Prédiction génomique en grande dimension : densité de marqueurs nécessaire pour une prédiction fiable en sélection génomique

Charles-Elie Rabier [1] & Simona Grusea [2]

[1] *IMAG, Université de Montpellier, CNRS, France*
charles-elie.rabier@umontpellier.fr
[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, INSA de Toulouse, France*
grusea@insa-toulouse.fr

Résumé. La sélection génomique (GS) consiste à sélectionner des individus sur la base de prédictions génomiques effectuées à l'aide d'une grande densité de marqueurs. Une question d'importance en GS est de déterminer le nombre de marqueurs nécessaires pour une prédiction fiable. Pour ce faire, nous introduisons de nouveaux proxies pour la précision de la prédiction. Ces proxies sont appropriés dans le cadre d'une carte génétique discrète, où il est fréquent d'observer du déséquibre de liaison incomplet, i.e. la situation où les allèles à l'emplacement d'un gène et à l'emplacement d'un marqueur à proximité, diffèrent. De plus, nos proxies suggérés sont utiles pour concevoir des puces SNPs basées sur une densité modérée de marqueurs. Nous analysons des données de riz des Philippines, et nous nous focalisons sur la date de floraison recueillie durant la saison sèche 2012. En utilisant différentes densités de marqueurs, nous montrons qu'au moins 1553 marqueurs sont nécessaires afin d'obtenir une prédiction fiable et de mettre en place la GS. Déterminer le nombre optimal de marqueurs est crucial afin d'optimiser le programme de sélection.

Mots-clés. Statistique en grande dimension, Prédiction, Régression Ridge, Sélection génomique, Déséquilibre de liaison incomplet, Données de riz

Abstract. Genomic selection (GS) consists in predicting breeding values of selection candidates, using a large number of genetic markers. An important question in GS is to determine the number of markers required for a good prediction. For this purpose, we introduce new proxies for the accuracy of the prediction. These proxies are suitable under sparse genetic map where it is likely to observe some imperfect linkage disequilibrium, i.e. the situation where the alleles at a gene location and at a marker located nearby vary. Moreover, our suggested proxies are helpful for designing cost-effective SNP chips based on a moderate density of markers. We analyze rice data from Los Banos, Philippines and focus on the flowering time collected during the dry season 2012. Using different densities of markers, we show that at least 1553 markers are required to implement GS. Finding the optimal number of markers is crucial in order to optimize the breeding program.

Keywords. High Dimensional Data Analysis, Prediction, Ridge Regression, Genomic Selection, Imperfect Linkage Disequilibrium, Rice Data Analysis

Cet acte de congrès se veut un résumé de la publication Rabier et Grusea (2021).

1 Introduction

La sélection génomique (GS) est une méthode extrêmement populaire en génétique (Meuwissen et al., 2001), consistant à sélectionner des individus sur la base de prédictions génomiques. Ces prédictions, effectuées à l'aide d'une grande densité de marqueurs génétiques couvrant le génome, se doivent d'être précises afin de pouvoir sélectionner les meilleurs candidats pour le programme de sélection. La GS a été tout d'abord appliquée aux animaux et est désormais largement employée chez les plantes. On peut citer par exemple, les études sur la pomme (Muranty et al., 2015), l'eucalyptus (Tan et al., 2017), les poires japonaises (Minamikawa et al., 2018), les fraises (Gezan et al., 2017), la banane (Nyine et al., 2018) et le café (Ferrao et al., 2018).

Rappelons que l'on nomme QTLs, les loci sur le génome codant pour un caractère (i.e. trait ou phénotype) quantitatif. D'un point de vue méthodologique, la GS repose sur l'espoir que chaque QTL sera étroitement corrélé avec au moins un des marqueurs, en raison de la grande densité de marqueurs. En génétique, cette corrélation entre un QTL et un marqueur est nommée Déséquilibre de Liaison (DL) : cela correspond à la non indépendance des allèles à 2 loci différents (cf. Durett, 2008). Contrairement aux études d'associations à l'échelle du génome (GWAS) qui recherchent des QTLs, l'objectif de la GS est d'effectuer des prédictions à l'aide d'un grand nombre de marqueurs, sans avoir à détecter les QTLs. L'avantage des prédictions génomiques en GS, par rapport aux prédictions basées sur les loci détectés par GWAS, réside dans le fait que les QTLs à petits effets sont très difficiles à détecter pour la plupart des traits, qualifiés de traits complexes car gouvernés par beaucoup de QTLs à petits effets.

