HAL
archives-ouvertes.fr

Selective genotyping pour la détection de QTL

Charles-Elie Rabier, Jean-Marc Azais

▶ **To cite this version:**

Charles-Elie Rabier, Jean-Marc Azais. Selective genotyping pour la détection de QTL. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. 2009. <inria-00386740>

HAL Id: inria-00386740
https://hal.inria.fr/inria-00386740

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

"Selective genotyping" pour la détection de QTL

Charles-Elie Rabier[1,2] & Jean-Marc Azais[1]

[1] *Laboratoire de Statistiques et Probabilités, Université Paul Sabatier Toulouse, Institut de Mathématiques de Toulouse*
118 route de Narbonne, F-31062 Toulouse Cedex 9
azais@cict.fr
[2] *Station d'Amélioration Génétique des Animaux*
INRA, Auzeville B.P. 52627, 31326 Castanet Tolosan
charles-elie.rabier@toulouse.inra.fr

Résumé : En général, dans une expérience visant à mettre en évidence des gènes à effets quantitatifs, tous les individus d'une population sont mesurés et génotypés. La stratégie, qui consiste à mesurer plus d'individus et à ne génotyper que les extrêmes pour le caractère quantitatif intéressant, est nommée "Selective genotyping". Ainsi, le nombre d'individus génotypés, afin d'obtenir une puissance donnée, est réduit considérablement, à condition que le nombre d'individus phénotypés ait été augmenté. On se propose ici d'étudier les propriétés statistiques du "Selective genotyping".

Abstract : Usually, in an experiment deigned for detecting a gene responsible for the variation of a quantitative trait, all the individuals are genotyped and the phenotypes of the individuals are measured. "Selective genotyping" is an experimental design which consists of genotyping only individuals whose phenotype of interest is extreme. It allows to reduce the costs due to genotyping and to keep a good power for the statistical test provided that the number of individuals has been increased. The work presented here is a study on the statistical properties of "Selective genotyping".

Mots-clés : Génome, Détection de QTL, Processus expérimentaux, Tests Statistiques.
Keywords : Genome, QTL detection, Experimental design, Statistical tests.

1 Motivation

Les nouvelles technologies en matière de génomique se révèlent être efficaces afin de percer les secrets de la variation génétique d'un caractère quantitatif. Ces technologies permettent la caractérisation moléculaire de marqueurs polymorphes (i.e. présentant plusieurs allèles) sur l'ensemble du génome. Ces derniers seront par la suite utilisés pour identifier et localiser les loci (i.e. emplacements physiques précis sur un chromosome) où la variation allélique est associée à la variation du caractère quantitatif considéré. On nomme QTL de tels loci. Néanmoins, les coûts dûs au génotypage demeurent très élevés.

C'est pourquoi l'optimisation du processus expérimental est primordiale. L'un de ces processus expérimentaux s'intitule selective genotyping. Il a été proposé par Lebowitz and al. (1987), et élaboré par Lander et Botstein (1989), Darvasi et Soller (1992), puis Muranty et Goffinet (1997). Le selective genotyping consiste à génotyper uniquement les individus dont la valeur du caractère quantitatif est extrême (plus grande ou plus petite qu'un seuil). Cela permet de réduire les coûts dûs au génotypage tout en gardant une bonne puissance pour le test statistique, à condition que le nombre d'individus ait été augmenté.

2 Analyse théorique du selective genotyping

2.1 L'étude dans sa globalité

Deux stratégies différentes pour l'analyse statistique en selective genotyping ont tout d'abord été étudiées. La première consistait à conserver dans l'analyse statistique, tous les phénotypes, même les phénotypes qui ne sont pas considérés comme extrêmes et pour lesquels nous ne disposons pas du génotype. La seconde stratégie était basée sur la conservation uniquement des phénotypes extrêmes. Pour ces deux stratégies, les tests de Wald associés, ainsi que la puissance de ces tests sous des alternatives contigues ont été considérés. Cependant, toutes ces statistiques de test ne sont pas explicites dû à la difficulté d'obtenir l'estimateur du maximum de vraisemblance (EMV) : l'EMV est obtenu par l'algorithme EM pour la stratégie 1 et la méthode de Newton pour la stratégie 2. Par conséquent, l'intérêt a été porté par la suite sur une troisième stratégie reposant sur l'utilisation d'une statistique de test (explicite) qui est une simple comparaison de moyenne basée uniquement sur les phénotypes extrêmes. Tous ces tests ont été comparés en terme d'efficacité au test oracle, celui où tous les génotypes sont connus.

