Apports méthodologiques Amélioration de l'estimation du mérite génétique et prise en compte, en détection de QTL des caractères à seuils

1 Prise en compte des performances mesurées sur les générations postérieures à la troisième génération en détection de QTL avec analyse de liaison

1.1 Introduction

La détection de QTL par l'analyse de liaison est aujourd'hui une technique largement documentée dans la littérature (Soller et Genizi, 1978 ; Lander et Botstein, 1989 ; Weller *et al.,* 1990). Le principe et les modèles statistiques utilisés sont décrits dans le chapitre I. Elle consiste à observer, intra-famille de père (ou mère, génération G1), si les performances des descendants (génération G2) sont significativement différentes entre groupes définis par la combinaison allélique reçue du père aux marqueurs informatifs flanquant la position testée. Dans l'approche par régression (Haley et Knott, 1992), les performances des descendants sont régressées par les probabilités de transmission des allèles parentales au QTL sachant les informations aux marqueurs flanquants. Avec ce même modèle, il est possible de regrouper le mérite génétique des descendants, calculé à partir des performances mesurées sur les petits-descendants (G3), à la place de leurs performances propres si celles-ci ne sont pas mesurables. Cette modélisation correspond au protocole « petites-filles » (Weller *et al.,* 1990). L'intérêt de ce protocole est multiple. Il permet notamment d'augmenter la puissance de détection de QTL (paragraphe 3.3 du chapitre I). Tribout *et al.* (2008) ont étendu ce dispositif au cas où les descendants possèdent des performances propres.

Dans le calcul du mérite génétique des descendants génotypés, il peut être envisagé de prendre en compte, en plus de leurs performances et celles des petits-descendants, les performances des générations suivantes (G4, G5,…), quand elles sont disponibles.

Nous proposons, dans un article accepté par *Journal of Animal Breeding and Genetics*, une méthode permettant de prendre en compte ces « nouvelles » performances dans la détection de QTL. Cette méthode renforce la puissance des dispositifs de détection de QTL. Cela est particulièrement vrai pour les protocoles où les effectifs par familles sont réduits, comme c'est souvent le cas en génétique humaine ou pour des espèces où l'insémination artificielle n'est pas totalement maitrisée. Toutefois, les intervalles entre générations doivent être suffisamment courts pour que la mise en place d'un tel dispositif soit pratique.

Nous montrons comment obtenir la meilleure combinaison linéaire de performances de la descendance (G2, G3, …) qu'il convient d'analyser dans un modèle de régression de type Haley/Knott comme « performance » étendue des individus de la deuxième génération (G2) génotypés. Nous analysons les relations entre ces combinaisons linéaires et l'estimation BLUP des valeurs génétiques de ces individus G2.

1.2 Article 1

QTL detection from regression analysis of “generalized de-regressed proof” information

M. Kileh Wais[1, 2], J.M. Elsen[1]

[1] INRA, UR 631, Station d’Amélioration Génétique des Animaux, 31326 Castanet Tolosan, France
[2] University of Djibouti, Centre de Recherche Universitaire de Djibouti, Djibouti

Email : Mohamed Kileh Wais – MohamedKileh@gmail.com ; Jean-Michel Elsen - Jean-Michel.Elsen@toulouse.inra.fr

Soumis et accepté par Journal of Animal Breeding and Genetics

Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

SHORT COMMUNICATION

QTL detection from regression analysis of 'generalized de-regressed proof' information

M. Kileh Wais[1,2] & J.M. Elsen[1]

1 INRA, UR 631, Station d'Amélioration Génétique des Animaux, Castanet Tolosan, France
2 Université de Djibouti, Centre de Recherche Universitaire de Djibouti, Djibouti

Keywords
De-regressed proof; genetic merit; QTL; granddaughter designs.

Correspondence
M.K. Wais, INRA, UR 631, Station d'Amélioration Génétique des Animaux, 31326 Castanet Tolosan, France. Tel: 33 561285376; Fax: 33 561285353; E-mail: mohamedkileh@gmail.com

Received: 4 July 2011;
accepted: 19 November 2011

Summary

QTL detection using the regression of phenotypes on transmission probability is largely used when large families are available. In three generations designs, the use of a 'de-regressed proof' as a phenotype to be analysed was proposed by Weller *et al.* (1990) and Tribout *et al.* (2008). Our work generalizes this approach. A score (that we define as a 'generalized de-regressed proof') is described, which combines performance phenotypes recorded in multigenerational offspring of genotyped individuals. Estimation of the QTL effect on this score with a simple regression is unbiased. The link between this score and the BLUP animal model of the polygenic effect is demonstrated. The theory is developed and two simple examples illustrate how this technique can be implemented.

Introduction

Within-family analysis of the linkage between quantitative performances and the transmissions of marker alleles is a commonly used technique for QTL detection (Soller *et al.* 1976; Soller & Genizi 1978; Lander & Botstein 1989; Weller *et al.* 1990; Haley *et al.* 1994). Even if, because of the availability of large-scale SNP genotyping techniques, linkage disequilibrium analysis is now possible and powerful in many species, the within-family linkage approach is still attractive both for its robustness and easiness. Various models were proposed, including regression techniques (Haley & Knott 1992), maximum likelihood mixture analyses (Elsen *et al.* 1999) and variance components analyses (e.g. Uimari & Hoeschele 1997; George *et al.* 2000).

In regression approaches, the quantitative performances of half-sibs or full-sibs are related to the transmission, from the parents to the progeny, of markers flanking the position at which a segregating QTL is hypothesized: progeny quantitative performances are regressed on the QTL alleles transmission probabilities, given the marker information.

The standard design thus includes two generations: the parents (G1) tested to be heterozygous for a QTL, and their progeny (G2), in which both the quantitative trait and marker genotypes are determined.

This design can be enriched by assessing a third generation (the grand-offspring, G3). In this case, the marker genotypes of G2 individuals are related to the mean performance of their own G3 offspring. This was proposed by Weller *et al.* (1990) and largely implemented, in particular in dairy cattle, in so-called granddaughter designs (e.g. Zhang *et al.* 1998; Farnir *et al.* 2002; Bennewitz *et al.* 2003; Boichard *et al.* 2003).

In these designs, the performances of the third generation are summarized rather than analysed independently. The most routinely employed practical solution is to use, as pseudo-phenotypes for the genotyped G2 individuals their 'daughter yield deviation' (DYD), which are generally available with their 'estimated breeding values' (EBV) (*cf.* the

doi:10.1111/j.1439-0388.2011.00981.x

granddaughter designs mentioned previously). A few alternative, but less powerful, solutions have been proposed such as direct analysis of EBVs or 'de-regressed proofs' that is EBVs corrected for their precision (Thomsen *et al.* 2001), as also proposed by Lien *et al.* (1995) who used a 'regenerated right-hand side of mixed model equations'.

More recently, Tribout *et al.* (2008) extended the procedure to situations where performances are available in both G2 and G3 generations, with repetitions. The authors defined a 'genetic merit' or 'unregressed summary of own and progeny performances', a generalization of the use of DYD in granddaughter designs, as the phenotype to be analysed.

In practice, the precision of such summarized phenotypes is variable between G2 individuals. Taking account this heterogeneity of residual variances, the regression approach seems obvious, and currently available software (e.g. QTLExpress, Seaton *et al.* 2002; QTLMap, Filangi *et al.* 2010) can be used directly. Typically, precision is quantified by the squared correlation between the estimated and true genetic values and referred to as the determination coefficient or reliability (VanRaden & Wiggans 1991; Georges *et al.* 1995).

