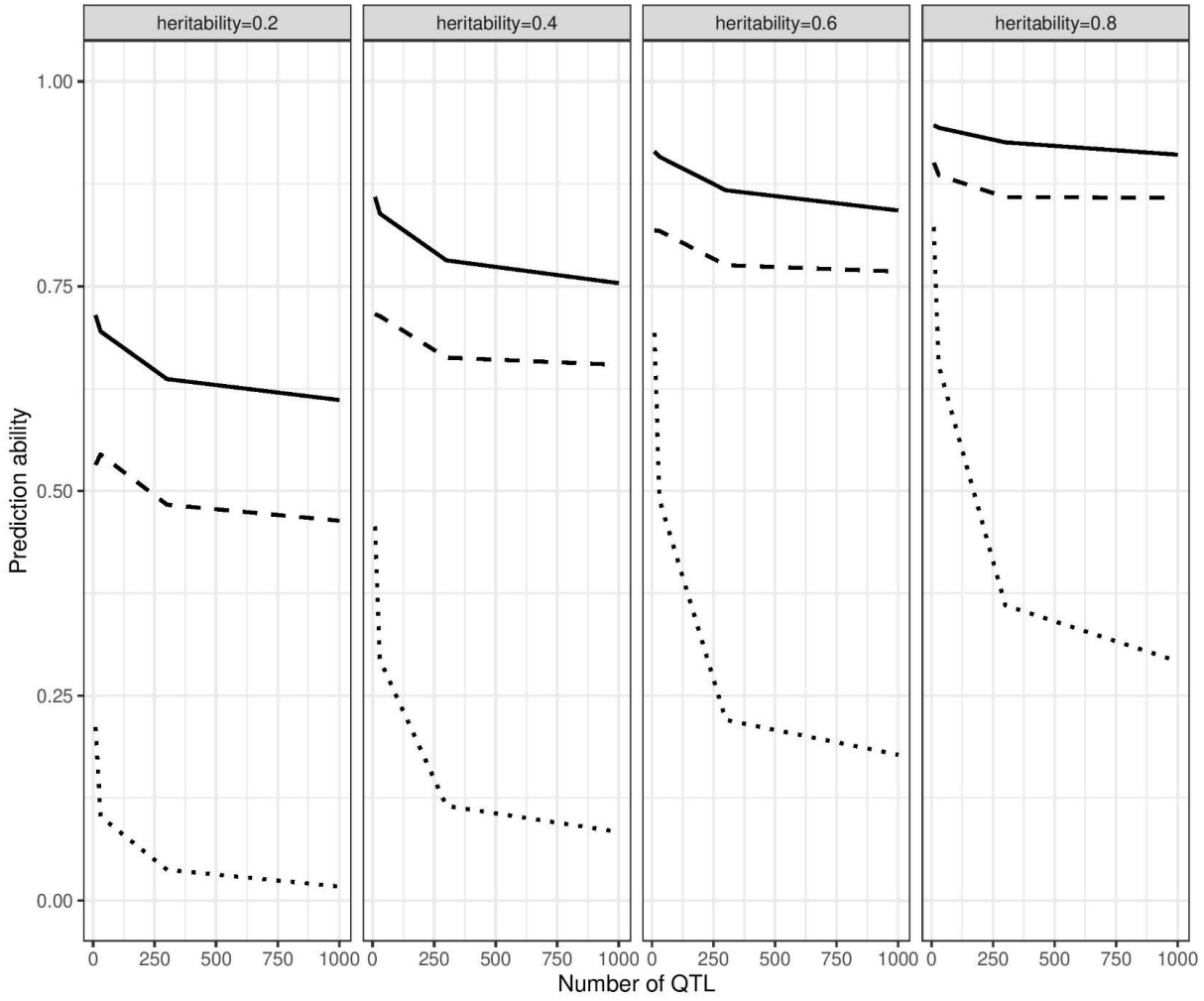
- **Phenotypes** (BLUE values):
  - Means of correlations between years: **0.69** (yield), **0.78** (grain protein content), **0.87** (plant height), **0.91** (heading date)

Trait	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_e$	$\hat{\sigma}^2_{(g\times e)}$	$\hat{\sigma}^2_r$	$\widehat{\text{rep}}_{\text{plot}}$	$\widehat{\text{rep}}_{\text{design}}$
Yield	18.9	237.1	24.8	15.4	0.32	0.91
Grain protein content	0.3	1.0	0.2	0.1	0.52	0.93
Plant height	32.6	56.5	6.5	5.6	0.73	0.97
Heading date	12.6	65.5	1.9	0.5	0.85	0.99

$$\widehat{\text{rep}}_{\text{plot}} = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \hat{\sigma}^2_{(g\times e)} + \hat{\sigma}^2_r}$$

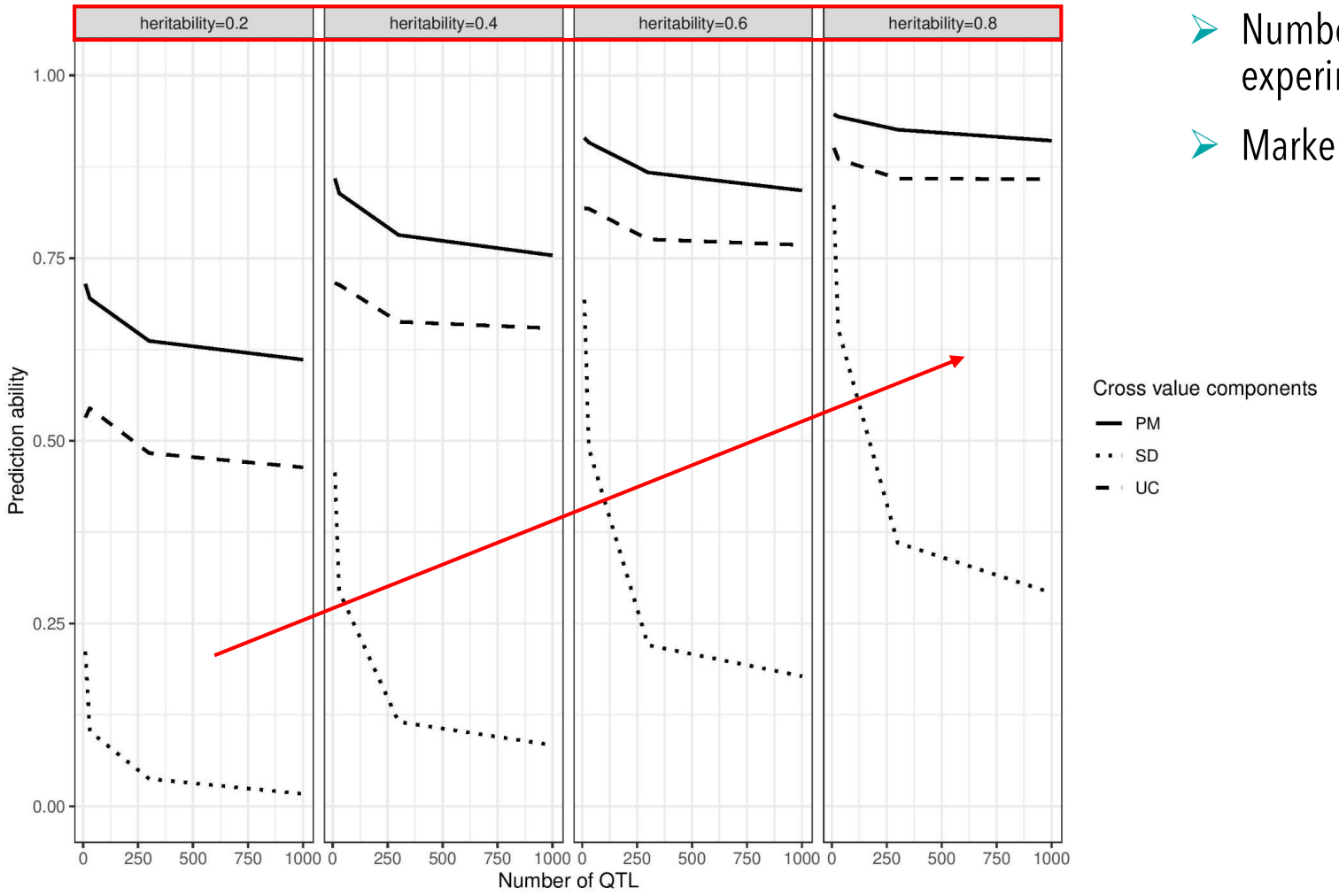
$$\widehat{\text{rep}}_{\text{design}} = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \frac{\hat{\sigma}^2_{(g\times e)}}{\text{nb\_env}} + \frac{\hat{\sigma}^2_r}{\text{nb\_rep}}}$$

Results: predictive ability and genetic architecture in simulations



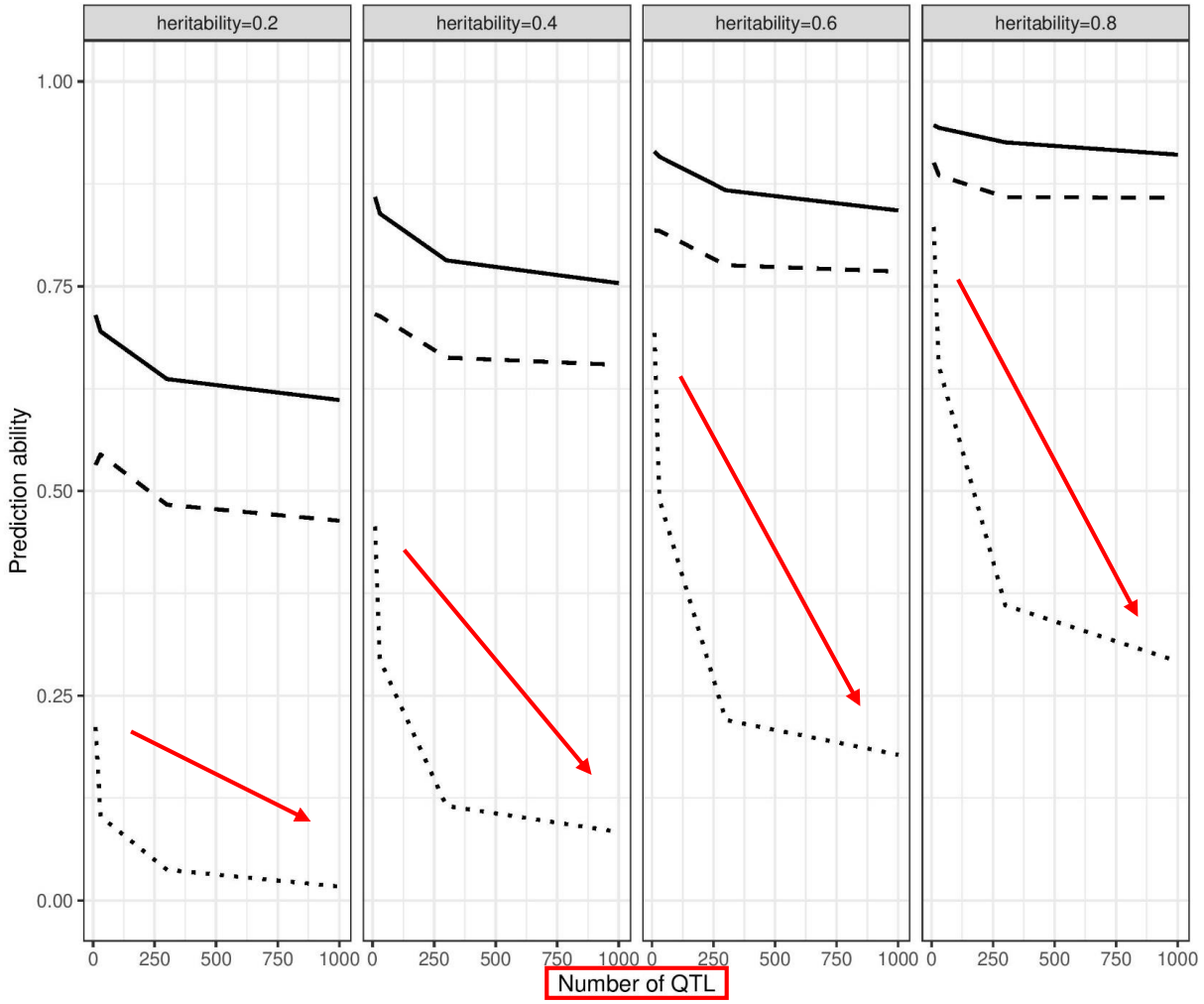
- Number of progenies per cross as in experimental data
- Marker effects estimated with BayesA

