

On Quantitative Trait Locus mapping with an interference phenomenon

Charles-Elie Rabier

Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., Toulouse, France.

INRA UR631, Station d'Amélioration Génétique des Animaux, Auzeville, France.

Summary. We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait) on the interval $[0, T]$ representing a chromosome. The observation is the trait and the composition of the genome at some locations called “markers”. As in Rebai et al. (95), we focus on the interference phenomenon : a recombination event inhibes the formation of another nearby. We give the asymptotic distribution of the LRT process under the null hypothesis that there is no QTL on $[0, T]$ and under local alternatives with a QTL at t^* on $[0, T]$. We show that the LRT process is asymptotically the square of a “linear interpolated and normalized process ” whereas the LRT process obtained recently by Azais et al., for a model without interference, was the square of a “non linear interpolated and normalized process ”. The computation of the supremum of our LRT process becomes easy due to the interpolation. Besides, we prove that we have asymptotically exactly the same thresholds for a model with or without interference. However, we also prove that the powers of detection are totally different between the two models.

Keywords: Gaussian process, Likelihood Ratio Test, Mixture models, Nuisance parameters present only under the alternative, QTL detection, MCQMC.

1. Introduction

We study a backcross population: $A \times (A \times B)$, where A and B are purely homozygous lines and we address the problem of detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured) on a given chromosome. The trait is observed on n individuals (progenies) and we denote by Y_j , $j = 1, \dots, n$, the observations, which we will assume to be Gaussian, independent and identically distributed (i.i.d.). The mechanism of genetics, or more precisely of meiosis, implies that among the two chromosomes of each individual, one is purely inherited from A while the other (the “recombined” one), consists of parts originated from A and parts originated from B , due to crossing-overs.

The chromosome will be represented by the segment $[0, T]$. The distance on $[0, T]$ is called the genetic distance, it is measured in Morgans. K genetic markers are located at fixed locations $t_1 = 0 < t_2 < \dots < t_K = T$. These markers will help us to find the QTL. $X(t_k)$ refers to the genetic information at marker k . For one individual, $X(t_k)$ takes the value $+1$ if, for example, the “recombined chromosome” is originated from A at location t_k and takes the value -1 if it is originated from B .

We use the Haldane modeling for the genetic information at marker locations. It can be represented as follows: $X(0)$ is a random sign and $X(t_k) = X(0)(-1)^{N(t_k)}$ where $N(\cdot)$ is a standard Poisson process on $[0, T]$. Due to the independence of increments of Poisson process, this model allow double recombinations between markers. For instance, if we consider 3 markers (ie. $K = 3$), we can have the scenario $X(t_1) = 1$, $X(t_2) = -1$ and $X(t_3) = 1$, which means that there has been a recombination between markers 1 and 2, and also a recombination between markers 2 and 3. Obviously, in the same way, we can have the scenario $X(t_1) = -1$, $X(t_2) = 1$ and $X(t_3) = -1$.

A QTL is lying at an unknown position t^* between two genetic markers. $U(t^*)$ is the genetic information at the QTL location. In the same way as for the genetic information at marker locations, $U(t^*)$ takes value $+1$ if the “recombined

chromosome” is originated from A at t^* , and -1 if it is originated from B . Due to Mendel law, $U(t^*)$ takes value $+1$ and -1 with equal probability. We assume an “analysis of variance model” for the quantitative trait :

$$Y = \mu + U(t^*) q + \sigma \varepsilon \tag{1}$$

where ε is a Gaussian white noise. The key point is that we will have to guess the value of $U(t^*)$, using only the information available, which is the information at genetic markers.

The originality of this paper is that we focus on the model introduced by Rebaï et al. (1995) in which double recombination between the QTL and its flanking markers is not allowed. For instance, if the QTL is lying between the first two markers (ie. $t^* \in]t_1, t_2[$), we can not have the scenario $X(t_1) = 1$, $U(t^*) = -1$ and $X(t_2) = 1$, which would have supposed that there had been a recombination between the first marker and the QTL, and also a recombination between the second marker and the QTL. In particular, the model consider that if we have a recombination between the QTL and one of its flanking marker, we could not have a recombination between the QTL and the other flanking marker. In other words, if $X(t_1) = 1$ and $U(t^*) = -1$, then we have automatically $X(t_2) = -1$. In the same way, if $X(t_2) = 1$ and $U(t^*) = -1$, then we have automatically $X(t_1) = -1$. We will explain in details this model in Section 2 and present the law of $U(t^*)$, given its flanking markers.

This way, inside the marker interval which contains the QTL, we model the interference phenomom : a recombination event inhibes the formation of another nearby. This phenomom was noticed by geneticists working on the *Drosophila* (Sturtevant (1915), Muller (1916)). In McPeck and Speed (1995), the authors study several interference models and also mention the importance of modeling interference. We focus here on the model proposed by Rebaï et al. (1995), and then extended to a whole chromosome in Rebaï et al. (1994). It will lead to original mathematical results with a real impact for geneticists.

So, since only the Quantitative trait and the genetic information at marker locations are available, one observation will be

$$(Y, X(t_1), \dots, X(t_K)).$$

We observe n observations $(Y_j, X_j(t_1), \dots, X_j(t_K))$ i.i.d. It can be proved that, conditionally to $X(t_1), \dots, X(t_K)$, Y obeys to a mixture model with known weights :

$$p(t^*) f_{(\mu+q, \sigma)}(\cdot) + \{1 - p(t^*)\} f_{(\mu-q, \sigma)}(\cdot), \tag{2}$$

where $f_{(m, \sigma)}$ is the Gaussian density with parameters (m, σ) and where the function $p(t^*)$ is the probability $\mathbb{P}\{U(t^*) = 1\}$ conditionally to the flanking markers (see Section 2) .

