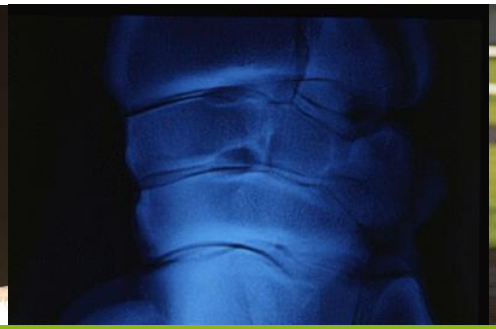
Rules & Tules, La rochelle, 23 Septembre 2010

Bayes C PI
GABAYES

A. Ricard, O. Filangi

et la collaboration de A. Legarra et R. Fernando qu'on a copié…

Institut français du cheval et de l'équitation
ECOLE NATIONALE D'EQUITATION Le Cadre Noir de Saumur
les Haras nationaux

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

Objectif

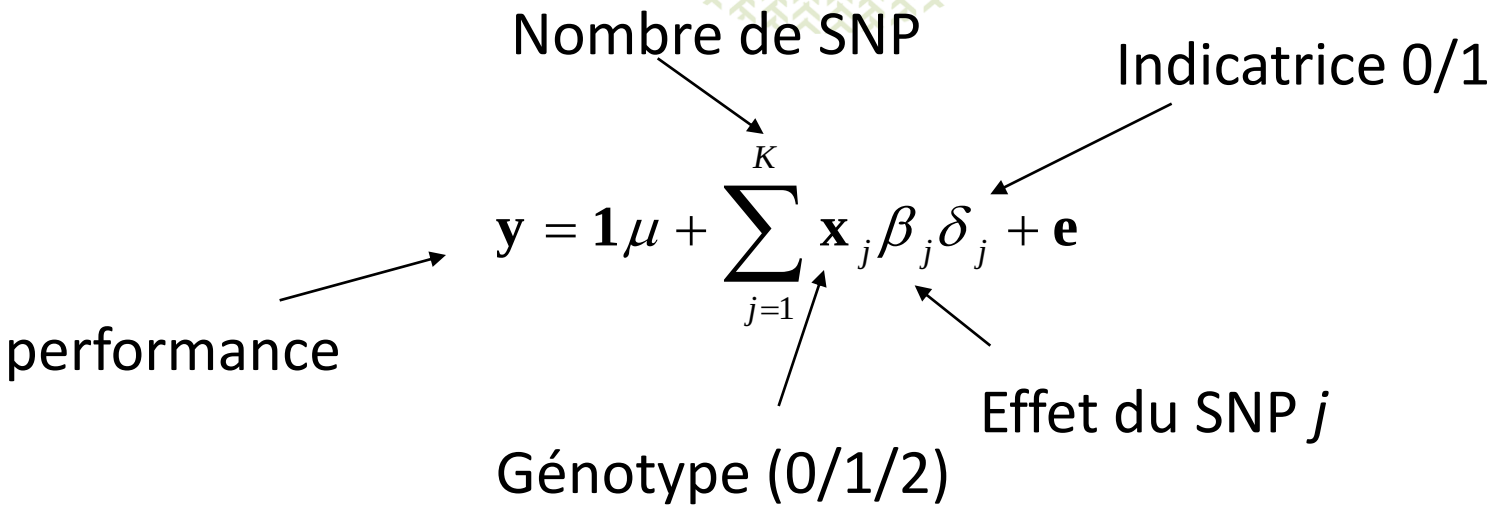
- Avoir un logiciel efficace pour calculer un BayesC (pi)
- Etat des lieux : programme R (Fernando/Legarra)
- Référence :

Kizilkaya, R., Fernando, R.L., Garrick., D.J., 2010. Genomic prediction of simulated multibreed and purebred performances using observed fifty thousand single nucleotide polymorphism genotypes. Journ. Anim. Sci., 88, 544-551.

Modèle de base (modifié dans Kizilkaya car on n'a plus qu'une distribution pour les SNP)

Meuvissen, T.H.E, Godard, M.E., 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel.Evol., 36, 261-279.

Modèle



$$\beta_j \quad N(0, \sigma^2_\beta), \forall j$$

$$\mathbf{e} \quad N(0, \mathbf{I}\sigma^2_e)$$

$$\delta_j \rightarrow \begin{cases} 0, \pi \\ 1, 1-\pi \end{cases}$$

Méthode : Gibbs Sampling

- ***Step 1 : Tirage*** de σ_e^2

$$\hat{S}_e^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{y} - \mathbf{1}\hat{\mu} - \sum_{j=1}^{K} \mathbf{x}_j \hat{\beta}_j \hat{\delta}_j)'(\mathbf{y} - \mathbf{1}\hat{\mu} - \sum_{j=1}^{K} \mathbf{x}_j \hat{\beta}_j \hat{\delta}_j)$$

$$\hat{\sigma}_e^2 = \frac{\hat{S}_e^2}{u} \longrightarrow \chi^2 n + 3ddl$$

- ***Step 2 : Tirage*** de μ

$$N\left(\frac{\mathbf{1}'(\mathbf{y} - \sum_{j=1}^{K} \mathbf{x}_j \hat{\beta}_j \hat{\delta}_j)}{n}, \frac{\hat{\sigma}_e^2}{n}\right)$$