Our objective in this paper is to generalize this approach to any type of pedigree descendant from the G2. We define a 'generalized de-regressed proof' (GDRP) to be used as G2 individuals pseudo-phenotypes in simple regression approach. We describe how this GDRP is linked to BLUP animal model estimations.

Methods

Notations

A set of n_1 unrelated sire families ($i = 1 \ldots n_1$) was considered. The progeny ($j = 1 \ldots n_{2i}$) of sire i were born from different dams. The probabilities of i to j transmission of alleles for a putative QTL located at a given position, conditional to marker information, were estimated: $t_{j,2i-1}$(resp.$t_{j,2i}$) is the probability that i transmitted its first (resp. second) QTL allele to j. It should be noted that $t_{j,2i} = 1 - t_{j,2i-1}$. The corresponding sire i QTL allele effects are designated by a'_i and a''_i.

The performance trait of the animal k, say y_{ik}, (belonging to G2 or G3+) depends on QTL allele inherited from the sire i and on its polygenic value u_k and a residual e_k, both two quantities being considered as random normal variables. The effect of QTL alleles that are not inherited from G1 contributes to the residual effect.

One step model

In the simplest situation ($k = j$ is a direct progeny of parent i), the regression model is:

$$y_{ij} = t_{j,2i-1} \cdot a'_i + t_{j,2i} \cdot a''_i + u_{ij} + e_{ij}$$

In the case of G3 progeny ($k = jl$), with only one parent belonging to G2, the performance y_{ijl} of the progeny l ($l = 1 \ldots n_{3ij}$) of the sire j born to grandsire i is:

$$y_{jl} = \frac{1}{2} t_{j,2i-1} \cdot a'_i + \frac{1}{2} t_{j,2i} \cdot a''_i + u_{ij} + e_{ij}$$

More generally, n individuals are considered, from which n_1 belong to the parental generation, in which the transmission of marker alleles is traced, n_2 belong to generation G2 which is genotyped, and n_3 are the progeny of subsequent generations (G3+). $n \geq n_1 + n_2 + n_3$.

Assuming there are no fixed effects, the description model is

$$y = ZPWTa + Zu + e \qquad (1)$$

where:

y is a vector of N performances from individuals belonging to various generations, including G2 and G3+

a a $2\,n_1$-vector of fixed QTL allele effects in G1 parents

u a n-vector of polygenic effects of pedigreed individuals, a random normal multi-variable with a 0 mean and D covariances matrix

e a N-vector of residual effects of pedigreed individuals, a random normal multi-variable with a 0 mean and $R = I \cdot \sigma_e^2$ covariances matrix

Z the $N \times n$ incidence matrix linking the N performances to the n individuals of the pedigree (Z_{ij} is 1 if the i th phenotype ($i = 1 \ldots N$) belongs to the j th individual ($j = 1 \ldots n$), 0 if not)

W the $n \times n_2$ incidence matrix which designates, among the n individuals, the n_2 individuals which belong to G2 (W_{ij} is 1 if the i th individual of the pedigree ($i = 1 \ldots n$) is the j th G2 individual ($j = 1 \ldots n_2$), 0 if not)

T the $n_2 \times 2n_1$ matrix of G1 to G2 transmission probabilities, conditional to marker information

P the $n \times n$ matrix of QTL allele transmission between generations. P is a lower triangular matrix that can be built by sorting the blocs by generation (First G1, G2 then...). It is constructed recursively as: $p_{ii} = 1$ and $p_{ij} = \frac{1}{2} p_{sj} + \frac{1}{2} p_{dj}$ where s and d are the indices of i's parents, with $j < i$ and $p_{ij} = 0$ for $j > i$.

The y covariance matrix is $var(Zu + e) = V = ZDZ' + R$, where $D = A\sigma^2_u$, A the additive relationship matrix and $R = I\sigma^2_e$.

It should be noted that A and P matrices are linked by $A = PCP'$, where C is a diagonal matrix in which the c_{kk} element is 1, ¾ or ½ (in the non-inbred situation) when none, 1 or 2 of the k parents are known (Quaas 1984).

Following this model [I], the generalized least squares estimation of QTL effects is

$$\hat{a} = (T'W'P'Z'V^{-1}ZPWT)^{-1}T'W'P'Z'V^{-1}y$$

Equivalent two steps model

This $\hat{a}$ estimation may be reformulated as

$$\hat{a} = (T'(QVQ')^{-1}T)^{-1}T'(QVQ')^{-1}z$$

where $Q = (W'P'Z'V^{-1}ZPW)^{-1}W'P'Z'V^{-1}$.

Thus, estimating the QTL effect from model [I] is equivalent to the analysis of a linear combination $z = Qy$ of the y phenotypes (called hereafter the 'score'), using the model:

$$z = Ta + e, with\ e \sim N(0, QVQ') \qquad (2)$$

which is the model currently used in QTL regression approaches (Seaton *et al.* 2002; Filangi *et al.* 2010).

Is this score consistent with the DYD, 'de-regressed proof' or 'genetic merit' used by others?

Meaning of the score

This quantity is linked to the BLUP animal model estimations of the polygenic values u, ignoring the QTL effect a, *that is*, following the model $y = Zu + e$.

Indeed, $z = (W'P'Z'V^{-1}\ ZPW)^{-1}W'(PC)^{-1}PCP'Z'V^{-1}y$

Noting that the BLUP of u is $\hat{u} = DZ'V^{-1}y = \sigma^2_u PCP'Z'V^{-1}y$

We get

$$z = (W'P'Z'V^{-1}ZPW)^{-1}W'(PC)^{-1}\hat{u}/\sigma^2_u$$

$$z = (W'P'Z'V^{-1}ZPW)^{-1}W'C^{-1}P^{-1}\hat{u}/\sigma^2_u$$

The P^{-1} inverse of the P matrix is quite simple: a lower triangular matrix in which diagonal elements are 1 and off-diagonal elements are 0, except for $(P^{-1})_{is}$ and $(P^{-1})_{id}$ which are $-1/2$, where s and d are the indices of i's parents (if they belong to the pedigree list).

Thus, the i th element of $P^{-1}\hat{u}$ is the BLUP $\hat{u}_i$ of u_i in deviation from the halved known parental values: $\hat{u}_i - \frac{1}{2}(\hat{u}_s + \hat{u}_d)$, $\hat{u}_i - \frac{1}{2}\hat{u}_s$, $\hat{u}_i - \frac{1}{2}\hat{u}_d$ or $\hat{u}_i$, when the pedigree list includes both parents, the sire only, the dam only or none of them, respectively.

Finally, $\tilde{u} = W'C^{-1}P^{-1}\hat{u}$ is the subset of standardized deviations $\tilde{u}_i$ of the n_2 G2 individuals.

The variance $V(\hat{u})$ of $\hat{u}$ is

$$V(\hat{u}) = \sigma^4_u PCP'Z'V^{-1}ZPCP$$

The variance $V(\tilde{u})$ of $\tilde{u}$ is $V(\tilde{u}) = W'C^{-1}P^{-1}V(\hat{u})P^{-1}C^{-1}W$, *that is*:

$$V(\tilde{u}) = \sigma^4_u W'P'Z'V^{-1}ZPW \qquad (3)$$

Thus,

$$z = \sigma^2_u V(\tilde{u})^{-1}\tilde{u} \qquad (4)$$

demonstrating that the score z generalizes the genetic merit defined by [2].