The challenge is that the true location t^* is not known. So, we test the presence of a QTL at each position t . $\Lambda_n(t)$ and $S_n(t)$ are the likelihood ratio test (LRT) statistic and the score test statistic (see Section 2 for a precise definition) of the null hypothesis “ $q = 0$ ”.

When t^* is unknown, considering the maximum of $\Lambda_n(t)$ still gives the LRT of “ $q = 0$ ”. This paper gives the exact asymptotic distribution of this LRT statistic under the null hypothesis and under contiguous alternatives. These distributions

have been given using some approximations under the null hypothesis, by Rebaï et al. (1995) and Rebaï et al. (1994). In Cierco (1998), Azaïs and Cierco-Ayrolles (2002), Azaïs and Wschebor (2009), Chang et al. (2009), Azaïs et al. (2011), the authors focus on another recombination model which does not model the interference phenomenon : recombination events occur independently from each other.

The main result of the paper (Theorems 1 and 2) is that the distribution of the LRT statistic is asymptotically that of the maximum of the square of a “linear normalized interpolated process”. It is a generalization of the results obtained by Rebaï et al. (1995), Rebaï et al. (1994), where the authors focused only on the null hypothesis and characterized the process only by its covariance function. The computation of such a maximum is easy due to the interpolation. Note that recently, for a model without interference, Azaïs et al. (2011) have proved that the LRT statistic is asymptotically that of the maximum of the square of a “non linear normalized interpolated process”. The second important result is that, under the null hypothesis, the maximum of the square of the “linear normalized interpolated process” is the same as those of the square of the “non linear normalized interpolated process” obtained by Azaïs et al. (2011). As a consequence, the Monte-Carlo Quasi Monte-Carlo method proposed by Azaïs et al. (2011) to compute thresholds is also suitable for our interference model. So, for our interference model, we have now a method to compute thresholds which is suitable whatever the genetic map is, which was not the case of the method proposed in Rebaï et al. (1994) based on Davies (1977). With the help of simulated data, we will see that, as expected, our method outperforms Rebaï’s method in terms of false positives. Finally, we will compare the theoretical power of QTL detection, for a model with interference (Azaïs et al. (2011)) and a model without interference (this paper). We will show that it is largely more powerful to detect a QTL under interference than without interference. To sum up, we prove that we have exactly the same threshold with or without interference, but we have a totally different power. This makes this paper original.

We refer to the book of Van der Vaart (1998) for elements of asymptotic statistics used in proofs.

2. Main results : two genetic markers

To begin, we suppose that there are only two markers ($K = 2$) located at 0 and $T : 0 = t_1 < t_2 = T$. Furthermore, a QTL is lying between these two markers at $t^* \in]t_1, t_2[$. Note that in order to make the reading easier, we consider that the QTL is not located on markers. However, the result can be prolonged by continuity at makers locations.

Let $r(t_1, t_2)$ be the probability that there is a recombination between the two markers. Calculation on the Poisson distribution show that :

$$r(t_1, t_2) = \mathbb{P}(X(t_1)X(t_2) = -1) = \mathbb{P}(|N(t_1) - N(t_2)| \text{ odd}) = \frac{1}{2} (1 - e^{-2|t_1 - t_2|}).$$

We will call $r_{t_1}(t^*)$ (resp. $r_{t_2}(t^*)$) the probability of recombination between the first (resp. second) marker and the QTL. So,

$$r_{t_1}(t^*) = \mathbb{P}(X(t_1)U(t^*) = -1) , r_{t_2}(t^*) = \mathbb{P}(X(t_2)U(t^*) = -1).$$

As explained in Section 1, only one recombination is allowed between the QTL and the two markers. We have :

$$\{X(t_1)X(t_2) = -1\} \Leftrightarrow \{X(t_1)U(t^*) = -1\} \cup \{X(t_2)U(t^*) = -1\}.$$

Indeed, $X(t_1)U(t^*) = -1$ means that there has been a recombination between the first marker and the QTL, so the second marker is not allowed to recombine with the QTL. As a consequence, $X(t_2) = U(t^*)$ and we have $X(t_1)X(t_2) = -1$. Same remark for $X(t_2)U(t^*) = -1$ but this time, it is the first marker which is not allowed to recombine with the QTL.

Note that since $\{X(t_1)U(t^*) = -1\} \cap \{X(t_2)U(t^*) = -1\} = \emptyset$, we have

$$r(t_1, t_2) = r_{t_1}(t^*) + r_{t_2}(t^*). \tag{3}$$

In the same way as in Rebaï et al. (1995), we consider :

$$r_{t_1}(t^*) = \frac{t^* - t_1}{t_2 - t_1} r(t_1, t_2), \quad r_{t_2}(t^*) = \frac{t_2 - t^*}{t_2 - t_1} r(t_1, t_2).$$

This way, the probability of recombination of the marker and the QTL is proportional to the probability of recombination of the two markers, and also proportional to the distance between between the QTL and the marker. Note that formula (3) stands with these expressions of $r_{t_1}(t^*)$ and $r_{t_2}(t^*)$.