Méthode : Gibbs Sampling

- ***Step 3.*** Pour tous les SNP *i* de 1 à K, tirage de δ_i β_i

On suppose que la probabilité pour que δ_i soit égale à zéro est le produit d'une probabilité a priori et d'une densité conditionnelle de l'estimation du coefficient de régression connaissant les autres paramètres, normale, soit (à une constante près)

Vecteur corrigé

$$t_0 = \frac{1}{\sqrt{\mathbf{x}_i'\mathbf{x}_i\hat{\sigma}_e^2}} \exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}_i'\mathbf{y}_i\right)^2}{\mathbf{x}_i'\mathbf{x}_i\hat{\sigma}_e^2}\right)\hat{\pi}$$

$\hat{\delta}_i = u$

$$t_1 = \frac{1}{\sqrt{\mathbf{x}_i'\mathbf{x}_i\sigma_e^2 + \left(\mathbf{x}_i'\mathbf{x}_i\right)^2\sigma_\beta^2}} \exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}_i'\mathbf{y}_i\right)^2}{\mathbf{x}_i'\mathbf{x}_i\sigma_e^2 + \left(\mathbf{x}_i'\mathbf{x}_i\right)^2\sigma_\beta^2}\right)(1-\hat{\pi})$$

$$f = \frac{t_0}{t_0 + t_1}$$

Et on tire donc δ_i dans une loi 0 ou 1 de probabilité *f* et 1-*f*

Méthode : Gibbs Sampling

- Tirage de β_i

  Si $\hat{\delta}_i = 0$, $\hat{\beta}_i = 0$

  Si $\hat{\delta}_i = 1$, on va tirer β_i dans une loi normale

$$N\left(\frac{\mathbf{x}_i'\mathbf{y}_i}{\mathbf{x}_i'\mathbf{x}_i + \frac{\hat{\sigma}_e^2}{\hat{\sigma}_\beta^2}}, \frac{\hat{\sigma}_e^2}{\mathbf{x}_i'\mathbf{x}_i + \frac{\hat{\sigma}_e^2}{\hat{\sigma}_\beta^2}}\right)$$

Méthode : Gibbs Sampling

- ***Step 4.*** Tirage de σ_β^2

$$\hat{S}_\beta^2 = \boldsymbol{\beta}'\boldsymbol{\beta} + (\nu_\beta - 2)\frac{\sigma_a^2}{K(1-\hat{\pi})2\overline{\overline{pq}}}$$

$$\hat{\sigma}_\beta^2 = \frac{\hat{S}_\beta^2}{u} \longleftarrow \chi^2 \grave{a} \sum_{j=1}^{K} \hat{\delta}_j + \nu_\beta ddl$$

- ***Step 5*** *: Tirage de* π *dans une loi beta de paramètres*

$$K - \left(\sum_{j=1}^{K} \hat{\delta}_j\right) + 1 \qquad \left(\sum_{j=1}^{K} \hat{\delta}_j\right) + 1$$

Détails

- Marche avec des génotypes manquants
  - ? Pour l'instant on tire les génotypes en fonction uniquement des fréquences., rajouter les perf ?
- Tirage de π facultatif
- A rajouter :
  - autres effets fixes (que la moyenne)
  - Autres effets aléatoires (genre génétique, environnement perm.)
  - Variances résiduelles hétérogènes (genre index de-regresses)

Voici un trac : http://pluton.toulouse.inra.fr/trac-gabayes

Construction du fichier de paramètre

in_numanim	Nombre d animaux
in_nummark	Nombre de marqueurs
in_pheno_and_genotype	Fichier de performance et de typage (Format 1)
in_trait	Fichier de performance (Format 2)
in_genotype	Fichier de genotype (Format 2)
out_output	Fichier de sortie resultat
in_vara	Variance genetique total initiale
in_vare	Variance residuelle initiale
in_pi	Valeur de Pi initiale
in_randompi	Si vrai tire une nouvelle valeur de Pi durant chaque cycle
in_nubeta	Value de nubeta
in_numberchain	Nombre de Chain a exécuter en parallèle
in_chainlength	Longueur de la chaine
in_numcycle	Pas de recolte des valeurs à estimer
in_burnin	Nombre de cycle de chauffe...
set_random_generator	1 -> random sur le PRNG implémenté par le compilateur fortran , 2 -> Tina's RNG

