Prédiction en sélection génomique

Charles-Elie Rabier

ISEM, Institut des Sciences de l'Evolution de Montpellier
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique

Brigitte Mangin

LIPM, Laboratoire des Interactions Plantes et Microorganismes, Toulouse

Plan

Plan

Introduction

Sélection Génomique

- Objectif = prédire les valeurs génétiques des candidats à la sélection
- Plus besoin de détecter les QTLs ! ! !

Nouvelles technologies de sequençage

- Milliers de marqueurs disponibles $\Rightarrow$ tous les QTLs fortement corrélés (fort Déséquilibre de liaison) avec au moins un marqueur

Outils statistiques

- Tous les marqueurs analysés simultanément $\Rightarrow$ Régression sur tout le génome

Problème de haute dimension

$K >> n$

Genomic Selection

Training population

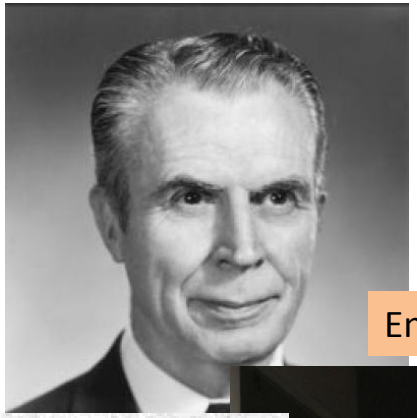
$$Y = \mu + \underbrace{f(M, \theta)}_{g} + \varepsilon$$ ← Model

$$\hat{\theta}$$ ← Estimation method

Test population (selection candidates)

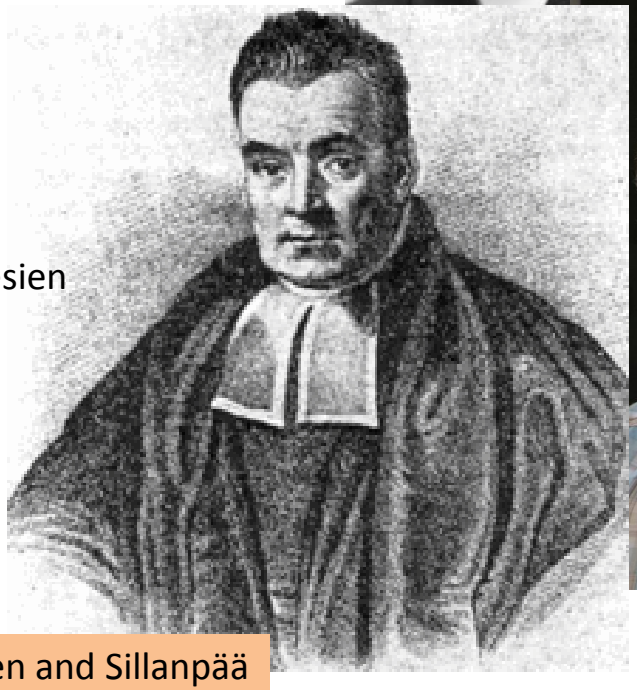
$$\underbrace{f(M, \hat{\theta})}_{\hat{g}}$$ ← Prediction

Cadres statistiques

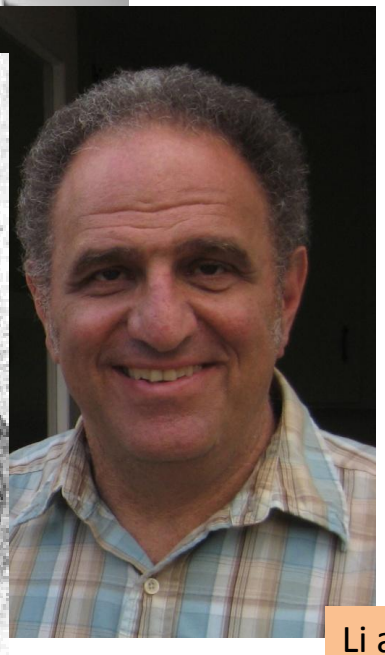


C.R. Henderson
Modèle mixte

Endelman



T. Bayes
Modèle bayésien



R. Tibshirabi
Régression pénalisée

Li and Sillanpää

Kärkkäinen and Sillanpää

Pour les bayésiens, un petit changement dans les lois = une nouvelle méthode

Nom	Référence	*A priori* effet size	Indicator	Hyperprior	Estimation
BayesA	Meuwissen et al. (2001)	Student	No	No	MCMC
BayesB	Meuwissen et al. (2001)	Student	Yes	No	MCMC
BAS/BayesA	Xu (2003)	Student	No	No	MCMC
SSVS	Yi et al. (2003)	Student	Yes	No	MCMC
	Yi and Xu (2008)	Student et Laplace	No	Yes	MCMC
fBayesB	Meuwissen et al. (2009)	Laplace	Yes	No	EM
BL	de los Campos et al. (2009)	Laplace	No	Yes	MCMC
R/BhGLM	Yi and Banerjee (2009)	Student	No	No	EM
BayesC	Verbyla et al. (2009)	Student	Yes	No	MCMC
BL	Xu (2010)	Laplace	No	No	EM
wBSR	Hayashi and Iwata (2010)	Student	Yes	No	EM
BAL	Sun et al. (2010)	Laplace	No	Yes	MCMC
IAL	Sun et al. (2010)	Laplace	No	Yes	ECM
emBayesB	Shepherd et al. (2010)	Laplace	Yes	Yes	EM
EBL	Mutshinda and Sillanpää(2010)	Laplace	No	Yes	MCMC
BayesCπ	Habier et al. (2011)	Student	Yes	Yes	MCMC

Classement

- En général, le classement des méthodes est Bayes $\geq$ Régressions pénalisées $>$ Modèle mixte
- mais les méthodes ont moins d'influence que la densité de marquage, la taille de la population Training, l'héritabilité
- ou que la "distance" entre la population Training et la population TEST

Cet exposé : focus sur GBLUP, RRBLUP, Ridge (Pénalité L2)

Modèle causal vs modèle de prédiction

Objectif : Prédiction en grande dimension

Modèle causal (Q QTLs)

$\theta^\star$ effets QTLs, $M^\star$ matrice de génotypes aux QTLs,
n individus d'entrainement

$$Y = M^\star \theta^\star + e$$

où $Y = (Y_1, ..., Y_n)'$, $\theta^\star = (\theta_1^\star, ..., \theta_Q^\star)'$, $e \sim N(0, \sigma_e^2 I_n)$

Modèle Bayésien de prédiction (K marqueurs, où $K >> n$)

θ effets marqueurs, M matrice de génotypes aux marqueurs des "Training"

$$Y = M\theta + \varepsilon$$

où $Y = (Y_1, ..., Y_n)'$, $\theta = (\theta_1, ..., \theta_K)' \sim N(0, \sigma_\theta^2 I_K)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$, $\varepsilon_j \perp\!\!\!\perp \theta_k$

Apprentissage $\hat{\theta} = \mathbb{E}(\theta \mid Y) = M' (MM' + \lambda I_n)^{-1} Y = (M'M + \lambda I_K)^{-1} M' Y$

Régression Ridge (Pénalité L2) avec $\lambda = \sigma_\varepsilon^2 / \sigma_\theta^2$

Introduction d'un individu TEST

- Soit un individu TEST numéroté $n+1$

$$Y_{n+1} = M^{\star}_{n+1}\ \theta^{\star} + e_{n+1} \quad \text{où} \quad e_{n+1} \sim N(0, \sigma^2_e)$$
$$\text{et} \quad M^{\star}_{n+1} \text{ génotypes aux QTLs de l'individu } n+1$$

- Prédiction de la valeur phénotypique

$$\hat{Y}_{n+1} = M_{n+1}\hat{\theta} = M_{n+1}M'\left(MM' + \lambda I_n\right)^{-1} Y$$
$$= M_{n+1}\left(M'M + \lambda I_K\right)^{-1} M'Y$$

$\Rightarrow$ Critère d'accuracy

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{n+1},\ Y_{n+1}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{n+1}\right)\text{Var}\left(Y_{n+1}\right)}} \quad \text{avec } M_{n+1} \text{ aléatoire et } M \text{ fixe}$$

Plan

Formule générale pour l'accuracy

Théorème (R., Barre, ..., Mangin, Plos One 2016)

Conditionnellement à M et $M^\star$,

$$\rho = \frac{\theta^{\star\prime}\ \mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right) M' V^{-1} M^\star \theta^\star}{\left\{\sigma_e^2 \mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right) + \theta^{\star\prime} M^{\star\prime} V^{-1} M \mathrm{Var}\left(M'_{n+1}\right) M' V^{-1} M^\star \theta^\star\right\}^{1/2} \Omega^{1/2}}$$

où $V = MM' + \lambda I_n$ et $\Omega = \mathrm{Var}\left(M_{n+1}^\star \theta^\star\right) + \sigma_e^2$

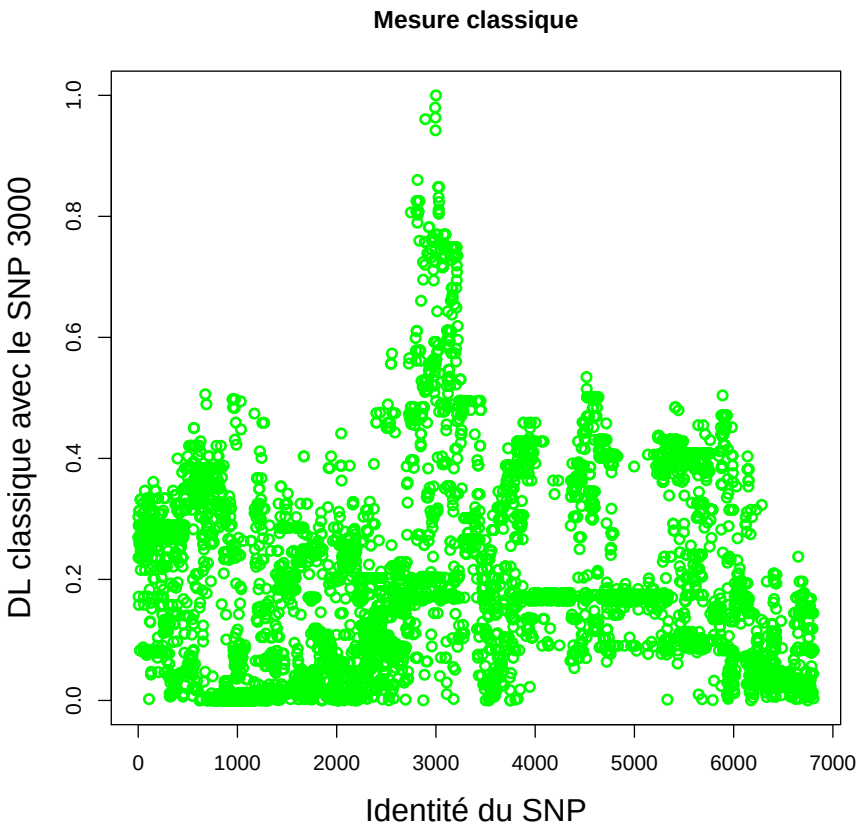
Facteurs agissant sur l'accuracy :

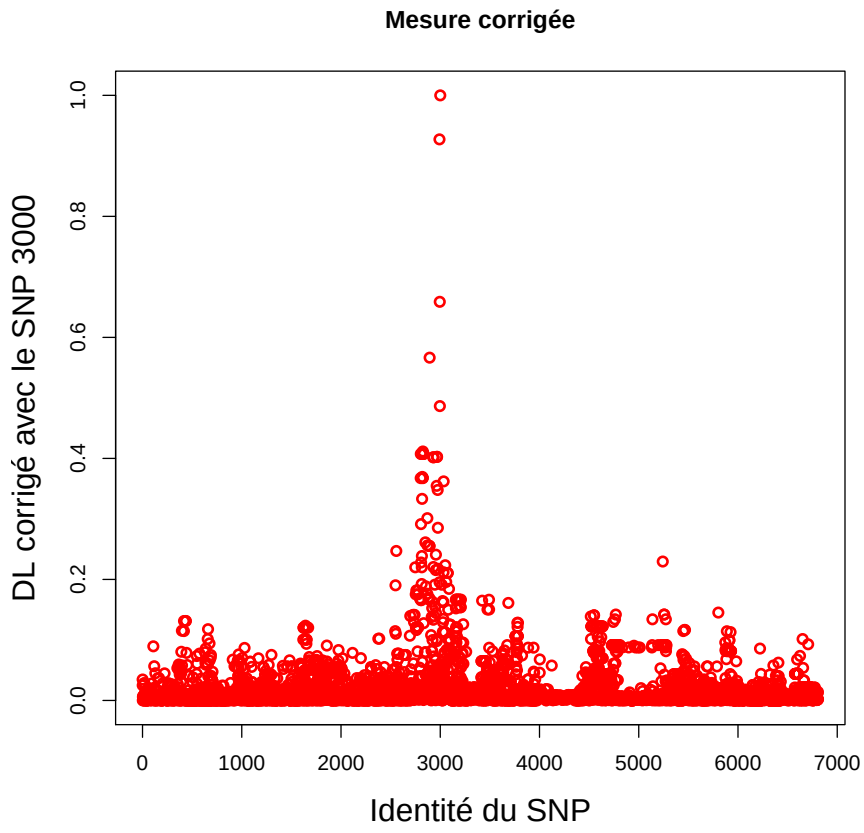
- Colonne q de $M' V^{-1} M^\star$: DL (utilisant la métrique V) entre chaque marqueur et le QTL q
- $\mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right)$: similarité entre TRN et TEST
- $\mathrm{Var}\left(M'_{n+1}\right)$: matrice de covariance de taille $K \times K$, contenant l'ensemble des DL classiques du TEST

Déséquilibre de liaison (tranche 100000-200000 SNPs du chromosome 6 du riz)

50 variétés, 6806 SNPs après filtrage 100000 SNPs ...