Let define now

$$p(t^*) = \mathbb{P}\{U(t^*) = 1 | X(t_1), X(t_2)\}.$$

Obviously, since $U(t^*)$ takes value +1 or -1, we have

$$1 - p(t^*) = \mathbb{P}\{U(t^*) = -1 | X(t_1), X(t_2)\}.$$

Since only one recombination is allowed between the QTL and its flanking markers, we have

$$\mathbb{P}\{U(t^*) = 1 | X(t_1) = 1, X(t_2) = 1\} = 1, \quad \mathbb{P}\{U(t^*) = 1 | X(t_1) = -1, X(t_2) = -1\} = 0.$$

Besides, according to Bayes rules

$$\begin{aligned} & \mathbb{P}\{U(t^*) = 1 | X(t_1) = 1, X(t_2) = -1\} \\ &= \frac{\mathbb{P}\{X(t_1) = 1 | U(t^*) = 1, X(t_2) = -1\} \mathbb{P}\{U(t^*) = 1, X(t_2) = -1\}}{\mathbb{P}\{X(t_1) = 1, X(t_2) = -1\}} \\ &= \frac{r_{t_2}(t^*)/2}{r(t_1, t_2)/2} = \frac{r_{t_2}(t^*)}{r(t_1, t_2)} = \frac{t_2 - t^*}{t_2 - t_1}. \end{aligned}$$

In the same way,

$$\mathbb{P}\{U(t^*) = 1 | X(t_1) = -1, X(t_2) = 1\} = \frac{r_{t_1}(t^*)}{r(t_1, t_2)} = \frac{t^* - t_1}{t_2 - t_1}.$$

As a consequence,

$$p(t^*) = 1_{X(t_1)=1}1_{X(t_2)=1} + \frac{t_2 - t^*}{t_2 - t_1} 1_{X(t_1)=1}1_{X(t_2)=-1} + \frac{t^* - t_1}{t_2 - t_1} 1_{X(t_1)=-1}1_{X(t_2)=1}. \tag{4}$$

So, as explained in Section 1, conditionally to $X(t_1)$ and $X(t_2)$, Y obeys to the mixture model of formula (2). Note that, using the formula above for $p(t^*)$, and using properties of conditional expectation, it is easy to check that we have $\mathbb{P}\{U(t^*)\} = 1/2$, so $U(t^*)$ takes values $+1$ and -1 with equal probability (as explained in Section 1). As explained previously, since the location t^* of the QTL is unknown, we will have to perform tests at each position t between the two genetic markers. We will consider only positions t distinct of the marker locations and the result can be prolonged by continuity on markers.

Let define now (with $p(t)$ given in formula (4))

$$u(t) = 2p(t) - 1 .$$

Let $\theta = (q, \mu, \sigma)$ be the parameter of the model at t fixed. The likelihood of the triplet $(Y, X(t_1), X(t_2))$ with respect to the measure $\lambda \otimes N \otimes N$, λ being the Lebesgue measure, N the counting measure on \mathbb{N} , is :

$$L_t(\theta) = [p(t)f_{(\mu+q,\sigma)}(y) + \{1 - p(t)\} f_{(\mu-q,\sigma)}(y)] g(t) \quad (5)$$

where the function

$$g(t) = \frac{1}{2} \{ \bar{r}(t_1, t_2) 1_{X(t_1)=1} 1_{X(t_2)=1} + r(t_1, t_2) 1_{X(t_1)=1} 1_{X(t_2)=-1} \} \\ + \frac{1}{2} \{ r(t_1, t_2) 1_{X(t_1)=-1} 1_{X(t_2)=1} + \bar{r}(t_1, t_2) 1_{X(t_1)=-1} 1_{X(t_2)=-1} \}$$

can be removed because it does not depend on the parameters. By a small abuse of notation we still denote $L_t(\theta)$ for the likelihood without this function. Thus we set

$$L_t(\theta) = [p(t)f_{(\mu+q,\sigma)}(y) + \{1 - p(t)\} f_{(\mu-q,\sigma)}(y)]$$

and $l_t(\theta)$ will be the loglikelihood. We first compute the Fisher information at a point θ_0 that belongs to H_0 .

$$\frac{\partial l_t}{\partial q} |_{\theta_0} = \frac{y - \mu}{\sigma^2} u(t) \quad (6)$$

$$\frac{\partial l_t}{\partial \mu} |_{\theta_0} = \frac{y - \mu}{\sigma^2} \quad , \quad \frac{\partial l_t}{\partial \sigma} |_{\theta_0} = -\frac{1}{\sigma} + \frac{(y - \mu)^2}{\sigma^3}$$

After some calculations, we find

$$I_{\theta_0} = \text{Diag} \left[\frac{\mathbb{E}\{u^2(t)\}}{\sigma^2} , \frac{1}{\sigma^2} , \frac{2}{\sigma^2} \right] . \quad (7)$$

Our main result is the following :

THEOREM 1. *Suppose that the parameters (q, μ, σ^2) vary in a compact and that σ^2 is bounded away from zero. Let H_0 be the null hypothesis $q = 0$ and define the following local alternative*

H_{at^*} : "the QTL is located at the position t^* with effect $q = a/\sqrt{n}$ where $a \neq 0$ ".

