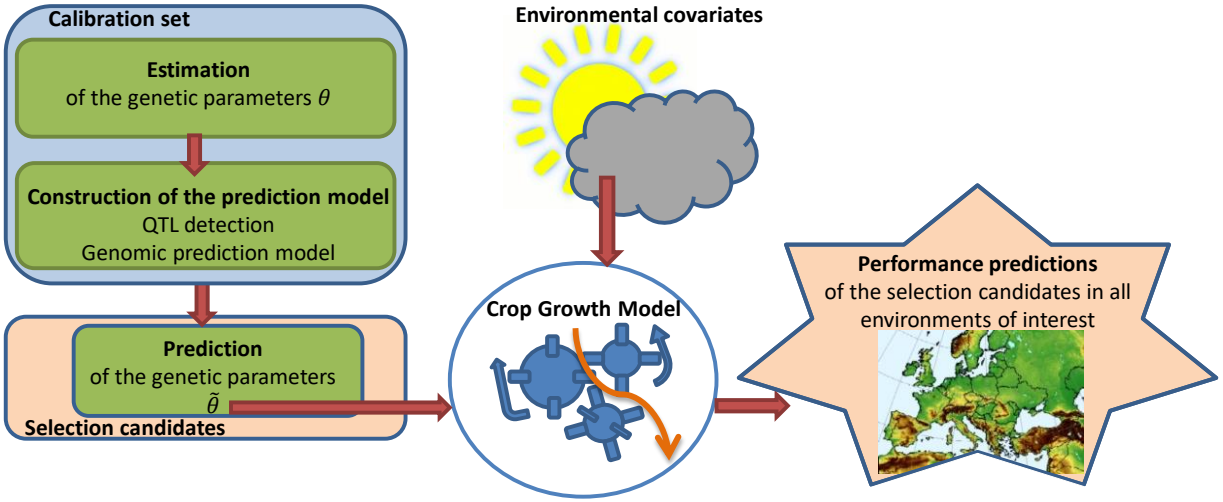
Optimization of multi-environment trials for genomic selection based on crop models

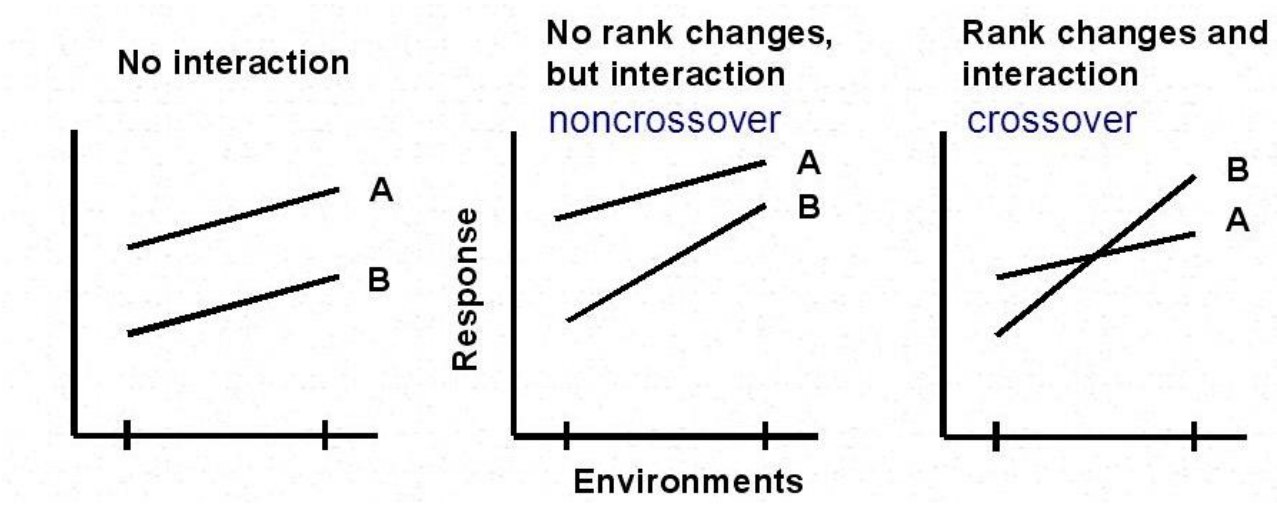
R. Rincent, E. Kuhn, H. Monod, V. Allard, J. Le Gouis