Mangin et al. (Heredity, 2012)





Descriptif des simulations

- Population panmictique évoluant pendant 30, 50 or 70 generations
- Taille de la population : 100 TESTS +
  - $n = 500$ Trainings
  - $n = 1000$ Trainings
- Espèce haploïde
- Recombination modélisée selon Haldane (processus de Poisson d'intensité 1)
- Longueur du génome $L = 1M$
- Marqueurs équidistants
- Densité de marqueurs : 100 SNPs, 1000 SNPs, 5000 SNPs ou 10000 SNPs
- Nombre de QTLs
  - 2 QTLs à 3cM et 80cM avec effet $+1$ et -2
  - 100 QTLs équidistants avec effet $+0.15$

Deux configurations étudiées ($n = 500$)

- 2 QTLs à 3cM et 80cM d'effets respectifs 1 et -2

nombre générations	héritabilité (h^2)	h
30	0.5415	0.7359
50	0.5310	0.7287
70	0.5059	0.7113

- 100 QTLs tous les centi Morgan, d'effets 0.15

nombre générations	héritabilité (h^2)	h
30	0.7450	0.8631
50	0.6488	0.8055
70	0.5716	0.7560

Deux configurations étudiées ($n = 1000$)

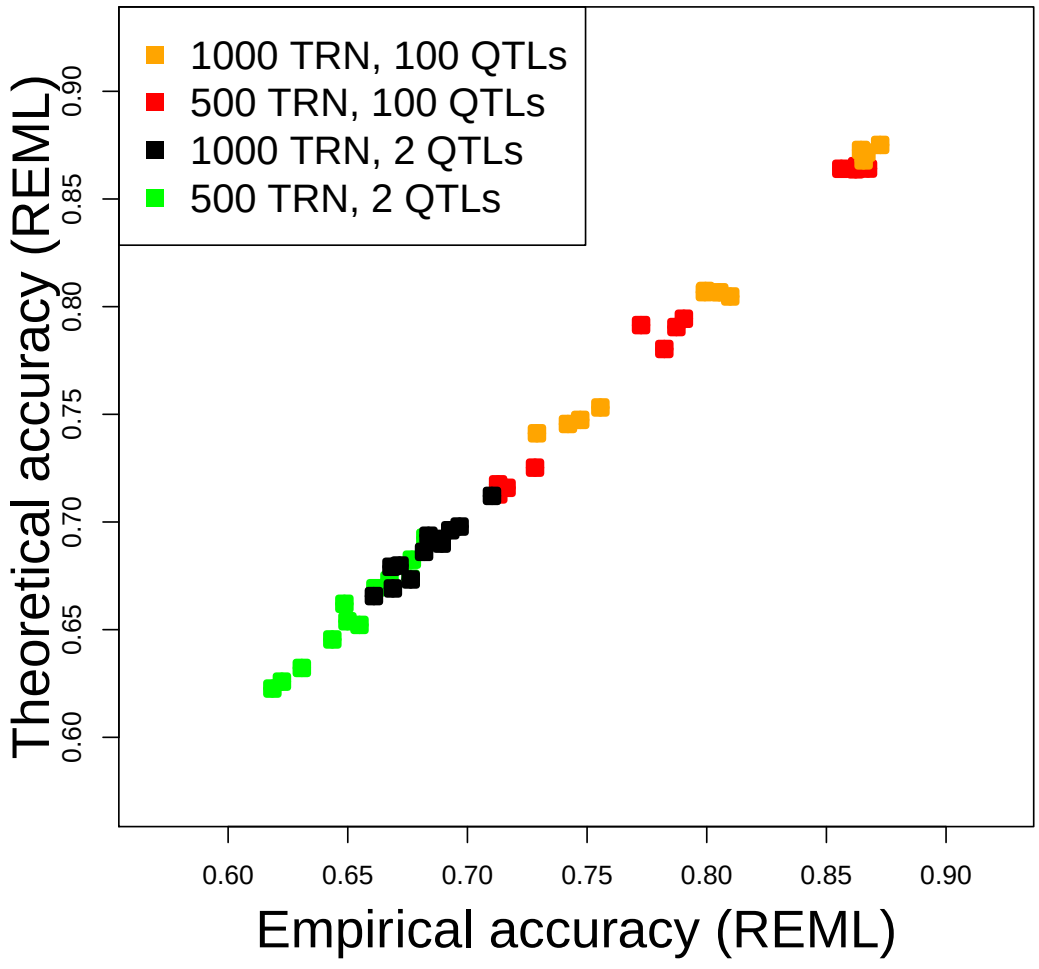
- 2 QTLs à 3cM et 80cM d'effets respectifs 1 et -2

nombre générations	héritabilité (h^2)	h
30	0.5481	0.7402
50	0.5473	0.7397
70	0.5324	0.7296

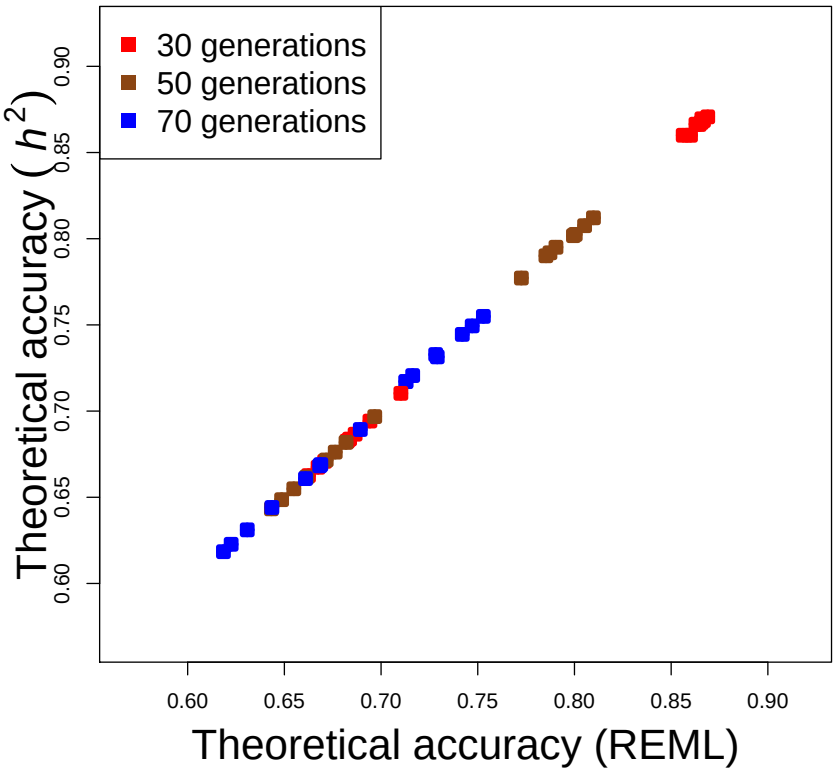
- 100 QTLs tous les centi Morgan, d'effets 0.15

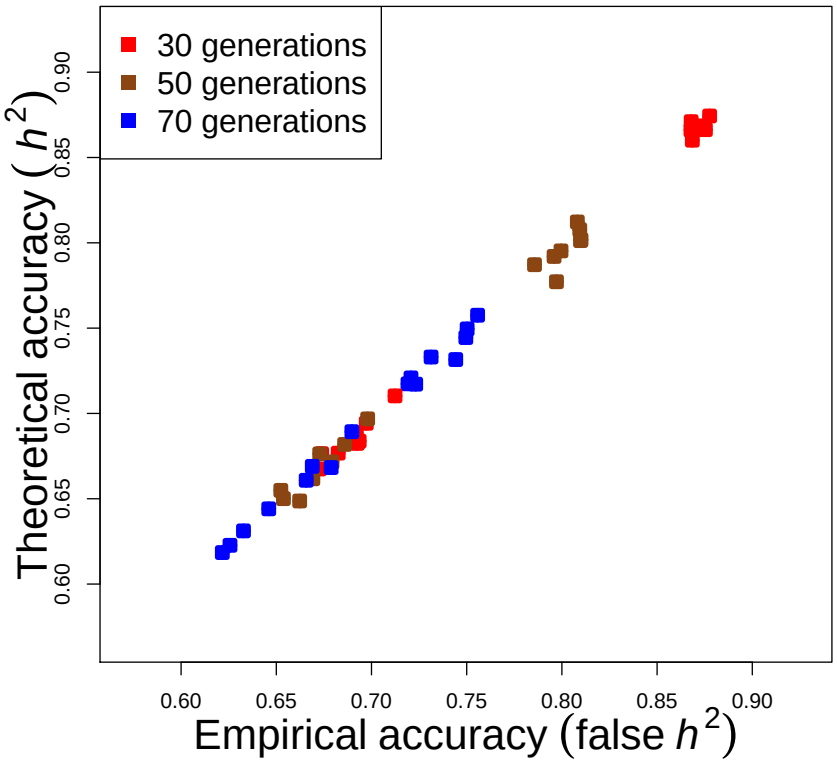
nombre générations	héritabilité (h^2)	h
30	0.7695	0.8772
50	0.6812	0.8253
70	0.5784	0.7605

Les QTLs sont situés sur quelques marqueurs (DL parfait)

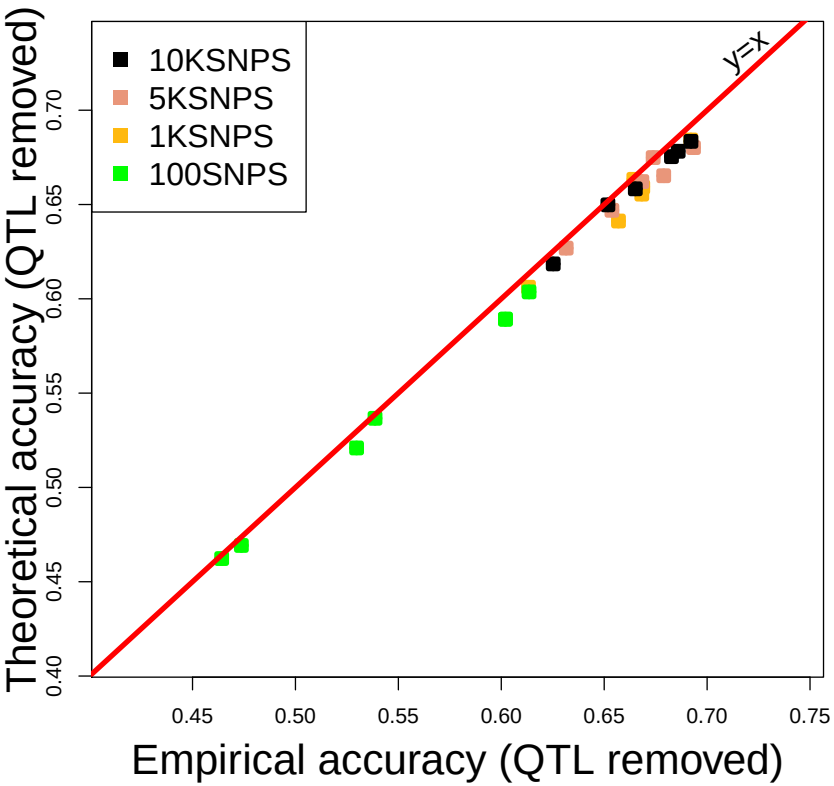


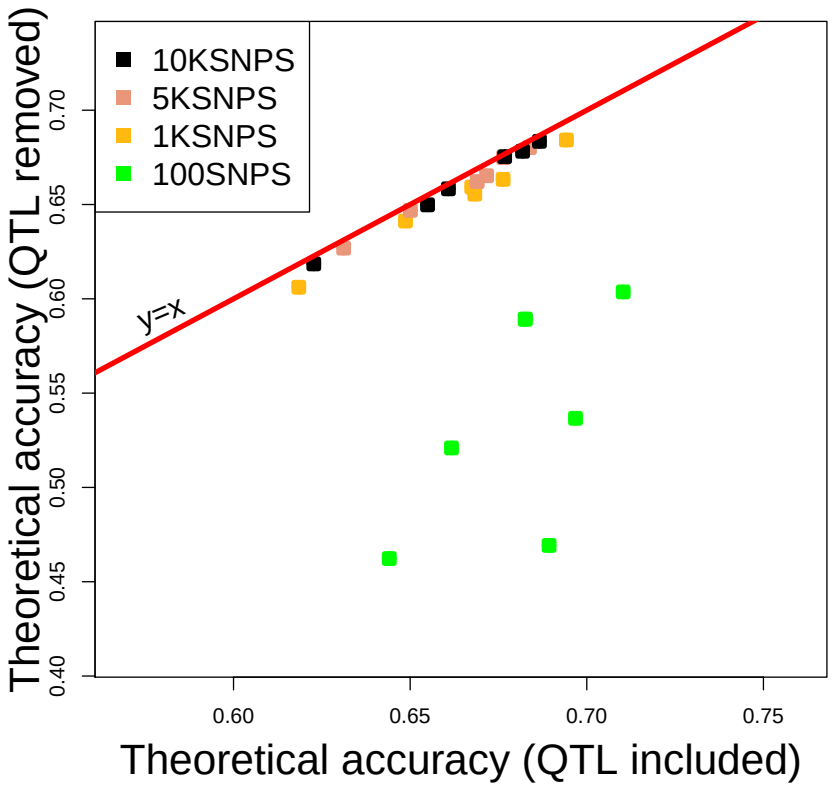
Plusieurs façon d'estimer le paramètre de régularisation λ