With the previous defined notations,

$$S_n(\cdot) \Rightarrow W(\cdot) \quad , \quad \Lambda_n(\cdot) \xrightarrow{F.d.} W^2(\cdot) \quad , \quad \sup \Lambda_n(\cdot) \xrightarrow{\mathcal{L}} \sup W^2(\cdot)$$

as n tends to infinity, under H_0 and H_{at^*} where :

- \Rightarrow is the weak convergence, $\xrightarrow{F.d.}$ is the convergence of finite-dimensional distributions and $\xrightarrow{\mathcal{L}}$ is the convergence in distribution
- $W(\cdot)$ is the Gaussian process with unit variance such as :

$$W(t) = \frac{\alpha(t)W(t_1) + \beta(t)W(t_2)}{\sqrt{\mathbb{V}\{\alpha(t)W(t_1) + \beta(t)W(t_2)\}}}$$

where

$$\begin{aligned} \text{Cov}\{W(t_1), W(t_2)\} &= \rho(t_1, t_2) = \exp(-2|t_1 - t_2|) \\ \alpha(t) &= \frac{t_2 - t}{t_2 - t_1}, \quad \beta(t) = \frac{t - t_1}{t_2 - t_1} \end{aligned}$$

and with expectation :

- under H_0 , $m(t) = 0$,
- under H_{at^*}

$$m_{t^*}(t) = \frac{\alpha(t) m_{t^*}(t_1) + \beta(t) m_{t^*}(t_2)}{\sqrt{\mathbb{V}\{\alpha(t)W(t_1) + \beta(t)W(t_2)\}}}$$

where

$$m_{t^*}(t_1) = \frac{a}{\sigma} \{\alpha(t^*) + \beta(t^*)\rho(t_1, t_2)\}, \quad m_{t^*}(t_2) = \frac{a}{\sigma} \{\alpha(t^*)\rho(t_1, t_2) + \beta(t^*)\}.$$

As a consequence, $W(\cdot)$ will be called a "linear normalized interpolated process".

In Azaïs et al. (2011), the authors present a lemma called Lemma 1, which is very useful to compute the supremum of the square of an interpolated process. So, the computation of the maximum of our process $W^2(\cdot)$ can be obtained easily using their Lemma 1, since $\frac{\beta(t)}{\alpha(t)+\beta(t)}$ takes every value in $[0, 1]$ (cf. Azaïs et al. (2011)).

On the other hand, we have this interesting result :

LEMMA 1. *With the previous defined notations, under H_0 ,*

$$\max_{t \in [t_1, t_2]} W^2(t) = \max_{t \in [t_1, t_2]} Z^2(t),$$

where $Z(\cdot)$ is the "non linear normalized interpolated process" obtained by Azaïs et al. (2011).

In other words, under the null hypothesis, our Lemma 1 says that the maximum of the square of the "non linear normalized interpolated process" is the same as the maximum of the square of the "linear normalized interpolated process".

In order to prove this lemma, we just have to remark that under H_0 at marker locations, we have $Z(t_1) = W(t_1)$ and $Z(t_2) = W(t_2)$. Indeed, under H_0 , the processes overlap at marker locations since there are no QTL affecting the processes and also because the recombination model (ie Haldane) is the same at marker locations. Then, using Lemma 1 of Azaïs et al. (2011), the computation of the maximum of $Z^2(\cdot)$ and $W^2(\cdot)$ is the same.

Note that our Lemma 1 stands only under the null hypothesis and not under the alternative.

Proof of Theorem 1 :

Introducing the score process

The log likelihood at t , associated to n observations will be denoted by $l_t^n(\theta)$. Since the Fisher Information matrix is diagonal, the score statistics of the hypothesis “ $q = 0$ ” will be defined as

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q} |_{\theta_0}}{\sqrt{\mathbb{V}\left(\frac{\partial l_t^n}{\partial q} |_{\theta_0}\right)}}.$$

Study of the score process under the null hypothesis

The study is based on the key lemma :

LEMMA 2.

$$u(t) = \alpha(t)X(t_1) + \beta(t)X(t_2)$$

$$\text{with } \alpha(t) = \frac{t_2-t}{t_2-t_1} \text{ and } \beta(t) = \frac{t-t_1}{t_2-t_1}.$$

To prove this lemma use formula (4) and check that both coincide whatever the value of $X(t_1)$, $X(t_2)$ is.

Now using (6), we have

$$\frac{\partial l_t^n}{\partial q} |_{\theta_0} = \sum_{j=1}^n \frac{Y_j - \mu}{\sigma^2} u_j(t) = 1/\sigma \sum_{j=1}^n \varepsilon_j u_j(t) = \frac{\alpha(t)}{\sigma} \sum_{j=1}^n \varepsilon_j X_j(t_1) + \frac{\beta(t)}{\sigma} \sum_{j=1}^n \varepsilon_j X_j(t_2) \quad (8)$$

this proves the interpolation.

On the other hand

$$S_n(t_k) = \sum_{j=1}^n \frac{\varepsilon_j X_j(t_k)}{\sqrt{n}} \quad k = 1, 2$$

and a direct application of central limit theorem implies that these two variables have a limit distribution which is Gaussian centered distribution with variance

$$\begin{pmatrix} 1 & \exp(-2|t_2 - t_1|) \\ \exp(-2|t_2 - t_1|) & 1 \end{pmatrix}.$$

This proves the expression of the covariance.