The covariances matrix of z is

$$V_z = \sigma^4_u V(\tilde{u})^{-1} \qquad (5)$$

When a QTL regression model ($z = Ta + e$) has to be applied to the GDRP z, the heterogeneity of the residual variances $var(\varepsilon)_{ii} = \sigma^4_u(V(\tilde{u})^{-1})_{ii}$ must be considered. Nothing that $[V(\tilde{u})]_{ii}/\sigma^2_u$ is the reliability coefficient CD_i of the BLUP of $[u_i - \frac{1}{2}(u_s + u_d)]/C_{ii}$, it must be emphasized that this residual variances heterogeneity is not measured by the inverse of this reliability coefficient but by $\sigma^2_u(V(\tilde{u})^{-1})_{ii}$. Nevertheless, when the off-diagonal elements of the $V(\tilde{u})$ matrix are negligible, both expressions are very close.

Complete treatment of the information should include the covariances between z_i which are not considered in QTL regression approaches (e.g. Haley & Knott 1992; Elsen *et al.* 1999).

Illustration: application to the classical situations

To illustrate the method developed previously, it will be applied to two standard situations: daughter and granddaughter designs.

Daughter design

In this basic situation, none of the G3+ information is used, and the method is expected to be reduced to the direct analysis of y information, without variance heterogeneity. No fixed effects are considered here.

Equations are provided for a single family to simplify the demonstration, and the sire i indice is omitted. Let d be the number of G2 daughters of this sire.

The matrices describing the designs follows:

$$Z=\begin{pmatrix}0&1&\dots&0\\ \vdots&\vdots&\ddots&\vdots\\ 0&0&\dots&1\end{pmatrix}\quad W=\begin{pmatrix}0&\dots&0\\ 1&\dots&0\\ \vdots&\ddots&\vdots\\ 0&\dots&1\end{pmatrix}\quad P=\begin{pmatrix}1&0&\dots&0\\ 1/2&1&\dots&0\\ \vdots&\vdots&\ddots&\vdots\\ 1/2&0&\dots&1\end{pmatrix}\quad C=\begin{pmatrix}1&0&\dots&0\\ 0&3/4&\dots&0\\ \vdots&\vdots&\ddots&\vdots\\ 0&0&\dots&3/4\end{pmatrix}$$

Let σ_u^2 and σ_e^2 be the genetic and residual variances, with $\sigma^2=\sigma_u^2+\sigma_e^2$, the additive relationship matrix and y covariance matrix come to be

$$D=A\sigma_u^2=PCP'\sigma_u^2=\begin{pmatrix}1&1/2&\dots&1/2\\ 1/2&1&\dots&1/4\\ \vdots&\vdots&\ddots&\vdots\\ 1/2&1/4&\dots&1\end{pmatrix}\sigma_u^2$$

$$V=ZDZ'+R=\begin{pmatrix}1&1/4&\dots&1/4\\ 1/4&1&\dots&1/4\\ \vdots&\vdots&\ddots&\vdots\\ 1/4&1/4&\dots&1\end{pmatrix}\sigma_u^2+\begin{pmatrix}1&0&\dots&0\\ 0&1&\dots&0\\ \vdots&\vdots&\ddots&\vdots\\ 0&0&\dots&1\end{pmatrix}\sigma_e^2$$

V is a linear combination $aI_d + bJ_d$, where I_d is the $d\times d$ identity matrix and J_d a $d\times d$ matrix in which all the elements are 1. The coefficients are given by $Ma=\sigma^2-1/4\sigma_u^2$ and $b=1/4\sigma_u^2$. Thus, $V^{-1}=\frac{1}{a}I_d-\frac{b}{a[a+db]}J_d$.

As expected we find that $z_j=y_j$ and $var(z_j)=\sigma_u^4[V(\vec{u})^{-1}]^{jj}=\sigma^2$, the elements usually analysed in QTL regression analysis of sire families.

Granddaughter design

In this design, the only recorded phenotypes belong to the G3 generation. To simplify the presentation, we consider a single grandsire in a balanced situation (p sires/grandsire and d daughters/sire, giving $n=1+p+pd$). The matrices describing the designs follow:

$$Z=\begin{pmatrix}0&0&\dots&0&1&\dots&0\\ \vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots\\ 0&0&\dots&0&0&\dots&1\end{pmatrix}\quad W=\begin{pmatrix}0&\dots&0\\ 1&\dots&0\\ \vdots&\ddots&\vdots\\ 0&\dots&1\\ 0&\dots&0\\ \vdots&\ddots&\vdots\\ 0&\dots&0\end{pmatrix}\quad C=\begin{pmatrix}1&0&\dots&0\\ 0&3/4&\dots&0\\ \vdots&\vdots&\ddots&\vdots\\ 0&0&\dots&3/4\end{pmatrix}$$

Noting $\lambda=\frac{\sigma_2-1/4\sigma_u^2}{1/4\sigma_u^2}$, we get $V^{-1}=\frac{4}{\sigma_u^2}[\frac{1}{\lambda}I_d-\frac{1}{\lambda(\lambda+d)}J_d]$.

Applying $\hat{u}=DZ'V^{-1}y$ and noting $\bar{y}$ the mean of the daughters' performances, the estimated polygenic values of the sire ($\hat{u}_0$) and its daughters ($\hat{u}_j$) are given by

$\hat{u}_0=2\frac{d}{d+2}\bar{y}$ and $\hat{u}_j=\frac{3}{2}(y_j-\frac{1}{2}\hat{u}_0)+\frac{1}{2}\hat{u}_0$

Thus, $\tilde{u}_j=\frac{4}{3}(\hat{u}_j-\frac{1}{2}\hat{u}_0)$.

On the other hand, $V(\vec{u})=\sigma_u^4\,W'P'Z'V^{-1}ZPW$ gives here $V(\vec{u})=\frac{4\sigma_u^2}{\lambda}[I_d-\frac{1}{\lambda+d}J_d]$ and $V(\vec{u})^{-1}=\frac{1}{\sigma_u^2}V$

$$P=\begin{pmatrix}1&0&\dots&0&0&\dots&0&\dots&0&\dots&0\\ 1/2&1&\dots&0&0&\dots&0&\dots&0&\dots&0\\ \vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots&\ddots&\vdots&\ddots&\vdots\\ 1/2&0&\dots&1&0&\dots&0&\dots&0&\dots&0\\ 1/4&1/2&\dots&0&1&\dots&0&\dots&0&\dots&0\\ \vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots&\ddots&\vdots&\ddots&\vdots\\ 1/4&1/2&\dots&0&0&\dots&1&\dots&0&\dots&0\\ \vdots&\vdots&&\vdots&\vdots&\ddots&\vdots&\ddots&\vdots&\ddots&\vdots\\ 1/4&0&\dots&1/2&0&\dots&0&\dots&1&\dots&0\\ \vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots&\ddots&\vdots&\ddots&\vdots\\ 1/4&0&\dots&1/2&0&\dots&0&\dots&0&\dots&1\end{pmatrix}$$

Using [III], [IV] and [V], we find that $z_j = 2\,y_{ij}$, *that is to say* twice the mean performances of sire j daughters and

$$var(z_j) = \sigma_a^4 (V(\vec{u})^{-1})_{jj} = \sigma^2 \frac{1 + \frac{1}{4}(d-1)h^2}{\frac{1}{4}d}$$

Discussion and Conclusion

We generalized to any number of generations the de-regressed proof to be analysed as a phenotype in QTL regression approaches as proposed for instance by Haley & Knott (1992) (QTL regression). In QTL regression, the performances of G2 progenies of a parent supposed to be heterozygous for a QTL are correlated with the genetic markers transmitted by this parent to its descendants. The GDRP used to replace these performances is based on the BLUP estimation of the polygenic value independently of the QTL. For a given progeny, this estimation combines its own performances and the performances of all its recorded progeny regardless of the generation number. From the simple algebra developed in this paper, it is easily possible to calculate both the GDRP [IV] and its precision [V] which should be considered as a factor of heterogeneity in the QTL regression model.