Les QTLs ne sont pas situés sur les marqueurs (DL imparfait)





Introduction d'une proxy

Notre Formule théorique depend des positions des QTLs (inconnues)

$\Rightarrow$ Une nouvelle proxy pour l'accuracy (DL parfait)

$$\rho = h \sqrt{\frac{h^2/(1-h^2)}{\mathbb{E}\left(\|M_{n+1}M'V^{-1}\|^2\right) + \frac{h^2}{1-h^2}}}$$

où h^2 est l'héritabilité du trait chez les TESTS

Notre proxy ne dépend pas des paramètres liés aux QTLs !!!

Hypothèses

- Densité de marqueurs suffisante pour que chaque QTL soit en DL complet avec un des SNPs
- Nombre de générations suffisamment important afin de casser les blocs de DL

Proxies existantes dans la littérature

La plupart son construites à partir de Daetwyler, PloS One 2008

Contexte de l'étude de Daetwyler :

- positions des Q QTLs connues
- QTLs indépendants
- effets QTLs inconnus
- $Q < n$

$\Rightarrow$ Considération du génome causal
$\Rightarrow$ Y_{n+1} estimé par Moindre Carrés (OLS)

$$\hat{Y}_{n+1}^{OLS} = M_{n+1}^{\star}(M^{\star\prime}M^{\star})^{-1}M^{\star\prime}Y$$

$\Rightarrow$ Formule de Daetwyler (2008)

$$\rho = \frac{h\ \sqrt{h^2/(1-h^2)}}{\sqrt{\frac{Q}{n}\ +\ \frac{h^2}{1-h^2}}}$$

Proxies existantes dans la littérature

Méthodes existantes = remplacer Q par le nombre de loci indépendants M_e

Cas $M_e > n$ peut même être traité en raison de l'indépendance !
(Daetwyler et al., Genetics 2010)

2 façons d'estimer M_e

- nombre de tests indépendants en étude d'association (Li and Ji 2005)
- taille de population efficace N_e
  (Goddard 2009, Goddard Hayes and Meuwissen 2011)

$$M_{e1} = \frac{2N_eL}{\log(4N_el)} \text{ , } M_{e2} = \frac{2N_eL}{\log(2N_el)} \text{ , } M_{e3} = \frac{2N_eL}{\log(N_el)}$$

L longueur du génome, l longueur moyenne du chromosome

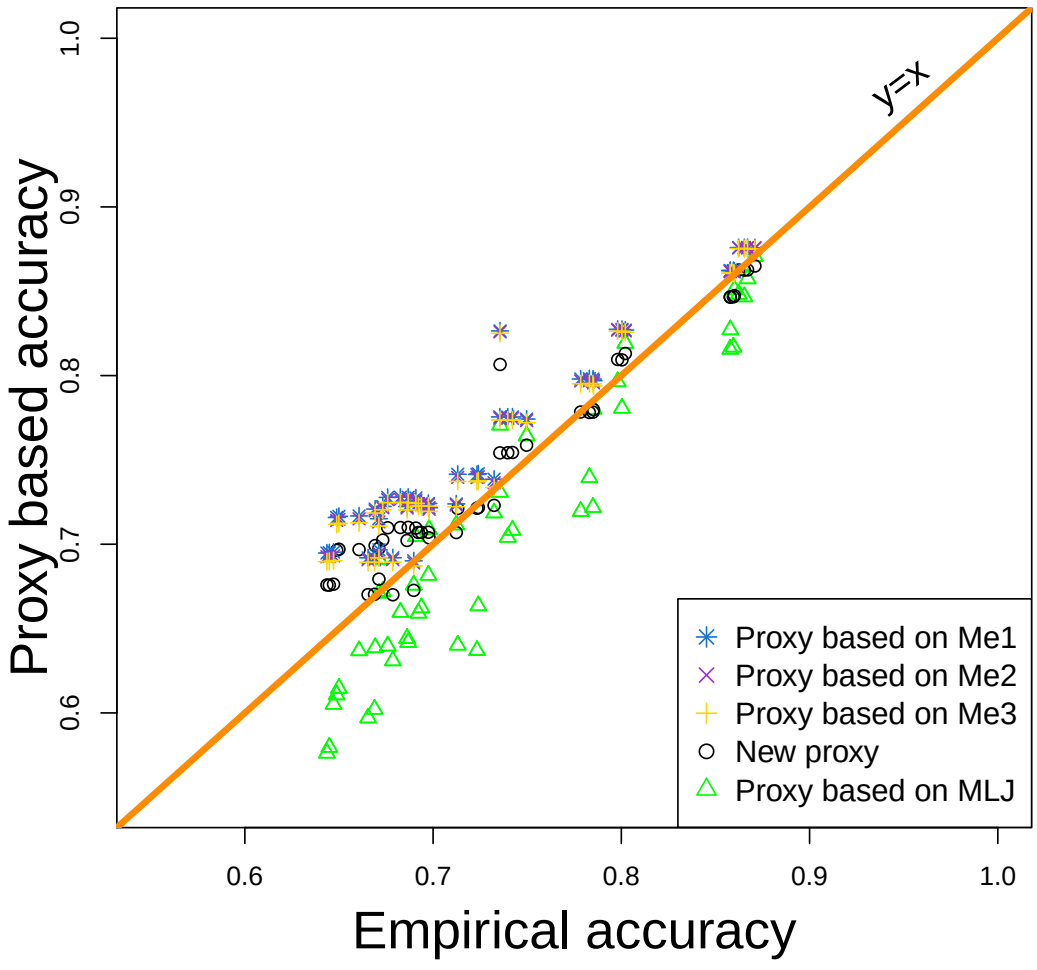
Erreur quadratique moyenne (relativement à l'accuracy empirique) correspondant à 5 proxies

- Mêmes simulations que précédemment (LD parfait)
- N_e estimé à partir de la formule de Hill and Weirr (Theoretical Population Biology, 1988)
- Moyenne sur toutes les architectures

Accuracy théorique	4.204×10^{-5}
Notre proxy	4.628×10^{-4}
Proxy basée sur M_{e1}	1.228×10^{-3}
Proxy basée sur M_{e2}	1.157×10^{-3}
Proxy basée sur M_{e3}	1.0669×10^{-3}
Proxy basée sur M_{LJ}	1.474×10^{-3}

Les proxy existantes ne gèrent pas la grande dimension correctement !

Erreur quadratique moyenne (relativement à l'accuracy empirique) correspondant à 5 proxies



Notre formule générale permet de retrouver la formule de Daetwyler

- Les positions des QTLs sont connus
- On effectue une régression Ridge sur ces causaux

Théorème

Conditionnellement à $M^\star$,

$$\rho = \frac{\theta^{\star\prime}\ \mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}^{\star}\right) M^{\star\prime} V^{\star -1} M^{\star} \theta^{\star}}{\left\{\sigma_e^2 \mathbb{E}\left(\left\|M_{n+1}^{\star} M^{\star\prime} V^{\star -1}\right\|^2\right) + \theta^{\star\prime} M^{\star\prime} V^{\star -1} M^{\star}\, Var\left(M_{n+1}^{\star\prime}\right) M^{\star\prime} V^{\star -1} M^{\star} \theta^{\star}\right\}^{1/2} \Omega}$$

où $V = M^\star M^{\star\prime} + \lambda I_n$ et $\Omega = \mathrm{Var}\left(M_{n+1}^{\star}\theta^{\star}\right) + \sigma_e^2$

Notre formule générale permet de retrouver la formule de Daetwyler

Si on suppose

- indépendance des loci
- $\lambda = 0$
- $M^{\star}_{1,q}, \ldots, M^{\star}_{n,q}, M^{\star}_{n+1,q}$ iid

alors, on obtient

$$\theta^{\star\prime}\, \mathbb{E}\left(M^{\star\prime}_{n+1} M^{\star}_{n+1}\right) M^{\star\prime} V^{\star -1} M^{\star} \theta^{\star} = \mathrm{Var}\left(M^{\star}_{n+1}\theta^{\star}\right) \quad , \quad \mathbb{E}\left(\left\| M^{\star}_{n+1} M^{\star\prime} V^{\star -1} \right\|^2\right) \approx \frac{Q}{n}$$

$$\theta^{\star\prime} M^{\star\prime} V^{\star -1} M^{\star} \mathrm{Var}\left(M^{\star\prime}_{n+1}\right) M^{\star\prime} V^{\star -1} M^{\star} \theta^{\star} = \mathrm{Var}\left(M^{\star}_{n+1}\theta^{\star}\right)$$

De plus, on sait que

$$\mathrm{Var}\left(M^{\star}_{n+1}\theta^{\star}\right) / \sigma^2_e = h^2/(1-h^2)$$

D'où la formule de Daetwyler (2008)

$$\rho = \frac{h \sqrt{h^2/(1-h^2)}}{\sqrt{\frac{Q}{n} + \frac{h^2}{1-h^2}}}$$

Un software intégrant différentes formules pour l'accuracy en selection génomique

ShinyGPAS par Morota (Université de Nebraska Lincoln)
disponible sur https ://chikudaisei.shinyapps.io/shinygpas/

Formules implémentées :

- Daetwyler et al. (Plos One 2008, Genetics 2010)
- Goddard et al. (Genetica 2009, Journal Of Animal Breeding And Genetics 2011)
- Nous :) (Plos One, 2016)
- de los Campos et al. (Plos Genetics, 2013)

$$\rho \leq h\sqrt{\{1-(1-b)^2\}\,h^2}$$

b : coeff de régression entre relations génomiques aux marqueurs et relations génomiques aux génomes causaux

- Karaman et al. (Plos One, 2016), $\tilde{h}^2$ proportion de variance expliquée par les marqueurs

$$\rho = \tilde{h}\sqrt{\frac{n\tilde{h}^2}{n\tilde{h}^2 + M_e}}$$

Comparaison entre les Me et notre suggestion

- Trainings et TESTS basés sur le même nombre de générations
- $n = 500$ Trainings, 100 TESTS
- 100 replicats, matrice M fixe
- 100 QTLs

Nb générations	Nb SNPs	Nous	M_{LJ}	M_{e1}	M_{e2}	M_{e3}
30	100	71.85	50.13	11.96	14.02	16.92
	5,000	74.83	177.45	11.76	13.78	16.66
50	100	66.94	59.73	17.71	20.43	24.12
	5,000	70.66	233.49	18.52	21.32	25.12
70	100	58.49	64.39	22.68	25.93	30.27
	5,000	66.68	258.63	22.46	25.70	30.02

Nous= $n\mathbb{E}\left(\left\|M_{n+1}M'V^{-1}\right\|^2\right)$, $M_{e1} = \frac{2N_eL}{\log(4N_el)}$, $M_{e2} = \frac{2N_eL}{\log(2N_el)}$, $M_{e3} = \frac{2N_eL}{\log(N_el)}$.