Sorties

- Fichier avec à chaque itération σ_e^2 σ_β^2 π
- Fichier avec pour tous les SNP, sur l'ensemble des itérations, pour chaque chaine :
  - Moyenne des β_i
  - Écart type β_i
  - Moyenne des δ_i

Vérifications

- Nous avons vérifié la concordance de GABAYES et du programme R sur des simulations <40000 SNP (compte tenu des temps de calcul du programme R)
- Tests de comparaison avec :
- $K = 100,\ \pi = 0$ $K = 1000,\ \pi = 0.10$ $K = 1000,\ \pi = 0.90$
- Nous avons utilisé 100000 itérations, burn-in de 5000 pour le programme d'Olivier et 10000 itérations, burn-in de 1000 pour le programme R mais la convergence est très rapide quand elle est bonne et aussi mauvaise dans les deux cas quand elle l'est
- La moyenne et les écart types des estimations de σ_e^2 σ_β^2 π sur l'ensemble des itérations après burn in sont les mêmes
- La corrélation entre les estimés des effets des SNP est supérieure à 0.999 pour les SNP ayant un effet simulé comme pour les autres.
- Nous avons conclu que les deux programmes faisaient bien la même chose.

Temps

- Je n’ai pas utilisé la parallélisation et pourtant
- / R : J'ai mis 40 minutes pour 1000 animaux/1000 SNP/100 000 itérations et 1H10 en R pour la même chose avec 10 fois moins d'itérations (10000 itérations)...

Il n’y a pas photo…

- Pour des tailles plus réelles (/ à mes données) j'ai mis 10 heures pour 1000 animaux/40 000 SNP et 60 000 itérations , 2jours 5 heures pour 200000 itérations

Validation

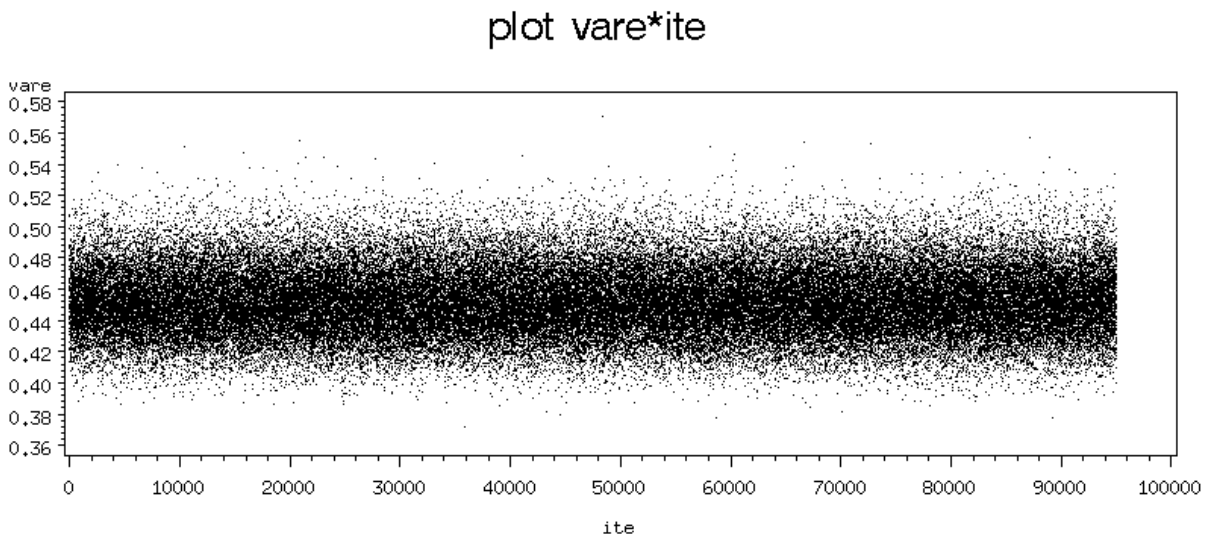
- J'ai voulu tester uniquement l'efficacité de l'algorithme pour estimer les paramètres du modèle qu'il postule.
  - L'objectif est de vérifier la possibilité d'estimer l'effet de chaque SNP en tant que QTL, non de mesurer la valeur génétique globale dans un but de sélection génomique.
- Je laisse la validation de la méthode en tant que chercheur de QTL ou d'évaluation génomique dans le cas de véritables typages (DL, apparentement, non explicites dans le modèle) aux autres (nombreux)
- D'autre part, je me suis située dans des ordres de grandeur propre à nos populations chevaux actuelles,

Simulations

- 1000 animaux
- *K* Marqueurs biallèlique, fréquence ½ (*K=100, 1000, 40000*)
- On tire la valeur de chaque SNP, d'abord 0/≠0 avec une probabilité π puis, le cas échéant dans une normale de variance $\sigma_\beta^2 = \frac{\sigma_a^2}{K(1-\pi)0.5}, \sigma_a^2 = 0.5$
- On calcule la performance comme la somme des effets des allèles de l'individu
- Les simulations n'ont pas été répétées…