Results: prediction ability and genetic architecture in simulations



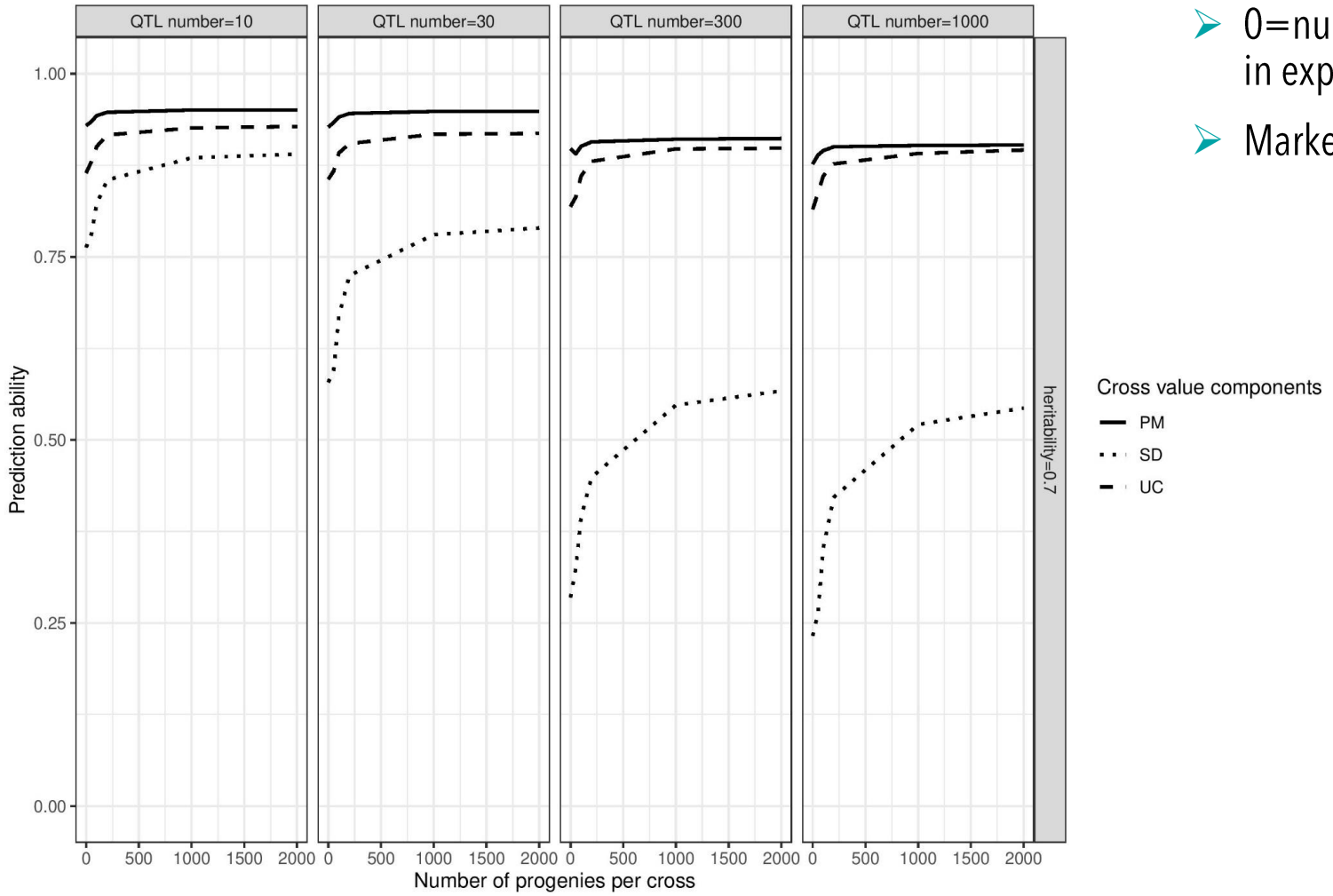
- Number of progenies per cross as in experimental data
- Marker effects estimated with BayesA

Results: prediction ability and genetic architecture in simulations



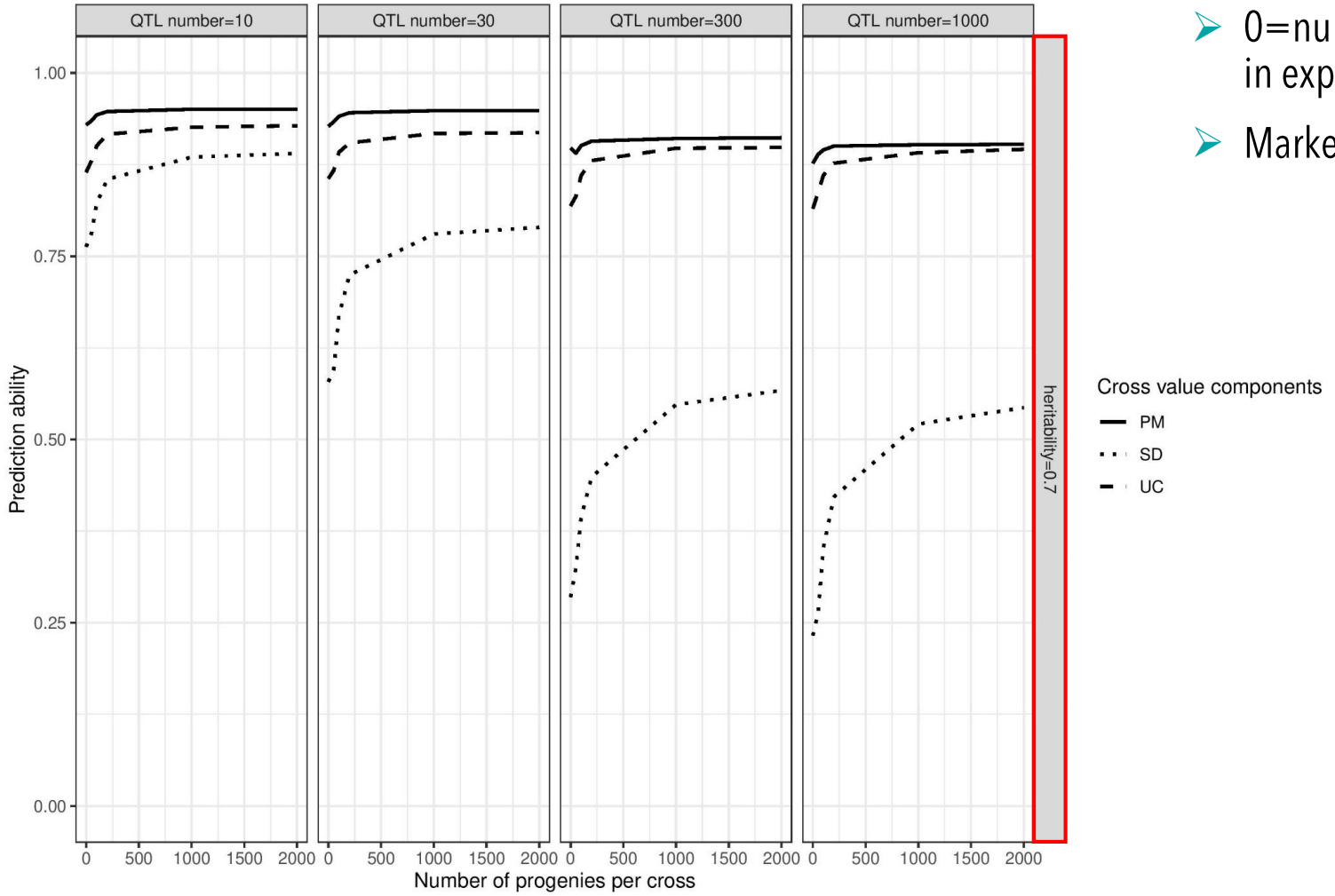
- Number of progenies per cross as in experimental data
- Marker effects estimated with BayesA

Results: prediction ability and progeny size in simulations



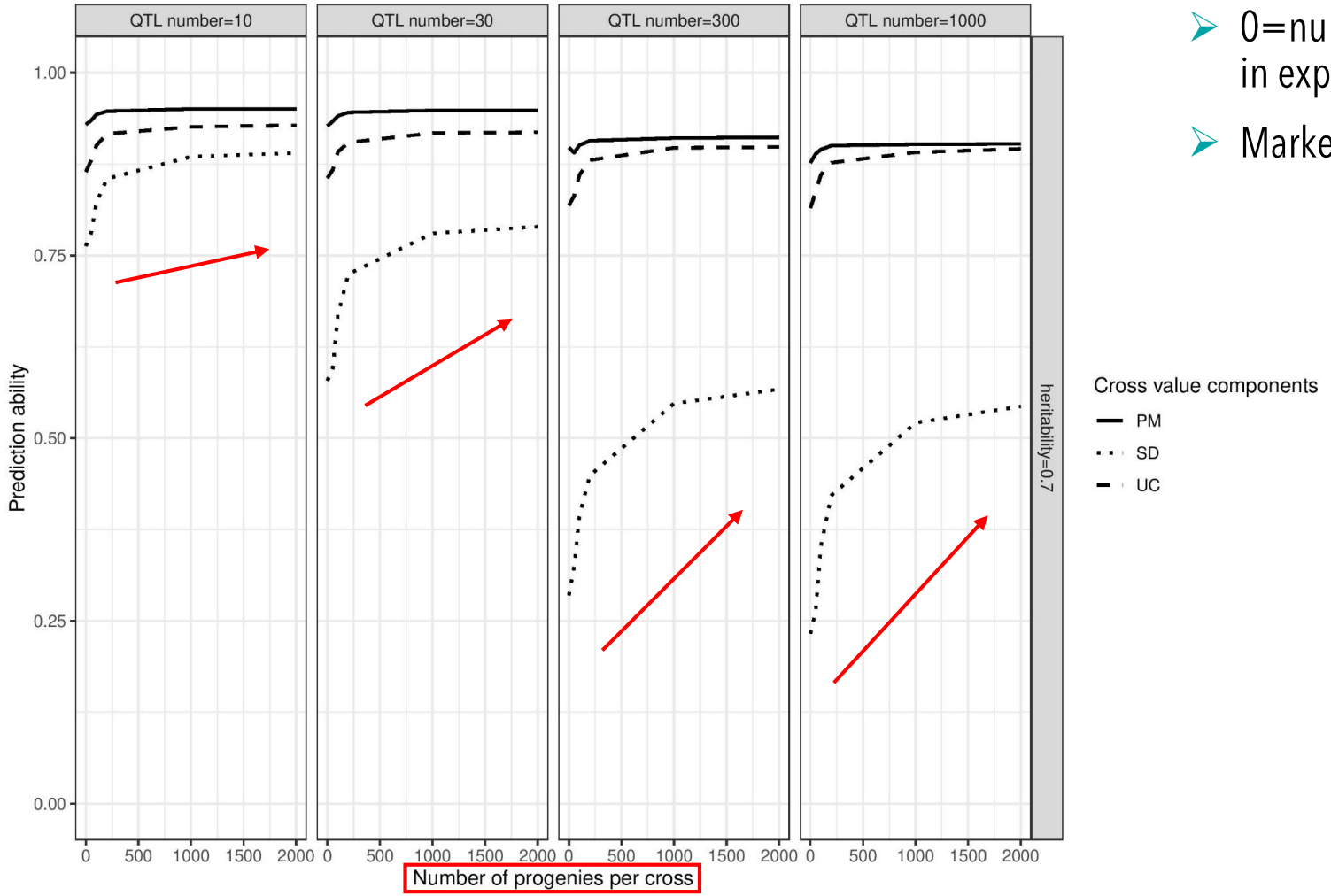
- 0=number of progenies per cross as in experimental data
- Marker effects estimated with BayesA