Nos collaborations sur CROPDL

- INRA Lusignan : Ray-grass anglais (Philippe Barre)
- INRA Clermont-Ferrand : Blé tendre (Gilles Charmet, François Balfourier, Delphine Ly)
- SupAgro Montpellier : Blé dur (Muriel Tavaud, Jacques David)

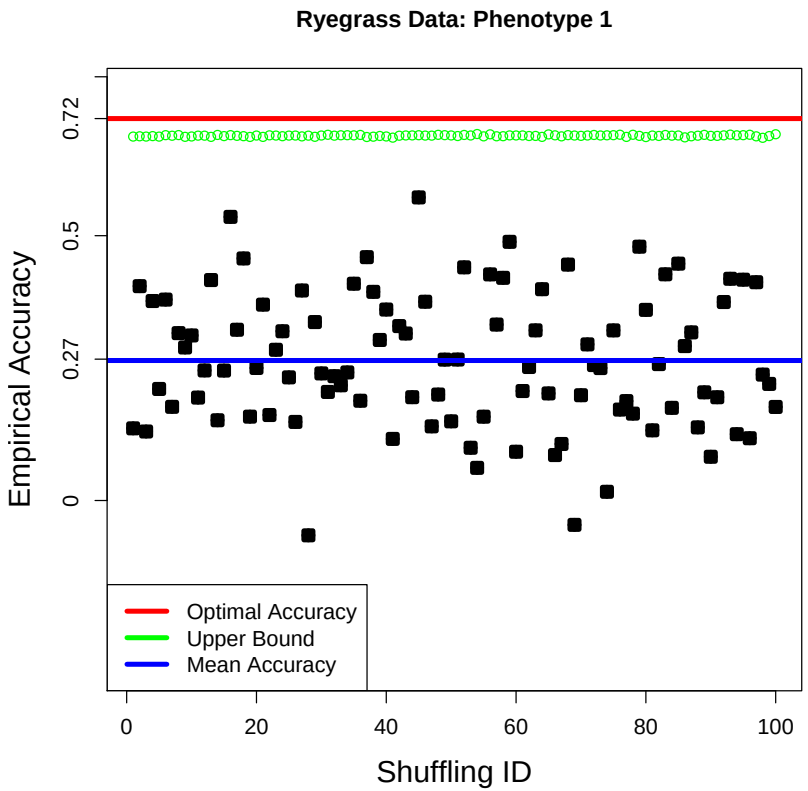
Données réelles Ray-grass anglais (INRA Lusignan)

Population Unlike

- 450 plantes (3 Générations), 8 fondateurs
- Génome du Ray-grass : 2.7 10^9 paires de bases
- Population très apparentées pour obtenir des blocs de DL
- 24957 SNPs
- 2 phénotypes corrélés de 0.35
  - longueur de la feuille entière
  - vitesse de repousse

Peut-on faire de la sélection génomique ? ? ?

Accuracy pour la longueur de la feuille entière



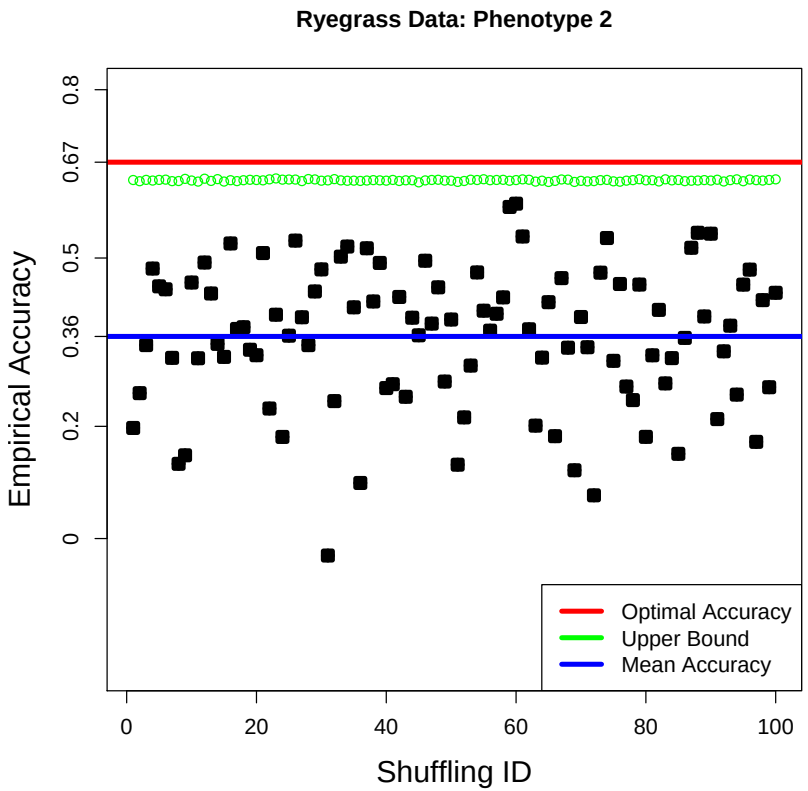
Données

- $h^2 = 0.52$
- 405 plantes
- 24957 SNPS

Analyse statistique

- Ridge avec un λ basé sur l'héritabilité
- Estimation de l'accuracy sur 100 jeux de données obtenus par permutation (360 Trainings/ 45 TESTS)

Accuracy pour la vitesse de repousse



Données

- $h^2 = 0.45$
- 367 plantes
- 24957 SNPs

Analyse statistique

- Ridge avec un λ basé sur l'héritabilité
- Estimation de l'accuracy sur 100 jeux de données obtenus par permutation (322 Trainings/ 45 TESTS)

Les différents M_e sur ce jeux de données

Paramètres :

- $N_e = 8$
- $L = 8.11$
- $l = L/7 = 1.16$

M_{e1}	M_{e2}	M_{e3}	M_{LJ}	Notre terme
35.92	44.45	58.29	12514	22.71

Données blé tendre

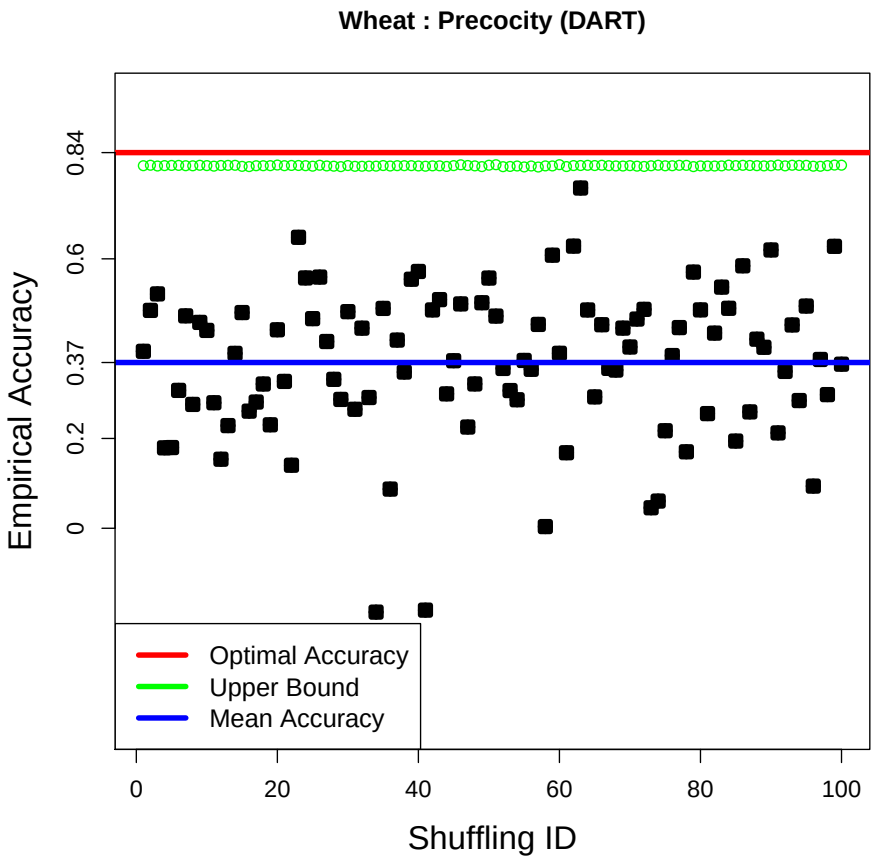
Données

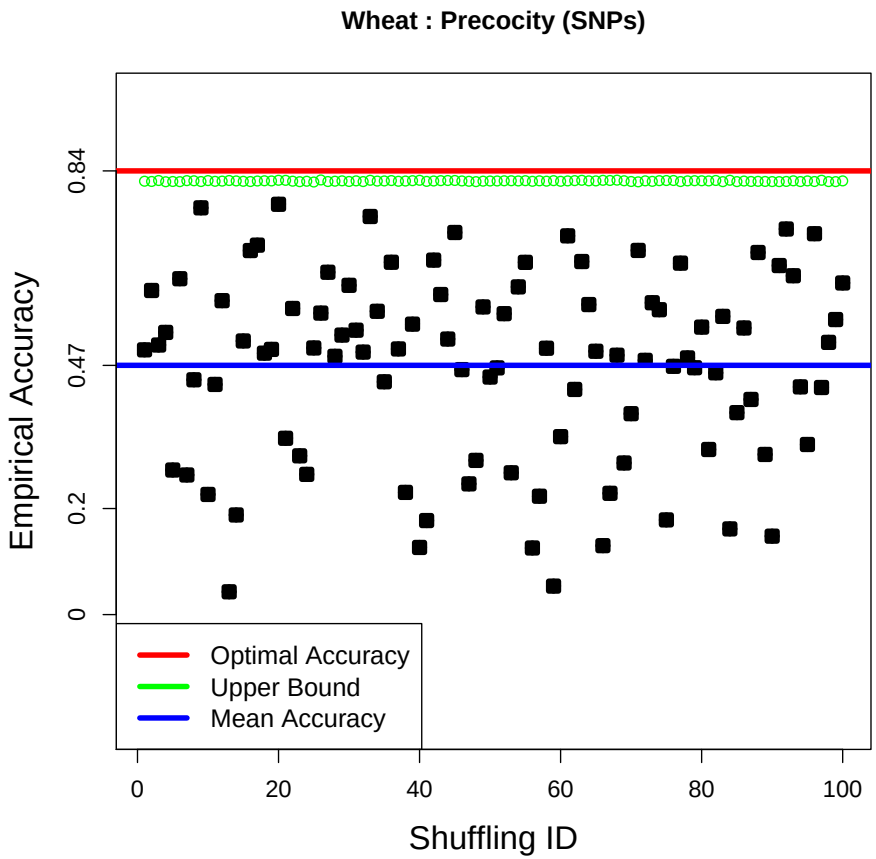
- 2236 DARTS / 122499 SNPs
- 2 caractères : Précocité / Rendement
- 339 blés (DART)/ 344 blés (SNPs)
- $h^2 = 0.70$
- BLUE : phénotypes corrigés par année, lieu, et type d'essai

Analyse statistique

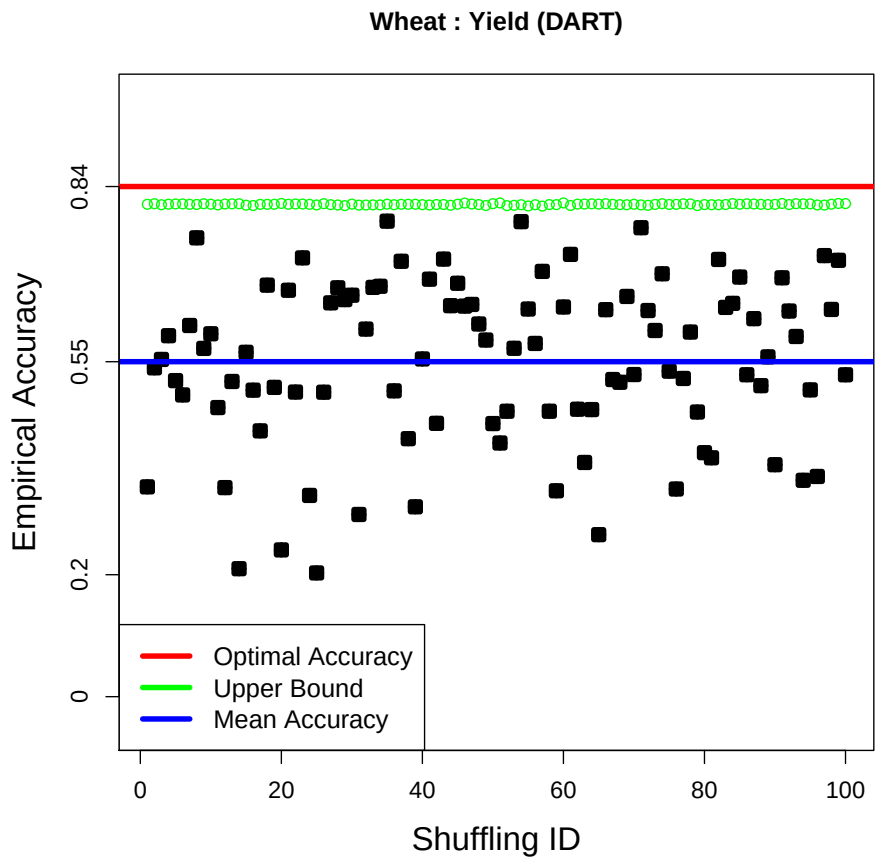
- Ridge avec un λ basé sur l'héritabilité
- Estimation de l'accuracy sur 100 jeux de données obtenus par permutation (90% Trainings / 10% TESTS)