Study of the score process under the local alternative

Under the alternative

$$S_n(t) = \frac{a}{n\sigma} \sum_{j=1}^n \frac{U_j(t^*)u_j(t)}{\sqrt{\mathbb{V}\{u(t)\}}} + \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j \frac{u_j(t)}{\sqrt{\mathbb{V}\{u(t)\}}}.$$

The second term has the same distribution as under the null hypothesis and the first one gives the expectation. We have

$$\mathbb{E}\{S_n(t)\} = \frac{a \mathbb{E}\{U(t^*)u(t)\}}{\sigma \sqrt{\mathbb{V}\{u(t)\}}}.$$

According to Lemma 2, we have :

$$\mathbb{E} \{U(t^*)u(t)\} = \alpha(t) \mathbb{E} \{X(t_1)U(t^*)\} + \beta(t) \mathbb{E} \{U(t^*)X(t_2)\}.$$

So, we need now to calculate $\mathbb{E} \{X(t_1)U(t^*)\}$ and $\mathbb{E} \{U(t^*)X(t_2)\}$. We have

$$\begin{aligned} \mathbb{P} \{X(t_1)U(t^*) = -1\} &= \mathbb{P} \{U(t^*) = 1 \mid X(t_1) = -1, X(t_2) = 1\} \mathbb{P} \{X(t_1) = -1, X(t_2) = 1\} \\ &\quad + \mathbb{P} \{U(t^*) = 1 \mid X(t_1) = -1, X(t_2) = -1\} \mathbb{P} \{X(t_1) = -1, X(t_2) = -1\} \\ &\quad + \mathbb{P} \{U(t^*) = -1 \mid X(t_1) = 1, X(t_2) = 1\} \mathbb{P} \{X(t_1) = 1, X(t_2) = 1\} \\ &\quad + \mathbb{P} \{U(t^*) = -1 \mid X(t_1) = 1, X(t_2) = -1\} \mathbb{P} \{X(t_1) = 1, X(t_2) = -1\} \\ &= \frac{\beta(t^*)r(t_1, t_2)}{2} + 0 + 0 + \frac{\beta(t^*)r(t_1, t_2)}{2} = \beta(t^*)r(t_1, t_2). \end{aligned}$$

As a consequence,

$$\mathbb{P} \{X(t_1)U(t^*) = 1\} = 1 - \beta(t^*)r(t_1, t_2).$$

It comes

$$\mathbb{E} \{X(t_1)U(t^*)\} = 1 - 2\beta(t^*)r(t_1, t_2) = \alpha(t^*) + \beta(t^*)\rho(t_1, t_2) \text{ with } \rho(t_1, t_2) = e^{-2|t_1-t_2|}.$$

In the same way, we obtain

$$\mathbb{E} \{U(t^*)X(t_2)\} = \alpha(t^*)\rho(t_1, t_2) + \beta(t^*).$$

This gives the result.

About the LRT process

The likelihood ratio statistic at t , for n independent observations, will be defined as

$$\Lambda_n(t) = 2 \left\{ l_t^n(\hat{\theta}) - l_t^n(\hat{\theta}_{|H_0}) \right\},$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE), and $\hat{\theta}_{|H_0}$ the MLE under H_0 .

Since the model with t fixed is regular, it is easy to prove that for fixed t

$$\Lambda_n(t) = S_n^2(t) + o_P(1) \tag{9}$$

under the null hypothesis.

Let us consider a local alternative defined by t^* and $q = a/\sqrt{n}$. The model with t^* fixed is differentiable in quadratic mean, this implies that the alternative defines a contiguous sequence of alternatives. By Le Cam's first Lemma, relation (9) remains true under the alternative. This gives the result for the convergence of finite-dimensional distribution. Concerning the study of the supremum of the LRT process, the proof is exactly the same as in Azaïs et al. (2011) which is based on recent results of Azaïs et al. (2006).

3. Several markers

In that case suppose that there are K markers $0 = t_1 < t_2 < \dots < t_K = T$. A QTL is lying at a position t^* . So, in order to find the QTL, we will perform tests at every positions t on the chromosome. We consider values t or t^* of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions. For $t \in [t_1, t_K] \setminus \mathbb{T}_K$ where $\mathbb{T}_K = \{t_1, \dots, t_K\}$, we define t^ℓ and t^r as :

$$t^\ell = \sup \{t_k \in \mathbb{T}_K : t_k < t\} \quad , \quad t^r = \inf \{t_k \in \mathbb{T}_K : t < t_k\} .$$

In other words, t belongs to the "Marker interval" (t^ℓ, t^r) .

As explained in Section 1, in order to infer the value of $U(t^*)$, we just need to keep the flanking markers. In others words, the information brought by the other markers is useless. So, we have

$$\mathbb{P} \{U(t^*) = 1 | X(t_1), \dots, X(t_K)\} = \mathbb{P} \{U(t^*) = 1 | X(t^{\ell}), X(t^{*r})\} .$$

As a consequence, our problem becomes the same as the one with two genetic markers (see Section 2). In order to perform our tests at every positions t , we simply have to consider all the different marker interval.

THEOREM 2. *We have the same results as in Theorem 1 except that the following functions must be redefined :*

- t_1 becomes t^ℓ and t_2 becomes t^r in all the expressions, except in the expressions $\alpha(t^*)$ and $\beta(t^*)$, where t_1 becomes t^{ℓ} and t_2 becomes t^{*r}
- $m_{t^*}(t^\ell) = \frac{a}{\sigma} \{ \alpha(t^*)\rho(t^\ell, t^{\ell}) + \beta(t^*)\rho(t^\ell, t^{*r}) \}$
- $m_{t^*}(t^r) = \frac{a}{\sigma} \{ \alpha(t^*)\rho(t^{\ell}, t^r) + \beta(t^*)\rho(t^r, t^{*r}) \}$.

Proof of Theorem 2 :

The proof of the theorem is the same the proof of Theorem 1 as soon as we can limit our attention to the interval (t^ℓ, t^r) when considering a unique instant t . So, under H_0 , the result is straightforward. However, under local the alternative, the proof is more complicated than the proof of Theorem 1. Indeed, the location t^* of the QTL and the location t , can belong to a different marker interval.