Results: prediction ability and progeny size in simulations



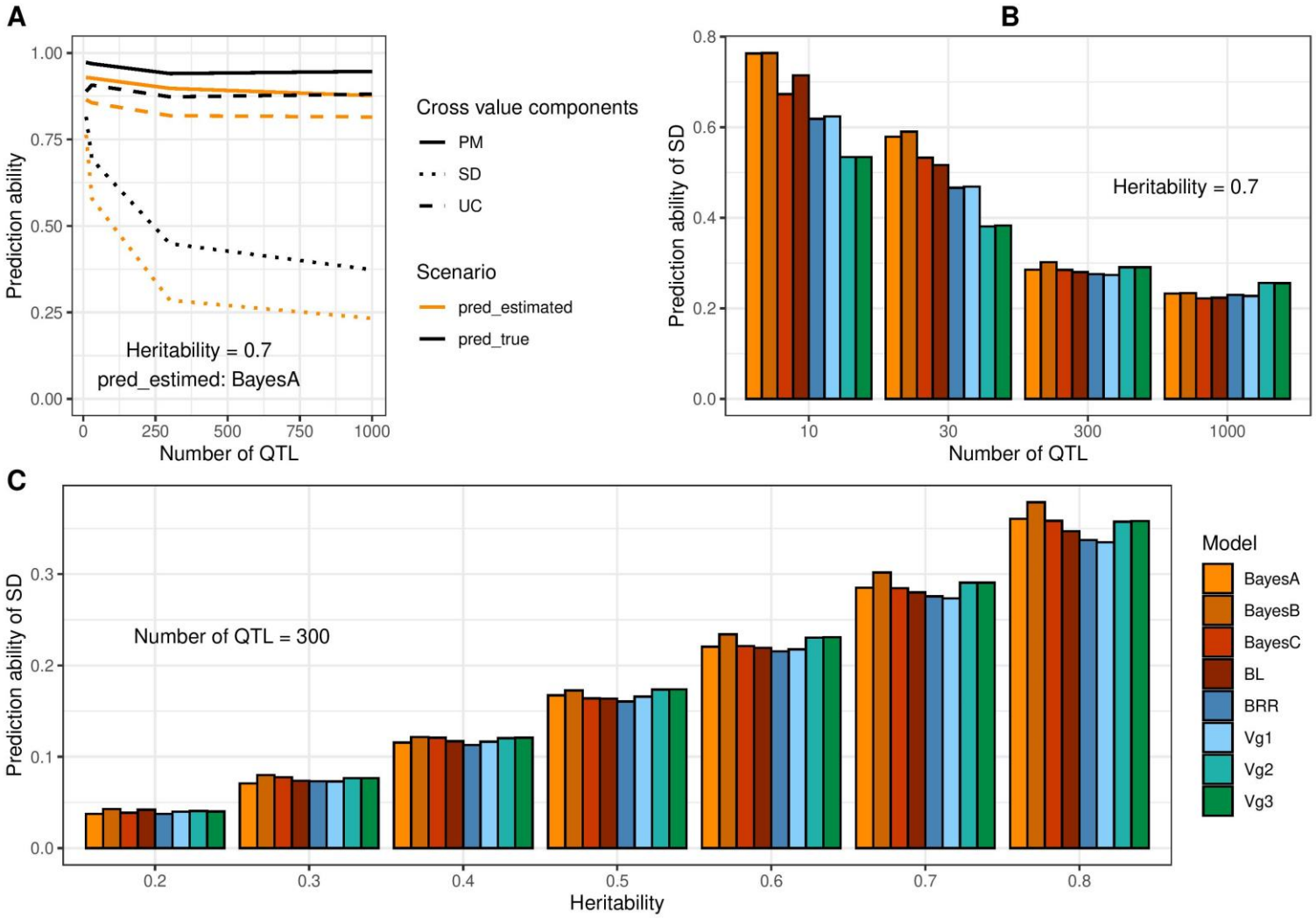
- 0=number of progenies per cross as in experimental data
- Marker effects estimated with BayesA

Results: prediction ability and progeny size in simulations



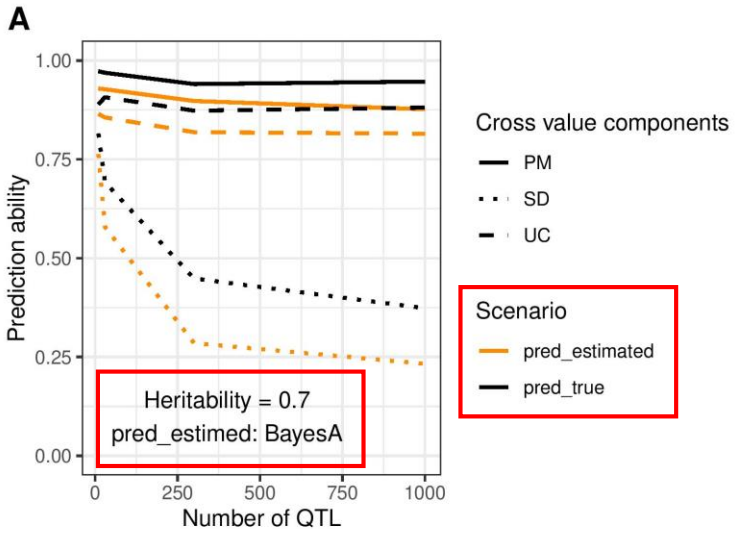
- 0=number of progenies per cross as in experimental data
- Marker effects estimated with BayesA

Results: prediction ability and knowledge of QTL in simulations



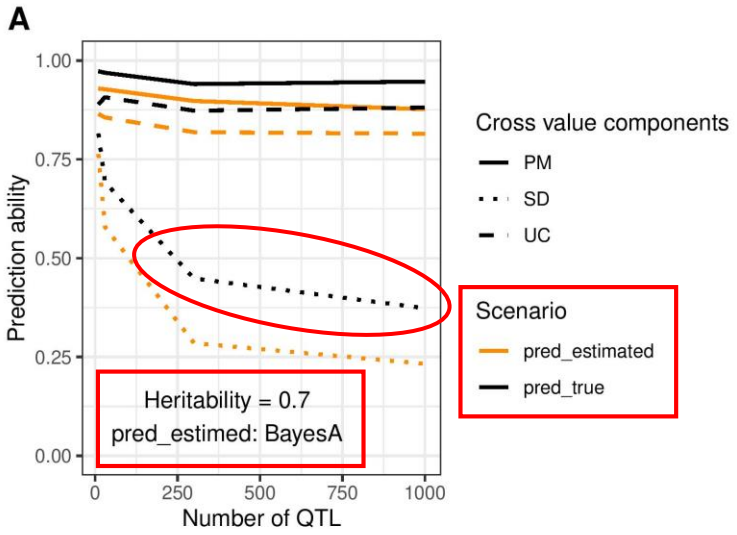
- Number of progenies per cross as in experimental data

Results: prediction ability and knowledge of QTL in simulations



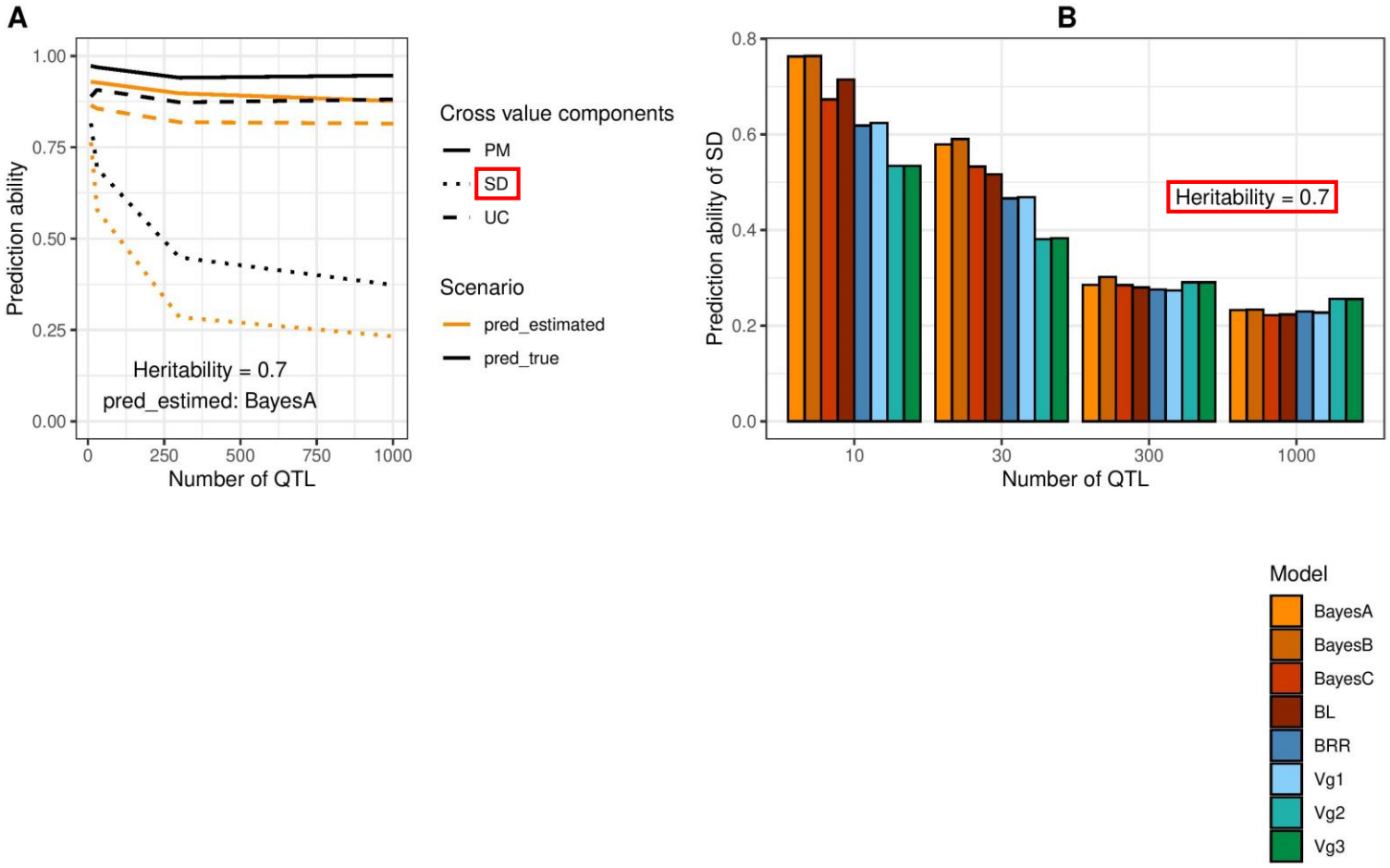
- Number of progenies per cross as in experimental data