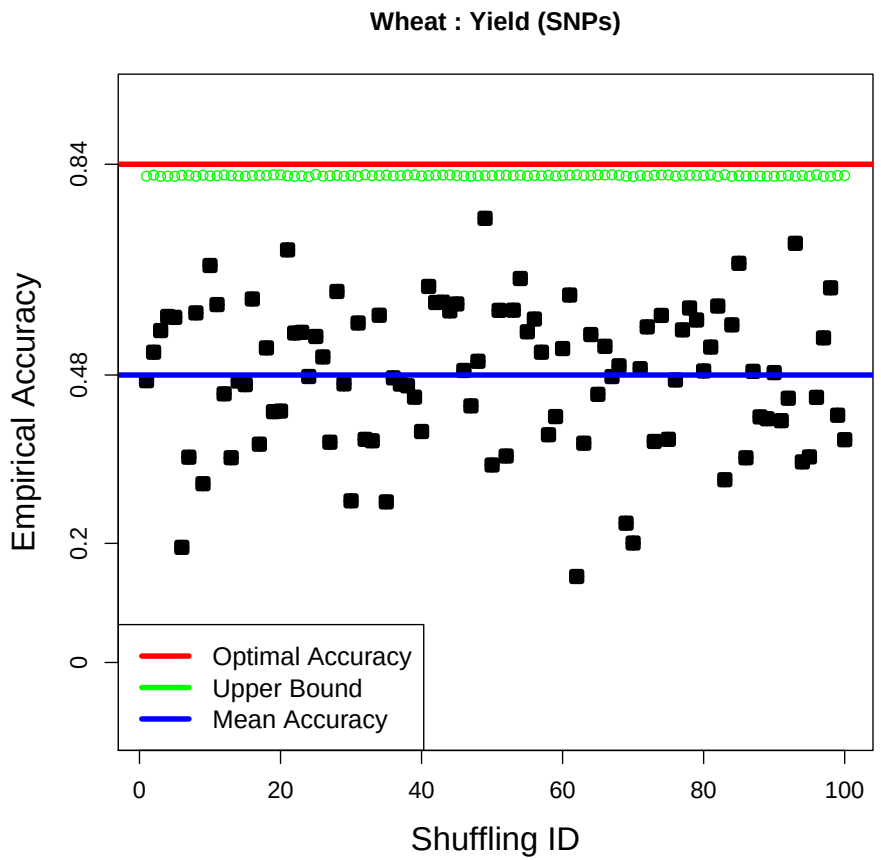
Accuracy pour la précocité (DARTS + SNPs)





Accuracy pour le rendement (DARTS + SNPs)





Peut-on capter la variablité due aux différents tirages ?

De nouvelles proxies utilisant les effets QTLs

En injectant l'estimation des causaux dans notre formule ...

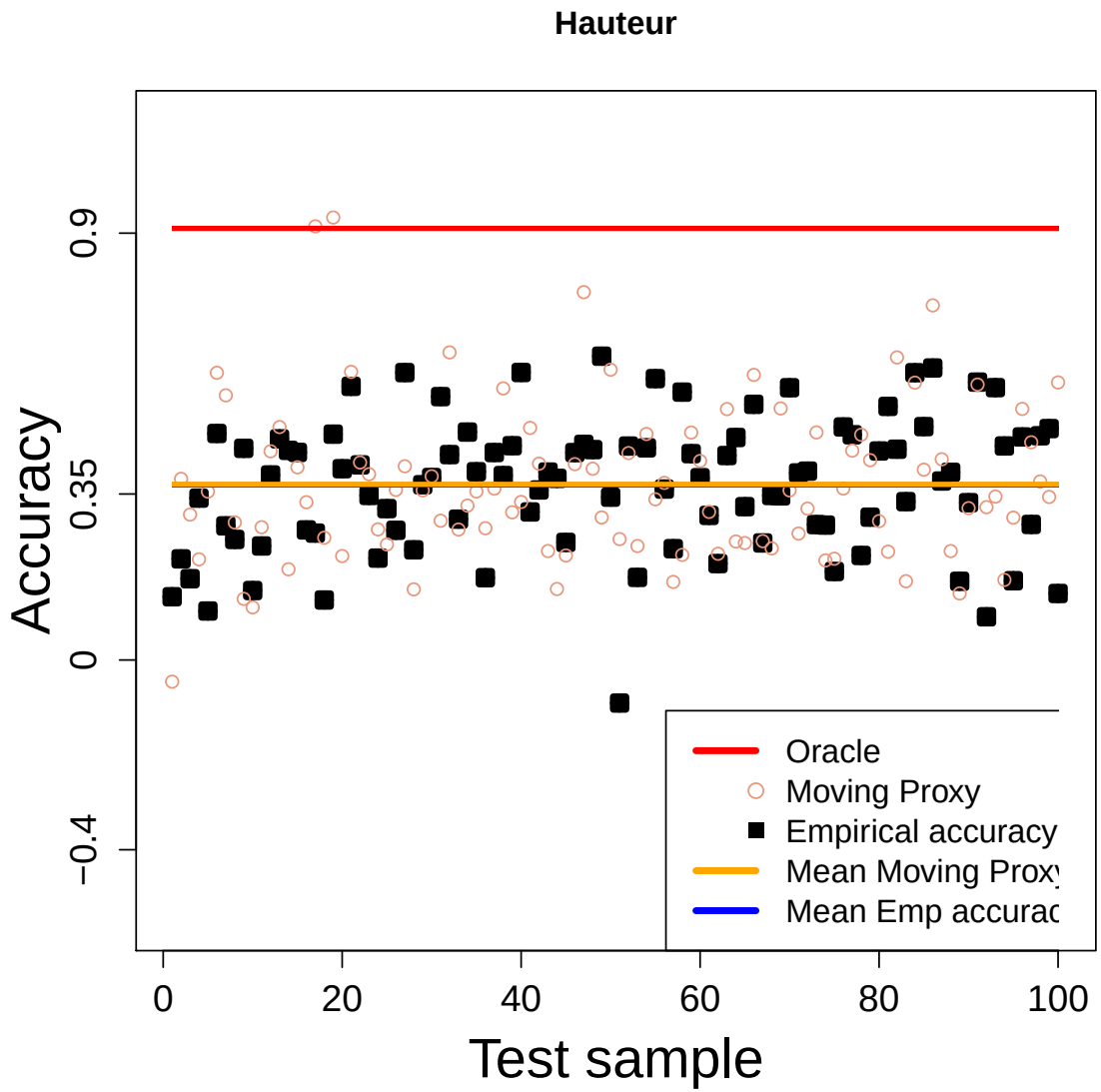
$$\hat{\rho} = \frac{\hat{\theta}^{\star\prime}\ \mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right) M' V^{-1} \hat{M}^{\star} \hat{\theta}^{\star}}{\left\{\sigma_e^2 \mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right) + \hat{\theta}^{\star\prime} \hat{M}^{\star\prime} V^{-1} M \mathrm{Var}\left(M'_{n+1}\right) M' V^{-1} \hat{M}^{\star} \hat{\theta}^{\star}\right\}^{1/2} \Omega^{1/2}}$$

où $V = MM' + \lambda I_n$ et $\Omega = \mathrm{Var}\left(M_{n+1}^{\star} \theta^{\star}\right) + \sigma_e^2$

$\mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right)$, $\mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right)$ et $\mathrm{Var}\left(M'_{n+1}\right)$ peuvent être estimés à l'aide :

- uniquement des Trainings $\rightarrow$ estimation de l'accuracy avant le génotypage des TESTS ... $\hat{\rho}_{before}$
- à la fois des Trainings et TESTS $\rightarrow$ estimation de l'accuracy après le génotypage des TESTS ... $\hat{\rho}_{after}$

La proxy varie désormais en fonction des tirages !!!



Régressions pénalisées

LASSO (Tibshirani, JRSSB 1996)

$$\hat{\theta}_{LASSO} = \text{argmin} \, \| Y - X\theta \|^2 + \lambda \sum_{k=1}^{K} |\theta_k|$$

Adaptative LASSO (Zou, JASA 2006)

$$\hat{\theta}_{ADLASSO} = \text{argmin} \, \| Y - X\theta \|^2 + \lambda \sum_{k=1}^{K} w_k \, |\theta_k|$$

Group LASSO (Yuan and Lin, JRSSB 2006)

sparsité par groupes, L nombre de groupes
n_ℓ nombre de marqueurs dans le groupe ℓ

$$\hat{\theta}_{GPLASSO} = \text{argmin} \left\| Y - \sum_{\ell=1}^{L} X_\ell \vec{\theta}_\ell \right\|^2 + \lambda \sum_{\ell=1}^{L} \sqrt{n_\ell} \left\| \vec{\theta}_\ell \right\|^2$$

Accuracy moyenne (100 simulations)

- LD parfait
- 100 QTLs avec le même effet $+0.15$
- 2 QTLs à 3cM et 80cM avec effets $+1$ and -2
- Training et TESTS basés sur 50 générations
- 1000 marqueurs, 500 Trainings, 100 TESTS

Méthode	100 QTLs	2 QTLs
Emp. Acc.	0.8143 (0.0036)	0.6683 (0.0053)
$\hat{\rho}_{before}(\theta^{\star})$	0.8086 (0.0001)	0.6597 (0.0001)
$\hat{\rho}_{before}(\hat{\theta}^{\star}_{LASSO})$	0.7635 (0.0017)	0.5354 (0.0031)
$\hat{\rho}_{before}(\hat{\theta}^{\star}_{ADLASSO})$	0.7627 (0.0012)	0.6488 (0.0017)
$\hat{\rho}_{before}(\hat{\theta}^{\star}_{GPLASSO})$	0.7581 (0.0014)	0.5471 (0.0029)
$\hat{\rho}_{after}(\theta^{\star})$	0.8045 (0.0019)	0.6576 (0.0021)
$\hat{\rho}_{after}(\hat{\theta}^{\star}_{LASSO})$	0.7502 (0.0026)	0.5347 (0.0037)
$\hat{\rho}_{after}(\hat{\theta}^{\star}_{ADLASSO})$	0.7489 (0.0023)	0.6454 (0.0027)
$\hat{\rho}_{after}(\hat{\theta}^{\star}_{GPLASSO})$	0.7479 (0.0024)	0.5495 (0.0034)

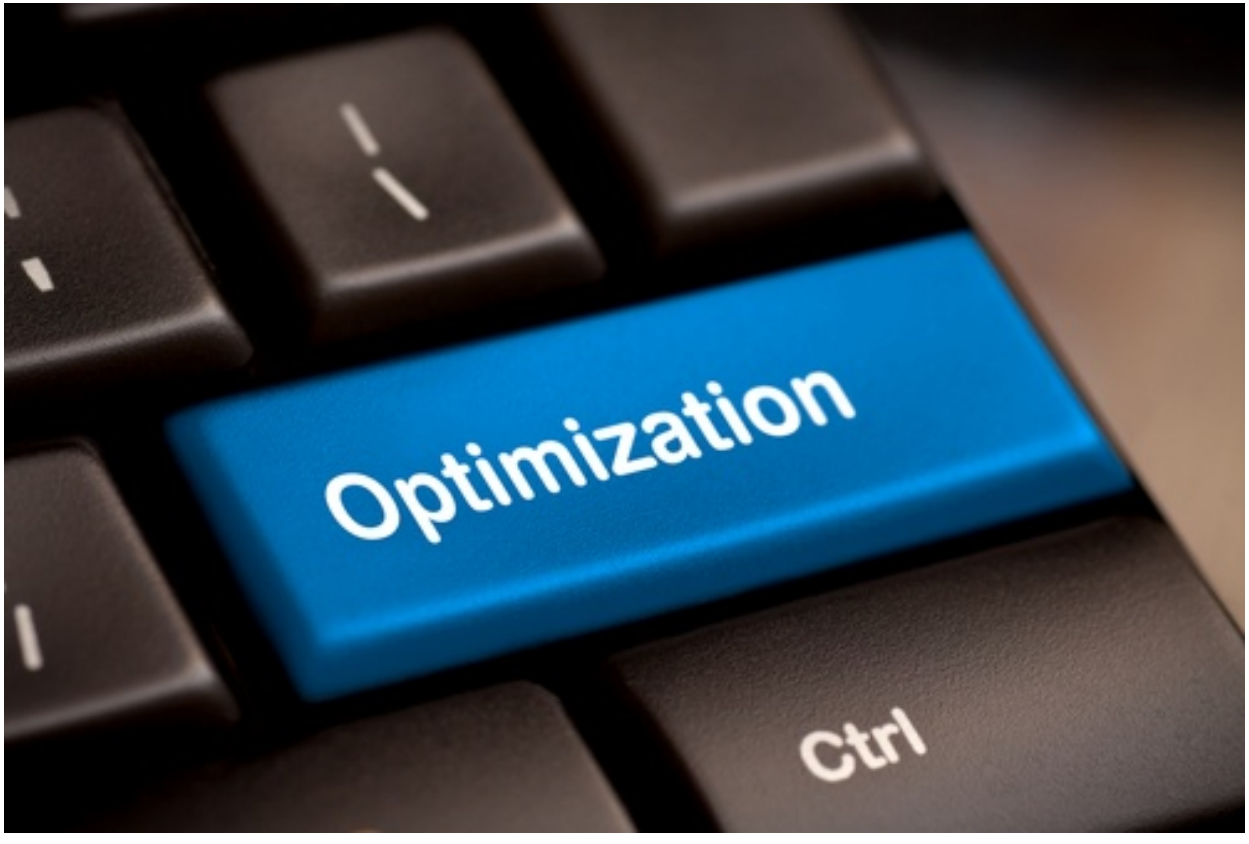
Erreur quadratique moyenne (relativement à l'accuracy empirique) correspondant à 7 proxies