According to the proof of Theorem 1, under the alternative

$$S_n(t) = \frac{a}{n\sigma} \sum_{j=1}^n \frac{U_j(t^*)u_j(t)}{\sqrt{\mathbb{V}\{u(t)\}}} + \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j \frac{u_j(t)}{\sqrt{\mathbb{V}\{u(t)\}}} .$$

As previously, the second term has the same distribution as under the null hypothesis and the first one gives the expectation. We have

$$\mathbb{E} \{S_n(t)\} = \frac{a \mathbb{E} \{U(t^*)u(t)\}}{\sigma \sqrt{\mathbb{V}\{u(t)\}}} .$$

We remark that we have $u(t^*) = \mathbb{E} \{U(t^*) | X(t^{\ell})X(t^{*r})\}$. Besides, $u(t)$ is a function of $X(t^\ell)$ and $X(t^r)$. As a consequence, by the properties of conditional expectancy, we have

$$\mathbb{E} \{U(t^*)u(t)\} = \mathbb{E} \{u(t^*)u(t)\} .$$

According to Lemma 2,

$$\begin{aligned} \mathbb{E}\{u(t^*)u(t)\} &= \alpha(t^*) \alpha(t) \mathbb{E}\{X(t^{*\ell})X(t^\ell)\} + \beta(t^*) \alpha(t) \mathbb{E}\{X(t^{*r})X(t^\ell)\} \\ &\quad + \alpha(t^*) \beta(t) \mathbb{E}\{X(t^{*\ell})X(t^r)\} + \beta(t^*) \beta(t) \mathbb{E}\{X(t^{*r})X(t^r)\} \\ &= \alpha(t^*) \alpha(t) \rho(t^\ell, t^{*\ell}) + \beta(t^*) \alpha(t) \rho(t^\ell, t^{*r}) \\ &\quad + \alpha(t^*) \beta(t) \rho(t^{*\ell}, t^r) + \beta(t^*) \beta(t) \rho(t^r, t^{*r}) . \end{aligned}$$

In order to obtain $\mathbb{E}\{u(t^*)u(t^\ell)\}$, we just have to use the dominated convergence theorem. It comes

$$\mathbb{E}\{u(t^*)u(t^\ell)\} = \alpha(t^*) \rho(t^\ell, t^{*\ell}) + \beta(t^*) \rho(t^\ell, t^{*r}) .$$

In the same way,

$$\mathbb{E}\{u(t^*)u(t^r)\} = \alpha(t^*) \rho(t^{*\ell}, t^r) + \beta(t^*) \rho(t^r, t^{*r}) .$$

This gives the result.

4. Application

In this Section, we present some applications of our study. We first focus on the null hypothesis and then we will move on to the alternative hypothesis.

4.1. Application to the computation of thresholds

In QTL detection, in order to conclude to the presence of a QTL or not, it is always important to use an appropriate threshold for the statistical test. Our aim is to show that with our theoretical study, we are now able to propose a threshold which gives better performances than the classical threshold proposed by Rebaï et al. (1995) and Rebaï et al. (1994) for the interference model.

To begin, we remind that $W(\cdot)$ is our "linear normalized interpolated process" whereas $Z(\cdot)$ is the "non linear normalized interpolated process" of Azaïs et al. (2011). According to Lemma 1, when we consider only two genetic markers, the maximum of $W^2(\cdot)$ is the same as the maximum of $Z^2(\cdot)$ under the null hypothesis. Since when we deal with several markers, we just have to consider the different marker intervals, it is easy to check that Lemma 1 is still true with several markers. This way, the threshold will be the same for a model with interference (this paper) and for a model without interference (Azaïs et al. (2011)). In order to compute the threshold, Azaïs et al. propose a Monte-Carlo Quasi Monte-Carlo (MCQMC) method, based on Genz (1992). This method is very fast, and the advantage of MCQMC is that it is more accurate than a simple Monte-Carlo method. We refer to Azaïs et al. (2011) and Genz (1992) for more details.

Let's explain now the method to compute thresholds, proposed by Rebaï et al. (1995) and Rebaï et al. (1994). In Rebaï et al. (1995), the authors consider only two markers. They propose to use results of Davies (1977) and Davies (1987). Indeed, in Davies, we can find an upper bound for a threshold corresponding to the supremum of a stochastic process (Gaussian process or Chi square process) which depends on a nuisance parameter only present under the alternative. In

QTL detection, the nuisance parameter is the position of the QTL. Note that in Rebaï et al. (1995), the authors use as a scale the recombination units whereas in this paper, we use the genetic distance. In other words, if we call $W'(\cdot)$ the process studied in Rebaï et al. (1995) with only two markers, we have the relationship $\forall t \in [t_1, t_2]$:

$$W(t) = W' \left\{ r(t_1, t_2) \frac{t - t_1}{t_2 - t_1} \right\} .$$

In their paper, they show that

$$\frac{\partial^2 \text{Cov} \{W'(t), W'(t')\}}{\partial t'^2} \Big|_{t'=t} = - \frac{4 r(t_1, t_2) \{1 - r(t_1, t_2)\}}{\left[r(t_1, t_2) - 4r^2(t_1, t_2) \frac{t-t_1}{t_2-t_1} + 4 \left\{ r(t_1, t_2) \frac{t-t_1}{t_2-t_1} \right\}^2 \right]^2} .$$