Cas 100 marqueurs

- Convergence OK



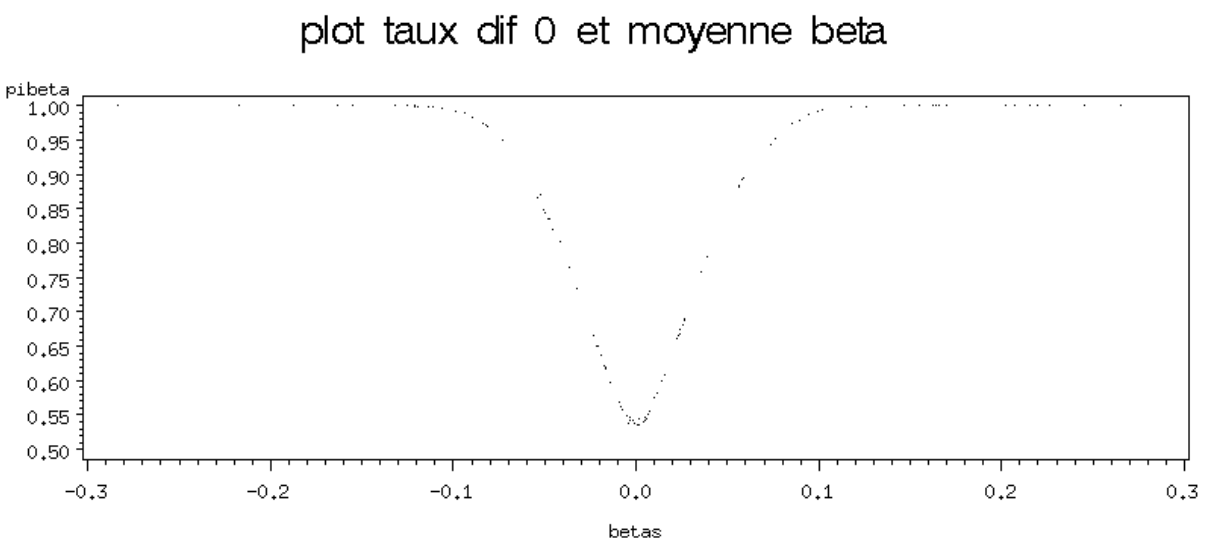
- Variance résiduelle et génétique OK

Cas 100 marqueurs

- π biaisé à la hausse quand proche de 0 : attendu

Simulé	0.00	0.10	0.50	0.90	1.00
estimé	0.196	0.151	0.456	0.930	0.989

Moyenne des $\hat{\delta}_i$



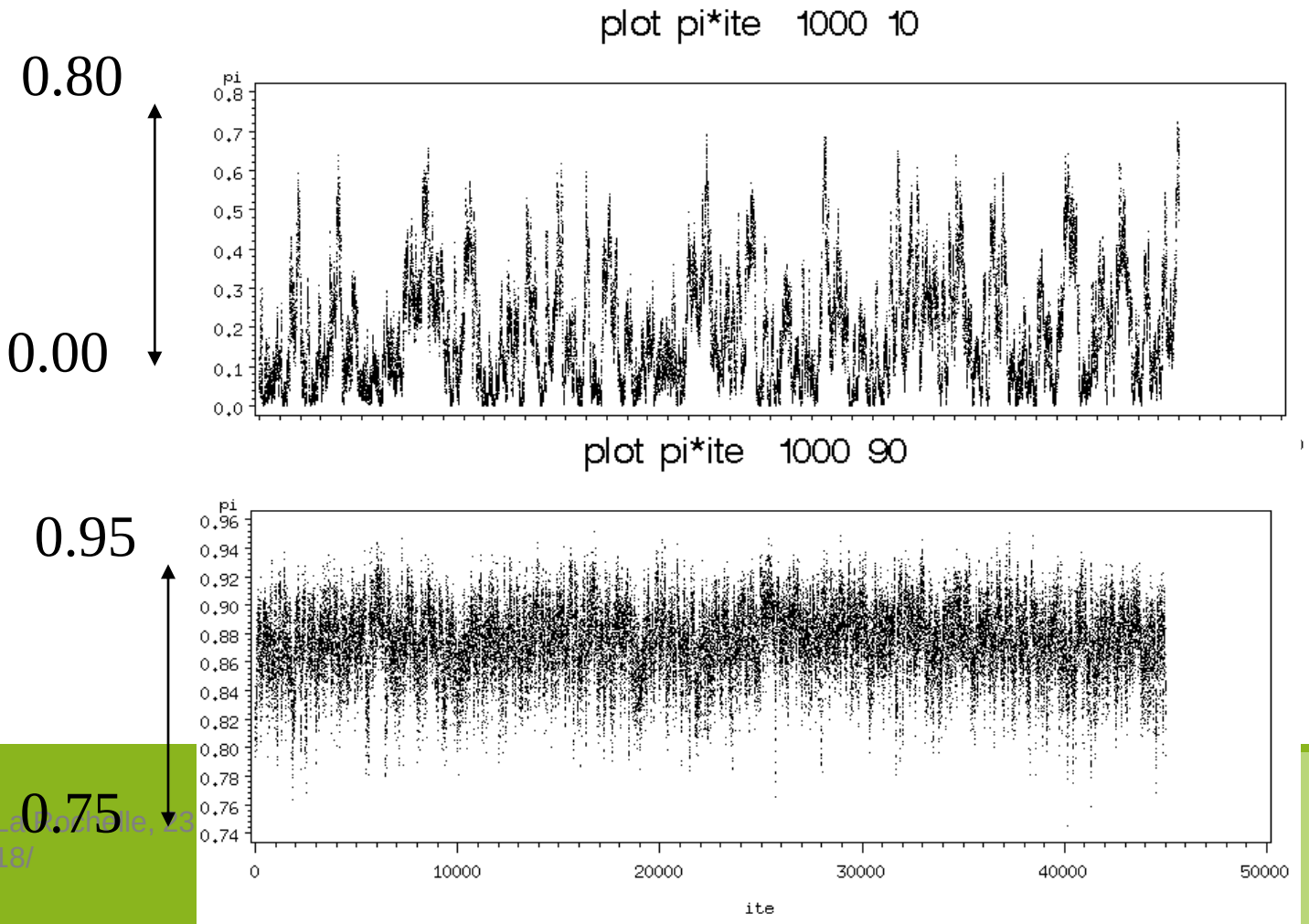
Valeur estimée des $\hat{\beta}_i$

Cas 100 marqueurs

- Valeurs des $\hat{\beta}_i$ OK $r(\hat{\beta}_i, \beta_i) > 0.94$
- Donc, OK, Peut être amélioré à la marge en tenant compte de la puissance du dispositif.

Cas 1000 marqueurs

- Convergence commence à être moyenne pour π faible (beaucoup de SNP avec effets)