- LD parfait
- Moyenne sur 15 architectures
  - 100, 1000, 2000 marqueurs
  - 2 QTLs, 100 QTLs et 3 autres scenarios
- 30 générations pour les Trainings

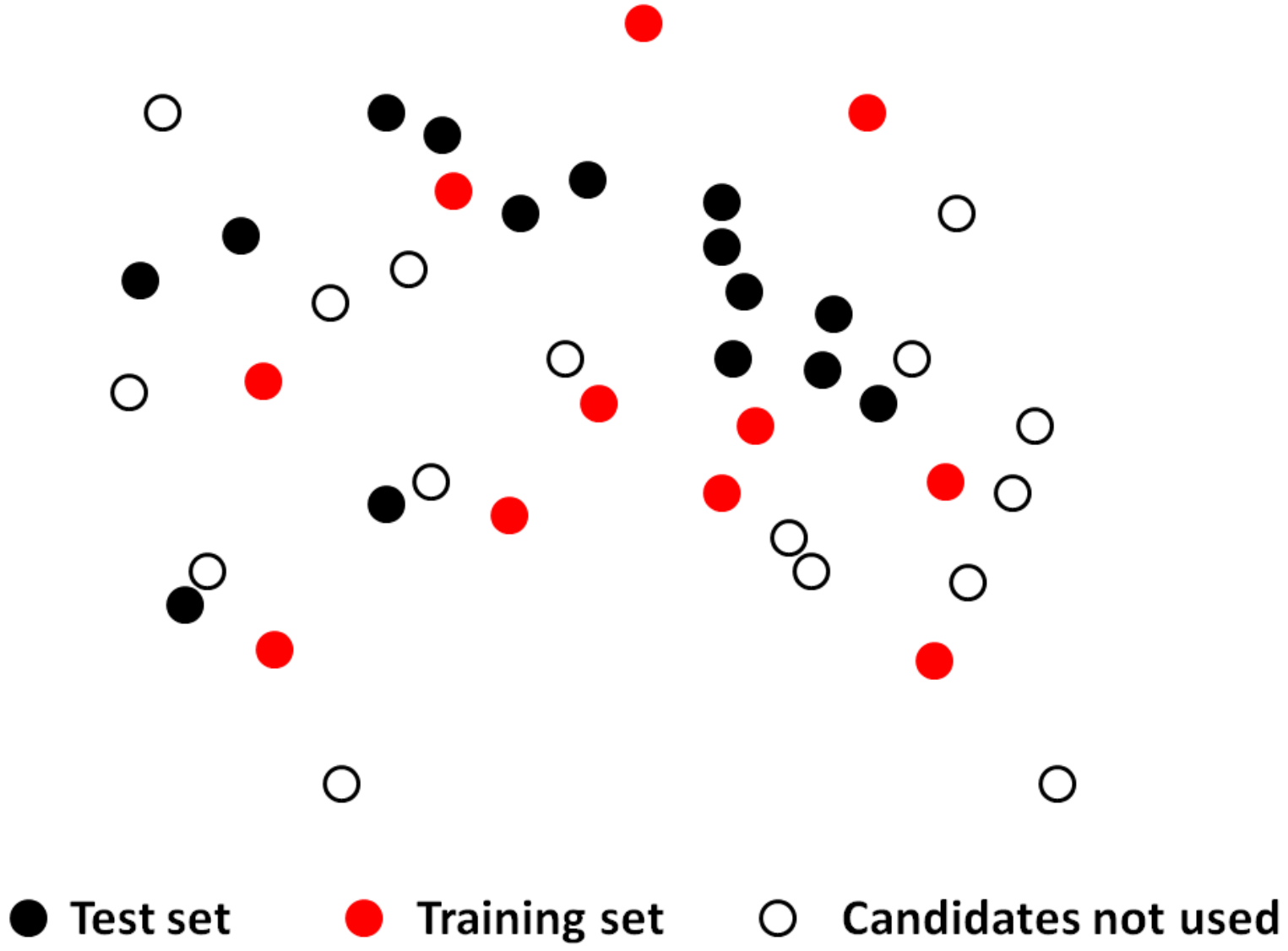
MSE based on	50 generations for TST	70 generations for TST
$\hat{\rho}_{after}(\theta^{\star})$	5.9685×10^{-5}	3.8455×10^{-5}
$\hat{\rho}_{after}(\hat{\theta}^{\star}_{ADLASSO})$	1.2108×10^{-3}	1.2118×10^{-3}
$\hat{\rho}_{before}(\hat{\theta}^{\star}_{ADLASSO})$	2.2677×10^{-3}	1.5168×10^{-3}
Proxy Plos One (2016)	3.3056×10^{-3}	1.007×10^{-2}
M_{e1}	3.7936×10^{-3}	1.3779×10^{-2}
M_{e2}	3.7508×10^{-3}	1.3518×10^{-2}
M_{e3}	3.6970×10^{-3}	1.3165×10^{-2}
M_{LJ}	5.5578×10^{-3}	6.1021×10^{-3}

Plan

Training set Optimization



Brigitte Mangin (LIPM)


Test set
Training set
Candidates not used

Predictive ability : ρ

There are several predictive ability estimators

- assuming the mixed model is correct
  - based on coefficient of determination (CD) :

$$\hat{\rho}^{\text{CD}} = \frac{h}{n_{\text{test}}} \sum_i \sqrt{\text{CD}(g_{\text{test},i})}$$

  - based on prediction error variance (PEV) :

$$\hat{\rho}^{\text{PEV}} = \frac{h}{n_{\text{test}}} \sum_i \sqrt{1 - \frac{\text{PEV}(g_{\text{test},i})}{\sigma_g^2}}$$

- assuming a fixed linear QTL model is known and correct and using the mixed model as ”instrumental” model : the theoritical accuracy Rabier et al., PlosOne, 2016

Drawback of the theoritical accuracy

QTLs have to be found : comparison of several multi-locus GWAS methods

- penalized regressions (Lasso.min, Lasso.1se, EN05.1se, EN01.1se) Waldmann et al., Frontiers in Genetics, 2013, (EN05.FDR) Yi et al., Genetics ,2015
- MLMM Segura et al., Nature Genetics, 2012

EthAcc (i.e. $\hat{\rho}_{after}$) comparison results

Ithaque est une ile Grecque

TABLE – **Mean Square Error of EthAcc using several methods to estimate causal QTLs on different traits (mean over 100 random test sets and 7 traits on sugar beet).**

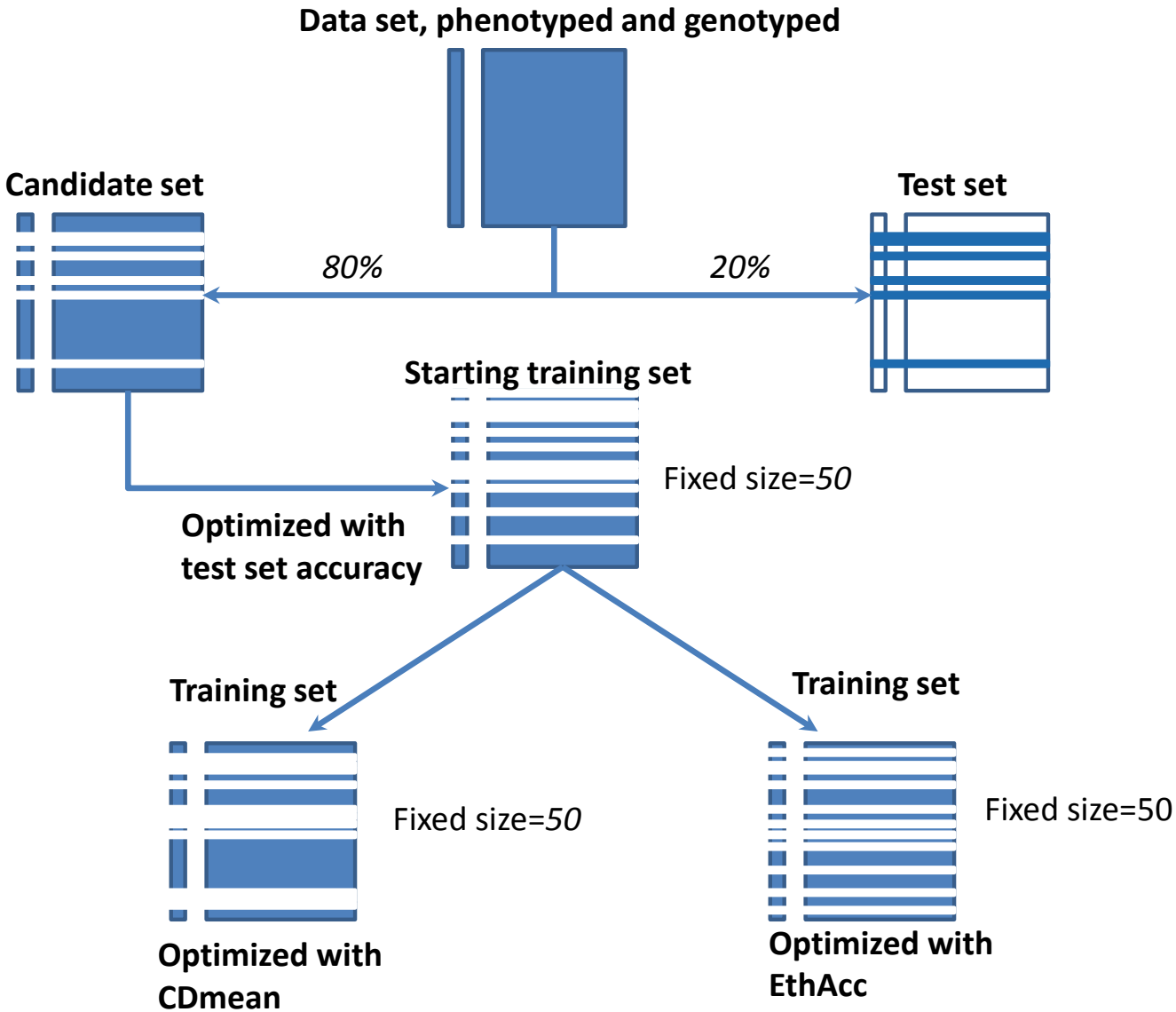
GWAS	MSE
MLMM	$1.22\ 10^{-3}$
Lasso.min	$3.25\ 10^{-3}$
Lasso.1se	$1.60\ 10^{-3}$
EN05.1se	$1.65\ 10^{-3}$
EN01.1se	$1.78\ 10^{-3}$
EN05.FDR	$8.54\ 10^{-3}$