Introduction

- Genotype x Environment Interactions

Most crops submitted to **Genotype x Environment Interactions** (GEI)



Challenge : The best varieties depend on the environment

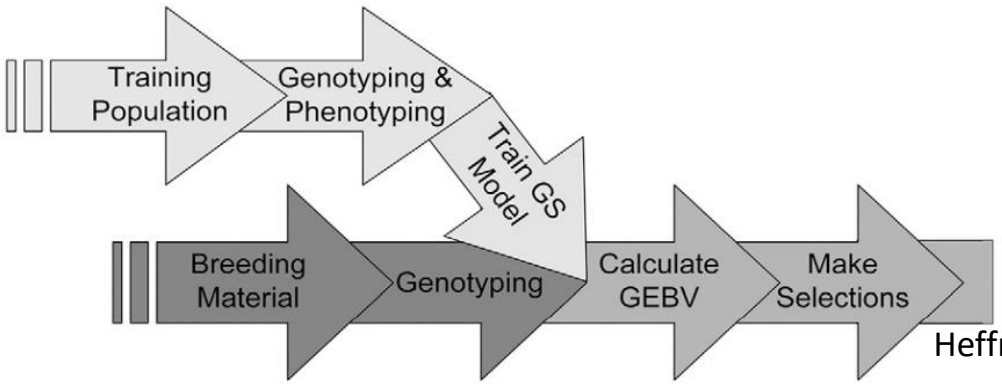
Opportunity : Local adaptation, tolerance to climate change...

Introduction

- Marker assisted prediction

It is not possible to evaluate all possible genotypes in all possible environments.

Molecular markers can be used to make **predictions** – Genomic Selection (GS).



Heffner *et al.* (2009)

GS would be of great interest if it could predict GEI (main limiting factor for the implementation of GS in plants).

Introduction

- Predicting GEI

- With environment specific marker effects (Burgueno et al. 2012, Schulz-Streeck et al. 2013).

- By including environmental covariates in the GS model (Heslot et al. 2014, Jarquin et al. 2014):

  Factorial regression on the environmental covariates (slope specific).