Then, since

$$\int_0^{r(t_1, t_2)} \sqrt{-\frac{\partial^2 \text{Cov} \{W'(t), W'(t')\}}{\partial t'^2} \Big|_{t'=t}} dt = 2 \arctan \left(\sqrt{\frac{r(t_1, t_2)}{1 - r(t_1, t_2)}} \right)$$

and using Davies formula, they find that

$$\mathbb{P} \left\{ \sup_{[0, r(t_1, t_2)]} W'(t) > c \right\} \leq \Phi(-c) + \frac{e^{-c^2/2}}{\pi} \arctan \left(\sqrt{\frac{r(t_1, t_2)}{1 - r(t_1, t_2)}} \right) ,$$

where Φ is the cumulative distributive function of a standardized normal distribution. Note that since

$$\mathbb{P} \left\{ \sup_{[t_1, t_2]} W(t) > c \right\} = \mathbb{P} \left\{ \sup_{[0, r(t_1, t_2)]} W'(t) > c \right\} ,$$

it gives also the threshold for our process $W(\cdot)$. In Rebaï et al. (1994), the authors generalize their approach to several markers. Their formula adapted to our process $W(\cdot)$ becomes :

$$\mathbb{P} \left\{ \sup_{[t_1, t_K]} W(t) > c \right\} \leq \Phi(-c) + \frac{e^{-c^2/2}}{\pi} \sum_{k=1}^{K-1} \arctan \left(\sqrt{\frac{r(t_k, t_{k+1})}{1 - r(t_k, t_{k+1})}} \right) . \quad (10)$$

In order to obtain the threshold, we just have to find for which value of c , the right-side of formula (10) is equal to $\alpha/2$, and we will obtain the threshold c^2 for the supremum of our process $W^2(\cdot)$. Note that this threshold c^2 will only correspond to a level lower or equal than α , due to the upper bound of formula (10).

In Figure 1, we propose to compare numerically, the two approaches to compute thresholds for the interference model : Azaïis et al. (2011) and Rebaï et al. (1994). For the genetic map, we consider the same configurations as in Table 1 of Rebaï et al. (1994), that is to say a chromosome of length $T = 1\text{M}$, different numbers of markers, and a level α equal to 5%. According to Figure 1, we can see that the two approaches give different thresholds. It was expected since

Rebai's threshold correspond only to a level lower or equal to 5%. Besides, the more markers there are, the more different the thresholds are. It is due to the fact that the derivative of the process $W(\cdot)$ has a jump at each markers location, and Davies (1977) formula is suitable when the derivative of the process has a finite number of jumps. In other words, the more markers there are, the less appropriate Rebai's threshold will be.

To conclude, since the two approaches are based on asymptotic results, we propose to check the asymptotic validity on simulated data. We simulated under the null hypothesis, 10000 samples of $n = 200$ individuals. We analyzed data using Lemma 1 of Azaïs et al. (2011) (still suitable here, cf. our Section 2), that is to say performing LRT on markers and performing only one test in each marker interval if the ratio of the score statistics on markers fulfilled a given condition. According to Figure 1, Azaïs' method always gives a percentage of false positives close to 5%, whereas Rebai's method is too conservative. So, for our interference model, we have now a method to compute thresholds which is suitable whatever the genetic map is, and which does not require the number of individuals n to be too large.

4.2. About the power

We focus now on the alternative hypothesis. In our paper, double recombination between the QTL and its flanking markers is not allowed. This way, we model the interference phenomenon. In Azaïs et al. (2011), since the authors don't model interference, double recombination between the QTL and its flanking markers is allowed. The main difference is that, for an interference model, the LRT process is asymptotically the square of a linear interpolated and normalized process (ie. $W(\cdot)$), whereas for a model without interference, the LRT process is asymptotically the square of a non linear interpolated and normalized process (ie. $V(\cdot)$). In Figures 2, 3, 4 and 5, we propose to compare the asymptotic power of the two approaches, using these asymptotic processes. We consider $a = 4$ (ie. the constant for the QTL effect) and 100000 paths of each process. First, in Figures 2, 3, 4, we consider some sparse maps. In Figure 2, we consider a chromosome of length $T = 1\text{M}$ and 2 markers are located at each extremity of the chromosome. We can see that when the QTL is located at $t^* = 30\text{cM}$ and $t^* = 60\text{cM}$, there are huge differences of power between the model with interference and the model without interference. For instance, we have 85.10% chances of detecting a QTL located at 30cM with interference, whereas we have only 49.77% chances of detecting the same QTL without interference. This is due to the fact that the mean functions are totally different between the two asymptotic processes. We obtain the same kind of conclusions in Figures 3 and 4 for other sparse maps. In Figure 5, we consider a more dense map : a chromosome of length $T = 1\text{M}$ and 6 markers equally spaced every 20cM. We can see that there is now only a little difference of power. To conclude, in the same way as what has been done in the previous section, we propose to check the asymptotic validity of our asymptotic results. So, in Figure 6, we consider the same configuration as in Figure 2 : a chromosome of length $T = 1\text{M}$ and 2 markers located at each extremity. We simulated 10000 samples of $n = 50$, $n = 100$, $n = 200$, $n = 1000$ individuals, according to the interference model. We can see that for $n = 200$, we are close to the asymptotic results. It validates

our asymptotic study.

Method \ number of markers	101	51	41	26	6
Rebai	9.74 2.69%	9.09 3.23%	8.88 3.77%	8.43 4.04%	6.92 4.83%
Azais et al.	8.41 5.03%	8.27 4.80%	8.16 5.32%	7.91 5.21%	6.76 5.19%

Fig. 1. Threshold and Percentage of False Positives as a function of the number of markers and the method considered. The chromosome is of length $T = 1M$ and the markers are equally spaced.