EthAcc, CD, PEV comparison results

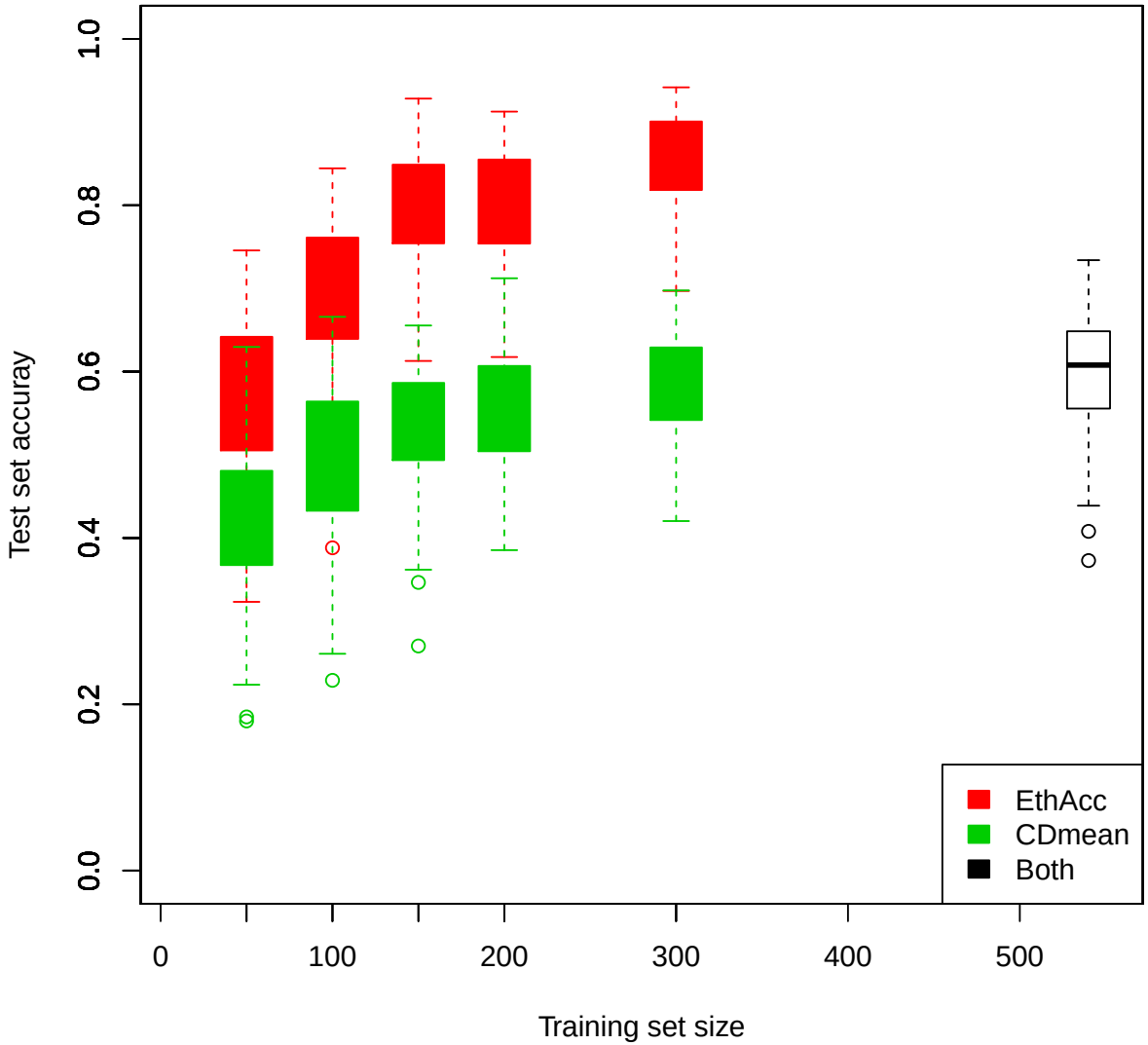
TABLE – **Accuracy and its estimate by EthAcc, CD and PEV using sugar beet structures in two panels on several traits (mean over 100 random test sets) for white sugar yield in t/ha.**

Test	Training	Accuracy	Estimated by		
			EthAcc	CD	PEV
Panel_A	Panel_A+B	0.583	0.562	0.735	0.735
Panel_A	Panel_A	0.597	0.603	0.702	0.703
Panel_B	Panel_A+B	0.640	0.644	0.865	0.865
Panel_B	Panel_B	0.536	0.527	0.678	0.680

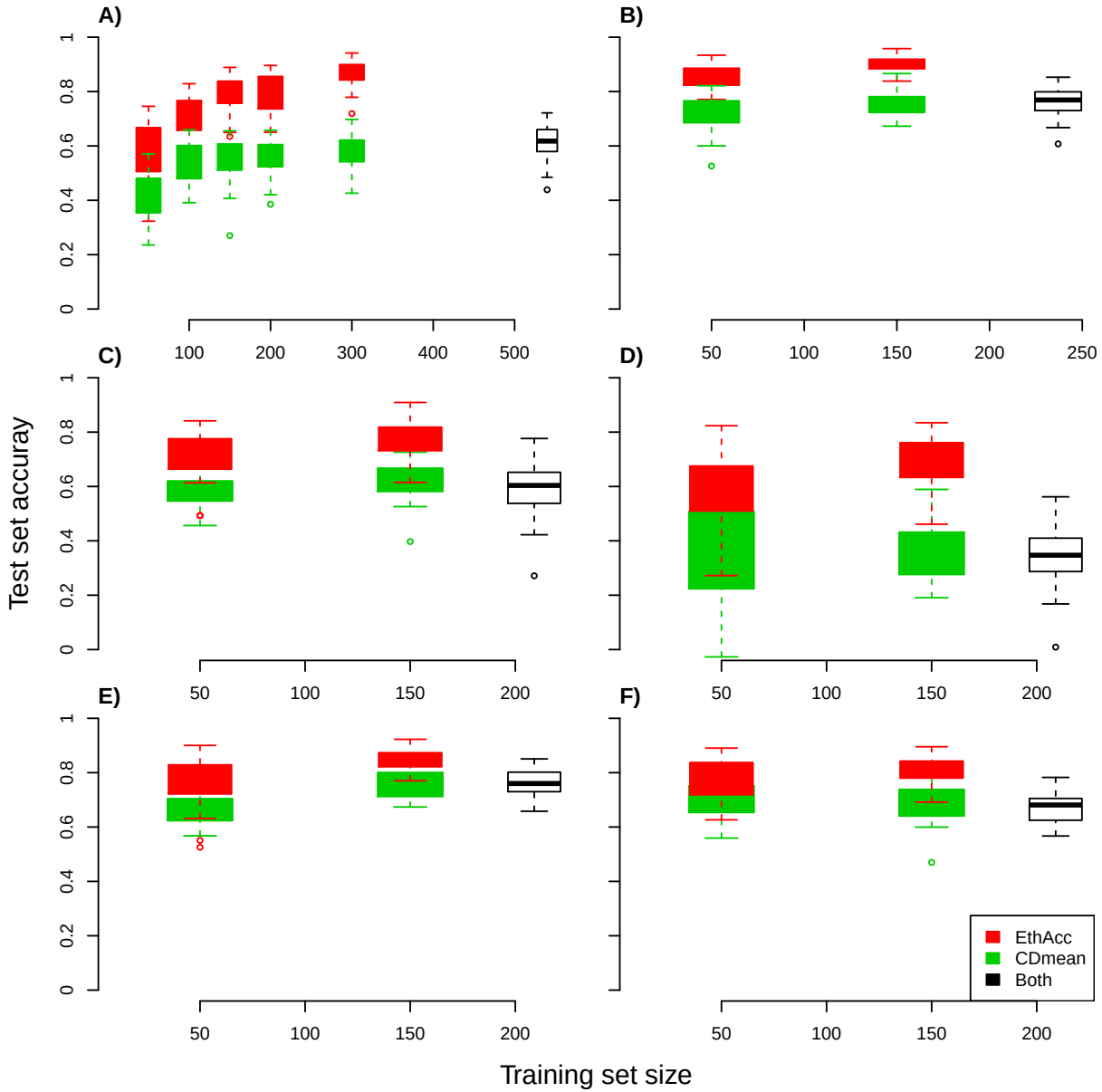
Training set optimization : EthAcc versus CDmean (Rincent et al, Genetics 2012)

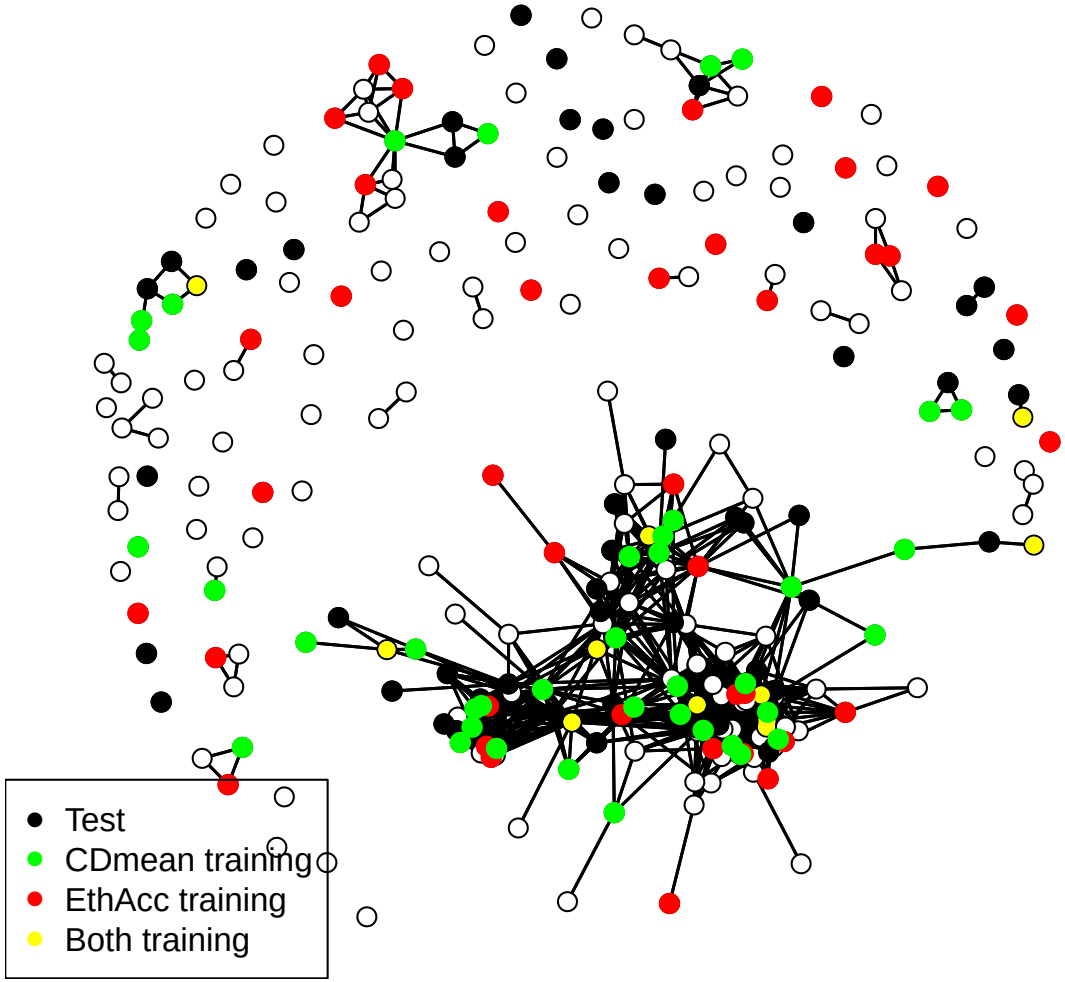


Training set optimization (Sugar beet) : EthAcc versus CDmean



Sugar beet, Wheat, Maize



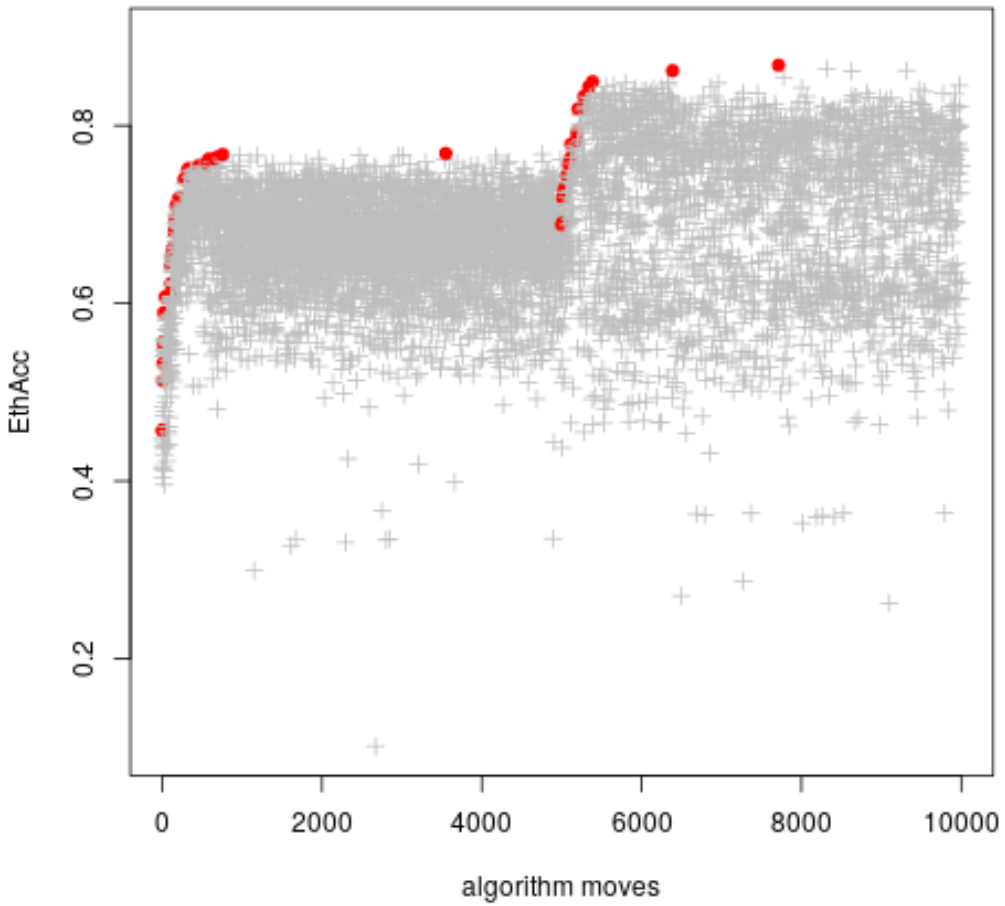


- CDmean optimization gave an accuracy of 0.07
- EthAcc optimization gave an accuracy of 0.76

Conclusion

- EthAcc outperforms CDmean for training panel optimisation
- EthAcc shows that ”bigger is not better”

Optimization problem



- 5 000 moves
- Left : random starting point
- Right : optimized starting point

Thank you for your attention.