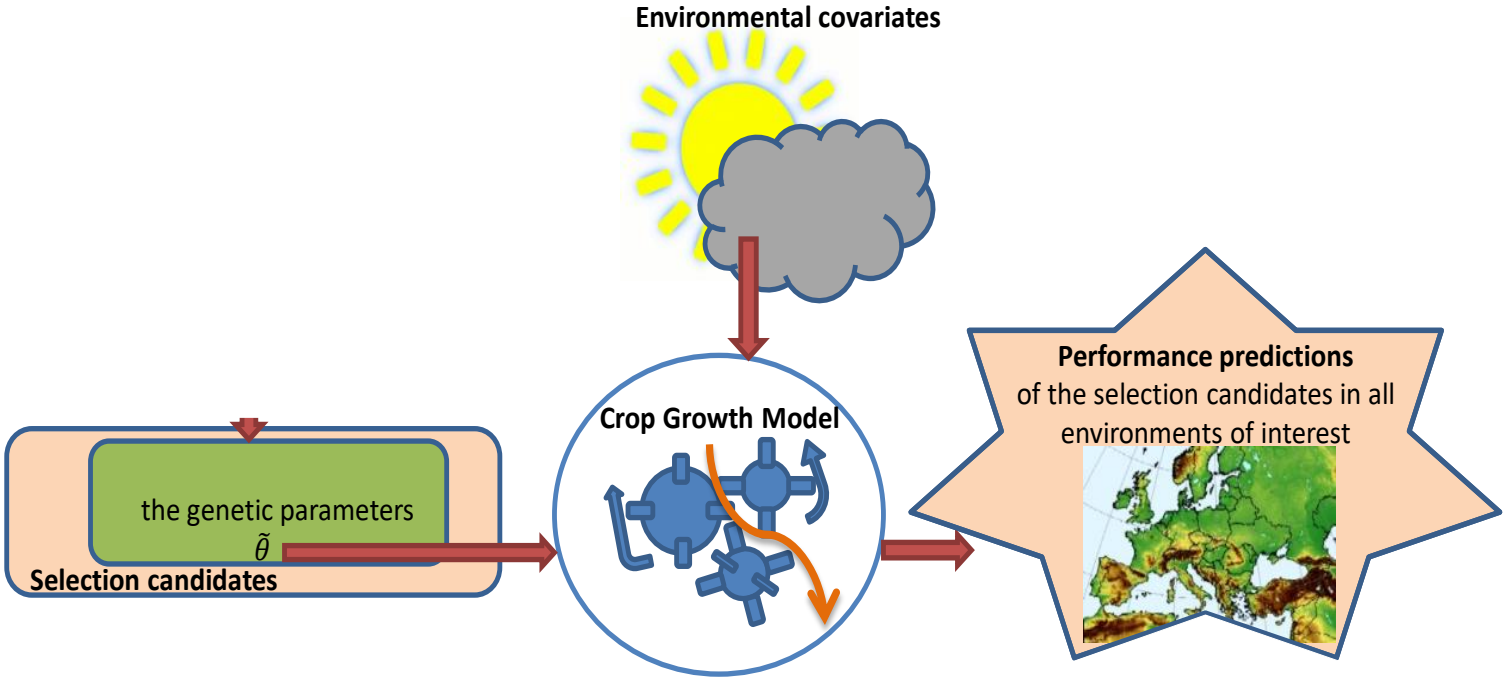
  Covariance between environments computed with the environmental covariates.

  -> Gains limited so far (at best gain of 15%)

- By combining Crop Growth Model (CGM) and genetic modelling (Reymond et al. 2003, Bogard et al. 2014, Cooper et al. 2016...)

Introduction

- Predicting GEI by combining CGM and genetics



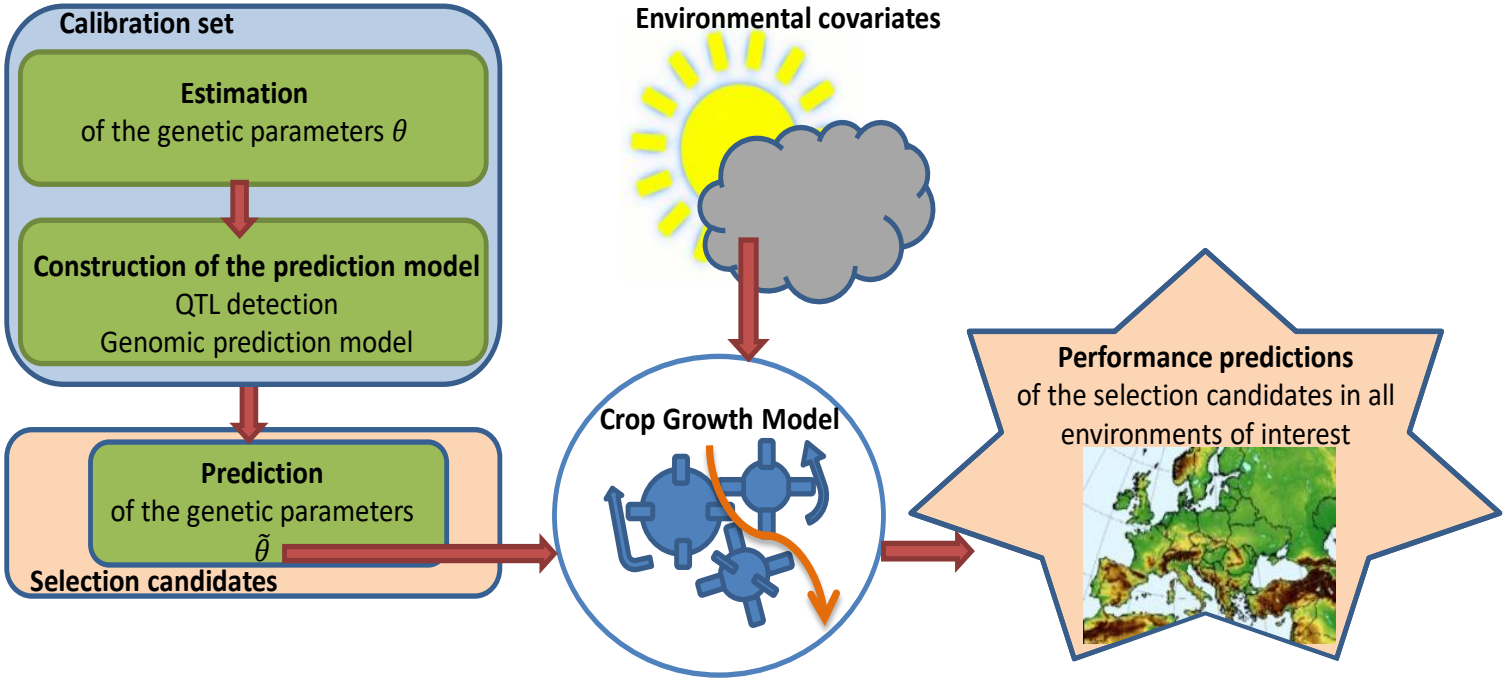
Example: sensitivity to photoperiod in a CGM simulating flowering time is a genetic parameter.