Extending this procedure to repeated measures is straightforward. In this case, a Zv term, where v is an individual random normal effect of 0 mean and $I \cdot \sigma_v^2$ covariances, is added to the equation (1) describing the trait. The covariance matrix becomes $V = ZDZ' + ZZ'\sigma_v^2 + R$, and the elements $(\vec{u}) = \sigma_a^4 W' P' Z' V^{-1} Z P W$, $z = \sigma_a^2 V(\vec{u})^{-1}\vec{u}$ and $\sigma_a^2 (V(\vec{u})^{-1})_{ii}$ are directly deduced from this new V matrix.

Similarly, considering fixed effects in the model [I] is performed easily by adding a $X\beta$ term. In this case, the inverse V^{-1} is replaced in the transformation matrix Q by $V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$. Noting that the BLUP estimation u is now $\hat{u} = DZ'(V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})y$, it is easy to demonstrate that the transformation Qy is the GDRP.

It must be emphasized that, in our presentation, the BLUP estimation of the polygenic effect u (i) did not include the QTL effect (the model used was $y = Zu + e$ and not $y = ZPWTa + Zu + e$) and (ii) was only based on the performances measured in the descendants (belonging to G2 or G3+) of genotyped animals.

In our algebra, the A relationship matrix between the polygenic effects u of the genotyped and/or phenotyped n individuals was only based on the bottom part of the pedigree, *that is*, did not include possible relations between G1 reproducers, which could share ancestors (the top part of the pedigree): the G1 parents were supposed unrelated, an hypothesis made in standard QTL regression approaches. It is not proven that the QTL effects estimations from GDRP obtained in this more general situation (the case often considered, e.g. Lien *et al.* 1995) equal $\hat{a}$, the QTL effect estimation from [I]. Even if including ancestral relationships between parents in QTL regression approach is straightforward in the mixed model framework, extending the GDRP to this situation needs new developments.

Weller *et al.* (1990) clearly demonstrated that the granddaughter design is more powerful than the daughter design, in particular for small QTL effects and moderate heritability. This extra power arises from the more precise estimation of the contrasts of G1 (grandsire) origins (here estimated by $\hat{a}$), which is because of the recording of large numbers of granddaughters, as compared to the estimations based only on the performances of the daughters. In their algebra, the power was estimated considering a simplified chi-squared analysis and depended on a non-central parameter inversely proportional to the contrast variances. In our approach, these variances are 'replaced' by the precision of $\hat{a}$estimations, *that is*, $V(\hat{a}) = (T'V(\vec{u})T)^{-1}$. When including more generations in the 'generalized de-regressed proof', the CD_i reliability coefficient, inversely proportional to $V(\vec{u})$ increases, as does the equivalent non-central parameter and the detection power.

However, this gain in power, which comes from extra information to the grand progeny (G3) information, may be reduced if G3 information is already abundant. In such a situation, the CD_i precision obtained using only the G3 data is high, and adding extra generation information will have a limited impact. Thus, working with generalized de-regressed proofs will be mostly useful for species in which artificial insemination is not used on a large scale, such as small ruminants or birds.

Inversely, in the case of lowly heritable traits, use of our generalized de-regressed proof will be advantageous. This is directly in the line with the conclusions of Weller *et al.* (1990) on the effect of heritability.

Finally, the G1-G2 DNA samples, or marker data from these samples, must still be available when phenotypes from extra generations are recorded, and this constraint may limit the practical use of generalized de-regressed proof to a limited number of generations.

The method has been implemented as one of the options available in QTLMap software (http://dga7.-jouy.inra.fr/qtlmap).

Acknowledgements

MKW was supported by the State of Djibouti and the INRA Animal division. This work was partially financed by the ANR GENECAN project. We thank Christèle Robert Granié for her helpful comments.

Competing interests

The authors declare they have no competing interests.

Authors contributions

JME drafted the manuscript. Both authors participated in the development of the method and read and approved the final manuscript

References

Bennewitz J., Reinsch N., Grohs C., Levéziel H., Malafosse A., Thomsen H., Xu N., Looft C., Kühn C., Brockmann G.A., Schwerin M., Weimann C., Hiendleder S., Erhardt G., Medjugorac I., Russ I., Förster M., Brenig B., Reinhardt F., Reents R., Averdunk G., Blümel J., Boichard D., Kalm E. (2003) Combined analysis of data from two granddaughter designs: A simple strategy for QTL confirmation and experimental power in dairy cattle. *Genet. Sel. Evol.*, **35**, 319–338.

Boichard D., Grohs C., Bourgeois F., Cerqueira F., Faugeras R., Neau A., Rupp R., Amigues Y., Boscher M.Y. (2003) Detection of genes influencing economic traits in three French dairy cattle breeds. *Genet. Sel. Evol.*, **35**, 77–101 (available at: http://www.ncbi.nlm.nih.gov/pubmed?term=%22Lev%C3%A9ziel%20H%22%5BAuthor%5D; last accessed 15 December 2011).

Elsen J.M., Mangin B., Goffinet B., Boichard D., Le Roy P. (1999) Alternative models for QTL detection in livestock - I General introduction. *Genet. Sel. Evol.*, **31**, 213–224.

Farnir F., Grisart B., Coppieters W., Riquet J., Berzi P., Cambisano N., Karim L., Mni M., Moisio S., Simon P., Wagenaar D., Vilkki J., Georges M. (2002) Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics*, **161**, 275–287.

Filangi O., Moreno C., Gilbert H., Legarra A., Le Roy P., Elsen J.M. (2010) QTLMap, a Software for QTL Detection in Outbred Populations. Proceedings of the 9th WCGALP, 1–6 August Leipzig, Germany.

George A.W., Visscher P.M., Haley C.S. (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, **156**, 2081–2092.

Georges M., Nielsen D., Mackinnon M., Mishra A., Okimoto R., Pasquino A.T., Sargeant L.S., Sorensen A., Steele M.R., Zhao X., Womack J.E., Hoeschele I. (1995) Mapping Quantitative Trait Loci Controlling Milk Production in Dairy Cattle by Exploiting Progeny Testing. *Genetics*, **139**, 907–920.

Haley C.S., Knott S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.

Haley C.S., Knott S.A., Elsen J.M. (1994) Mapping Quantitative Trait Loci in Crosses between Outbred Lines Using Least Squares. *Genetics*, **136**, 1195–1207.

Lander E.S., Botstein D. (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

Lien S., Gomez-Raya L., Steine T., Fimland E., Rogne S. (1995) Associations Between Casein Haplotypes and Milk Yield Traits. *J. Dairy Sci.*, **78**, 2047–2056.

Quaas R.L (1984) Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics*, **32**, 949–953.

Seaton G., Haley C.S., Knott S.A., Kearsey M., Visscher P.M. (2002) QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics*, **18**, 339–340.

Soller M., Genizi A. (1978) The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics*, **34**, 47–55.