Results: prediction ability and knowledge of QTL in simulations



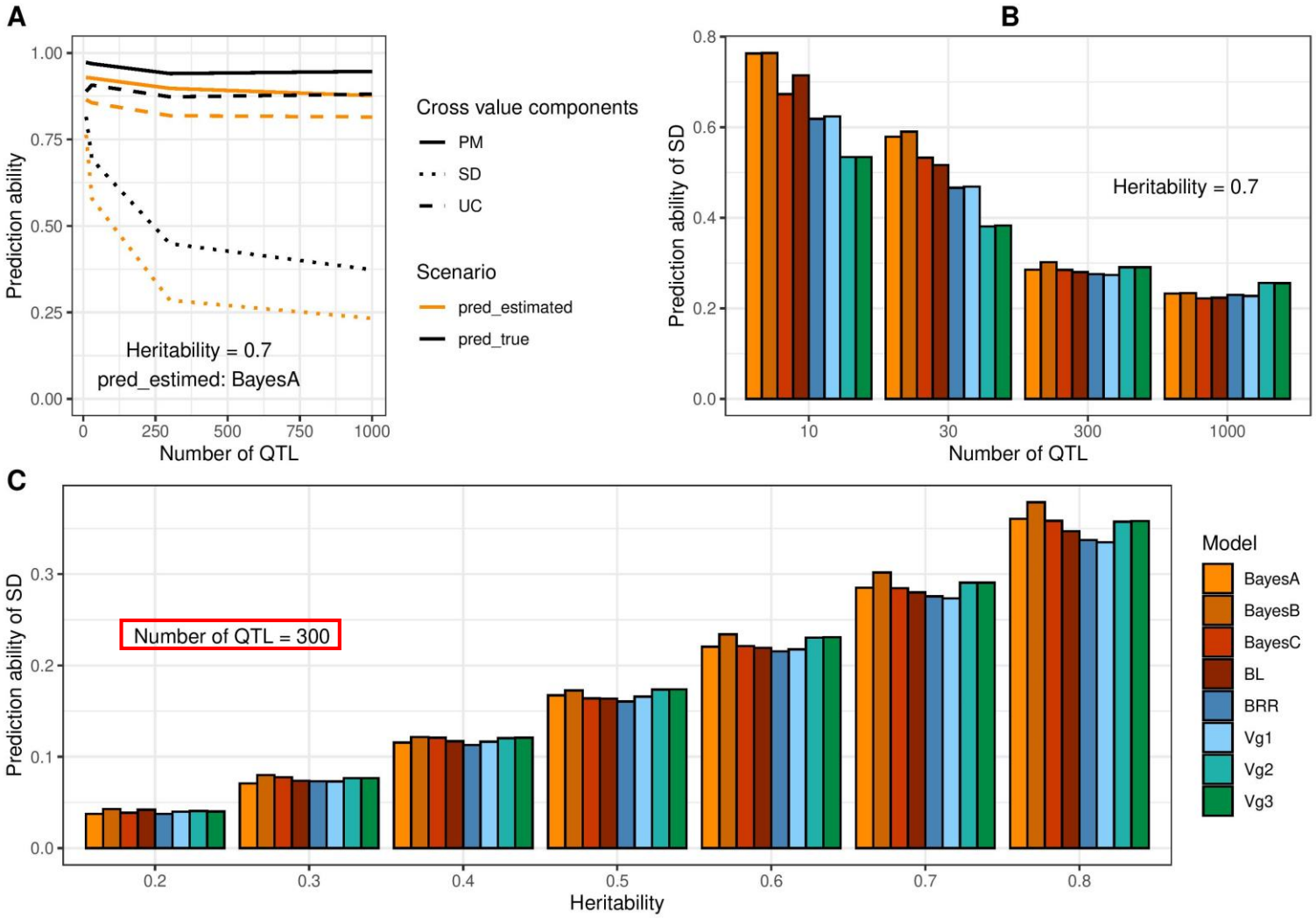
- Number of progenies per cross as in experimental data

Results: prediction ability and knowledge of QTL in simulations



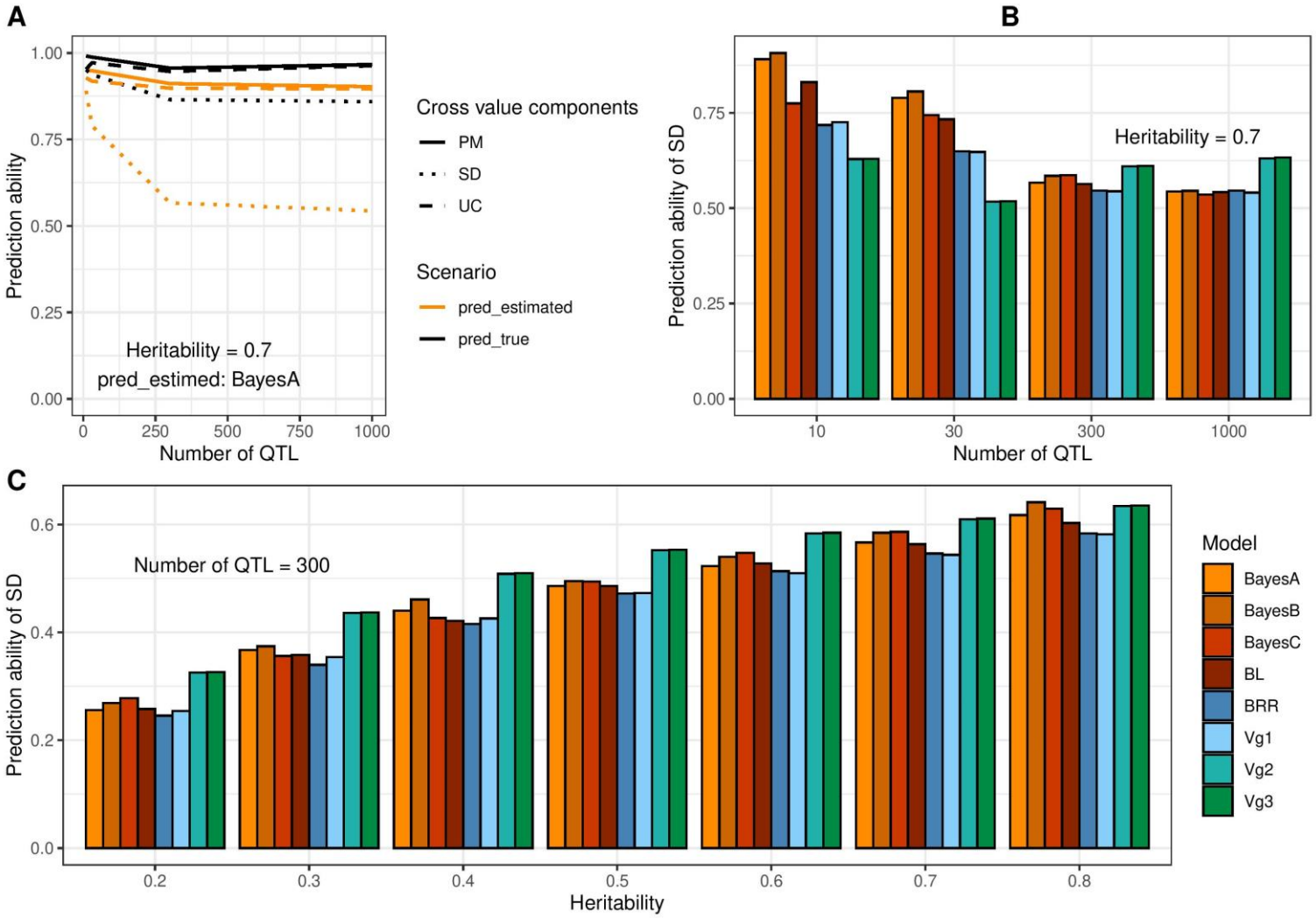
- Number of progenies per cross as in experimental data

Results: prediction ability and knowledge of QTL in simulations



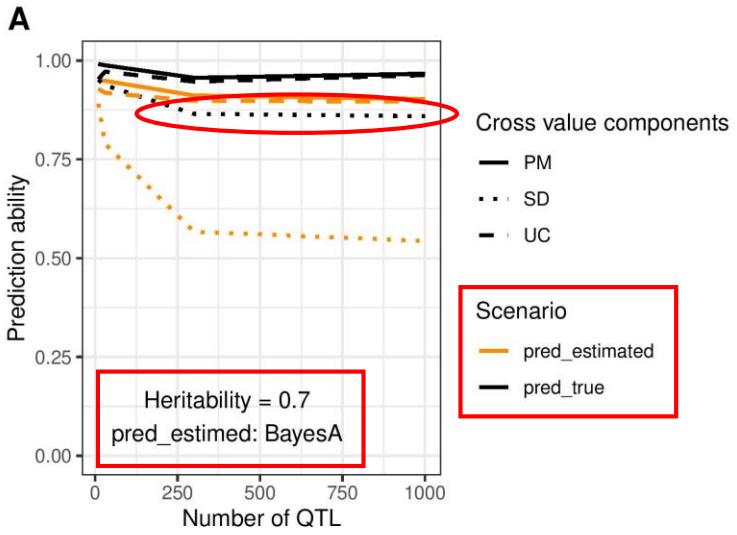
- Number of progenies per cross as in experimental data