2.2 Modèle correspondant au test oracle

Soit N le nombre total d'individus. Soit X la variable aléatoire correspondant au génotype au QTL. On considérera 2 génotypes possibles au QTL pour une observation k :

$$X_k = \begin{cases} -1 & \text{avec probabilité } p_1^- \\ 1 & \text{avec probabilité } p_1 \end{cases}$$

Les variables aléatoires $(X_k)_{1 \leq k \leq N}$ seront considérées indépendantes et équidistribuées.

Soit Y la variable aléatoire correspondant au phénotype. Le modèle pour une observation

Y_k s'écrit :

$$Y_k = \mu + qX_k + \epsilon_k$$

les variables aléatoires $(\epsilon_k)_{1\leq k\leq N}$ étant indépendantes, de loi normale de moyenne 0 et de variance σ^2. On supposera μ et σ^2 connus.

2.3 Modèle correspondant au selective genotyping

On adopte le même modèle que pour le test oracle. Seule différence, on n'observe plus X_k mais $\overline{X}_k$ définie de la manière suivante :

$$\overline{X}_k = \begin{cases} X_k & \text{si } Y_k \notin [S_-\ ,\ S_+] \\ 0 & \text{sinon} \end{cases}$$

où S_- et S_+ sont deux réels tels que $S_- \leq S_+$.

2.4 Efficacité des différents tests

Afin de tester la présence de QTL, on confronte les 2 hypothèses suivantes :

$$H_0 :\ q = 0\ \ vs\ \ H_1 :\ q \neq 0$$

Il peut être intéressant de regarder comment évolue la puissance du test lorsque l'on augmente le nombre d'individus, tout en gardant la même valeur de l'effet qtl q. N_{sg} désignera le nouveau nombre d'individus et on définit le ratio $\eta = \frac{N_{sg}}{N}$. On définit l'efficacité d'un test 2 relativement à un test 1, $\kappa = \frac{1}{\eta_{eff}}$ où η_{eff} désigne la valeur de η pour lequel la puissance du test 2 est égale à celle du test 1. Ici, le test oracle, qui consiste en une simple comparaison de moyenne, servira de test de référence. On nommera κ_1, κ_2 et κ_3 les efficacités correspondant respectivement aux stratégies 1, 2 et 3 énoncées en section 2.1. On notera φ la densité d'une nomale centrée réduite.

Théorème : Supposons $P_{H_0}\left(Y_k \notin [S_-,\ S_+]\right) = \gamma\ ,\quad P_{H_0}\left(Y_k > S_+\right) = \gamma_+$

$P_{H_0}\left(Y_k < S_-\right) = \gamma_-\ ,\quad z_{\gamma_-} = \dfrac{S_- - \mu}{\sigma}\ ,\quad z_{\gamma_+} = \dfrac{S_+ - \mu}{\sigma}$

alors,

$\kappa_1 = \gamma\ +\ z_{\gamma_+}\varphi(z_{\gamma_+})\ -\ z_{\gamma_-}\varphi(z_{\gamma_-})\ +\ (p_1 - p_1^-)^2\ \left\{1 - \gamma - z_{\gamma_+}\varphi(z_{\gamma_+}) + z_{\gamma_-}\varphi(z_{\gamma_-})\right\}$

$\kappa_2 = \gamma\ +\ z_{\gamma_+}\varphi(z_{\gamma_+})\ -\ z_{\gamma_-}\varphi(z_{\gamma_-})\ +\ \dfrac{(p_1 - p_1^-)^2}{1\ -\ \gamma}\left\{\varphi(z_{\gamma_-})\ -\ \varphi(z_{\gamma_+})\right\}^2\quad \forall\gamma \neq 1$

$\kappa_3 = 4\, p_1\, p_1^-\ [\gamma\ -\ z_{\gamma_-}\ \varphi(z_{\gamma_-})\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})]$

$\kappa_1 = \kappa_2 = \kappa_3 \Leftrightarrow p_1 = p_1^- = \dfrac{1}{2}$

En conclusion, lorsque $p_1 = p_1^- = \frac{1}{2}$, les 3 stratégies sont équivalentes et par conséquent, l'apport des phénotypes non extrêmes dans l'analyse statistique n'apporte alors aucun gain de puissance.

Bibliographie

[1] Botstein E.S., Lander D. (1989), Mapping mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics.
[2] Darvasi A., Soller M. (1992), Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus, Theoretical and Applid Genetics.
[3] Cierco C. (1996), Problèmes statistiques liés à la détection et à la localisation d'un gène à effet quantitatif, Thèse de l'université Paul Sabatier, Toulouse.
[4] Muranty H., Goffinet B. (1997), Selective genotyping for location and estimation of the effect of a quantitative trait locus, Biometrics, vol. 53.
[5] Wu R., Ma C.X., Casella G. (2007), Statistical genetics of quantitative traits, Springer.
[6] van der Vaart, A.W. (1998), Asymptotic Statistics, Cambridge series in statistical and probabilistic mathematics.