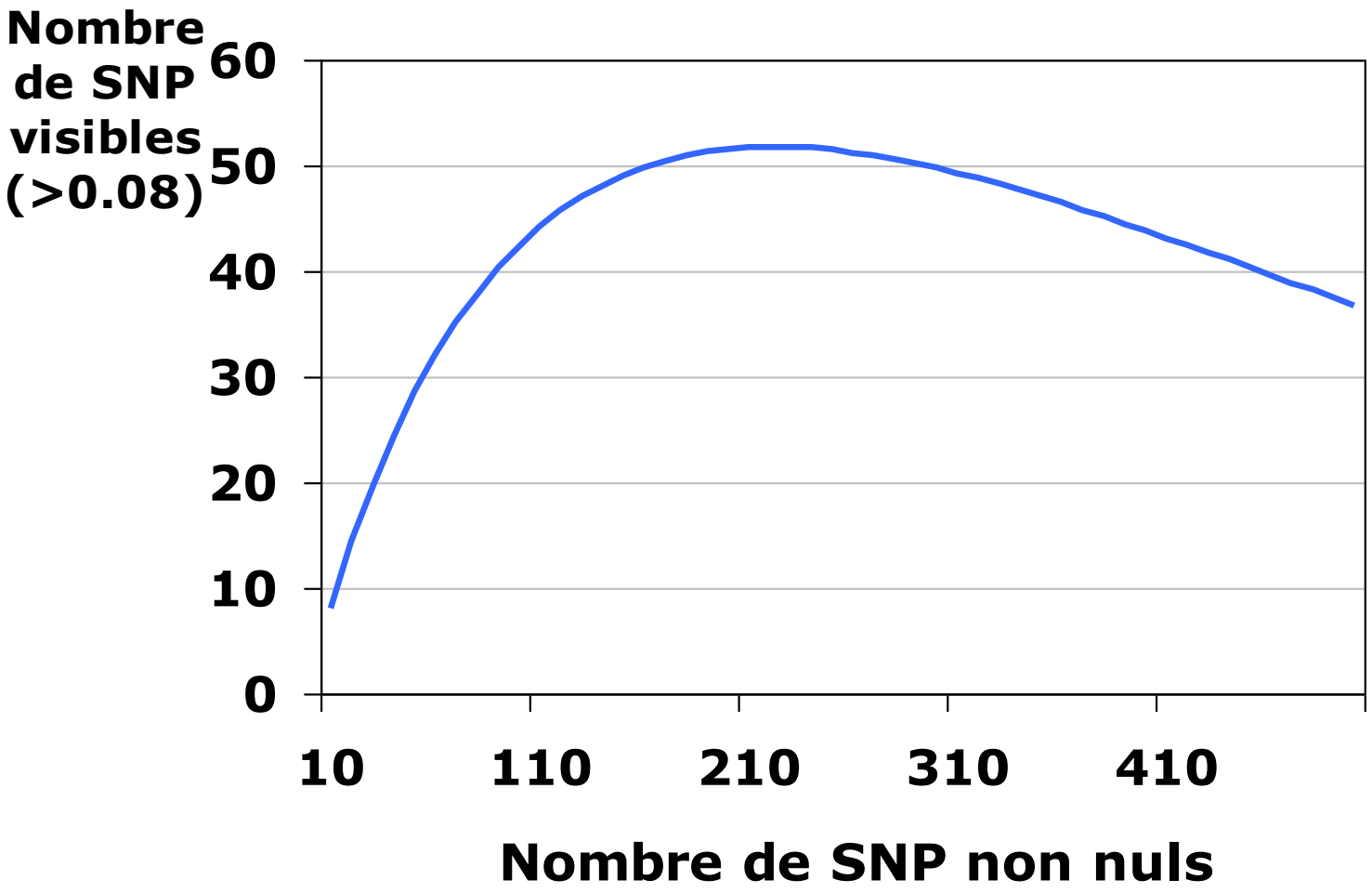
Cas 1000 marqueurs

- Avec π =0.10 ou π =0.50, les estimations des effets sont assez mauvaises (corrélation avec les vraies valeurs inférieure à 0.67) et pour =0.10, l'écart types des solutions des effets de SNP vrai et nulle est le même.
- Maintenant, ce n'est pas très étonnant. Avec π =0.10
  - nombreux effets et donc une variance d'effet très faible (0.0011)
  - Dans ce contexte, la proportion d'effets repérables avec notre faible effectif de chevaux est faible (si on fixe un seuil de visibilité de l'effet à 0.08, cela ne fait qu'une quinzaine de SNP potentiellement visibles sur notre échantillon). Sur l'échantillon simulé, 10 effets de SNP ont une valeur absolue supérieure à 0.08 et dans ce cas la corrélation estimation/vraie valeur est de 0.94.