Results: prediction ability and knowledge of QTL in simulations



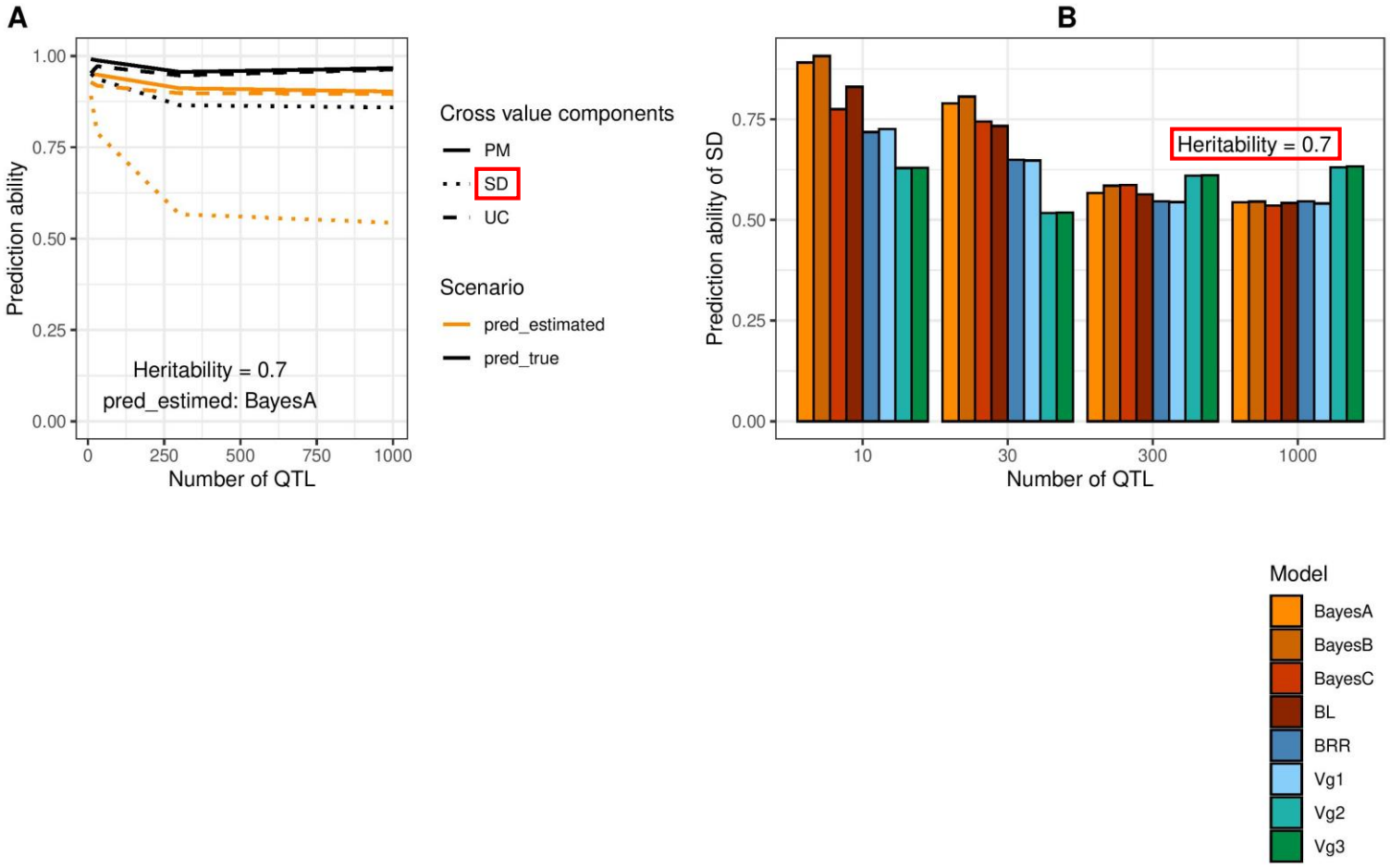
- Number of progenies per cross = 2,000

Results: prediction ability and knowledge of QTL in simulations



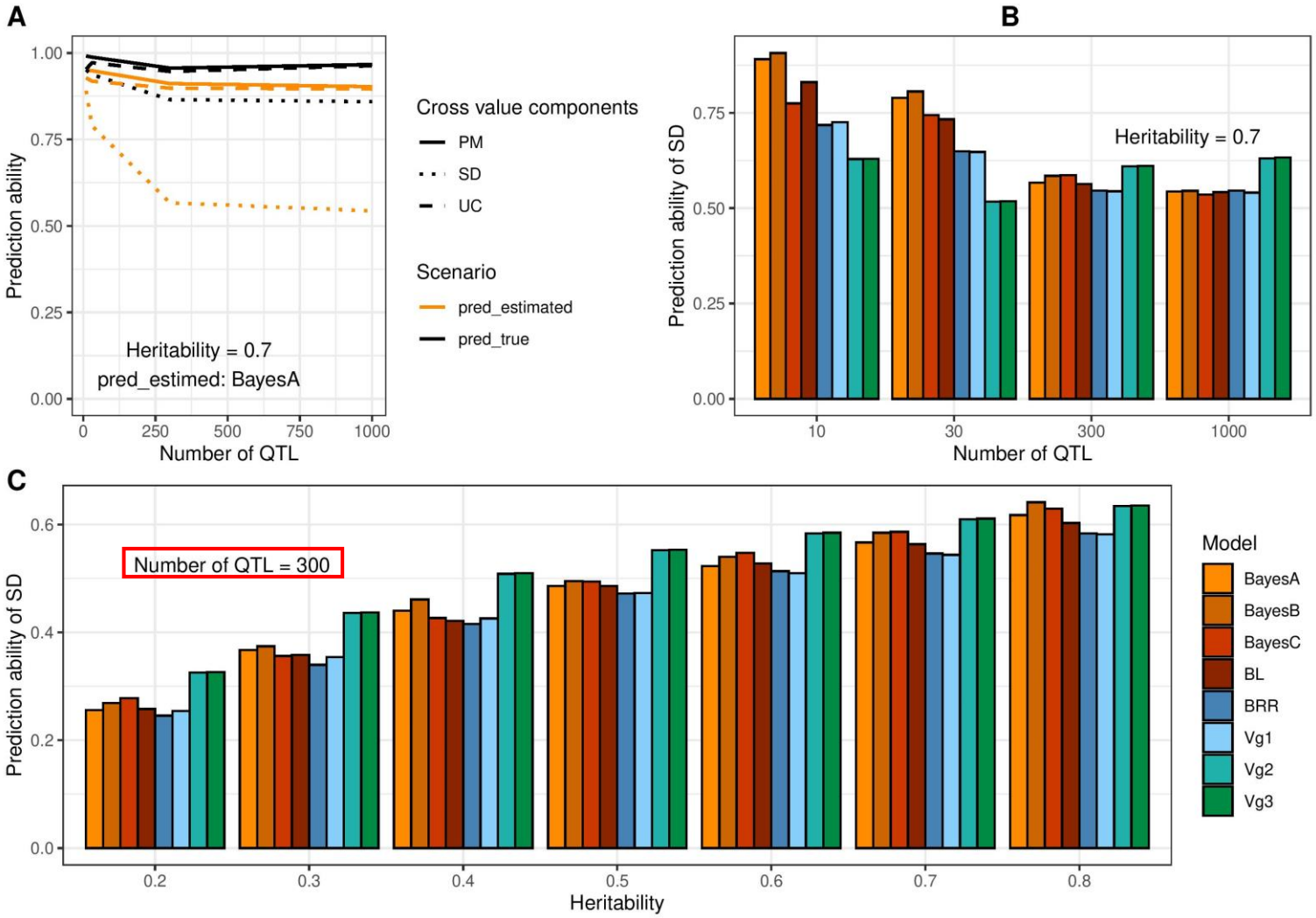
- Number of progenies per cross = 2,000

Results: prediction ability and knowledge of QTL in simulations



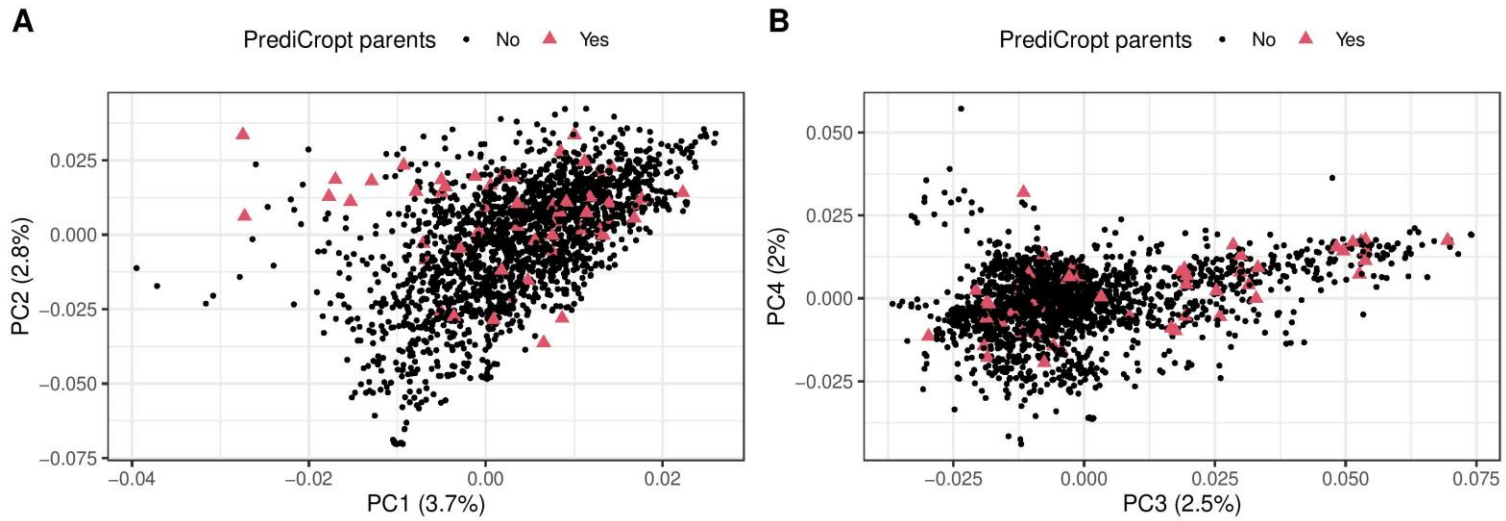
- Number of progenies per cross = 2,000

Results: prediction ability and knowledge of QTL in simulations

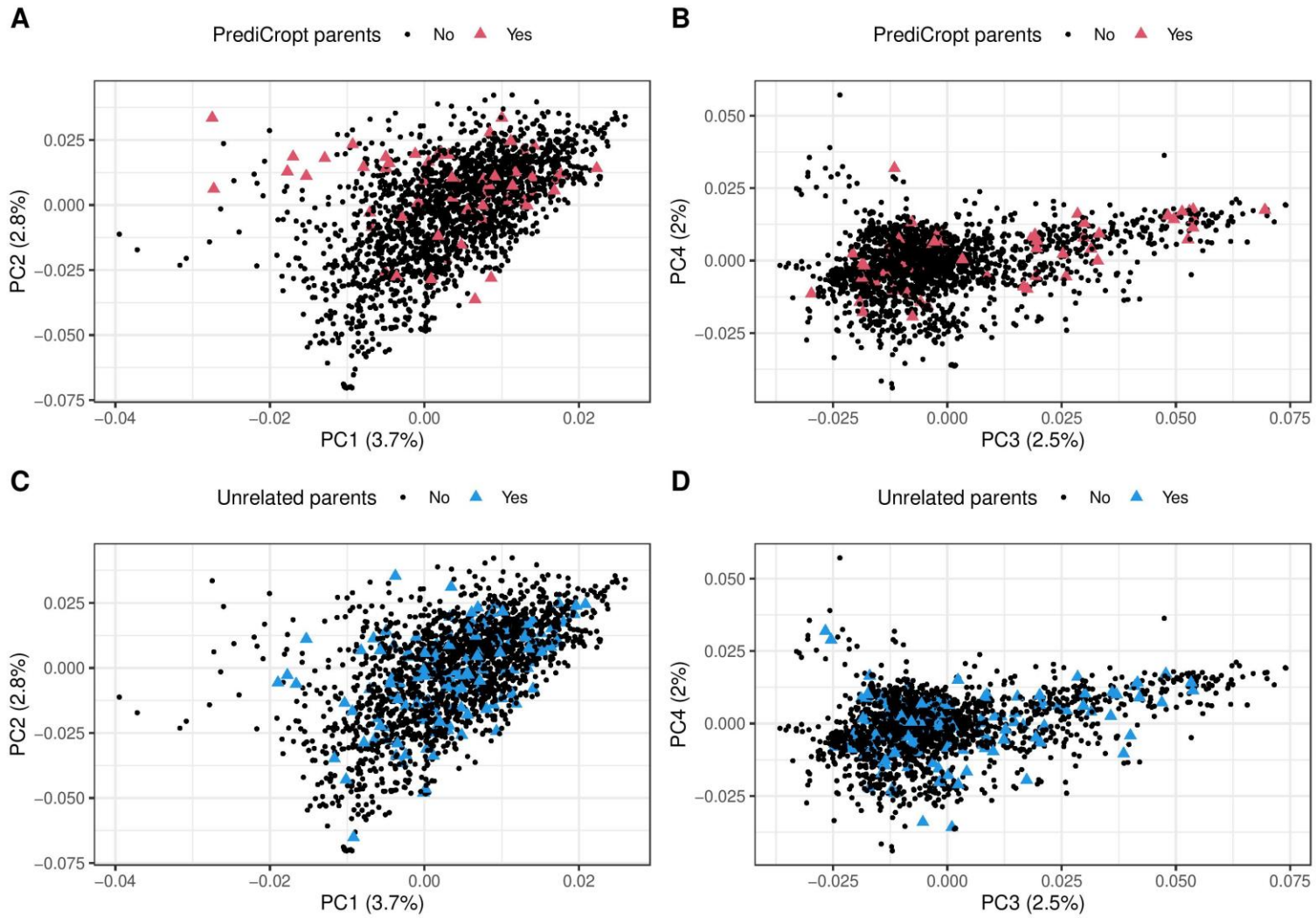


- Number of progenies per cross = 2,000

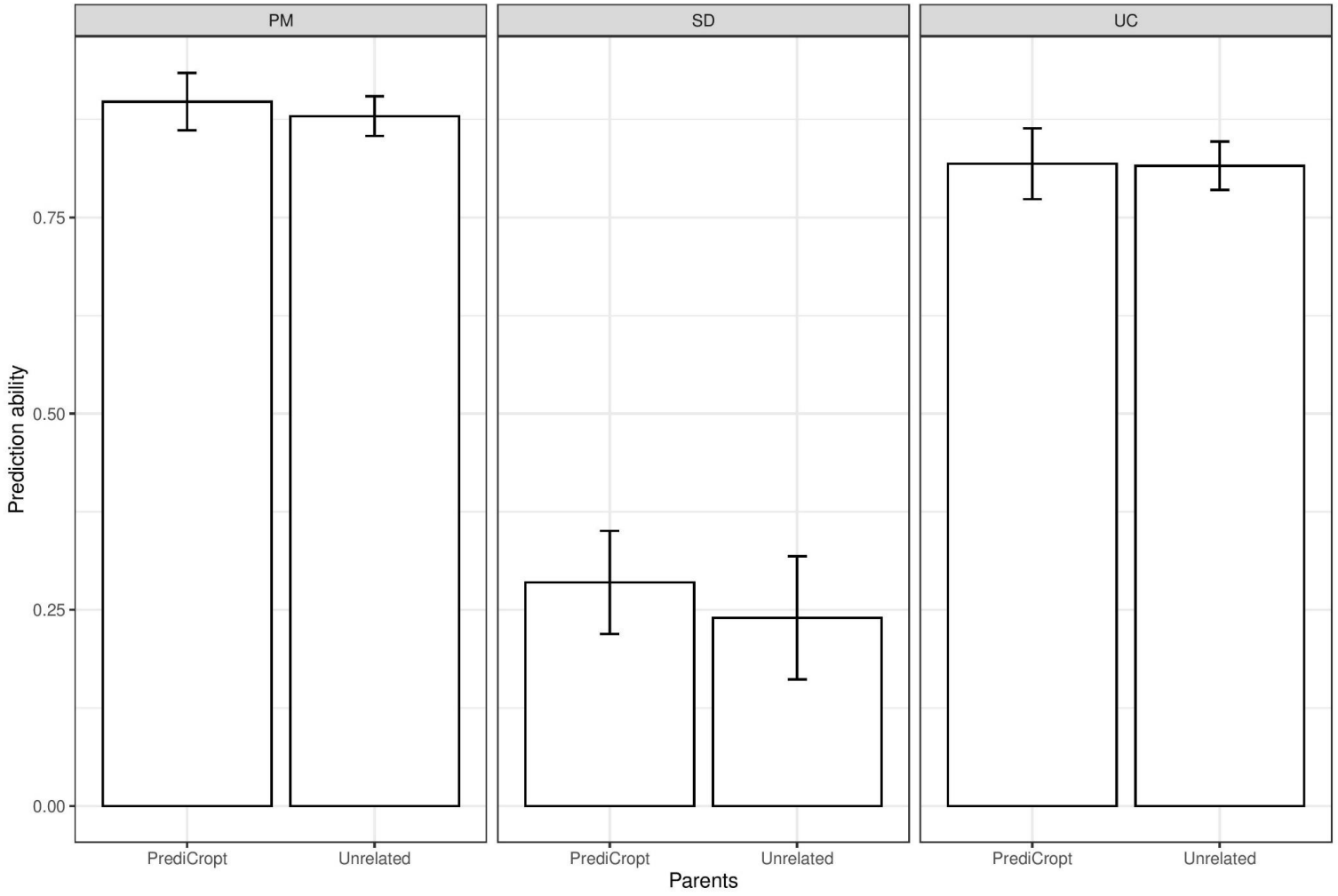
Results: genetic constitution of parents in simulations



Results: genetic constitution of parents in simulations

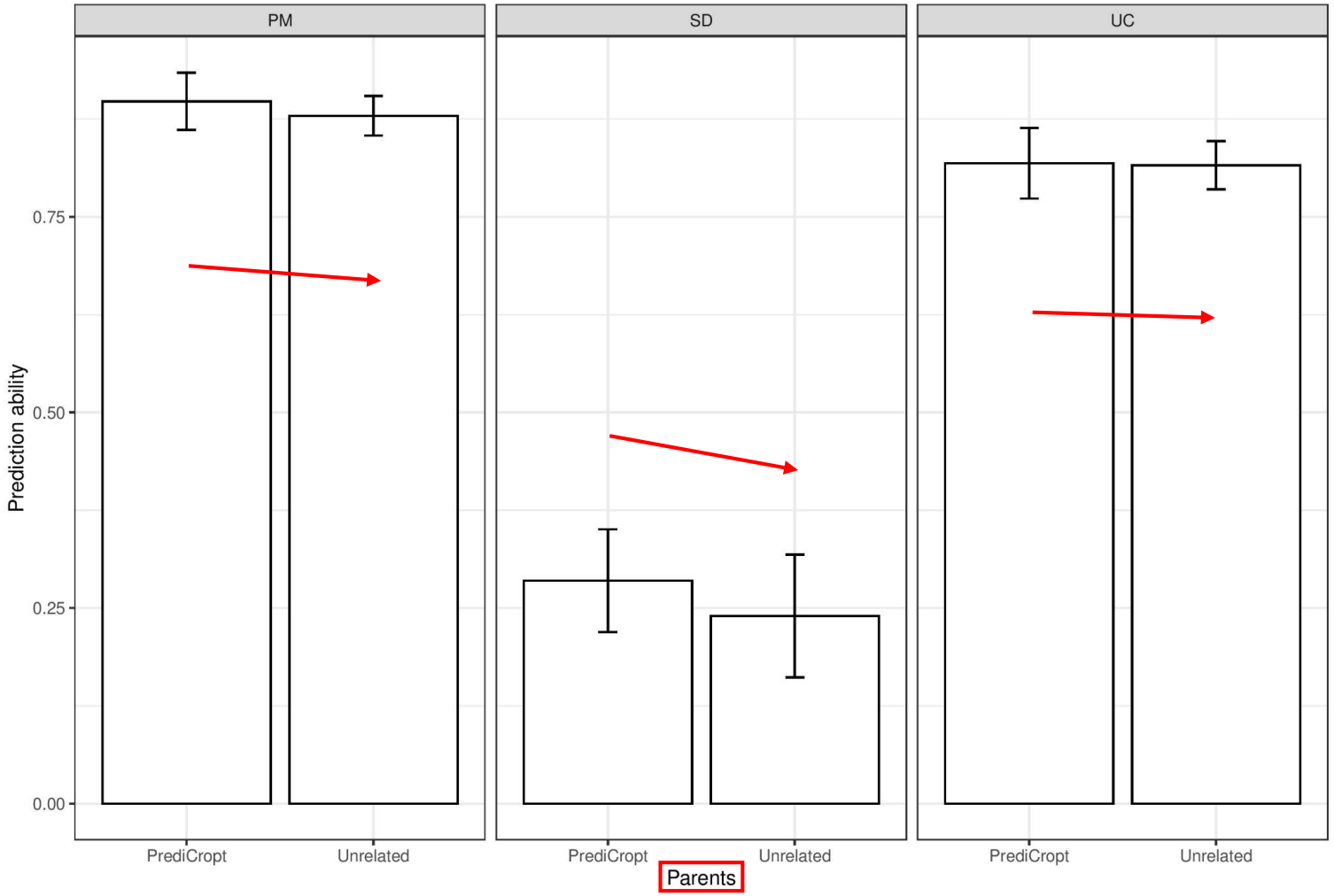


Results: prediction ability and genetic constitution of parents in simulations



- Number of progenies per cross as in experimental data
- Heritability = 0.7
- Number of QTL = 300
- Marker effects estimated with BayesA

Results: prediction ability and genetic constitution of parents in simulations



- Number of progenies per cross as in experimental data
- Heritability = 0.7
- Number of QTL = 300
- Marker effects estimated with BayesA

Results: experimental data quality

Number of	Yield	Grain protein content	Plant height	Heading date
BLUE values (lines)	**5,658**	**3,596**	**5,684**	**5,683**

- 3 years, 6 locations → 15 environments (year*location) except for grain protein content (11)

Results: experimental data quality

Number of	Yield	Grain protein content	Plant height	Heading date
BLUE values (lines)	**5,658**	**3,596**	**5,684**	**5,683**

- 3 years, 6 locations → 15 environments (year*location) except for grain protein content (11)

- Means of correlations between locations:
  - Yield: **0.48**
  - Grain protein content: **0.43**
  - Plant height: **0.79**
  - Heading date: **0.88**