De nombreux facteurs sont connus pour influencer la qualité des prédictions en GS : la taille du jeu d'entrainement sur lequel le modèle est appris, l'héritabilité du caractère liée au rapport signal/bruit, la densité de marqueurs, le DL, le lien entre jeux d'entrainement et de validations ... Dans Rabier et al (2018), nous nous étions interessés uniquement au DL complet : les QTLs étaient localisés uniquement sur quelques marqueurs. Cependant, lorsque les QTLs ne coïncident pas avec les positions des marqueurs, les allèles au QTL et au marqueur à proximité, diffèrent. On parle dès lors de DL incomplet. Ainsi, l'objectif de ce travail est d'étudier la prédiction génomique dans le cadre d'un DL incomplet.

Une thématique de recherche connexe est la détermination du nombre de marqueurs requis pour mettre en place la GS. Dans leur étude sur le maïs, Zhang et al (2015) ont montré que la prédiction d'un trait complexe nécessitait un grand nombre de marqueurs génétiques (environ 58000 marqueurs par GBS), alors que 200 marqueurs étaient suffisants pour prédire un trait simple. D'autre part, dans une étude sur le café, Ferrao et al (2019) ont montré que les prédictions basées sur 4000 marqueurs donnaient les mêmes résultats que celles basées sur 35000 marqueurs. Pour finir, en aquaculture, Kriaridou et al. (2020)

ont obtenu des résultats similaires en comparant 1000 marqueurs et une haute densité de marqueurs. Un sujet également proche s'avère la conception de puces LD (e.g. Bolormaa et al. 2015, Corbin et al., 2014). Notre travail pourrait s'avérer utile pour concevoir des puces LD qui réduisent les coûts dus au génotypage en considérant une densité modérée de marqueurs, contrairement aux puces haute densité (Wu et al, 2016).

1.1 Modèle linéaire causal en présence de DL incomplet

Introduisons tout d'abord le modèle statistique causal. Nous nous intéressons à un trait quantitatif (i.e. un phénotype) qui est observé sur un échantillon d'entraînement de taille n. $Y_1, \ldots, Y_n$ désignent les variables aléatoires relatives à ce trait quantitatif. On considère qu'il existe m QTLs sur le génome, ayant une influence sur le trait quantitatif.

Pour $1 \leq j \leq m$, on note $\beta_j^\star$ le $j$ème effet QTL. $X^\star$ désigne la matrice de taille $n \times m$ contenant les allèles aux QTLs des n individus. La $i$ème ligne de $X^\star$, qui s'écrit $\boldsymbol{x_i}' = \left(X_{i,1}^\star, ..., X_{i,p}^\star\right)$, correspond au $i$ème individu, en utilisant $'$ pour symboliser la transposée.

On considère le modèle linéaire causal suivant pour le trait quantitatif :

$$\boldsymbol{Y} = X^\star \boldsymbol{\beta^\star} + \boldsymbol{\varepsilon}, \tag{1}$$

où $\boldsymbol{Y} = (Y_1, ..., Y_n)'$, $\boldsymbol{\beta^\star} = (\beta_1^\star, ..., \beta_m^\star)'$, $\boldsymbol{\varepsilon} \sim N(0, \sigma_e^2 I_n)$, I_n est la matrice identité de taille n et σ_e^2 représente la variance environnementale.

On suppose que $X^\star$ et $\boldsymbol{\varepsilon}$ sont indépendants. De plus, l'information génomique est disponible à p marqueurs, avec $p > n$ (contexte de grande dimension). A noter que les QTLs **ne coïncident pas forcément** avec les marqueurs (cadre du DL incomplet).

Soit X la matrice de taille $n \times p$ contenant l'information génomique aux p marqueurs pour les n individus. Rappelons que dans notre étude précédente, nous avions considéré le DL complet, où les QTLs se trouvent parmi les marqueurs. Dans ce cas, le modèle considéré s'écrivait :

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

avec $1 \leq j \leq p$, $\beta_j = 0$ si le marqueur n'est pas un QTL, et $\beta_j \neq 0$ si le jème marqueur est un QTL.