Varieties with different sensitivities will react differently to change in photoperiods -> The CGM will automatically generate GEI.

Literature: White and Hoogenboom 1996, Reymond 2003, Quilot 2005, Bogard et al. 2014, Technow et al. 2015...

Introduction

- Predicting GEI by combining CGM and genetics



Calibration set used to train the prediction model for the genetic parameters

No GEI for the genetic parameters θ (QTL and prediction formula stable for any environment).

Major difficulty : estimate the genetic parameters for the individuals composing the calibration set.

Introduction

- Estimate the genetic parameters by direct observations

INRA Montpellier

- In high-throughput phenotyping platforms
- In semi-controlled platforms
- In the fields

INRA Clermont - Ferrand

-> Not possible for many traits (costs, no high-throughput approach...)

Introduction

- Estimate the genetic parameters by statistical inference

Crop model

$$Y_{ijk} = f(\theta_i, E_j) + e_{ijk}$$

Phenotypes
(easily observable)

Genetic
parameters

Environmental covariates

Residuals

-> Estimate θ_i using Bayesian inference, non linear mixed models...

The precision of the estimation of θ_i is a key point because it will affect QTL detection power and the accuracy of GS and thus genetic progress.

In which environments should the varieties be observed (*Y*) to get the most precise estimate of θ_i ?

I/ OptiMET – Theory and practice

II/ Evaluating OptiMET with simulations (Wheat flowering time)

III/ Evaluating OptiMET with real data (Wheat flowering time)

Conclusions and perspectives

I/ OptiMET - Theory

- Inspired from Leube et al. (2012).

Basic principle: A Multi-Environment Trial (MET) in which the CGM generates distant outputs for distant genetic parameter vectors. In this MET we suppose that two different varieties will have different phenotypes, and so the statistical model will be able to distinguish between these two varieties.

For this we consider a huge finite number (*m*) of possible a priori values of genetic parameters. These *m* genetic parameter vectors can be chosen based on expert knowledge or on literature.

$$OptiMET^d = \sum_{u,v=1}^{m} (\mathrm{dist}(\theta_u, \theta_v) \times W_{uv}^d).$$

Sum on each pair of parameter vector

Distance btw the 2 parameter vectors

Linked to the likelihood of θ_u given the synthetic noise-free data $f(\theta_v, E_j)$

We want MET with low OptiMET

I/ OptiMET – In practice

To compute OptiMET, you need :

- *m* representative parameter vectors (at least bounds of each parameters known)

- To run your CGM on the candidate environments (likelihood)
  - Climate of past years
  - Statistics (average) over past years
  - Climate simulator (complex covariate, climate change...)