PA = Prediction Ability
CI = Confidence Interval

Results: prediction ability in real experimental data

Trait	Model	PM					
		PA	CI				
Yield	BayesA	0.37	0.18, 0.55				
	BayesB	0.38	0.19, 0.56				
	BayesC	0.38	0.19, 0.56				
	BL	0.38	0.20, 0.56				
	BRR	0.38	0.20, 0.56				
	Vg1	0.38	0.20, 0.56				
	Vg2						
	Vg3						
Grain protein content	BayesA	0.63	0.44, 0.81				
	BayesB	0.62	0.44, 0.80				
	BayesC	0.63	0.45, 0.81				
	BL	0.63	0.45, 0.81				
	BRR	0.63	0.45, 0.81				
	Vg1	0.63	0.45, 0.81				
	Vg2						
	Vg3						

Trait	Model	PM					
		PA	CI				
Plant height	BayesA	0.51	0.34, 0.68				
	BayesB	0.51	0.34, 0.68				
	BayesC	0.51	0.34, 0.68				
	BL	0.51	0.34, 0.68				
	BRR	0.50	0.32, 0.67				
	Vg1	0.50	0.33, 0.67				
	Vg2						
	Vg3						
Heading date	BayesA	0.92	0.85, 1.00				
	BayesB	0.93	0.85, 1.00				
	BayesC	0.82	0.71, 0.93				
	BL	0.92	0.84, 1.00				
	BRR	0.90	0.82, 0.99				
	Vg1	0.90	0.82, 0.99				
	Vg2						
	Vg3						

PA = Prediction Ability
CI = Confidence Interval

Results: prediction ability in real experimental data

Trait	Model	PM		UC	
		PA	CI	PA	CI
Yield	BayesA	0.37	0.18, 0.55	0.45	0.28, 0.63
	BayesB	0.38	0.19, 0.56	0.47	0.29, 0.64
	BayesC	0.38	0.19, 0.56	0.46	0.28, 0.63
	BL	0.38	0.20, 0.56	0.45	0.28, 0.63
	BRR	0.38	0.20, 0.56	0.45	0.28, 0.63
	Vg1	0.38	0.20, 0.56	0.46	0.28, 0.63
	Vg2			0.45	0.27, 0.62
	Vg3			0.45	0.27, 0.62
Grain protein content	BayesA	0.63	0.44, 0.81	0.52	0.32, 0.72
	BayesB	0.62	0.44, 0.80	0.52	0.33, 0.72
	BayesC	0.63	0.45, 0.81	0.51	0.31, 0.71
	BL	0.63	0.45, 0.81	0.51	0.31, 0.71
	BRR	0.63	0.45, 0.81	0.51	0.31, 0.71
	Vg1	0.63	0.45, 0.81	0.52	0.32, 0.72
	Vg2			0.53	0.33, 0.72
	Vg3			0.53	0.33, 0.72

Trait	Model	PM		UC	
		PA	CI	PA	CI
Plant height	BayesA	0.51	0.34, 0.68	0.66	0.52, 0.81
	BayesB	0.51	0.34, 0.68	0.66	0.51, 0.81
	BayesC	0.51	0.34, 0.68	0.65	0.50, 0.80
	BL	0.51	0.34, 0.68	0.60	0.44, 0.75
	BRR	0.50	0.32, 0.67	0.47	0.30, 0.64
	Vg1	0.50	0.33, 0.67	0.48	0.31, 0.65
	Vg2			0.45	0.27, 0.62
	Vg3			0.45	0.27, 0.62
Heading date	BayesA	0.92	0.85, 1.00	0.70	0.56, 0.84
	BayesB	0.93	0.85, 1.00	0.71	0.57, 0.85
	BayesC	0.82	0.71, 0.93	0.64	0.49, 0.79
	BL	0.92	0.84, 1.00	0.71	0.58, 0.85
	BRR	0.90	0.82, 0.99	0.77	0.64, 0.90
	Vg1	0.90	0.82, 0.99	0.77	0.64, 0.90
	Vg2			0.78	0.65, 0.90
	Vg3			0.78	0.65, 0.90

PA = Prediction Ability
CI = Confidence Interval

Results: prediction ability in real experimental data

Trait	Model	PM		SD		UC	
		PA	CI	PA	CI	PA	CI
Yield	BayesA	0.37	0.18, 0.55	0.02	-0.18, 0.22	0.45	0.28, 0.63
	BayesB	0.38	0.19, 0.56	0.05	-0.15, 0.24	0.47	0.29, 0.64
	BayesC	0.38	0.19, 0.56	0.02	-0.18, 0.21	0.46	0.28, 0.63
	BL	0.38	0.20, 0.56	0.01	-0.19, 0.21	0.45	0.28, 0.63
	BRR	0.38	0.20, 0.56	0.00	-0.20, 0.20	0.45	0.28, 0.63
	Vg1	0.38	0.20, 0.56	-0.01	-0.20, 0.19	0.46	0.28, 0.63
	Vg2			-0.17	-0.36, 0.02	0.45	0.27, 0.62
	Vg3			-0.17	-0.36, 0.02	0.45	0.27, 0.62
Grain protein content	BayesA	0.63	0.44, 0.81	0.17	-0.05, 0.40	0.52	0.32, 0.72
	BayesB	0.62	0.44, 0.80	0.26	0.03, 0.48	0.52	0.33, 0.72
	BayesC	0.63	0.45, 0.81	0.07	-0.16, 0.30	0.51	0.31, 0.71
	BL	0.63	0.45, 0.81	0.10	-0.14, 0.33	0.51	0.31, 0.71
	BRR	0.63	0.45, 0.81	0.08	-0.15, 0.32	0.51	0.31, 0.71
	Vg1	0.63	0.45, 0.81	0.05	-0.19, 0.28	0.52	0.32, 0.72
	Vg2			0.16	-0.07, 0.39	0.53	0.33, 0.72
	Vg3			0.16	-0.07, 0.39	0.53	0.33, 0.72

Trait	Model	PM		SD		UC	
		PA	CI	PA	CI	PA	CI
Plant height	BayesA	0.51	0.34, 0.68	0.64	0.48, 0.79	0.66	0.52, 0.81
	BayesB	0.51	0.34, 0.68	0.63	0.48, 0.79	0.66	0.51, 0.81
	BayesC	0.51	0.34, 0.68	0.64	0.49, 0.79	0.65	0.50, 0.80
	BL	0.51	0.34, 0.68	0.64	0.49, 0.79	0.60	0.44, 0.75
	BRR	0.50	0.32, 0.67	0.53	0.36, 0.70	0.47	0.30, 0.64
	Vg1	0.50	0.33, 0.67	0.54	0.37, 0.70	0.48	0.31, 0.65
	Vg2			0.41	0.23, 0.59	0.45	0.27, 0.62
	Vg3			0.41	0.23, 0.59	0.45	0.27, 0.62
Heading date	BayesA	0.92	0.85, 1.00	0.46	0.28, 0.63	0.70	0.56, 0.84
	BayesB	0.93	0.85, 1.00	0.46	0.29, 0.64	0.71	0.57, 0.85
	BayesC	0.82	0.71, 0.93	0.49	0.32, 0.66	0.64	0.49, 0.79
	BL	0.92	0.84, 1.00	0.41	0.23, 0.59	0.71	0.58, 0.85
	BRR	0.90	0.82, 0.99	0.33	0.14, 0.52	0.77	0.64, 0.90
	Vg1	0.90	0.82, 0.99	0.34	0.16, 0.53	0.77	0.64, 0.90
	Vg2			0.33	0.14, 0.51	0.78	0.65, 0.90
	Vg3			0.33	0.14, 0.51	0.78	0.65, 0.90

Take-home message

- Parameters impacting the prediction abilities of PM, SD and UC (simulations):

Take-home message

- Parameters impacting the prediction abilities of PM, SD and UC (simulations):
  - **Genetic architecture** = factor with the strongest impact
  - Very strong impact of the **number of progenies per cross**
  - **SD difficult to predict** even in "true" scenarios ~ infinite TP size (with small number of progenies per cross)
  - Very **little impact** of marker effect estimation **model, except Vg3 when the number of QTL >300, especially when heritability is low**
  - Little/**no** impact of **selected crosses: try with higher germplasm diversity (genetic ressources)**

Take-home message

- Parameters impacting the prediction abilities of PM, SD and UC (simulations):
  - **Genetic architecture** = factor with the strongest impact
  - Very strong impact of the **number of progenies per cross**
  - **SD difficult to predict** even in "true" scenarios ~ infinite TP size (with small number of progenies per cross)
  - Very **little impact** of marker effect estimation **model**
  - Taking into account the **error in marker effect estimates** improved SD predictions for quantitative traits with low heritability
  - Little/**no** impact of **selected crosses**

- Application on real experimental data:

Take-home message

- Parameters impacting the prediction abilities of PM, SD and UC (simulations):
  - **Genetic architecture** = factor with the strongest impact
  - Very strong impact of the **number of progenies per cross**
  - **SD difficult to predict** even in "true" scenarios ~ infinite TP size (with small number of progenies per cross)
  - Very **little impact** of marker effect estimation **model**
  - Taking into account the **error in marker effect estimates** improved SD predictions for quantitative traits with low heritability
  - Little/**no** impact of **selected crosses**

- Application on real experimental data:
  - **PM** and **UC** were **correctly** predicted for the 4 traits
  - **SD correctly** predicted for **plant height** and **heading date** and **badly** predicted for **yield** and **grain protein content** (GxE?)

Take-home message

- Parameters impacting the prediction abilities of PM, SD and UC (simulations):
  - **Genetic architecture** = factor with the strongest impact
  - Very strong impact of the **number of progenies per cross**
  - **SD difficult to predict** even in "true" scenarios ~ infinite TP size (with small number of progenies per cross)
  - Very **little impact** of marker effect estimation **model**
  - Taking into account the **error in marker effect estimates** improved SD predictions for quantitative traits with low heritability
  - Little/**no** impact of **selected crosses**

- Application on real experimental data:
  - **PM** and **UC** were **correctly** predicted for the 4 traits
  - **SD correctly** predicted for **plant height** and **heading date** and **badly** predicted for **yield** and **grain protein content** (GxE?)

- Recommendations for crossbreeding plans: **selecting on UC** and generating a **very large number of progenies** for a very **small number of crosses** would make it possible to obtain the most promising offspring?

Validation of cross progeny variance genomic prediction using simulations and experimental data in winter elite bread wheat

Claire Oget-Ebrad, Emmanuel Heumez, Laure Duchalais, Ellen Goudemand-Dugué, François-Xavier Oury, Jean-Michel Elsen, Sophie Bouchet

doi: https://doi.org/10.1101/2023.09.26.558758

Posted September 28, 2023.

Question

Est-ce que cette capacité de prédiction est suffisante pour apporter un gain génétique significatif dans un programme de sélection (matériel élite) et dans un programme de pre-breeding?

Optimiser le plan de croisement

Vecteur d'UC + **contraintes : budget, seuil de diversité génétique, Allèles imposés pour certains caractères**

Algorithm to optimize several objective functions(Danguy et al., 2023)

Optimisation du plan de croisement

croisement	UC	nb de descendants
$P_1 * P_2$	UC_1	D_{12}
$P_i * P_j$	UC_j	D_{ij}
….		