- -> nombre maximum de SNP identifiables en fonction de la taille de l'échantillon

dépend de la taille de l'échantillon

$$SNPvisible = 2\left(1-\Phi\left[\frac{0.08}{\sqrt{\sigma_a^2/\left(0.5K(1-\pi)\right)}}\right]\right)K(1-\pi)$$

Nombre de SNP détectables (>0.08)

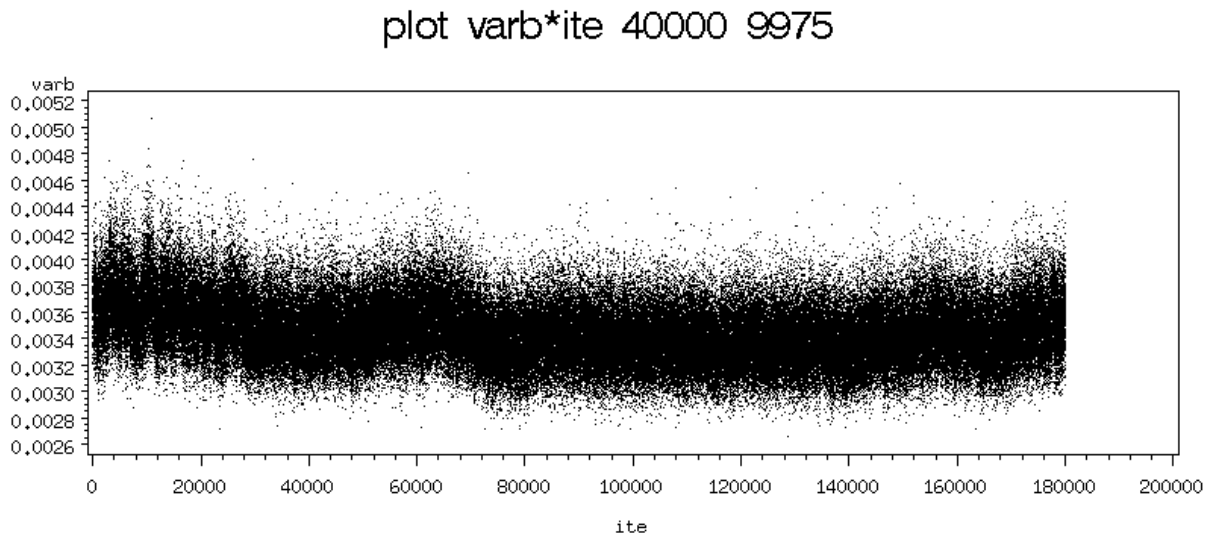


Cas 1000 SNP

- Fonction qui a un maximum autour de 50 pour un $K(1-\pi)$ autour 200 soit quand K=1000, $\pi = 0.8$ En dessous et au dessus, il y a finalement moins de SNP visibles (42 pour $\pi = 0.9$) Il serait donc inutile de tester l'efficacité de la méthode dans les cas ou π est nettement inférieur à 0.8 car on chercherait des choses introuvables.
- Ce raisonnement se tient sur ce cas de 1000 SNP avec $\pi >= 0.8$ qui donne des résultats convenables