- Algorithm to find the MET minimizing OptiMET: exchange algorithm.

I/ OptiMET – Case study

- Target trait: Wheat flowering time

- CGM: Sirius (Jamieson et al. 1998)
  - 3 genetic parameters: VAI (vernalization), SLDL (photoperiod) and Phyllocron
  - Env. Covariate: daily temperatures and day length

<u>OptiMET computation</u>:

- *m* parameter vectors defined to sample homogeneously in the space defined by the boundaries of the 3 parameters: VAI: 0 to 0.01, SLDL: 0 to 1, Phyll: form 80 to 120. *m* = 10 x 10 x 10 = 1000 (virtual genotypes).

- Env. Covariates approximated by the local daily average over 12 past years.

- The potential environments proposed to OptiMET are defined by a combination of a sowing date and location.

II/ Evaluating OptiMET with simulations

Simulation of the genetic parameters

A

For each of the three parameters:
- Sample 25 SNP, simulate effects.
- Compute the parameter values of each of the 370 varieties by multipliying the genotypes by the QTL effects.

Split dataset

Sampling of the MET

1

***J*=156 sowing date x location combinations** characterised by daily temperature and daylength averaged over 12 years

Sample METs composed of *Z*=4 sowing date x location combinations:
- Randomly
- By minimizing OptiMET
- Expert MET

Simulation of the phenotypes

B

Calibration set (*I*=100 varieties)

For each environment (year x location x sowing date combinations = 12 x 156 = 1872 environments)
- Use Sirius to generate flowering time for each variety, using as inputs the parameter values generated in A and the daily temperatures and daylength of the specific year at the specific location.
- Add a random error.

Validation set (270 varieties)

C

For 28 independent environments:
- Use Sirius to generate flowering time for each variety, using as inputs the parameter values generated in A and the daily temperatures and daylength of the specific year at the specific location.
- Add a random error.

Estimate parameters

2

Use the Bayesian algorithm to estimate the parameters of each of the *I*=100 varieties for each MET and each year using the phenotypes simulated in B.

Evaluate the METs

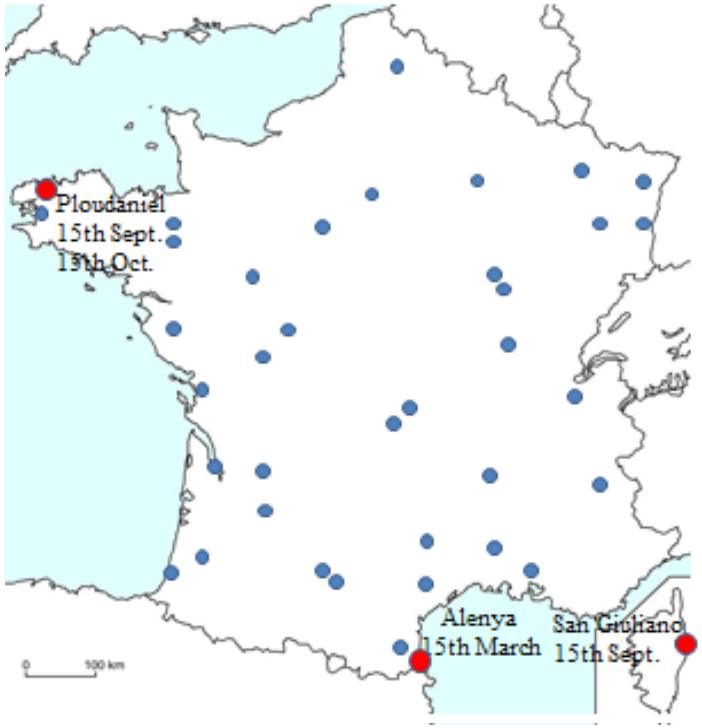
3

For each combination year x MET, use the parameter estimates (from 2) and the simulated values (from A) to:
- Compute root mean square error.
- Compute detection power of association tests.
- Compute prediction accuracy of the genetic parameters of the individuals in validation.
- Compute prediction accuracy of the flowering time of the individuals in validation in the independent environments (C).