Soller M., Brody T., Genizi A. (1976) On the power of experiments to detect marker-linked quantitative effects in crosses between inbred lines. *Theor. Appl. Genet.*, **47**, 35.

Thomsen H., Reinsch N., Xu N., Looft C., Grupe S., Kühn C., Brockmann G.A., Schwerin M., Leyhe-Horn B., Hiendleder S., Erhardt G., Medjugorac I., Russ I., Förster M., Brenig B., Reinhardt F., Reents R., Blümel J., Averdunk G., Kalm E. (2001) Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for QTL. *J. Anim. Breed. Genet.*, **118**, 357–370.

Tribout T., Iannuccelli N., Druet T., Gilbert H., Riquet J., Gueblez R., Mercat M.J., Bidanel J.P., Milan D., Le Roy P. (2008) Detection of quantitative trait loci for reproduction and production traits in Large White and French Landrace pig populations. *Genet. Sel. Evol.*, **40**, 61–78.

Uimari P., Hoeschele I. (1997) Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics*, **146**, 735–743.

VanRaden P.M., Wiggans G.R. (1991) Derivation, Calculation, and Use of National Animal Model Information. *J. Dairy Sci.*, **71**, 2737–2746.

Weller J.I., Kashi Y., Soller M. (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J. Dairy Sci.*, **73**, 2525–2537.

Zhang Q., Boichard D., Hoeschele I., Ernst C., Eggen A., Murkve B., Pfister-Genskow M., Witte L.P., Grignola F.E., Uimari P., Thaller G., Bishop M.D. (1998) Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics*, **149**, 1959–1973.

1.3 Conclusion et perspective

La méthode décrite dans l'article permet donc d'étendre celle proposée par Weller *et al*. (1990), puis repris par Tribout *et al*. (2008) à un nombre quelconque de générations. L'article présente comment obtenir « le mérite génétique dé-régressé » et le coefficient de détermination associé, pour chacun des individus de la deuxième génération génotypée. Des extensions pour les mesures répétées et la prise en compte des effets fixes dans le modèle sont proposées.

L'apport de nouvelles mesures phénotypiques permettra d'améliorer la puissance de détection. Toutefois, lorsque les tailles des familles sont suffisamment larges (comme chez les bovins laitiers, par exemple) pour garantir un bon niveau d'estimation du mérite génétique pour les individus marqués, l'intérêt peut être limité.

Un certain nombre d'améliorations peuvent être apportées aux développements proposés. Notamment, dans nos développements algébriques, nous considérons que les individus de la première génération sont indépendants, hypothèse assez répandue en détection de QTL par analyse de liaison, ce qui peut ne pas correspondre à la réalité. La prise en compte d'éventuels liens de parenté entre les pères fondateurs peut être une des améliorations à apporter à ce modèle.

2 Etude comparative des deux modèles dans la détection de QTL des caractères à seuils : considération de la sous-jacente normale ou utilisation du modèle de mélange de distributions normales

2.1 Introduction

Comme il est dit ci-dessus (paragraphe 3, chapitre I), certains caractères d'intérêt économique ou biologique pour la production animale ont une distribution discrète, et donc non gaussienne. Comme les caractères continus, ces caractères discrets présentent en général un déterminisme complexe et leur variabilité est contrôlée par de nombreux gènes. Une hypothèse fréquente est que la variabilité de ce type de caractère (disons Y) est sous l'influence d'une variable gaussienne sous-jacente (disons Z), dont l'échelle est jalonnée de seuils (s_1, s_2, …, s_n) : quand Z est dans l'intervalle [s_i, s_{i+1}], la variable Y vaut i+1 : on parle de caractères à seuils (Falconer, 1958). Le premier à aborder le modèle à seuil en génétique fut Wright (1934). Depuis, plusieurs auteurs ont analysé des caractères de ce type, le plus souvent dans le cadre polygénique classique, i.e. sans prendre en compte l'information moléculaire, pour estimer des paramètres génétiques tels que l'héritabilité, ou des effets de facteurs connus tels que le sexe de l'animal (Fouley *et al.,* 1983 ; Gianola et Foulley, 1983 ; Harville et Mee, 1984 ; Weller *et al.,* 1988 ; Weller et Gianola, 1989 ; Jamrozik *et al.,* 1991 ; Weller et Ron, 1992 ; Weller *et al.,* 1992). D'autres auteurs (Hackett et Weller, 1995 ; Xu et Atchley, 1996) intègrent, dans leurs études sur les caractères à seuils, l'information moléculaire, dans le but de cartographier des QTLs. Leurs méthodes sont développées pour des populations « inbred ». Elles ont été étendues pour des dispositifs « outbred » par Yi et Xu (1999). Plus récemment, Xu *et al.* (2005) ont proposé une méthode multi-caractères de détection de QTL pour les caractères binaires. Tous ces auteurs considèrent une distribution normale sous-jacente à la distribution réelle de la variable discontinue étudiée.

Plusieurs études concernant des caractères à seuils (Hackett et Weller, 1995 ; Xu et Atchley, 1996 ; Rao et Xu, 1998 ; Kadarmideen *et al*., 2000) comparent les précisions des estimations de certains paramètres (notamment la position estimée du QTL et le puissance de détection du modèle) obtenues en mettant en œuvre soit un modèle non-linéaire, soit la méthode usuellement utilisée pour les caractères à distributions normales. La différence entre ces deux modèles est donc la fonction de pénétrance considérée. Tous ces auteurs ont considéré des lignées croisées, et concluent que les avantages de l'utilisation d'un modèle discret pour l'analyse des caractères à seuils sont négligeables par rapport à l'utilisation du modèle normal. Hackett et Weller (1995) obtiennent des estimations de positions légèrement meilleures pour le modèle discret alors que Rao et Xu (1998) observent une faible supériorité en termes de puissance de détection pour le modèle discret comparé au modèle normal. Toutefois, certains auteurs (Visscher *et al.,* 1996 ; Rebai, 1997 ; Kadarmideen *et al.,* 2000) rapportent que les deux méthodes d'analyse présentent des caractéristiques équivalentes. Pour les caractères binaires, le modèle discret est même moins robuste qu'un modèle normal (Xu et Atchley, 1996 ; Yi et Xu, 1999).

Cette différence de robustesse entre les deux modèles serait liée, selon ces auteurs, à des pertes d'informations occasionnées par le passage de la variable observée à la variable sous-jacente à celle-ci.

Notre travail vient en complément de ceux des auteurs précédemment cités. En effet, nous proposons d'élargir les cas de figure simulés et de comparer les 2 approches à la fois sur la précision de la position estimée du QTL et sur la puissance de détection. Ainsi, nous simulons 16 cas de figures (certains cas n'ont fait l'objet d'aucune étude), un nombre important comparé à la littérature. Les deux modèles comparés de détection de QTL par analyse de liaison, utilisent la méthode des marqueurs flanquants (Lander et Botstein, 1989). Un protocole simple est simulé et différents cas sont considérés en jouant sur le nombre de modalités du caractère considéré, leurs fréquences dans la population ainsi que l'effet du QTL simulé et sa position.