Cas 40000 SNP

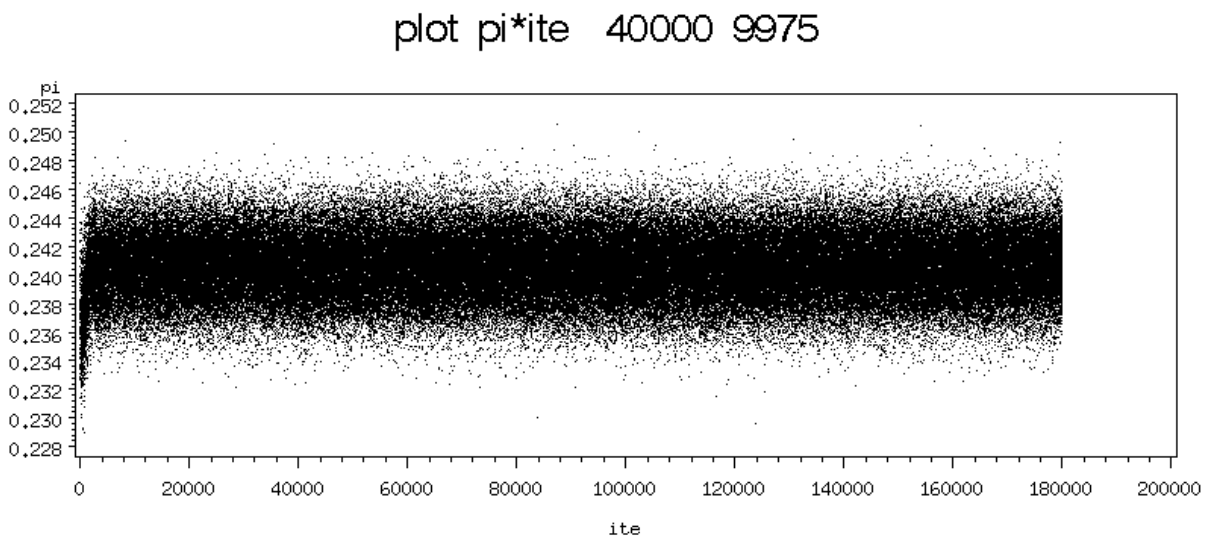
- Compte tenu de ce qui précède on a simulé avec $\pi = 0.9975$
- (soit 100 SNP avec effet espéré, 103 dans notre simulation)
- Convergence OK



Cas 40000 SNP

- MAIS en fait on ne converge pas vers les valeurs simulées (exemple π valeur de départ 0.60)