II/ Evaluating OptiMET with simulations

Potential environments

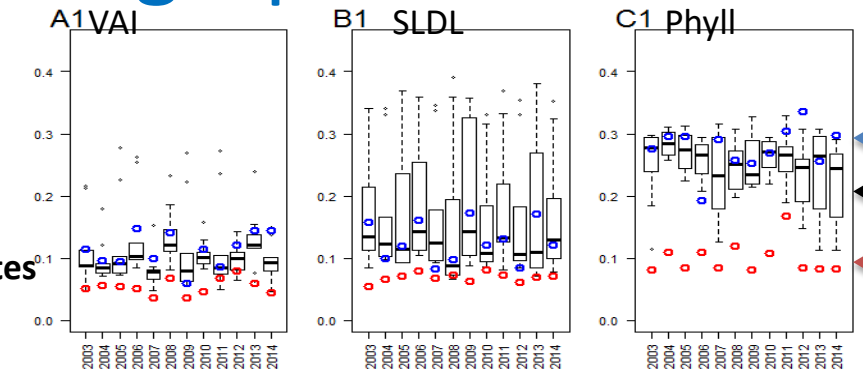
OptiMET MET



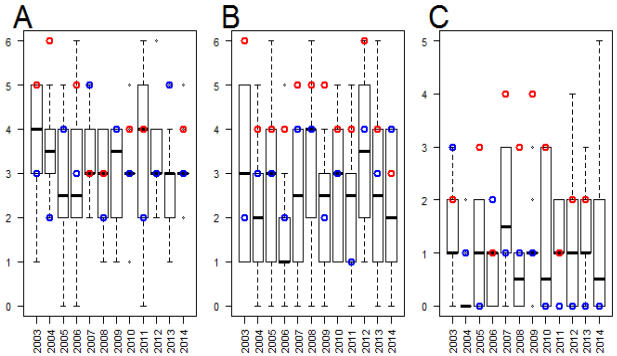
At each location, 4 possible sowing dates:
15th of Sept., Oct., Nov. or March

II/ Evaluating OptiMET with simulations

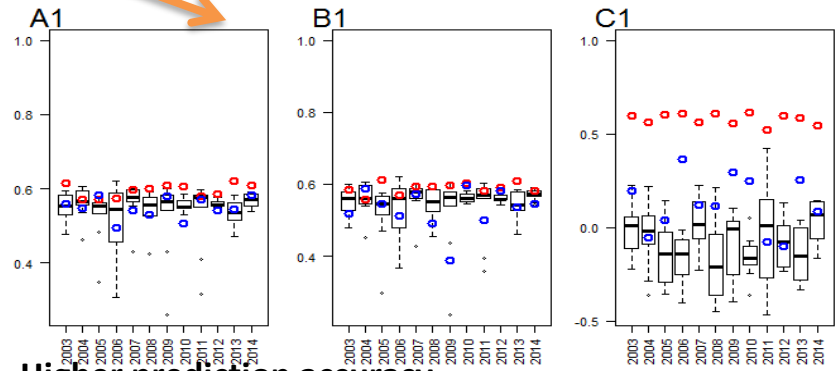
Lower NRMSE of the parameter estimates



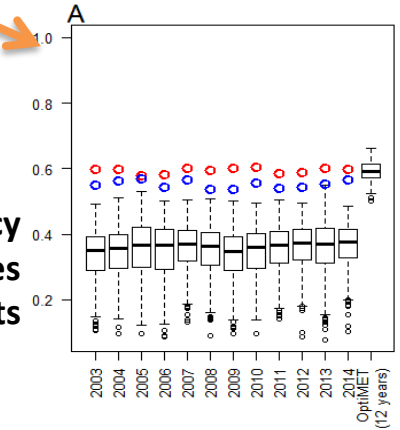
Expert MET
Random MET
OptiMET MET



Higher QTL detection power



Higher prediction accuracy of the parameters of the test set



Higher prediction accuracy of flowering time of independent varieties in independent environments