A special thank to Fanny Bonnafous, Prune Pegot-Espagnet, Charles-Elie Rabier, Ellen Goudemand, Renaud Rincent

Plan

Pour aller plus loin ...

Théorème (R., Barre, ..., Mangin, Plos One 2016)

Conditionnellement à M et $M^\star$,

$$\rho = \frac{\theta^{\star\prime}\ \mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right) M' V^{-1} M^\star \theta^\star}{\left\{\sigma_e^2 \mathbb{E}\left(\|M_{n+1} M' V^{-1}\|^2\right) + \theta^{\star\prime} M^{\star\prime} V^{-1} M \mathrm{Var}\left(M_{n+1}'\right) M' V^{-1} M^\star \theta^\star\right\}^{1/2} \Omega^{1/2}}$$

où $V = MM' + \lambda I_n$ et $\Omega = \mathrm{Var}\left(M_{n+1}^\star \theta^\star\right) + \sigma_e^2$.

Décomposition SVD de M

$$M = U\ D\ W'$$

où

- D matrice diagonale de taille $r \times r$, de plein rang
- U matrice de taille $n \times r$, telle que $U'U = I_r$
- W matrice de taille $p \times r$, telle que $W'W = I_r$

En injectant la décomposition SVD dans notre formule

Par abus de notation :

- $\theta^\star$ vecteur sparse de dimension K contenant les effets QTLs

Théorème (R., Mangin, Grusea (en révision pour Scand J. Statistics))

$$\hat{\rho}_{before} \geq \frac{\|WW'\theta^\star\|^2 \ \min \frac{d_s^4}{d_s^2+\lambda}}{\sqrt{\sigma_e^2 \ r + \ \|WW'\theta^\star\|^2 \ \max d_s^2} \quad \sqrt{\|WW'\theta^\star\|^2 \ \max d_s^2 + \sigma_e^2}}$$

$$\hat{\rho}_{before} = \frac{\sum_{s=1}^{r} \frac{d_s^4}{d_s^2+\lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2}{\left(\sigma_e^2 \ \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} + \sum_{s=1}^{r} \frac{d_s^6}{(d_s^2+\lambda)^2} \|W^{(s)} W^{(s)\prime} \theta^\star\|^2 \right)^{1/2} \left(\sum_{s=1}^{r} d_s^2 \|W^{(s)} W^{(s)\prime} \theta^\star\|^2 + \sigma_e^2 \right)}$$

On aimerait avoir $\sigma_e^2 \ \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2}$ petit, et $\sum_{s=1}^{r} \frac{d_s^4}{d_s^2+\lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2$ grand

Vers une amélioration du RRBLUP, GBLUP, Ridge

Idée : considérer un espace de dimension plus faible

Rappel : $U = \left(U^{(1)}, \ldots, U^{(r)}\right)$ base orthonormale de l'espace engendré par les colonnes de M.

On choisit $\tilde{r}$ colonnes de U.

Soit l'estimateur

$$\tilde{\theta} = M' V^{-1} \tilde{U} \tilde{U}' Y \quad \text{où} \quad \tilde{U} = \left(U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right)$$

où $\tilde{U}\tilde{U}' Y$ est la projection de Y sur $Vect\left\{U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right\}$.

$\Rightarrow$ Prédictions a l'aide de $\tilde{\theta}$

Accuracy basée sur ce nouvel estimateur

Théorème (R., Mangin, Grusea (en révision pour Scand J. Statistics))

$$\tilde{\rho}_{before} \geq \frac{\left\| \tilde{W}\tilde{W}'\theta^\star \right\|^2 \min_{1\leq s\leq \tilde{r}} \frac{d^4_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}}{\sqrt{\sigma_e^2\,\tilde{r} + \left\| \tilde{W}\tilde{W}'\theta^\star \right\|^2 \max_{1\leq s\leq \tilde{r}} d^2_{\sigma(s)}}\ \sqrt{\left\| \tilde{W}\tilde{W}'\theta^\star \right\|^2 \max_{1\leq s\leq \tilde{r}} d^2_{\sigma(s)} + \sigma_e^2}}$$

$$\tilde{\rho}_{before} = \frac{\sum_{s=1}^{\tilde{r}} \frac{d^4_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda} \left\| \tilde{W}^{(\sigma(s))}\tilde{W}^{(\sigma(s))\prime}\theta^\star \right\|^2}{\left(\sigma_e^2 \sum_{s=1}^{\tilde{r}} \frac{d^4_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2} + \sum_{s=1}^{\tilde{r}} \frac{d^6_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2} \left\| \tilde{W}^{(\sigma(s))}\tilde{W}^{(\sigma(s))\prime}\theta^\star \right\|^2 \right)^{1/2} (\Omega)^{1/2}}$$

$$\text{avec } \Omega = \sum_{s=1}^{r} d^2_{\sigma(s)} \left\| W^{(\sigma(s))}W^{(\sigma(s))\prime}\theta^\star \right\|^2 + \sigma_e^2$$

Dans quelles conditions améliore-t-on l'accuracy ?

- Estimateur Ridge $\hat{\theta}$ basé sur toutes les colonnes de U
  - accuracy $\hat{\rho}$, prédiction $\hat{Y}_{n+1}$
- Nouvel estimateur $\tilde{\theta}$ basé sur $\tilde{r}$ colonnes de $U \Rightarrow \tilde{\beta}$
  - accuracy $\tilde{\rho}$, prédiction $\tilde{Y}_{n+1}$
- Complémentaire $\vec{\theta}$ de notre nouvel estimateur basé sur les $r - \tilde{r}$ colonnes restantes de U
  - accuracy $\vec{\rho}$, prédiction $\vec{Y}_{n+1}$

$\tilde{\rho}_{before} \geq \hat{\rho}_{before}$ si et seulement si

$$\frac{\widehat{\text{Cov}}\left(\tilde{Y}_{n+1}, Y_{n+1}\right)}{\widehat{\text{Cov}}\left(\vec{Y}_{n+1}, Y_{n+1}\right)} \geq \frac{\widehat{\text{Var}}\left(\tilde{Y}_{n+1}\right)}{\widehat{\text{Var}}\left(\vec{Y}_{n+1}\right)} \left(1 + \sqrt{1 + \frac{\widehat{\text{Var}}\left(\vec{Y}_{n+1}\right)}{\widehat{\text{Var}}\left(\tilde{Y}_{n+1}\right)}}\right).$$

Un exemple où θ appartient à l'espace engendré par les lignes de M

- Chromosome de longueur 1M
- 1000 marqueurs
- $\theta = 0.3W^{(1)} + 0.3W^{(2)} + 0.3W^{(3)}$
- $\tilde{r}$ et les colonnes de U choisies par validation croisée
- 100 TESTS

σ_e^2	n	Méthode	200 générations	400 générations
1	500	$\hat{\rho}_{before}$	0.7550	0.6721
		$\hat{\tilde{\rho}}_{before}$	0.7810	0.7041
	800	$\hat{\rho}_{before}$	0.7487	0.7037
		$\hat{\tilde{\rho}}_{before}$	0.7728	0.7312
9	500	$\hat{\rho}_{before}$	0.3370	0.2623
		$\hat{\tilde{\rho}}_{before}$	0.3809	0.3079
	800	$\hat{\rho}_{before}$	0.3317	0.2904
		$\hat{\tilde{\rho}}_{before}$	0.3734	0.3330

Conclusion

Quelques pistes envisageables pour la suite :

- comparaison MLMM et Adaptative LASSO pour EthAcc
- utilisation d'un autre algorithme afin de determiner un bon point de départ pour l'optimisation de l'ensemble de training
- est-ce possible d'exploiter cet estimateur amelioré ?

Remerciements

Simona Grusea (INSA), Brigitte Mangin (LIPM), Celine Scornavacca (ISEM)

Collaborateurs sur CROPDL

- Muriel Tavaud / Jacques David (SupAgro Montpellier)
- Gilles Charmet / Delphine Ly / François Balfourier (INRA Clermont Ferrand)
- Philippe Barre (INRA Lusignan)/ Torben Asp (Aarhus university, Danemark)

Choix du λ de la ridge (Génome Causal)

Contexte

- QTLs connus
- Effets QTLs inconnus

Comment choisir le λ pour la Ridge ?

- Par REML, mais nécessite la connaissance des phénotypes
- En utilisant l'héritabilité

Choix du λ de la ridge (Génome Causal)

On doit supposer les QTLs indépendants pour passer du modèle aléatoire au modèle à effet fixes

- Modèle causal à effets fixes

$$\sigma^2_G = \text{Var}\left(\sum_{q=1}^{Q} M^\star_{i,q}\theta^\star_q\right) = \text{Var}(M^\star_{i,q}) \sum_{q=1}^{Q}(\theta^\star_q)^2$$

- Modèle causal Bayésien

$$\tilde{\sigma}^2_G = \text{Var}\left(\sum_{q=1}^{Q} M^\star_{i,q}\theta^\star_q\right) = Q\,\sigma^2_\theta\,\text{Var}(M^\star_{i,q})$$

- On égale les variances génétiques

$$\text{Var}(M^\star_{i,q}) \sum_{q=1}^{Q}(\theta^\star_q)^2 = Q\,\sigma^2_\theta\,\text{Var}(M^\star_{i,q}) \quad , \quad \sigma^2_\theta = \frac{1}{Q}\sum_{q=1}^{Q}(\theta^\star_q)^2$$

Choix du λ de la ridge (Génome Causal)

On rappelle que

$$\lambda = \sigma_\varepsilon^2 / \sigma_\theta^2$$

alors,

$$\lambda^\star = \frac{\sigma_\varepsilon^2 \; Q}{\sum_{q=1}^{Q} (\theta_q^\star)^2} = \frac{(1 - h^2) \; Q \; \mathrm{Var}(M_{i,q}^\star)}{h^2}$$

Choix du λ de la ridge (Génome basé sur les SNPs)

Modèle Bayésien utilisant K SNPS

$$\tilde{\sigma}^2_G = \text{Var}\left(\sum_{k=1}^{K} M_{i,k}\theta_k\right) = K\ \sigma^2_\theta\ \text{Var}(M_{i,k})$$

On égale cette variance génétique avec la variance génétique causale

$$\text{Var}(M^\star_{i,q})\ \sum_{q=1}^{Q}(\theta^\star_q)^2 = K\ \sigma^2_\theta\ \text{Var}(M_{i,k})\ \ ,\ \ \sigma^2_\theta = \frac{1}{K}\sum_{q=1}^{Q}(\theta^\star_q)^2$$

alors,

$$\lambda = \frac{(1-h^2)\ K\ \text{Var}(M_{i,k})}{h^2}$$

Modèle mixte

$$Y_i = \mu + \sum_{k=1}^{K} M_{i,k}\theta_k + \varepsilon_i$$

Effet marqueur : θ_k

Effet variance : σ^2_θ

Postulats pour l'estimation

les θ_k sont indépendants et identiquement distribués (iid)
$\theta_k \sim \mathcal{N}(0, \sigma^2_\theta)$ pour tout k

P1

ε_i indépendant des autres effets, iid
$\varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$

RR-BLUP versus Kinship-BLUP = G-BLUP

RR-BLUP

$Y_i = \mu + \sum_{k=1}^{K} M_{i,k}\theta_k + \varepsilon_i$

Postulats : P1
θ_k iid, $\theta_k \sim \mathcal{N}(0, \sigma^2_\theta Id)$

Kinship-BLUP

$Y_i = \mu + g_i + \varepsilon_i$

Postulats : P1
$g \sim \mathcal{N}(0, \sigma^2_g W)$

Conditionnellement à M, (M est fixe) si $W = MM'/K$, les deux modèles sont identiques (espérance et variance de Y sont les mêmes dans les 2 modèles)

On a alors $\sigma^2_\theta = \sigma^2_g / K$

Vers la régression pénalisée : RR-BLUP , ridge regression BLUP

En notation matriciel, le modèle est $Y = \mathbf{1}\mu + M\theta + \varepsilon$

On estime les paramètres des composantes de la variance σ^2_θ et σ^2_ε avec le REML.

On prédit θ par le BLUP en résolvant les équations du modèle mixte :

$$\hat{\theta} = (M'M + \lambda Id)^{-1} M'\tilde{Y}$$

C'est une "ridge régression" avec le paramètre : $\lambda = \hat{\sigma}^2_\varepsilon / \hat{\sigma}^2_\theta$