2.2 Méthode

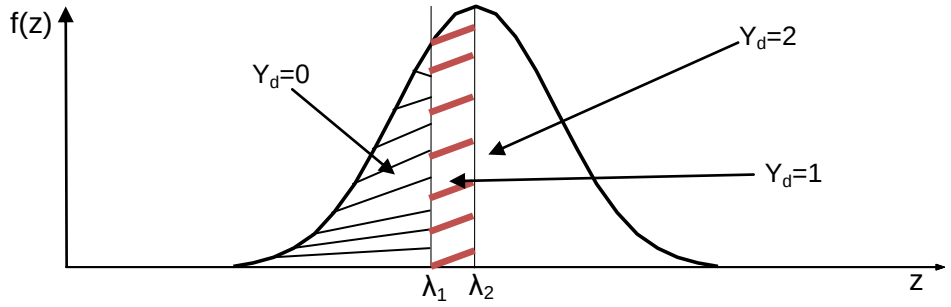
Les modèles

Le modèle normal est utilisé, classiquement, pour les caractères quantitatifs à distribution normale. Son application dans la détection de QTL dans les populations d'élevage est décrite dans le premier chapitre. Dans le modèle discret on fait l'hypothèse d'une variable aléatoire distribuée normalement, sous-jacente au caractère à seuils étudié. Cette variable aléatoire Z se caractérise ainsi :

Si $\lambda_{i-1} < Z < \lambda_i$ alors Y=i avec $i \in \{1, \dots, m\}$ et $\lambda_0 = -\infty$ et $\lambda_m = +\infty$

Les λ_i sont les seuils fixés, définis pour l'échelle sous-jacente grâce aux catégories observées sur la variable discrète Y et son nombre des modalités. La figure 2.1 représente un caractère à seuils présentant trois modalités (0, 1 et 2) avec 30% de la population totale présentant le phénotype 0, 20% le phénotype 1 et 50% le phénotype 2. m=3 est le nombre des modalités du caractère.

Figure 2.1 : Illustration d'un caractère à seuils avec 3 modalités



Dans ce cas particulier, la probabilité qu'un individu d pris au hasard dans la population soit de phénotype exprimé y_d=0 est de 30%.

Dans le cas plus général (avec un nombre quelconque m de modalités), la transformation de Z en Y peut être résumée en termes de probabilité, de la façon suivante :

$$\begin{cases} P(Y=1) = P(-\infty < Z < \lambda_1) \\ P(Y=2) = P(\lambda_1 < Z < \lambda_2) \\ \vdots \\ P(Y=m) = P(\lambda_{m-1} < Z < +\infty) \end{cases}$$

Par ailleurs, la variable aléatoire Z suit une loi normale. Cette hypothèse est notamment compatible avec celle d'un contrôle du caractère par plusieurs gènes, chacun ayant un petit effet. Pour une position donnée du génome, la performance Z_d, d'un individu d, peut s'écrire comme suit :

$$Z_d = \mu + x_{jd}\alpha_j + e_d$$

μ est la moyenne générale, x_{jd} la probabilité que l'individu d ait reçu l'allèle j au QTL sachant l'information aux marqueurs, α_j est l'effet QTL et e_j la résiduelle du modèle d'espérance nulle.

Les probabilités précédentes, pour un individu d, peuvent être calculées de la façon suivante :

$$P(Y_d = 1) = P(-\infty < Z_d < \lambda_1) = \int_{-\infty}^{\lambda_1} f(Z_d)dz$$

$$P(Y_d = 2) = P(\lambda_1 < Z_d < \lambda_2) = \int_{\lambda_1}^{\lambda_2} f(Z_d)dz$$

$$\vdots$$

$$P(Y_d = m) = P(\lambda_{m-1} < Z_d < +\infty) = \int_{\lambda_{m-1}}^{+\infty} f(Z_d)dz = 1 - \int_{-\infty}^{\lambda_{m-1}} f(Z_d)dz$$

$$\Leftrightarrow$$

$$P(Y_d = 1) = \phi(\lambda_1 - E(Z_d)) = \boldsymbol{\phi(\lambda_1 - \mu - x_{jd}\alpha_j)}$$

$$P(Y_d = 2) = \phi(\lambda_2 - E(z_d)) - \phi(\lambda_1 - E(Z_d)) = \boldsymbol{\phi(\lambda_2 - \mu - x_{jd}\alpha_j) - \phi(\lambda_1 - \mu - x_{jd}\alpha_j)}$$

$$\vdots$$

$$P(Y_d = m) = 1 - \phi(\lambda_{m-1} - E(Z_d)) = \boldsymbol{1 - \phi(\lambda_{m-1} - \mu - x_{jd}\alpha_j)}$$

f est la densité de la loi normale standard et ϕ sa fonction de répartition. On suppose ici, que Z_d est une variable aléatoire réduite.

Simulations et détection de QTL

Les simulations des populations et l'estimation des paramètres de détections de QTL (position, effet de substitution,…) ont été effectuées avec le logiciel QTLMap (Filangi *et al.,* 2010). Un module prenant en compte la sous-jacente normale aux caractères à seuils a été implémenté à cette occasion.

La méthode du maximum de vraisemblance détaillé dans le paragraphe 2.3.2.2 du chapitre I a été mise en œuvre. Seule la fonction de pénétrance est modifiée et s'écrit maintenant de la façon suivante :

$$\sum_{i=1}^{m} \delta_i(\phi(\lambda_i - \boldsymbol{\mu - x_j\alpha_j}) - \phi(\lambda_{i-1} - \boldsymbol{\mu - x_j\alpha_j})$$

δ_i est un indicateur prenant la valeur 1 si l'individu d présente un phénotype visible y_d=i, et 0 sinon.

Le rapport des vraisemblances est maximisé pour estimer les paramètres du modèle, il s'agit des seuils λ, des effets des locus α et de la distance entre la position estimée du QTL et les marqueurs flanquants.

2.3 Simulations

Pour des questions de facilité, on simule un protocole de type F2 avec sept pères fondateurs hétérozygotes aux marqueurs, chacun ayant été accouplé avec dix mères. Chacune des mères possède six produits. On a donc au total, 420 descendants. Pour chacun d'entre eux, on considère une carte génétique composée d'un chromosome de 1 Morgan, couvert de 11 marqueurs, régulièrement espacés de 10cM.

Dans toutes les situations étudiées, l'héritabilité est fixée à h^2=0,35. Le tableau 2.1 présente les caractéristiques du QTL (sa position et son effet de substitution) et du caractère (le nombre de modalités et leurs fréquences dans la population) dans les différentes simulations effectuées.

Les positions du QTL

Dans les différents dispositifs simulés, le QTL est tantôt placé à l'extrémité du chromosome (à 20cM), tantôt placé au milieu du chromosome (à 50cM), cela afin de vérifier si la position du QTL a un effet sur la précision de l'estimation des paramètres de détection de QTL.

Tableau 2.1 : Les caractéristiques des différentes populations simulées

N° Cas simulé	Position du QTL simulé (cM)	Effet du QTL simulé	Nombre de modalités[1] (fréquences)
1	0,5	fort (0,8)	2 (50% et 50%)
2	0,2	fort (0,8)	2 (50% et 50%)
3	0,5	faible (0,2)	2 (50% et 50%)
4	0,2	faible (0,2)	2 (50% et 50%)
5	0,5	fort (0,8)	2 (10% et 90%)
6	0,2	fort (0,8)	2 (10% et 90%)
7	0,5	faible (0,2)	2 (10% et 90%)
8	0,2	faible (0,2)	2 (10% et 90%)
9	0,5	fort (0,8)	3 (30%, 30% et 40%)
10	0,2	fort (0,8)	3 (30%, 30% et 40%)
11	0,5	faible (0,2)	3 (30%, 30% et 40%)
12	0,2	faible (0,2)	3 (30%, 30% et 40%)
13	0,5	fort (0,8)	3 (10%, 10% et 80%)
14	0,2	fort (0,8)	3 (10%, 10% et 80%)
15	0,5	faible (0,2)	3 (10%, 10% et 80%)
16	0,2	faible (0,2)	3 (10%, 10% et 80%)

[1]Nombre de modalités du caractère discret simulé

Les effets de substitution du QTL

Dans nos simulations, l'effet de substitution du QTL est fixé à 80% de l'écart phénotypique (donc très fort) ou à 20% (donc faible).