Remark 1 *A noter que le DL complet peut être vu comme un cas particulier du DL incomplet. C'est le cas lorsque les m QTLs sont localisés sur quelques marqueurs et dans ce cas, $\boldsymbol{\beta}$ représente le vecteur sparse de taille p, contenant les composantes de $\boldsymbol{\beta^\star}$.*

1.2 Individu Test (TST)

On suppose que pour un individu aléatoire supplémentaire, dit individu test (TST), noté *new*, on dispose de son information génomique (i.e. ses allèles aux différents marqueurs)

mais pas de son phénotype. Soit $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$ le vecteur colonne contenant les allèles aux m QTLs de l'individu new. Le trait quantitatif Y_{new} vérifie :

$$Y_{new} = \boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\, \boldsymbol{\beta}^{\star} + \varepsilon_{new},$$

où $\varepsilon_{new} \sim N(0, \sigma_e^2)$, et $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$, ε_{new} et $\boldsymbol{\varepsilon}$ sont tous indépendants.

De plus, $\boldsymbol{x}_{\boldsymbol{new}}$ fait référence au génome aléatoire aux marqueurs. A noter que $\boldsymbol{x}_{\boldsymbol{new}}$ et $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$ sont corrélés en raison de la liaison génétique due à la taille fixe du génome.

Dans ce qui suit, nous assumerons que $\boldsymbol{Y}$, Y_{new}, $\boldsymbol{x}_{\boldsymbol{new}}$, $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$, les colonnes de X et les colonnes de $X^{\star}$ sont centrés.

1.3 Modèle de prédiction et précision

Construisons un estimateur $\hat{Y}_{new}$ pour la valeur phénotypique de l'individu new, à l'aide de l'information génomique disponible pour les n individus d'entrainement.

Prédicteur basé sur la régression Ridge
Nous utiliserons comme modèle instrumental, la régression Ridge qui peut être vue comme une régression Bayésienne. L'estimateur de la régression Ridge pour les effets marqueurs $\boldsymbol{\beta}$ (Tikhonov 1963, Hoerl et Kennard 1970) s'écrit :

$$\hat{\boldsymbol{\beta}} = \left(X'X + \lambda I_p\right)^{-1} X'\boldsymbol{Y} = X'V^{-1}\boldsymbol{Y}, \tag{3}$$

où $V = XX' + \lambda I_n$, et où λ et I_p désignent respectivement le paramètre de régularisation, et la matrice identité de taille $p \times p$.

A noter que l'estimateur Ridge est basé sur l'information génomique aux marqueurs et est approprié dans le cadre de la grande dimension (i.e. $p > n$, cf. e.g. Shao and Deng 2012, et Bühlmann 2013). La prédicteur s'écrit $\hat{Y}_{new} = \boldsymbol{x}_{\boldsymbol{new}}{}'\hat{\boldsymbol{\beta}}$.

Critère de précision
En génétique, la capacité prédictive est quantifiée selon la *précision phénotypique*, ρ_{ph} (e.g. Visscher et al., 2010) ou la *précision génotypique*, ρ_g (e.g. Daetwyler et al., 2008 et 2010):

$$\rho_{ph} := \frac{\mathrm{Cov}\left(\hat{Y}_{new},\, Y_{new}\right)}{\sqrt{\mathrm{Var}\left(\hat{Y}_{new}\right)\mathrm{Var}\left(Y_{new}\right)}}\,,\ \rho_g := \frac{\mathrm{Cov}\left(\hat{Y}_{new},\, \boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\boldsymbol{\beta}^{\star}\right)}{\sqrt{\mathrm{Var}\left(\hat{Y}_{new}\right)\mathrm{Var}\left(\boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\boldsymbol{\beta}^{\star}\right)}}\,. \tag{4}$$

En particulier, cette corrélation est une composante essentielle dans l'équation du sélectionneur (cf. Lynch and Walsh, 1998). A noter que dans Rabier et Grusea (2021), nous nous intéressons également à l'erreur quadratique, plus commune dans la communauté statistique.

Héritabilité
Comme $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$, ε_{new} et $\boldsymbol{\varepsilon}$ sont tous supposés indépendants, nous avons la relation $\rho_{ph}/\rho_g = h$, où h est la racine de carré de l'héritabilité du trait :

$$h^2 := \frac{\mathrm{Var}\left(\boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\,\boldsymbol{\beta}^{\star}\right)}{\mathrm{Var}\left(Y_{new}\right)}\,. \tag{5}$$

Avec la notation $\sigma^2_G = \mathrm{Var}\left(\boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\,\boldsymbol{\beta}^{\star}\right)$, nous avons la relation $h^2 = \sigma^2_G/\left(\sigma^2_G + \sigma^2_e\right)$.