Model \ t^*	10cM	30cM	60cM	80cM
interference	92.89%	85.10%	82.12%	89.16%
without interference	86.01%	49.77%	47.46%	70.90%

Fig. 2. Asymptotic power of the Interval Mapping as a function of the model considered and the location of the QTL t^* . The chromosome is of length $T = 1M$ and 2 markers are located at each extremity ($a = 4, \sigma = 1$).

Model \ t^*	20cM	70cM	90cM	1.2M
interference	88.26%	80.11%	59.18%	22.29%
without interference	74.82%	64.57%	28.68%	9.52%

Fig. 3. Asymptotic power of the Interval Mapping as a function of the model considered and the location of the QTL t^* . The chromosome is of length $T = 1.5M$ and 3 markers are located at $t_1 = 0cM, t_2 = 50cM, t_3 = 1.5M$ ($a = 4, \sigma = 1$).

Model \ t^*	40cM	90cM	1.2M	1.7M
interference	76.62%	88.75%	79.46%	83.27%
without interference	49.49%	81.26%	59.12%	73.30%

Fig. 4. Asymptotic power of the Interval Mapping as a function of the model considered and the location of the QTL t^* . The chromosome is of length $T = 2M$ and 4 markers are located at $t_1 = 0cM, t_2 = 80cM, t_3 = 1.5M, t_4 = 2M$ ($a = 4, \sigma = 1$).

Model \ t^*	18cM	44cM	70cM
interference	93.52%	92.03%	90.45%
without interference	92.59%	91.34%	89.18%

Fig. 5. Asymptotic power of the Interval Mapping as a function of the model considered and the location of the QTL t^* . The chromosome is of length $T = 1\text{M}$ and 6 markers are equally spaced every 20cM ($a = 4, \sigma = 1$).

n \ t^*	10cM	30cM	60cM	80cM
1000	92.76%	85.10%	81.20%	88.12%
200	92.18%	83.81%	80.29%	87.94%
100	91.50%	81.89%	78.45%	86.75%
50	89.70%	78.90%	74.46%	83.53%
theoretical power	92.89%	85.10%	82.12%	89.16%

Fig. 6. Asymptotic power and Empirical power as a function of n and the location of the QTL t^* (interference model). The chromosome is of length $T = 1\text{M}$ and 2 markers are located at each extremity ($a = 4, \sigma = 1$).

5. Acknowledgements

The author thank Jean-Marc Azaïs, Céline Delmas and Jean-Michel Elsen for fruitful discussions. This work has been supported by the Animal Genetic Department of the French National Institute for Agricultural Research, SABRE, and the National Center for Scientific Research.

Charles-Elie Rabier (rabier@stat.wisc.edu)

Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., F-31062 Toulouse Cedex 9, France.

INRA UR631, Station d'Amélioration Génétique des Animaux, BP 52627-31326 Castanet-Tolosan Cedex, France.

References

Azaïs, J. M. and Cierco-Ayrolles, C. (2002). An asymptotic test for quantitative gene detection. *Ann. I. H. Poincaré*, **38**, **6**, 1087-1092.

Azaïs, J. M., Gassiat, E., Mercadier, C. (2006). Asymptotic distribution and local power of the likelihood ratio test for mixtures. *Bernoulli*, **12**(5), 775-799.

Azaïs, J. M., Gassiat, E., Mercadier, C. (2009). The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM*, To appear.

Azaïs, J. M. and Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. Wiley, New-York.

- Azaïs, J. M., Delmas C., Rabier, C.E. (2011). *Likelihood ratio test process for Quantitative Trait Locus detection*. Submitted to *ESAIM*.
- Billingsley, P. (1999). *Convergence of probability measures*. Wiley, New-York.
- Chang, M. N., Wu, R., Wu, S. S., Casella, G. (2009). Score statistics for mapping quantitative trait loci. *Statistical Application in Genetics and Molecular Biology*, **8**(1), 16.
- Cierco, C. (1998). Asymptotic distribution of the maximum likelihood ratio test for gene detection. *Statistics*, **31**, 261-285.
- Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247-254.
- Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33-43.
- Feingold, E., Brown, P.O., Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Human. Genet.*, **53**, 234-251.
- Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. I. H. Poincaré*, **6**, 897-906.
- Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Stat.*, 141-149.
- Ghosh, J.K., Sen, P.K. (1984). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. *Inst. Statistics Mimeo Series*, 1467.
- Haldane, J.B.S (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*, **8**, 299-309.
- Lander, E.S., Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138**, 235-240.
- Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer.
- McPeck, M. S., Speed, T. P. (1995). Modeling interference in genetic recombination. *Genetics*, **139**, 1031-1044.
- Muller, H.J. (1916). The mechanism of crossing-over. *Am. Nat.*, **50**, 193-221, 284-305, 350-366, 421-434.
- Rabier, C-E. (2010). *PhD thesis*, Université Toulouse 3, Paul Sabatier.
- Rebaï, A., Goffinet, B., Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138**, 235-240.
- Rebaï, A., Goffinet, B., Mangin, B. (1995). Comparing power of different methods for QTL detection. *Biometrics*, **51**, 87-99.
- Siegmund, D. (1985). *Sequential analysis : tests and confidence intervals*. Springer, New York.

- Sturtevant, A.H. (1915). The behavior of the chromosomes as studied through linkage. *Z. Indukt. Abstammungs. Vererbungsl.*, **13**, 234-287.
- Van der Vaart, A.W. (1998) *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.
- Wu, R., MA, C.X., Casella, G. (2007) *Statistical Genetics of Quantitative Traits*, Springer