Pour une combinaison donnée de ces paramètres (soit un des 16 cas simulés répertoriés dans le tableau 2.1), 500 simulations permettent d'obtenir les distributions des estimations des modèles. Par ailleurs, les seuils de rejet de l'hypothèse d'absence de QTL sont obtenus par 500 simulations sous l'hypothèse d'absence de QTL, afin d'estimer la puissance de détection de QTL des deux modèles. Concrètement cette puissance est estimée par la proportion de LRT dépassant le seuil empirique de 5%.

Les critères de comparaison

Les critères pris en compte sont la puissance de détection de QTL et la précision de la position estimée du QTL. A cette fin sont comparés les écart-types de ces positions estimées et leurs erreurs moyennes quadratiques par rapport à la position simulée. Par ailleurs, le temps de calcul est aussi considéré comme un des éléments de l'efficacité des modèles.

2.4 Résultats

Nous détaillons ici, les résultats de 4 des 16 cas simulées. Elles correspondent aux quatre types de caractères simulés (cas simulé N°1, 5, 9 et 13). Les résultats des 12 autres cas simulés sont présentés dans l'Annexe A.

Cas simulé N° 1

Dans ce cas, le caractère discret simulé possède deux modalités (50%/50%), le QTL simulé a un effet fort (0,8) et est situé au milieu (50cM) du chromosome. Dans cette situation, le QTL est bien localisé par les deux modèles (tableau 2.2). Le modèle normal présente toutefois une puissance de détection plus élevée, avec un temps de calcul deux fois plus faible que pour le modèle discret. Quand le caractère étudié présente deux modalités, avec une fréquence équilibrée dans la population, le modèle normal est donc plus attractif pour la détection de QTL. Si on diminue l'effet du QTL simulé (cas simulé N°3 – annexe A), la puissance de détection reste toujours meilleure dans le modèle normal, 9,2% contre 8,3% pour le modèle discret. Toujours pour ce type de caractère, la position du QTL (qu'il soit à l'extrémité ou au milieu du chromosome) ne change rien quant à l'efficacité du modèle normal par rapport au modèle discret. Toutefois, quand le QTL est à l'extrémité du chromosome (20cM) et que le QTL présente un effet faible (cas simulé N°4 – annexe A), on note une très mauvaise estimation, par les deux modèles, de la position du QTL.

Tableau 2.2 : paramètres estimés pour les deux modèles comparés grâce aux données simulées du cas 1

Paramètres	Vraie valeur	Modèle discret	Modèle normal
Position du QTL (M)	0,5	0,49	0,49
Ecart-type	-	0,18	0,15
Erreur moyenne quadratique	-	0,03	0,02
puissance (%)	-	68,9	76,0

Cas simulé N°5

Le caractère a deux modalités en fréquences déséquilibrées (10%/90%). Le QTL simulé a un effet fort (0,80) et est situé au milieu (50cM) du chromosome (tableau 2.3). Dans ce cas, les précisions relatives à l'estimation de la position du QTL sont similaires et satisfaisantes (assez proche de la valeur vraie) pour les deux modèles. La puissance de détection du modèle normal est supérieure à celle du modèle discret. Si on place le QTL à l'extrémité du chromosome (20cM), tous les autres paramètres étant inchangés (cas simulé N°6 – annexe A), la puissance reste sensiblement la même, et ce pour les deux modèles. La position du QTL est assez bien estimée pour les deux modèles, tout comme le cas N°5. Le modèle normal reste toujours privilégié pour sa puissance de détection supérieure à celle du modèle discret, ce dernier quant à lui présente une meilleure précision pour l'estimation de la position du QTL. Par ailleurs, si on diminue l'effet de substitution du QTL (0,2 au lieu de 0,8), les autres paramètres étant inchangés (cas simulé N°7 – annexe A), la position du QTL est beaucoup moins bien estimée (erreur moyenne quadratique=0,10 pour les deux modèles). Enfin si le QTL présente un effet faible (0,20) et est positionné à l'extrémité du chromosome (cas simulé N°8 – annexe A), on obtient des précisions assez médiocres de l'estimation de la position du QTL : dans ce cas, le modèle discret présente un léger avantage (7,4% contre 6,3%) en termes de puissance de détection de QTL par rapport au modèle normal.

Tableau 2.3 : paramètres estimés pour les deux modèles comparés grâce aux données simulées du cas 5

Paramètres	Vraie valeur	Modèle discret	Modèle normal
Position du QTL (M)	0,5	0,45	0,48
Ecart-type	-	0,22	0,19
Erreur moyenne quadratique	-	0,05	0,04
puissance (%)	-	46,6	52,6

Cas simulé N°9

Dans ce cas, on simule un caractère en fréquences déséquilibrées à trois modalités (30%/30%/40%). Le QTL simulé présente un effet fort (0,80) et est situé au milieu (50cM) du chromosome (tableau 2.4). La précision de l'estimation de la position du QTL est assez satisfaisante pour les deux modèles. En effet, l'erreur moyenne quadratique est seulement de 0,03 pour le modèle normal et 0,04 pour le modèle discret. Le modèle normal présente une puissance supérieure de 5 points par rapport au modèle discret. Avec le coût en temps de calcul élevé du modèle discret par rapport au modèle normal (quasiment, deux fois plus de temps), le modèle normal peut être préféré au premier pour ce type de caractères. Les puissances de détection des deux modèles sont quasiment similaires (66,2% pour le modèle discret et 65,2% pour le modèle normal) si le QTL est positionné à l'extrémité du chromosome (cas simulé N°10 – annexe A), le reste des paramètres étant inchangé par rapport au cas 9. On remarque même un changement de tendance, dans la mesure où le modèle discret est plus puissant que le modèle normal dans le cas 9, l'écart restant néanmoins assez minime. Comme pour les autres types de caractères précédemment décrit, si le QTL est placé à 20cM et a un effet faible (cas simulé N°12 – annexe A), la précision de l'estimation de la position est assez mauvaise. En conclusion, les deux modèles présentent des précisions d'estimation et des puissances de détection comparables, on note toutefois une légère supériorité pour le modèle normal quand le QTL recherché est suffisamment fort et qu'il n'est pas à l'extrémité du chromosome.