Optimiser le plan de croisement élite sur 1 génération

Article Navigation

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Comparison of genomic-enabled cross selection criteria for the improvement of inbred line breeding populations

Alice Danguy des Déserts, Nicolas Durand, Bertrand Servin, Ellen Goudemand-Dugué, Jean-Marc Alliot, Daniel Ruiz, Gilles Charmet, Jean-Michel Elsen, Sophie Bouchet Author Notes

G3 Genes|Genomes|Genetics, jkad195, https://doi.org/10.1093/g3journal/jkad195
Published: 25 August 2023 **Article history**

PDF Split View Cite Permissions Share

Abstract

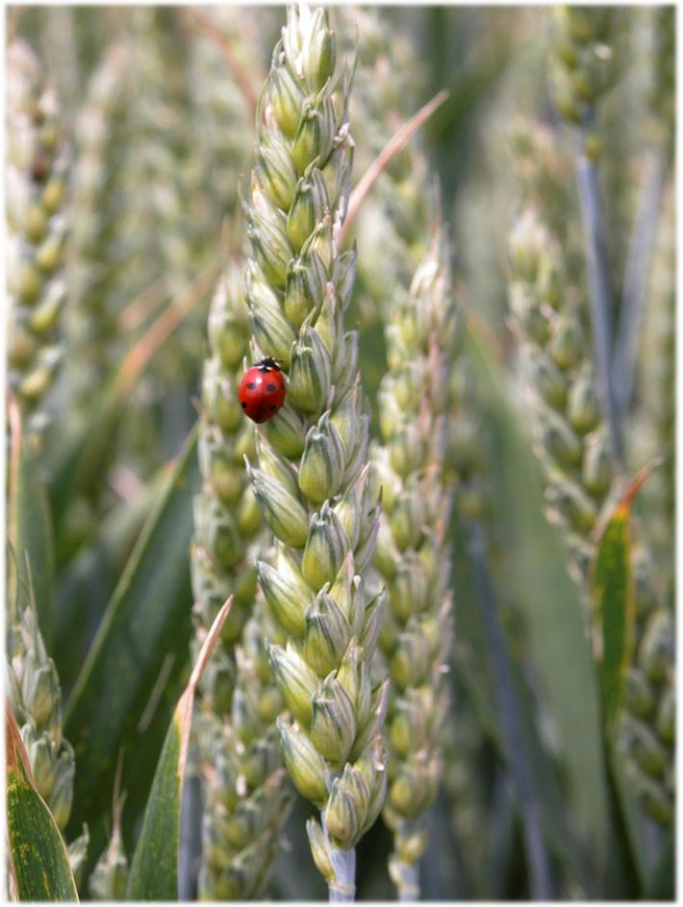
A crucial step in inbred plant breeding is the choice of mating design to derive high-performing inbred varieties while also maintaining a competitive breeding population to secure sufficient genetic gain in future generations. In practice, the mating design usually relies on crosses involving the best parental inbred lines to ensure high mean

Perspective

- Plusieurs générations
- Le gain génétique dépend du ratio t = var(σ)/var(PM)
- Matériel de départ plus divers (pre-breeding)
- Contraintes sur allèles importants pour d'autres caractères (qualité boulangère, maladies…)

- post-doc optimisation de plans de croisements (Sophie Bouchet)
- Post-doc optimisation prédictions GxE avec crop models (Justin Blancon)

Thank you for your attention



RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

FSOV

FLORIMOND DESPREZ

GDEC
Génétique Diversité Ecophysiologie des Céréales

GenPhySE
Génétique, Physiologie et Systèmes d'Elevage

AO = Agri-Obtentions
GEVES = *Groupe d'Etude et de contrôle des Variétés Et des Semences* (Variety and Seed Study and Control Group)
VATE = *Valeur Agronomique Technologique et Environnementale* (Agronomic, technological and environmental value)

Materials: Training Population (TP)

- 2 datasets * 2 geographical areas:
  - **INRAE-AO**: F8-F9 winter bread wheat lines developed by INRAE-AO (2000-2022)
    157-169 (North) and 26-42 (South) lines evaluated each year
  - **GEVES**: VATE winter bread wheat data from the evaluation of varieties for national registration (2000-2022)
    44-46 (North) and 26-27 (South) lines evaluated each year
- Crop management methods = **high yield objectives** (optimized pesticide, fungicide and nitrogen amount)

Trait	Number of environments (location*year)	Number of spatially adjusted means	Number of lines with BLUE values	Number of lines with phenotypes + genotypes (23K markers)
Yield	1,031	57,226	3,241	**2,146**
Grain protein content	714	30,901	2,933	2,062
Plant height	551	29,445	3,008	2,126
Heading date	832	44,614	3,237	2,145

Materials: real experimental data (2/2)

- Crop management methods = **high yield objectives**
- Observed phenotypes: **yield**, **grain protein content** (not available for INRAE 2022 crosses: 74 crosses), **plant height** & **heading date**

Organism	Number of	Yield	Grain protein content	Plant height	Heading date
FD	Parents/Controls	29	29	29	29
	Progenies	1,139	1,139	1,139	1,139
INRAE	Parents/Controls	74	52	74	74
	Progenies	4,426	2,383	4,452	4,451
Both	Adjusted phenotypes	7,867	4,638	7,907	7,905
	BLUE values (lines)	**5,658**	**3,596**	**5,684**	**5,683**

- $\left(\hat{\sigma}^2_{P_1\times P_2}\right)_{\mathrm{Vg}_1} = \widehat{\boldsymbol{\beta}}'\boldsymbol{V}_{\boldsymbol{P_1}\times\boldsymbol{P_2}}\widehat{\boldsymbol{\beta}}$
- $\left(\hat{\sigma}^2_{P_1\times P_2}\right)_{\mathrm{Vg}_2} = \widehat{\boldsymbol{\beta}}'\boldsymbol{V}_{\boldsymbol{P_1}\times\boldsymbol{P_2}}\widehat{\boldsymbol{\beta}} + \mathrm{trace}\{\boldsymbol{V}_{\boldsymbol{P_1}\times\boldsymbol{P_2}}\mathrm{var}(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y})\}$

$\left(\hat{\sigma}^2_{P_1\times P_2}\right)_{\mathrm{Vg}_2}$ is an algebraic version of the Posterior Mean Variance (PMV) method described by (Lehermeier et al. 2017), i.e. the average of the variances obtained over L successive samplings in the Markov chain:

$\sigma^{2\,(\mathrm{PMV})} = \frac{1}{L}\sum_{s=1}^{L}\boldsymbol{\beta}'_{(s)}\boldsymbol{V}\boldsymbol{\beta}_{(s)}$ where $\boldsymbol{\beta}_{(s)}$ is the value obtained at the s^{th} sampling.

For L large $\sigma^{2\,(\mathrm{PMV})} \sim \mathrm{E}_{\boldsymbol{\beta}}\left[\mathrm{var}\left(\boldsymbol{X}_{\boldsymbol{P_1}\times\boldsymbol{P_2}}\boldsymbol{\beta}|\mathrm{data}\right)\right]$

- $\left(\hat{\sigma}^2_{P_1\times P_2}\right)_{\mathrm{Vg}_3} = \widehat{\boldsymbol{\beta}}'\boldsymbol{V}_{\boldsymbol{P_1}\times\boldsymbol{P_2}}\widehat{\boldsymbol{\beta}} + \mathrm{trace}\{\boldsymbol{V}_{\boldsymbol{P_1}\times\boldsymbol{P_2}}\mathrm{var}(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y})\} + 0.25\boldsymbol{X}'_{\boldsymbol{P_1}\times\boldsymbol{P_2}}\mathrm{var}(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y})\boldsymbol{X}_{\boldsymbol{P_1}\times\boldsymbol{P_2}}$

$\left(\hat{\sigma}^2_{P_1\times P_2}\right)_{\mathrm{Vg}_3}$ considers the fact that the uncertainty of the estimation of marker effects is modulated for each cross by its own genomic constitution

Materials: cross value components prediction ability

Analytic formulae (genomic predictions) (2/2)

- Different **prediction models** were tested to **estimate marker effects**: BayesA, BayesB, BayesC, Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Ridge Regression BLUP (Vg1)

- We developed **2 alternative approaches** to compute **gametic variance** (tested only with Ridge Regression BLUP):
  - Taking into account the **marker effect estimation error** (algebraic version of the PMV of Lehermeier *et al.* 2017):

$$\left(\widehat{SD}^2_{P_1\times P_2}\right)_{Vg_2} = \boxed{\hat{\beta}' V_{P_1\times P_2}\hat{\beta}}_{Vg_1} + \text{trace}\left\{V_{P_1\times P_2}\,\text{var}(\beta|X,y)\right\}$$

$$\text{var}(\beta|X,y) = \hat{\sigma}^2_\beta\left(I - X'\left(XX' + I\frac{\hat{\sigma}^2_r}{\hat{\sigma}^2_\beta}\right)^{-1}X\right)$$

$\hat{\sigma}^2_\beta, \hat{\sigma}^2_r$ = markers and residuals estimated variances

X = vector of TP's genotypes

  - Considering that the uncertainty of the estimation of marker effects is **modulated by the genomic constitution of each parent**:

$$\left(\widehat{SD}^2_{P_1\times P_2}\right)_{Vg_3} = \boxed{\hat{\beta}' V_{P_1\times P_2}\hat{\beta} + \text{trace}\left\{V_{P_1\times P_2}\,\text{var}(\beta|X,y)\right\}}_{Vg_2} + 0.25X'_{P_1\times P_2}\,\text{var}(\beta|X,y)X_{P_1\times P_2}$$

$X_{P_1\times P_2}$ = vector of genotypes for the F1 of cross $P_1 \times P_2$

Results: Training Population (TP) quality

- **Genotypes**:
  - After quality control (MAF & call rate) and imputation of missing genotypes → **2,146** lines and **23,140** markers

- **Phenotypes** (BLUE values):
  - Means of correlations between years: **0.69** (yield), **0.78** (grain protein content), **0.87** (plant height), **0.91** (heading date)

Trait	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_e$	$\hat{\sigma}^2_{(g\times e)}$	$\hat{\sigma}^2_r$	$\widehat{\text{rep}}_{\text{plot}}$	$\widehat{\text{rep}}_{\text{design}}$	$\hat{h}^2_{\text{genomic}}$	Cross-validation accuracy
Yield	18.9	237.1	24.8	15.4	0.32	0.91	0.53	0.66 (±0.018)
Grain protein content	0.3	1.0	0.2	0.1	0.52	0.93	0.51	0.60 (±0.018)
Plant height	32.6	56.5	6.5	5.6	0.73	0.97	0.56	0.52 (±0.021)
Heading date	12.6	65.5	1.9	0.5	0.85	0.99	0.73	0.67 (±0.016)

$$\widehat{\text{rep}}_{\text{plot}} = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \hat{\sigma}^2_{(g\times e)} + \hat{\sigma}^2_r}$$

$$\widehat{\text{rep}}_{\text{design}} = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \frac{\hat{\sigma}^2_{(g\times e)}}{\text{nb\_env}} + \frac{\hat{\sigma}^2_r}{\text{nb\_rep}}}$$

Results: variance parameters in simulations

Trait	$\widehat{h}^2_{genomic}$
Yield	0.53
Grain protein content	0.51
Plant height	0.56
Heading date	0.73