III/ Evaluating OptiMET with real data

Dataset: 110 varieties, 26 environments with various sowing dates:

2008/2009		2009/2010	
sowing date	**location**	**sowing date**	**location**
17/10/2008	Allonnes	23/10/2009	Mons-en-Chaussée
20/10/2008	Mons-en-Chaussée	28/10/2009	Clermont-Ferrand
20/10/2008	Le Moulon	28/10/2009	Louville
22/10/2008	Auchy	29/10/2009	Clermont-Ferrand
23/10/2008	Villiers-le-Bâcle	29/10/2009	Maule
29/10/2008	Montroy	29/10/2009	Caussade
12/11/2008	Clermont-Ferrand	30/10/2009	La Minière
20/11/2008	Clermont-Ferrand	25/11/2009	Villiers-le-Bâcle
12/12/2008	La Miniere	14/12/2009	Clermont-Ferrand
24/12/2008	Mons-en-Chaussée	15/12/2009	Clermont-Ferrand
05/01/2009	Clermont-Ferrand	23/02/2010	Clermont-Ferrand
25/02/2009	Clermont-Ferrand	04/03/2010	Mons-en-Chaussée
16/03/2009	Mons-en-Chaussée		
14/04/2009	Mons-en-Chaussée		

Sampling strategies for MET of size 4, 6 or 8:

- OptiMET
- random
- reasoned (various sowing dates)

We suppose that the climate of these specific years is unknown. Again, we use the average of temperature and day length over 12 other years.

III/ Evaluating OptiMET with real data

Criterion used to compare the efficiency of the different MET to estimate the 3 genetic parameters:

The reference values θ_{is}^* are the parameter estimates obtained with the full dataset (26 envts)

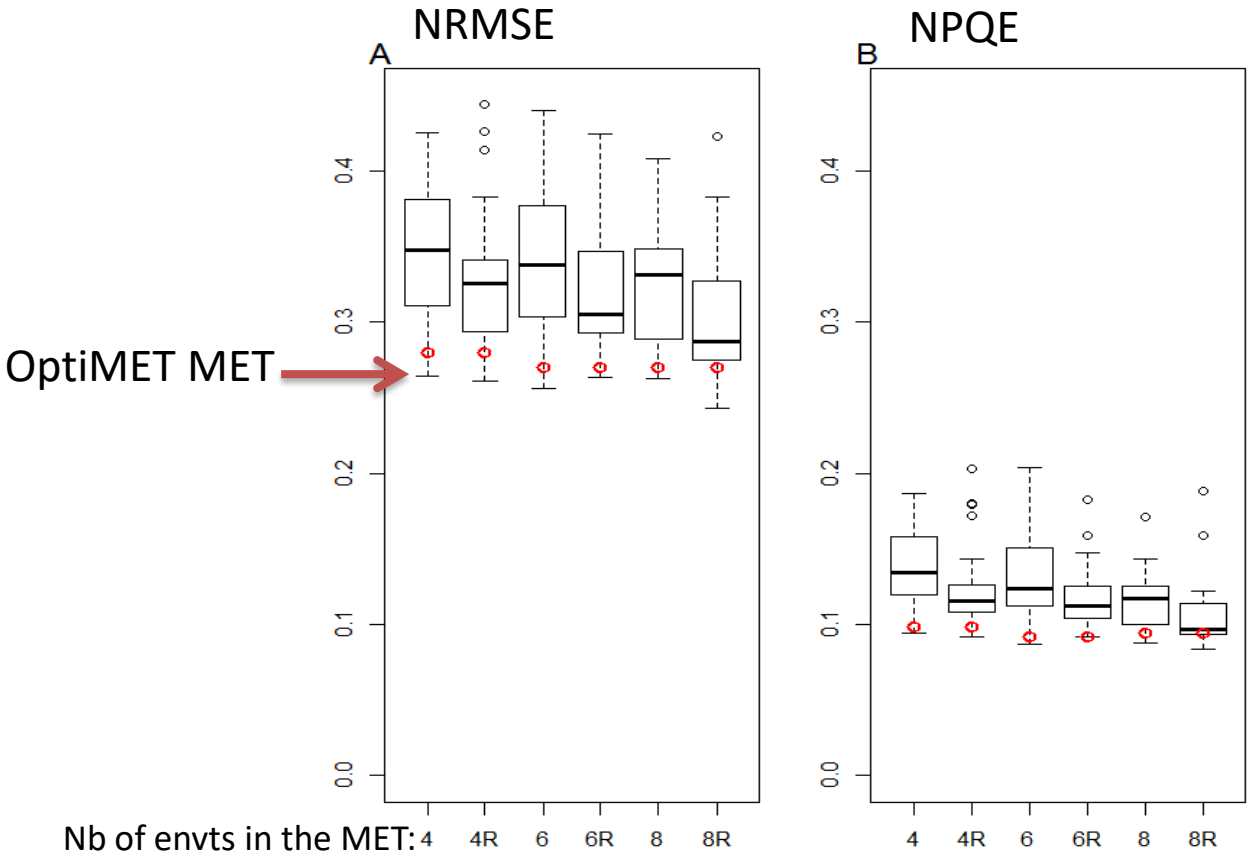
Normalized Root Mean Square Error:

$$NRMSE^*(\theta_s) = \left[\frac{1}{I}\sum_{i=1}^{I}\left(\frac{\hat{\theta}_{is} - \theta_{is}^*}{M_s^* - m_s^*}\right)^2\right]^{1/2}$$

Normalized Posterior Quadratic Error:

$$NPQE(\theta_s) = \frac{1}{I}\sum_{i=1}^{I}\mathbb{E}\left[\left(\frac{\pi(\theta_{is}/y) - \theta_{is}^*}{M_s^* - m_s^*}\right)^2\right]$$

III/ Evaluating OptiMET with real data



OptiMET MET composed of 4 environments performed better than the average of the reasoned MET composed of 8 environments.

Conclusions and perspectives

- Conclusions:

OptiMET was able to increase the precision of the parameter estimates, resulting in an increase of QTL detection power and GS-CGM prediction accuracy.

The optimized MET performed better than expert/reasoned MET.

- Limits:

OptiMET requires a reliable CGM (in the environments of interest).

OptiMET cannot be used to define the optimal size of the MET.

- Perspectives:

Apply to other traits/CGM. High-throughput phenotyping platforms.

Adapt OptiMET to compare MET of different sizes.

Acknowledgment

Genotypic data
E. Paux- BreedWheat

Dataset FSOV Precocite
FX Oury
M Rousset

Metaprogramme SelGen

Thank you for your attention

$L_{uv}^{d} = \frac{1}{(4\pi\sigma_e^2)^{n_y/2}} exp\left(-\frac{1}{4\sigma_e^2}\left(\Delta_{uv}^{d}\right)^t \Delta_{uv}^{d}\right)$,

where n_y is the number of observations for a given variety ($n_y = Z \times K$), and $\Delta_{uv}^{d} = (f\left(\theta_u, E_j\right) - f\left(\theta_v, E_j\right), j \in \mathrm{J_d})$, The quantity L_{uv}^{d} corresponds to the likelihood of the parameter vector θ_u given the synthetic noise-free data ($f\left(\theta_v, E_j\right), j \in \mathrm{J_d}$) (for more details see Leube et al. 2012, Appendix B).

The matrix (L_{uv}^{d}) is normalized by computing the weight matrix $W_{uv}^{d} = \frac{L_{uv}^{d}}{\sum_u L_{uv}^{d}}$.

We define the value of the criterion OptiMET for a given MET *d* by:

$OptiMET^{d} = \sum_{u,v=1}^{m}(\mathrm{dist}(\theta_u, \theta_v) \times W_{uv}^{d})$.