1.4 Contributions

Dans Rabier et Grusea (2021), nous montrons que la projection de la fonction de régression $X^{\star}\boldsymbol{\beta}^{\star}$ sur le sous-espace linéaire engendré par les colonnes de X, noté $\mathrm{CS}(X)$, est un élément clé pour la précision génotypique. A partir de ces résultats théoriques, il est possible de retouver les résultats obtenus sous le DL complet : le facteur clé devient la projection du signal $\boldsymbol{\beta}$ sur le sous-espace linéaire engendré par les lignes de X, noté $\mathrm{RS}(X)$. Assumant que le signal $\boldsymbol{\beta}^{\star}$ est réparti uniformément sur chaque sous-espace, nous montrons également au travers d'un lemme que la précision optimale (oracle) est atteinte dès que la limite d'un facteur de perte, dû au DL incomplet, est égal à 0.

Par la suite, nous introduisons un prédicteur modifié, inspiré de la régression Ridge, et qui améliore la qualité de la prédiction. Il repose sur la projection de $\boldsymbol{Y}$ sur un sous-espace bien choisi de $\mathrm{CS}(X)$. Nous étudions la précision de ce prédicteur : comme attendu, en présence de DL incomplet, la qualité de la prédiction dépend de la projection de la fonction de régression $X^{\star}\boldsymbol{\beta}^{\star}$ sur le sous-espace choisi. Nous présentons des résultats qui permettent de comparer la précision liée à l'estimateur Ridge et celle liée au prédicteur modifié, dans un contexte de DL incomplet.

D'autre part, nous proposons de nouveaux proxies reposant sur nos résultats théoriques. Les performances de ces proxies seront illustrés sur des données simulées et réelles. Nos proxies dépendent uniquement des phénotypes et des marqueurs des individus d'entrainement, mais prennent en compte la sparsité de la carte génétique des individus TST. Ainsi, afin de construire ces proxies, nous nous attaquons à un nouveau sujet en GS : la précision de la prédiction des individus TST lorsque la carte génétique des individus d'entrainement diffère de celle des TST. En particulier, nous suggérons de considérer une carte plus dense pour les individus d'entrainement que pour les TST : la carte dense nous aidera à estimer les paramètres de nuisance $X^{\star}$ et $\boldsymbol{\beta}^{\star}$ requis pour calculer nos proxies. Ce concept repose sur l'espoir que les QTLs se trouveront en DL complet sous la carte dense d'entrainement, ce qui n'est pas le cas pour la carte TST (DL incomplet). Contrairement à notre étude portant sur le DL complet (Rabier et al, 2018) où l'Adaptive LASSO (Zou, 2006) était le meilleur substitut pour $\boldsymbol{\beta}$, nous avons trouvé ici que le LASSO (Tibshirani, 1996) s'avérait le meilleur substitut pour $\boldsymbol{\beta}^{\star}$ lorsque la carte discrète TST était considérée.

Pour finir, nous analysons des données de riz des Philippines (Spindel et al., 2015), et nous nous focalisons sur la date de floraison recueillie durant la saison sèche 2012.

En considérant différentes densités de marqueurs et nos nouveaux proxies appropriés en présence de DL incomplet, nous montrons que les généticiens peuvent évaluer la précision des futures prédictions. Si la précision n'est satisfaisante, nous leur conseillons de re-densifier leur carte génétique à l'aide de marqueurs supplémentaires, afin d'améliorer la fiabilité des prédictions.

Bibliographie

[1] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), 1212-1242.

[2] Ferrao, L.F.V., Ferrao, R.G., Ferrao, M.A.G., Fonseca, A., Carbonetto, P., Stephens, M., and Garcia, A.A.F. (2019). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity*, 122(3), 261-275.

[3] Gezan, S.A., Osorio, L.F., Verma, S., and Whitaker, V.M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Horticulture research*, (4), 16070.

[4] Hoerl AE, and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

[5] Minamikawa, M.F, Takada, N., Terakami, S., Saito, T., Onogi, A., Kajiya-Kanegae, H. ..., and Iwata, H. Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (Pyrus pyrifolia Nakai). *Scientific reports*, 8(1), 11994.

[6] Rabier, C.E., Grusea, S. (2021), Prediction in high-dimensional linear models and application to genomic selection under imperfect linkage disequilibrium. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(4), 1001-1026.

[7] Rabier, C. E., Mangin, B., Grusea, S. (2019). On the accuracy in high-dimensional linear models and its application to genomic selection. *Scandinavian journal of statistics*, 46(1), 289-313.

[8] Shao, J., Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of statistics*, 40(2), 812-831.

[9] Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., and Redoña, E., et al (2015). Genomic Selection and Association Mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics*, 11(2), e1004982.

[10] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 267-288.

[11] Tikhonov, A.N. (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk.*, (151), 501-504..

[12] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.