Tableau 2.4 : paramètres estimés pour les deux modèles comparés grâce aux données simulées du cas 9

Paramètres	Vraie valeur	Modèle discret	Modèle normal
Position du QTL (M)	0,5	0,49	0,50
Ecart-type	-	0,19	0,16
Erreur moyenne quadratique	-	0,04	0,03
puissance (%)	-	59,2	64,8

Cas simulé N°13

Cette fois, un caractère à trois modalités en fréquences très inégales (10%/10%/80%) est simulé. Le QTL simulé présente un effet fort (0,8) et est situé au milieu (50cM) du chromosome (tableau 2.5). Ici encore, les paramètres de détection sont bien estimés par les deux modèles, les erreurs moyennes quadratiques étant négligeables. La puissance de détection du modèle discret est presque deux fois plus forte par rapport au modèle normal. Le modèle discret est donc le plus adapté pour ce type de caractère. Toutefois le temps de calcul (environ 2heures pour les 500 simulations ie autour de 20 secondes par analyse) peut constituer une barrière à l'utilisation de manière systématique de ce modèle. Si le QTL est positionné à l'extrémité du chromosome (cas simulé N°14 – annexe A), les précisions des estimations restent similaires et l'intérêt en termes de puissance de détection du modèle discret s'accentue. En effet, le modèle discret présente dans ce cas, une puissance de 67,1% contre 34,5% pour le modèle normal. Si, en plus d'être situé à

l'extrémité du chromosome, le QTL présente un effet de substitution faible (cas simulé N°16 – annexe A), les estimations de la position ne sont pas précises. L'erreur moyenne quadratique par rapport à la position vraie est de 0,16 pour le modèle discret, elle est de 0,18 pour le modèle normal. Dans ce dernier cas, la puissance de détection du modèle discret est 2,7 fois supérieure à la puissance du modèle normal. En conclusion, pour des caractères de ce type (caractères à trois modalités et de distribution déséquilibrée) le modèle discret semble plus approprié pour la détection de QTL et ce, quelque soit les caractéristiques du QTL simulé.

Tableau 2.5 : paramètres estimés pour les deux modèles comparés grâce aux données simulées du cas 13

Paramètres	Vraie valeur	Modèle discret	Modèle normal
Position du QTL (M)	0,5	0,50	0,51
Ecart-type	-	0,16	0,23
Erreur moyenne quadratique	-	0,03	0,05
puissance (%)	-	67,4	36,7

2.5 Discussion et conclusion

Précisions des paramètres

Dans la plupart des cas, les précisions des estimations des paramètres par les deux modèles sont satisfaisantes. Quand le QTL simulé présente un effet de substitution élevé sa position est toujours bien estimée. En effet, l'erreur moyenne quadratique (EMQ) est inférieure à 0,05 dans 75% des cas. Si le QTL n'est responsable que d'un faible pourcentage de la variabilité du caractère et qu'il est positionné au milieu du chromosome, les simulations montrent que les précisions des estimations restent acceptables (l'EMQ ne dépasse guère 0,06), et ce pour les deux modèles étudiés. Si l'effet de substitution du QTL est faible et qu'il est situé à l'extrémité du chromosome, sa position n'est jamais bien estimée.

Temps de calcul

Le modèle discret est beaucoup plus gourmand en temps de calcul que le modèle normal. Cette observation étant faite pour un code informatique faiblement optimisé, il faut en moyenne deux fois plus de temps, pour l'analyse par rapport au modèle normal. Insistons sur le fait que le modèle normal a été mis en place et programmé dans le logiciel QTLMap (Filangi *et al.,* 2010) bien avant le modèle discret, et a donc bénéficié d'un certain nombre d'améliorations qui optimisent les temps de calcul. A moyen terme, des améliorations de ce genre seront opérées pour le modèle discret. Quoiqu'il en soit, les temps de calcul resteront plus importants par la nature même de ce modèle non linéaire.

Modèle normal ou modèle discret ?

En oubliant le temps de calcul qui est amené à évoluer, la précision de la position du QTL estimée et la puissance de détection sont des éléments solides,

pour comparer les deux modèles étudiés. Pour ce qui est de l'estimation de la position, les deux modèles présentent des précisions intéressantes et assez similaires : dans 88% des cas, l'écart des EMQ entre ces 2 modèles sont inférieurs à deux points. Ces observations sont conformes à la littérature (Visscher *et al.,* 1996 ; Rebai, 1997 ; Kadarmideen *et al.,* 2000). La différence des précisions peut être jugée significative dans deux (cas simulés N°8 et N°14) des 16 cas simulés, et c'est le modèle discret qui s'avère le plus efficace. Il s'agit de cas où les modalités du caractère sont déséquilibrées dans la population et le QTL placé à l'extrémité du chromosome.

La puissance de détection est le critère le plus discriminant entre les deux modèles. Si on ne s'intéresse qu'à ce critère, le modèle normal est plus puissant pour les caractères binaires. Cette conclusion est vraie quelque soit les caractéristiques du QTL simulés. Quand la fréquence des modalités est déséquilibrée, le modèle normal reste néanmoins efficace sauf quand l'effet du QTL est très faible et que le QTL est placé à l'extrémité du chromosome. Il est fait état dans la littérature de ce faible apport du modèle discret, en termes de précision d'estimation et de puissance de détection par rapport au modèle normal, pour des caractères binaires avec des modalités en fréquences déséquilibrées. Yi et Xu (1999) montrent même que l'utilisation d'un modèle à seuils a plus d'apport en termes de précision d'estimation et de puissance de détection pour un caractère binaire avec des modalités en fréquences équilibrées, ces apports étant moindres voire inexistants pour un caractère binaire avec des modalités en fréquences déséquilibrées dans la population. A noter que dans ce dernier cas, les estimations des positions sont assez moyennes (EMQ=0,16 pour le modèle discret et 0,19 pour le modèle normal).

Les caractères à trois modalités se comportent différemment. En effet, si les modalités sont distribuées de façon homogène (30%/30%/40%) dans la population, le modèle normal est toujours le plus puissant, excepté quand le QTL est situé à l'extrémité du chromosome où l'on observe des puissances du même ordre pour les deux modèles. Par contre, si les modalités sont distribuées de façon déséquilibrée le modèle discret présente la puissance de détection la plus élevée, et ce quelque soit les caractéristiques du QTL.

Conclusion

Nous montrons donc que le modèle discret est plus précis et plus puissant quand le caractère étudié possède trois modalités distribuées de façon déséquilibrée dans la population. Ce résultat n'a pour l'heure jamais été publié. Toutefois, Hackett et Weller (1995) montrent l'intérêt du caractère discret en termes de précision des estimations de paramètres pour un caractère à cinq modalités, avec une hétérogénéité de fréquences dans la population. Selon ces mêmes auteurs, si les 5 modalités du caractère sont représentées de façon équilibrée dans la population, les deux modèles étudiés présentent des précisions similaires.

Pour les caractères binaires, le modèle normal présente des précisions et une puissance de détection similaires, voire légèrement meilleures, que le modèle discret, exception faite du cas où le QTL est placé à l'extrémité du chromosome avec un effet de substitution faible. Dans ce cas particulier, les précisions des estimations sont médiocres, indépendamment des modèles, et les puissances très faibles.

Perspectives

Même si nous apportons quelques éclaircissements sur l'intérêt d'utiliser le modèle discret, utilisant la sous-jacente normale aux caractères discrets, nos résultats doivent être confirmés et complétés par des simulations supplémentaires. En effet, l'effectif devra être accru dans ces nouvelles simulations pour mesurer l'influence de la taille du protocole sur l'efficacité de l'un ou l'autre des deux modèles : cela peut être envisagé en doublant le nombre de pères fondateurs ou des descendants par mère.

Une autre piste à explorer serait de quantifier l'effet de l'héritabilité du caractère sur la robustesse et la précision des estimations pour les deux modèles considérés. En effet considérant une héritabilité fixe pour tous nos simulations, nous ne sommes pas en mesure de dire si oui ou non, ce paramètre a une quelconque influence dans l'appréciation des modèles. Cependant, ces nouvelles simulations ne peuvent être réalisées que si le temps de calcul est significativement réduit, notamment pour le modèle discret qui, en l'état, est assez lourd à mettre en œuvre.