0.25

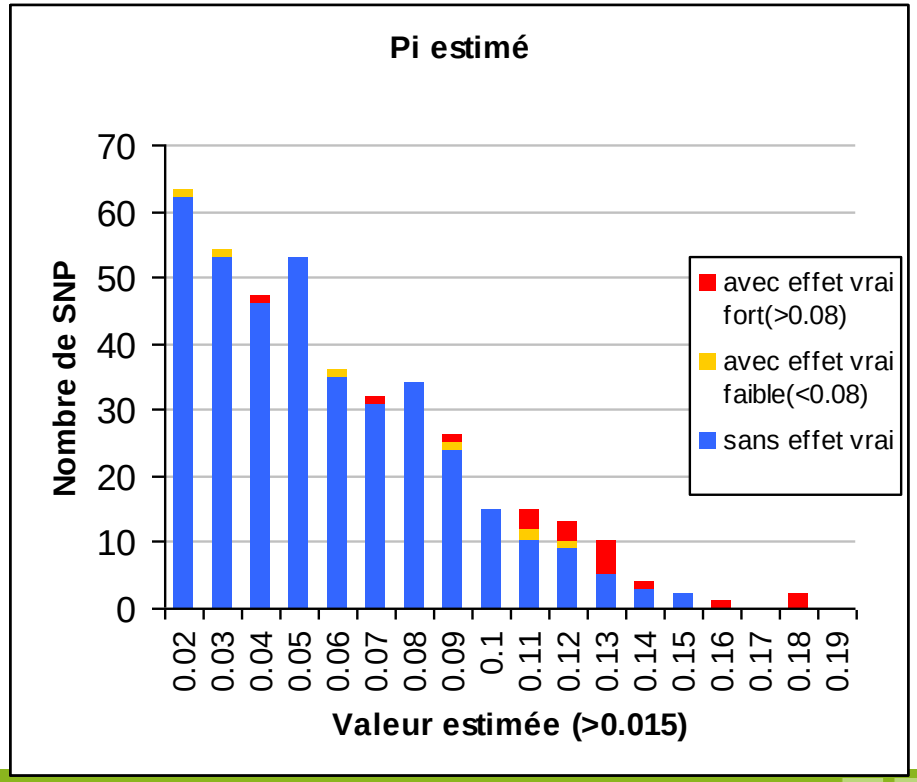


0.23

Cas 40000 SNP

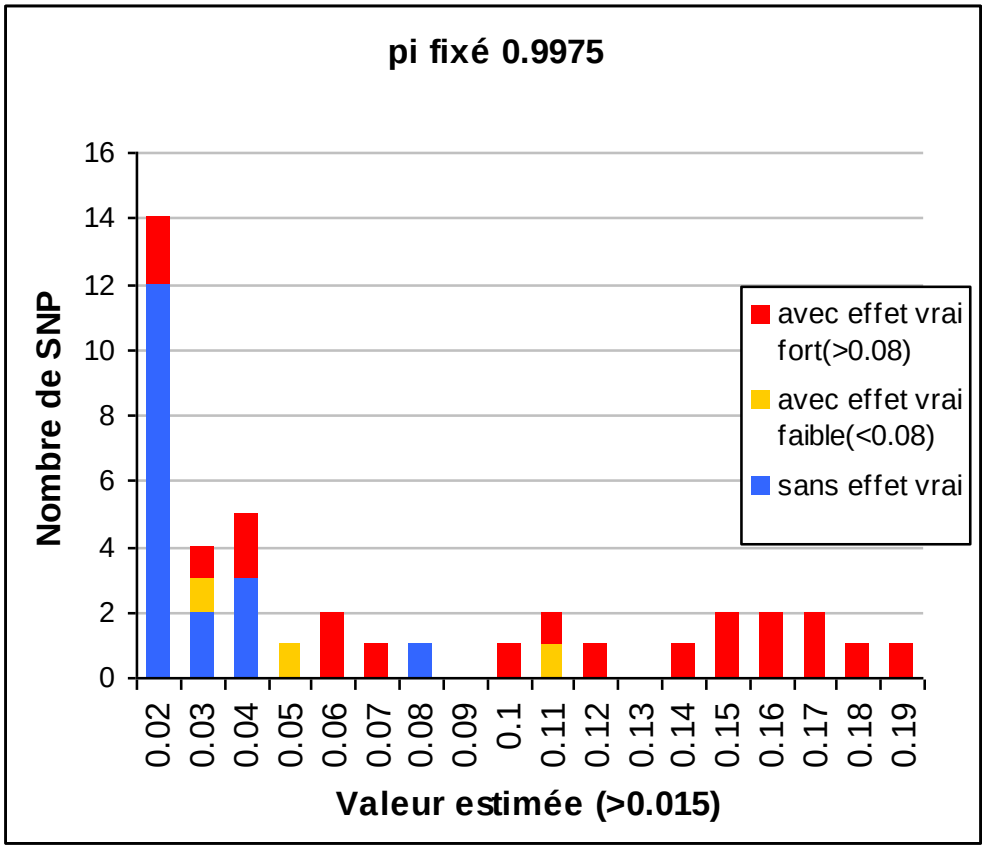
- Le système cherche à remplir complètement la variance expliquée par les effets des SNP et finit toujours par trouver la combinaison miracle (même en partant de la « vraie » valeur 0.9975 on converge vers 0.9860) -> la variance résiduelle est nulle

Histogramme des effets estimés >0.015



Cas 40000 SNP

- Il faut donc fixer π



ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

- Fait à peine mieux qu'une régression simple SNP/SNP ?
- Si je prends les meilleures 1‰ valeurs estimées avec une régression simple (les meilleures p-values) et les 1‰ valeurs moyennes les plus élevées pour le Bayes C on a :
  - 53% (regression SNP) de faux positifs contre 43% (Bayes)
  - 18% de trouvés vrais (regression) contre 22%(Bayes)

Conclusion

- Logiciel simple d'utilisation et rapide
- Ne pas chercher à estimer π
  - Le fixer à une valeur « raisonnable » en fonction de la proportion max de SNP identifiables dans notre échantillon (partage correct de l'héritabilité)
- Nos conclusions sur l'efficacité de la méthode ne portent pas sur une évaluation génomique qui cherche autre chose (la ressemblance entre apparenté, même si on se trompe sur les SNP), ne testent pas le principal intérêt de la méthode (la redondance entre SNP)…

La suite dans : http://pluton.toulouse.inra.fr/trac-gabayes

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA