On the asymptotic robustness of the Likelihood Ratio Test in Quantitative Trait Locus detection

Charles-Elie Rabier

On the asymptotic robustness of the Likelihood Ratio Test in Quantitative Trait Locus detection

C-E. Rabier[a,*]

[a] *University of Wisconsin-Madison, Statistic and Botany departments, Medical Science Center, 1300 University Avenue, Madison, WI 53706-1532, USA.*

Abstract

We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL (i.e. a gene with quantitative effect on a trait) on a chromosome. We consider two different recombination models. We prove that even if the LRT is constructed from the false recombination model (i.e. the model which does not correspond to the one of the data), the maximum of the LRT process converges asymptotically to the maximum of the LRT process constructed from the true recombination model. We also prove that under some conditions, the arg max of the LRT processes will be different.

Keywords: QTL detection, Likelihood Ratio Test, Gaussian process, Chi-Square process, Interference Phenomenon

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 $Y_j,\ j = 1, ..., n$, 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

*Corresponding author. Tel.:+1 608 265 9876; fax.:+1 608 262 0032
Email address: rabier@stat.wisc.edu (C-E. Rabier)

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 (see for instance Wu et al. [19] or Siegmund and Yakir [16]). K genetic markers are located at fixed locations $t_1 = 0 < t_2 < ... < t_K = T$. 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 [9] 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 $N(.)$ is a standard Poisson process on $[0, T]$. A QTL is lying at an unknown position $t^\star$ between two genetic markers. $U(t^\star)$ is the genetic information at the QTL location. In the same way as for the genetic information at marker locations, $U(t^\star)$ takes value $+1$ if the "recombined chromosome" is originated from A at $t^\star$, and -1 if it is originated from B. Due to Mendel law, $U(t^\star)$ takes value $+1$ and -1 with equal probability. We assume an "analysis of variance model" for the quantitative trait :

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

where ε is a Gaussian white noise.

The originality of this paper is that, inside the marker interval which contains the QTL, we consider two different recombination models. Indeed, it is always difficult for geneticists to know which model to use when they analyze real data. Obviously, for a given data set, geneticists usually try to use the most appropriate recombination model. However, it can happen that we do not choose the correct recombination model. As a consequence, the main question is : how does it affect QTL detection ? This way, in this paper, the focus is on the robustness of statistical tests in QTL detection. For the following, we will call "true recombination" model, the recombination model of the data whereas we will call "false recombination" model, the recombination model which is not the one of the data.

In particular, we will consider that our true recombination model, inside the marker interval which contains the QTL, is the Haldane model (i.e. the same model as the one at marker locations). Due to the independence of

increments of Poisson process, this model allows double recombinations between the QTL and its flanking markers. For instance, if the QTL is lying between the first two markers (i.e. $t^\star \in]t_1, t_2[$, we can have the scenario $X(t_1) = 1$, $U(t^\star) = -1$ and $X(t_2) = 1$, which means that there has been a recombination between the first marker and the QTL, and also a recombination between the QTL and the second marker. Obviously, in the same way, we can have the scenario $X(t_1) = -1$, $U(t^\star) = 1$ and $X(t_2) = -1$.

The false recombination model that we use in this paper (and which is chosen by geneticists) is the one proposed by Rebaï et al. [15] for which double recombination between the QTL and its flanking markers is not allowed (see in particular their Section 2). With this model, the focus is on the interference phenomenon : a recombination event inhibits the formation of another recombination event nearby. This phenomenon was noticed by geneticists working on the Drosophila (Sturtevant [17], Muller [12]). In McPeek and Speed [11], the authors study several interference models and also mention the importance of modeling interference. Let $\tilde{U}(t^\star)$ be the analogue of $U(t^\star)$ of formula (1) but for the interference model. Due to Mendel law, $\tilde{U}(t^\star)$ still takes value $+1$ and -1 with equal probability. The "analysis of variance model" for the quantitative trait is now :

$$Y = \mu \, + \, \tilde{U}(t^\star) \, q \, + \, \sigma\varepsilon \tag{2}$$

So, under the interference model, if the QTL is lying between the first two markers (i.e. $t^\star \in]t_1, t_2[$), we can not have the scenario $X(t_1) = 1$, $\tilde{U}(t^\star) = -1$ 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 considers 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 $\tilde{U}(t^\star) = -1$, then we have automatically $X(t_2) = -1$. In the same way, if $X(t_2) = 1$ and $\tilde{U}(t^\star) = -1$, then we have automatically $X(t_1) = -1$. We will explain in details this model in Section 2 and present the law of $\tilde{U}(t^\star)$, given its flanking markers. Note that in Rebaï et al. [15], the focus is only on one marker interval. In Rebaï et al. [14], this model was extended to a whole chromosome.

As said previously, only the quantitative trait and the genetic information at marker locations are available. As a consequence, one observation will be

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

We observe n observations (Y_j, $X_j(t_1)$, ..., $X_j(t_K)$) i.i.d. Under the model without interference (cf. formula 1), and conditionally to $X(t_1), \ldots, X(t_K)$, Y obeys to a mixture model with known weights :

$$p(t^*) f_{(\mu+q,\sigma)}(.) + \{1 - p(t^*)\} f_{(\mu-q,\sigma)}(.), \tag{3}$$

where $f_{(m,\sigma)}$ is the Gaussian density with parameters (m, σ) and where the function $p(t^\star)$ is the probability $\mathbb{P}\{U(t^\star) = 1\}$ conditionally to the flanking markers (see Azaïs et al. [2] and in particular their formula 3).

Furthermore, under the interference model (cf. formula 2) and conditionally to $X(t_1), \ldots, X(t_K)$, Y obeys to a mixture model with known weights :

$$\tilde{p}(t^*) f_{(\mu+q,\sigma)}(.) + \{1 - \tilde{p}(t^*)\} f_{(\mu-q,\sigma)}(.), \tag{4}$$

where the function $\tilde{p}(t^\star)$ is the probability $\mathbb{P}\left\{\tilde{U}(t^\star) = 1\right\}$ conditionally to the flanking markers (see Section 2) .

A challenge in QTL detection is that the true location t^* is not known. So, we test the presence of a QTL at each position $t \in [0, T]$. For the model without interference, $\Lambda_n(t)$ and $S_n(t)$ are respectively the likelihood ratio test (LRT) statistic and the score test statistic of the null hypothesis "$q = 0$" in formula (3). In the same way, for the interference model, $\tilde{\Lambda}_n(t)$ and $\tilde{S}_n(t)$ are respectively the likelihood ratio test (LRT) statistic and the score test statistic of the null hypothesis "$q = 0$" in formula (4). When t^* is unknown, considering the maximum of $\Lambda_n(t)$ (resp. $\tilde{\Lambda}_n(t)$) still gives the LRT of "$q = 0$" for the model without (resp. with) interference.

Under the model without interference, the distributions of the LRT, $\sup \Lambda_n(.)$, have been given using some approximations by Cierco [6], Azaïs and Cierco-Ayrolles [1], Azaïs and Wschebor [4], Chang et al. [5]. Recently, Azaïs et al. [2] have given the exact distribution under the null hypothesis and contiguous alternatives : the distribution of the LRT statistic is asymptotically that of the maximum of the square of a "non linear normalized interpolated Gaussian process". Under the interference model, I have proved in Rabier [13] that the distribution of the LRT statistic, $\sup \tilde{\Lambda}_n(.)$, 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. [15], Rebaï et al. [14], where the authors focused only on the null hypothesis and characterized the process only by its covariance function.

In this paper, we propose to study the distribution of $\sup \tilde{\Lambda}_n(.)$ under the model without interference. In other words, we focus on a test statistic constructed from the false recombination model (i.e. with interference), and we study its distribution under the true recombination model (i.e. without interference). The goal is to compare its distribution with the distribution of $\sup \Lambda_n(.)$ (given in Azaïs et al. [2]), constructed from the true recombination model.

The main result of the paper (Theorems 1 and 3) is that, under the true model (i.e. without interference), the distribution of the LRT statistic, $\sup \tilde{\Lambda}_n(.)$, is asymptotically that of the maximum of the square of a "linear normalized interpolated process". The second important result (Theorems 2 and 4) is that, under the null hypothesis and contiguous alternatives, the maximum of the square of this "linear normalized interpolated process" is the same as those of the square of the "non linear normalized interpolated process" obtained by Azaïs et al. [2]. That is to say, under the model without interference, we have the following relationship :

$$\sup \tilde{\Lambda}_n(.) = \sup \Lambda_n(.) + o_P(1) \ , \tag{5}$$

where $o_P(1)$ denotes a sequence of random vectors which tend to 0 in probability. As a consequence, there is "an asymptotic robustness of the likelihood ratio test" : even if we choose the false model in order to construct our LRT statistic, we will get asymptotically the optimal power for the detection of the QTL. On the other hand, Lemma 1 gives asymptotic results about $\arg\sup \tilde{\Lambda}_n(.)$ and $\arg\sup \Lambda_n(.)$. According to Lemma 1, under some given conditions, if we choose the false model, the location of the QTL will be estimated differently.

At the end of the paper, the focus is on the reverse configuration : now the true recombination model is the model with interference and the false model is the one without interference. We prove that formula (5) is still true under the model with interference. As a result, we can really use the terminology "asymptotic robustness of the likelihood ratio test" in QTL detection. This is a result which could be of interest for geneticists.

Note that a direct consequence of the results presented in this paper, is that in order to compute thresholds, the Monte-Carlo Quasi Monte-Carlo method proposed by Azaïs et al. [2] and based on Genz [8], is suitable for any model. We refer to the book of Van der Vaart [18] for elements of asymptotic statistics used in proofs.

2. Main results : two genetic markers

To begin with, 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 $t^\star \in]t_1, t_2[$. 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 marker locations. Let's suppose that we are under the interference model (cf. Section 1).

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

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

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

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

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

$$\left\{X(t_1)X(t_2) = -1\right\} \Leftrightarrow \left\{X(t_1)\tilde{U}(t^\star) = -1\right\} \cup \left\{X(t_2)\tilde{U}(t^\star) = -1\right\}.$$

Indeed, $X(t_1)\tilde{U}(t^\star) = -1$ 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, $X(t_2) = \tilde{U}(t^\star)$ and we have $X(t_1)X(t_2) = -1$. Same remark for $X(t_2)\tilde{U}(t^\star) = -1$ but this time, it is the first marker which is not allowed to recombine with the QTL.

Note that since $\left\{X(t_1)\tilde{U}(t^\star) = -1\right\} \cap \left\{X(t_2)\tilde{U}(t^\star) = -1\right\} = \oslash$, we have

$$r(t_1, t_2) = r_{t_1}(t^\star) + r_{t_2}(t^\star). \tag{6}$$

In the same way as in Rebaï et al. [15], we consider :

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

This way, the probability of recombination between the marker and the QTL is proportional to the probability of recombination between the two markers,

and also proportional to the distance between the QTL and the marker. Note that formula (6) stands with these expressions of $r_{t_1}(t^\star)$ and $r_{t_2}(t^\star)$.

Let's define now

$$\tilde{p}(t^\star) = \mathbb{P}\left\{\tilde{U}(t^\star) = 1 \middle| X(t_1), X(t_2)\right\}.$$

Obviously, since $\tilde{U}(t^\star)$ takes value $+1$ or -1, we have

$$1 - \tilde{p}(t^\star) = \mathbb{P}\left\{\tilde{U}(t^\star) = -1 \middle| X(t_1), X(t_2)\right\}.$$

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

$$\mathbb{P}\left\{\tilde{U}(t^\star) = 1 \middle| X(t_1) = 1, X(t_2) = 1\right\} = 1\ , \quad \mathbb{P}\left\{\tilde{U}(t^\star) = 1 \middle| X(t_1) = -1, X(t_2) = -1\right\} = 0.$$

Besides, according to Bayes rules

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

In the same way,

$$\mathbb{P}\left\{\tilde{U}(t^\star) = 1 \middle| X(t_1) = -1, X(t_2) = 1\right\} = \frac{r_{t_1}(t^\star)}{r(t_1,t_2)} = \frac{t^\star - t_1}{t_2 - t_1}.$$

As a consequence,

$$\tilde{p}(t^\star) = 1_{X(t_1)=1} 1_{X(t_2)=1} \ + \ \frac{t_2 - t^\star}{t_2 - t_1}\, 1_{X(t_1)=1} 1_{X(t_2)=-1} \ + \ \frac{t^\star - t_1}{t_2 - t_1}\, 1_{X(t_1)=-1} 1_{X(t_2)=1}\ . \tag{7}$$

Note that, using properties of conditional expectation, it is easy to check that we have $\mathbb{P}\left\{\tilde{U}(t^\star) = 1\right\} = 1/2$, so $\tilde{U}(t^\star)$ takes values $+1$ and -1 with equal probability (as explained in Section 1).

As explained previously, since the location $t^\star$ 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 $\theta = (q,\ \mu,\ \sigma)$ 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 $\forall t \in]t_1, t_2[$:

$$\tilde{L}_t(\theta) = \left[\ \tilde{p}(t)\ f_{(\mu+q,\sigma)}(y) + \{1-\tilde{p}(t)\}\ f_{(\mu-q,\sigma)}(y)\right]\ g(t) \tag{8}$$

where the function

$$\begin{aligned} g(t) = &\frac{1}{2}\left\{\bar{r}(t_1,t_2)\ 1_{X(t_1)=1}1_{X(t_2)=1}\ +\ r(t_1,t_2)\ 1_{X(t_1)=1}1_{X(t_2)=-1}\right\} \\ &+ \frac{1}{2}\left\{r(t_1,t_2)\ 1_{X(t_1)=-1}1_{X(t_2)=1}\ +\ \bar{r}(t_1,t_2)\ 1_{X(t_1)=-1}1_{X(t_2)=-1}\right\} \end{aligned}$$

can be removed because it does not depend on the parameters. Note that, for $t = t^\star$ we find our formula (4) of the introduction where $\tilde{p}(t^\star)$ is described in formula (7). As explained in Section 1, for the interference model, $\tilde{\Lambda}_n(t)$ and $\tilde{S}_n(t)$ are respectively the likelihood ratio test (LRT) statistic and the score test statistic at t of the null hypothesis "$q = 0$" in formula (8).

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*

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

With the previous defined notations and under the model without interference

$$\tilde{S}_n(.) \Rightarrow \tilde{Z}(.)\ \ ,\ \ \tilde{\Lambda}_n(.) \xrightarrow{F.d.} \tilde{Z}^2(.)\ \ ,\ \ \sup \tilde{\Lambda}_n(.) \xrightarrow{\mathcal{L}} \sup \tilde{Z}^2(.)$$

as n tends to infinity, under H_0 and $H_{at^\star}$ 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*

- *$\tilde{Z}(.)$ is the Gaussian process with unit variance such as :*

$$\tilde{Z}(t) = \frac{\tilde{\alpha}(t)\tilde{Z}(t_1) + \tilde{\beta}(t)\tilde{Z}(t_2)}{\sqrt{\mathbb{V}\left\{\tilde{\alpha}(t)\tilde{Z}(t_1) + \tilde{\beta}(t)\tilde{Z}(t_2)\right\}}}$$

where

$$Cov\left\{\tilde{Z}(t_1), \tilde{Z}(t_2)\right\} = \rho(t_1, t_2) = \exp(-2|t_1 - t_2|) \ ,$$
$$\tilde{\alpha}(t) = \frac{t_2 - t}{t_2 - t_1} \quad , \quad \tilde{\beta}(t) = \frac{t - t_1}{t_2 - t_1}$$

and with expectation :

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

$$\tilde{m}_{t^\star}(t) = \frac{\tilde{\alpha}(t)\ \tilde{m}_{t^\star}(t_1) + \tilde{\beta}(t)\ \tilde{m}_{t^\star}(t_2)}{\sqrt{\mathbb{V}\left\{\tilde{\alpha}(t)\tilde{Z}(t_1) + \tilde{\beta}(t)\tilde{Z}(t_2)\right\}}}$$

where

$$\tilde{m}_{t^\star}(t_1) = a\rho(t_1, t^\star)/\sigma \ , \ \tilde{m}_{t^\star}(t_2) = a\rho(t^\star, t_2)/\sigma \ .$$

Before interpreting this theorem, we remind that the LRT statistic, $\sup \tilde{\Lambda}_n(.)$ is constructed from the false recombination model (i.e. with interference). Theorem 1 gives the asymptotic distribution of the LRT statistic under the null hypothesis H_0 and under the local alternative $H_{at^\star}$ of one QTL located at $t^\star$ without interference (cf. formula 1). So, according to Theorem 1, the LRT statistic, $\sup \tilde{\Lambda}_n(.)$, converges to the maximum of the square of a "linear normalized interpolated process" called $\tilde{Z}(.)$.

In Theorem 2.1 of Azaïs et al. [2], is presented the asymptotic distribution of the LRT statistic, $\sup \Lambda_n(.)$, constructed from the true recombination model (i.e. without interference). It converges in distribution to the maximum of the square of a "non linear normalized interpolated process" called $Z(.)$:

$$Z(t) = \frac{\alpha(t)Z(t_1) + \beta(t)Z(t_2)}{\sqrt{\mathbb{V}\left\{\alpha(t)Z(t_1) + \beta(t)Z(t_2)\right\}}} \ . \tag{9}$$

We refer to Theorem 2.1 of Azaïs et al. [2] for the expressions of $\alpha(.)$ and $\beta(.)$. Note that $\alpha(t_1) = 1$, $\beta(t_1) = 0$, $\alpha(t_2) = 0$, $\beta(t_2) = 1$. Besides, since the models are exactly the same for the the genetic information at marker locations, we have $\tilde{Z}(t_1) = Z(t_1)$ and $\tilde{Z}(t_2) = Z(t_2)$.

Let's define the following quantity :

$$h(t_1, t_2) = \frac{\tilde{Z}^2(t_1) + \tilde{Z}^2(t_2) - 2\rho(t_1, t_2)\tilde{Z}(t_1)\tilde{Z}(t_2)}{1 - \rho^2(t_1, t_2)} 1_{\frac{\tilde{Z}(t_2)}{\tilde{Z}(t_1)} \in]\rho(t_1,t_2), \frac{1}{\rho(t_1,t_2)}[} \ .$$

Another important result of this paper is the following :

Theorem 2. *With the previous defined notations, under H_0 and $H_{at^\star}$,*

$$\max_{t\in[t_1,t_2]} \tilde{Z}^2(t) = \max_{t\in[t_1,t_2]} Z^2(t) = \max\left\{\tilde{Z}^2(t_1),\ h(t_1, t_2),\ \tilde{Z}^2(t_2)\right\}.$$

In other words, under the null hypothesis and under the alternative, the maximum of the square of the "non linear normalized interpolated process", $Z(.)$, is the same as the maximum of the square of the "linear normalized interpolated process", $\tilde{Z}(.)$. As a consequence, $\sup \tilde{\Lambda}_n(.) = \sup \Lambda_n(.) + o_P(1)$, where $o_P(1)$ denotes a sequence of random vectors which tend to 0 in probability. In other words, there is "an asymptotic robustness of the likelihood ratio test" : even if we choose the false model in order to construct our LRT statistic, we will get asymptotically the optimal power for the detection of the QTL. This result provides new tools to be used in the data analysis for geneticists. Let us introduce now the following lemma which focus on the $\arg\max$ of our processes :

Lemma 1. *With the previous defined notations, under H_0 and $H_{at^\star}$,*

- $\arg\max \tilde{Z}^2(.) = \arg\max Z^2(.) = t_2$ *when* $\tilde{Z}(t_2)/\tilde{Z}(t_1) = 1/\rho(t_1, t_2)$
- $\arg\max \tilde{Z}^2(.) = \arg\max Z^2(.) = t_1$ *when* $\tilde{Z}(t_2)/\tilde{Z}(t_1) = \rho(t_1, t_2)$
- $\arg\max \tilde{Z}^2(.) = \tilde{\xi}$ *and* $\arg\max Z^2(.) = \xi$ *when* $\tilde{Z}(t_2)/\tilde{Z}(t_1) \in]\rho(t_1, t_2), 1/\rho(t_1, t_2)[$ *with*

$$\tilde{\xi} = \frac{(t_2 - t_1)\left\{\rho(t_1, t_2)\tilde{Z}(t_1) - \tilde{Z}(t_2)\right\}}{\{\rho(t_1, t_2) - 1\}\left\{\tilde{Z}(t_1) + \tilde{Z}(t_2)\right\}} + t_1 \ , \ \frac{(t_2 - t_1)\beta(\xi)}{\alpha(\xi) + \beta(\xi)} + t_1 = \tilde{\xi} \ .$$

According to Lemma 1, when the ratio $\tilde{Z}(t_2)/\tilde{Z}(t_1)$ is equal to $1/\rho(t_1,t_2)$, the arg max of $Z^2(.)$ and $\tilde{Z}^2(.)$ are the same : t_2. That is to say, in both cases, the QTL location is estimated to be on the second marker. In the same way, when $\tilde{Z}(t_2)/\tilde{Z}(t_1)$ is equal to $\rho(t_1,t_2)$, the QTL location in both cases, is estimated to be on the first marker. However, when $\tilde{Z}(t_2)/\tilde{Z}(t_1)$ belongs to the interval $]\rho(t_1,t_2), 1/\rho(t_1,t_2)[$, the arg max ξ of $Z^2(.)$ and $\tilde{\xi}$ of $\tilde{Z}^2(.)$ are no the same anymore. As a result, the QTL location is not estimated at the same location.

To sum up Theorem 2 and Lemma 1, if we choose the false recombination model in order to construct our LRT statistic, we will keep asymptotically the optimal power, but the location of the QTL can be estimated differently than if we used the true recombination model.

Proof of Theorem 1

Fisher Information matrix

As said previously, we consider values of t distinct of marker locations and the result can be prolonged by continuity on markers. We first compute the Fisher information at a point θ_0 that belongs to H_0. Let $\tilde{l}_t(\theta)$ be the log likelihood and let define the quantity $\tilde{u}(t)$ such as :

$$\tilde{u}(t) = 2\tilde{p}(t) - 1 \ .$$

We have

$$\frac{\partial \tilde{l}_t}{\partial q} \mid_{\theta_0} = \frac{y-\mu}{\sigma^2}\, \tilde{u}(t) \tag{10}$$

$$\frac{\partial \tilde{l}_t}{\partial \mu} \mid_{\theta_0} = \frac{y-\mu}{\sigma^2} \quad , \quad \frac{\partial \tilde{l}_t}{\partial \sigma} \mid_{\theta_0} = -\frac{1}{\sigma} + \frac{(y-\mu)^2}{\sigma^3}$$

After some calculations, we find

$$I_{\theta_0} = Diag\left[\frac{\mathbb{E}\left\{\tilde{u}^2(t)\right\}}{\sigma^2}\ ,\ \frac{1}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right] \ . \tag{11}$$

Study of the score process under the null hypothesis

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

$$\tilde{S}_n(t) = \frac{\frac{\partial \tilde{l}_t^n}{\partial q} \mid_{\theta_0}}{\sqrt{\mathbb{V}\left(\frac{\partial \tilde{l}_t^n}{\partial q} \mid_{\theta_0}\right)}} .$$

The study is based on the key lemma :

Lemma 2.

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

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

To prove this lemma, use formula (7) and check that both coincide whatever the value of $X(t_1)$, $X(t_2)$ is.

Now using formula (10), we have

$$\frac{\partial \tilde{l}_t^n}{\partial q} \mid_{\theta_0} = \sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma^2} \tilde{u}_j(t) = 1/\sigma \sum_{j=1}^{n} \varepsilon_j \tilde{u}_j(t) = \frac{\tilde{\alpha}(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j X_j(t_1) + \frac{\tilde{\beta}(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j X_j(t_2) \tag{12}$$

this proves the interpolation.
On the other hand

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

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. The weak convergence of the score process, $\tilde{S}_n(.)$, is then a direct consequence of (12), the convergence of $(\tilde{S}_n(t_1), \tilde{S}_n(t_2))$ and the Continuous Mapping Theorem.

Study under the local alternative

Let us consider a local alternative defined by t^* and $q = a/\sqrt{n}$. We consider values of t and $t^\star$ distinct of marker locations and the result can be prolonged by continuity on markers. Since we consider that the true model is the model without interference, we have to consider the "analysis of variance model" for the quantitative trait, described in formula (1). Under the alternative

$$\tilde{S}_n(t) = \frac{a}{n\sigma} \sum_{j=1}^{n} \frac{U_j(t^*)\tilde{u}_j(t)}{\sqrt{\mathbb{V}\left\{\tilde{u}(t)\right\}}} + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \frac{\tilde{u}_j(t)}{\sqrt{\mathbb{V}\left\{\tilde{u}(t)\right\}}}$$

The second term has the same distribution as under the null hypothesis and the first one gives the expectation. As said in Section 1, we use the Haldane [9] modeling for the genetic information at marker locations t_1 and t_2. Besides, since we consider a model without interference, we also consider Haldane [9] modeling inside the marker interval. In other words, $X(0)$ is a random sign and $X(t) = X(0)(-1)^{N(t)}$ where $N(.)$ is a standard Poisson process on $[0, T]$ (here $t_1 = 0$ and $t_2 = T$). This way, the genetic information $U(t^\star)$ at the QTL location $t^\star$ is exactly the quantity $X(t^\star)$. We refer to Azaïs et al. [2] for more details about Haldane [9] modeling. As a consequence, using Lemma 2, it comes

$$\begin{aligned}\mathbb{E}\left\{\tilde{S}_n(t)\right\} &= \frac{a\,\mathbb{E}\left\{U(t^*)\tilde{u}(t)\right\}}{\sigma\sqrt{\mathbb{V}\left\{\tilde{u}(t)\right\}}} = \frac{a\left[\tilde{\alpha}(t)\mathbb{E}\left\{X(t^\star)X(t_1)\right\} + \tilde{\beta}(t)\mathbb{E}\left\{X(t^\star)X(t_2)\right\}\right]}{\sigma\sqrt{\mathbb{V}\left\{\tilde{u}(t)\right\}}} \\ &= \frac{a\,\tilde{\alpha}(t)\rho(t_1, t^\star)}{\sigma\sqrt{\mathbb{V}\left\{\tilde{u}(t)\right\}}} + \frac{a\,\tilde{\beta}(t)\rho(t^\star, t_2)}{\sigma\sqrt{\mathbb{V}\left\{\tilde{u}(t)\right\}}}\,.\end{aligned}$$

This gives the result.

About the LRT process

For the interference model, the likelihood ratio statistic at t, corresponding to n independent observations, will be defined as

$$\tilde{\Lambda}_n(t) = 2\left\{\tilde{l}_t^n(\widehat{\theta}) - \tilde{l}_t^n(\widehat{\theta}_{|H_0})\right\}\ ,$$

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

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

$$\tilde{\Lambda}_n(t) = \tilde{S}_n^2(t) + o_P(1) \tag{13}$$

under the null hypothesis.

Let us consider the local alternative defined by t^* and $q = a/\sqrt{n}$. Since we consider that in reality we are under a model without interference, conditionnally to $X(t_1)$ and $X(t_2)$, the quantitative trait Y follows the mixture model described in formula (3). We refer to formula (3) of Azaïs et al. [2] for the details about the weights $p(t^\star)$ of the mixture model. As mentioned in Azaïs et al. [2], 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 (13) 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. [2] which is based on results of Azaïs et al. [3] and Gassiat [7].

Proof of Theorem 2 and Lemma 1

We consider the process $W(.)$ on $[0,1]$ such as $\forall t' \in [0,1]$:

$$W(t') = \frac{(1-t')\, Z(t_1) \;+\; t'\, Z(t_2)}{\sqrt{(1-t')^2 + t'^{\,2} + 2\,\rho(t_1,t_2)\, t'\, (1-t')}}$$

We can remark that $W(0) = Z(t_1) = \tilde{Z}(t_1)$ and $W(1) = Z(t_2) = \tilde{Z}(t_2)$. Besides, we can apply Lemma 2.2 of Azaïs et al. [2] by taking $\gamma_1(t') = 1 - t'$, $\gamma_2(t') = t'$, $\tilde{\rho} = \rho(t_1,t_2)$, $C_1 = Z(t_1)$, $C_2 = Z(t_2)$ since $\frac{\gamma_1(t')}{\gamma_1(t')+\gamma_2(t')}$ and $\frac{\gamma_2(t')}{\gamma_1(t')+\gamma_2(t')}$ take every values in $[0,1]$. It comes according to Lemma 2.2 of Azaïs et al. [2]

$$\max_{t'\in[0,1]} W^2(t') = \max\left\{Z^2(t_1),\; h(t_1,t_2),\; Z^2(t_2)\right\}.$$

Besides, according to the proof of Lemma 2.2 of Azaïs et al. [2] :

$$\begin{aligned}
&\arg\max W^2(.) = t_2 \quad \text{when} \quad Z(t_2)/Z(t_1) = 1/\rho(t_1,t_2)\ ,\\
&\arg\max W^2(.) = t_1 \quad \text{when} \quad Z(t_2)/Z(t_1) = \rho(t_1,t_2)\ ,\\
&\arg\max W^2(.) = \xi' \quad \text{when} \quad Z(t_2)/Z(t_1) \in]\rho(t_1,t_2), 1/\rho(t_1,t_2)[\ ,
\end{aligned}$$

where

$$\xi' = \frac{\rho(t_1,t_2)Z(t_1) - Z(t_2)}{\{\rho(t_1,t_2)-1\}\{Z(t_1)+Z(t_2)\}} \ .$$

By construction, we have $\forall t \in [t_1, t_2]$

$$\tilde{Z}(t) = W\left(\frac{t-t_1}{t_2-t_1}\right) \text{ and } Z(t) = W\left\{\frac{\beta(t)}{\alpha(t)+\beta(t)}\right\} \ .$$

Besides, the functions $\frac{t-t_1}{t_2-t_1}$ and $\frac{\beta(t)}{\alpha(t)+\beta(t)}$ take every values in $[0,1]$. It comes

$$\max_{t\in[t_1,t_2]} Z^2(t) = \max_{t\in[t_1,t_2]} \tilde{Z}^2(t) = \max_{t'\in[0,1]} W^2(t') = \max\left\{Z^2(t_1),\ h(t_1,t_2),\ Z^2(t_2)\right\} \ .$$

Furthermore,

$$W(\xi') = \tilde{Z}(\tilde{\xi}) = Z(\xi) \text{ where } \tilde{\xi} = \xi'(t_2-t_1)+t_1$$
$$\text{and } \xi' = \frac{\alpha(\xi)}{\alpha(\xi)+\beta(\xi)} \ .$$

As a consequence,

$$\tilde{\xi} = \frac{(t_2-t_1)\beta(\xi)}{\alpha(\xi)+\beta(\xi)} + t_1 \ .$$

It concludes the proof.

3. Several markers : the "interval mapping" of Lander and Botstein [10]

We suppose now that there are K markers $0 = t_1 < t_2 < ... < t_K = T$. A QTL is lying at a position $t^\star$. In the same way as in Section 2, we consider that the QTL is lying between its two flanking markers without interference (cf. formula 1). In order to find the QTL, we will perform tests at every positions t on the chromosome, using a model with interference (since we consider the false recombination model). We consider values t or $t^\star$ of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions. For $t \in [t_1, t_K]\backslash\mathbb{T}_K$ where $\mathbb{T}_K = \{t_1, ..., t_K\}$, we define t^ℓ and t^r as :

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

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

As explained in Section 1, since we consider Haldane [9] modeling for the genetic information at marker locations, we just need to keep the flanking markers in order to infer the value of $\tilde{U}(t^\star)$. It is a direct consequence of the independance of the increments of Poisson process. In others words, the information brought by the other markers is useless. So, we have

$$\mathbb{P}\left\{\tilde{U}(t^\star) = 1 \middle| X(t_1), ..., X(t_K)\right\} = \mathbb{P}\left\{\tilde{U}(t^\star) = 1 \middle| X(t^{\star\ell}), X(t^{\star r})\right\} .$$

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 intervals.

Theorem 3. *We have the same result as in Theorem 1, provided that we make some adjustments and that we redefine $\tilde{Z}(.)$ in the following way :*

- *in the definition of $\tilde{\alpha}(t)$ and $\tilde{\beta}(t)$, t_1 becomes t^ℓ and t_2 becomes t^r*
- *under the null hypothesis, the process $\tilde{Z}(.)$ considered at marker positions is the "squeleton" of an Ornstein-Uhlenbeck process: the stationary Gaussian process with covariance $\rho(t_k, t_{k'}) = \exp(-2|t_k - t_{k'}|)$*
- *at the other positions, $\tilde{Z}(.)$ is obtained from $\tilde{Z}(t^\ell)$ and $\tilde{Z}(t^r)$ by interpolation and normalization using the functions $\tilde{\alpha}(t)$ and $\tilde{\beta}(t)$*
- *at the marker positions, the expectation is such as $\tilde{m}_{t^\star}(t_k) = a\rho(t_k, t^\star)/\sigma$*
- *at other positions, the expectation is obtained from $\tilde{m}_{t^\star}(t^\ell)$ and $\tilde{m}_{t^\star}(t^r)$ by interpolation and normalization using the functions $\tilde{\alpha}(t)$ and $\tilde{\beta}(t)$.*

Under the null hypothesis, the proof of the theorem is the same as the proof of Theorem 1 since we can limit our attention to the interval (t^ℓ, t^r) when considering a unique instant t. Under the alternative, the proof is extactly the same as the proof of Theorem 1 when t and $t^\star$ belong to the same marker interval (t^ℓ, t^r). When t and $t^\star$ belong to two different marker intervals, since we consider that the true model is the one without interference, we have $U(t^\star) = X(t^\star)$, and the expectation can be obtained in the same way as in the proof of Theorem 1.

Let us generalize our Theorem 2 to the case of several markers :

Theorem 4. *With the previous defined notations, under H_0 and $H_{at^\star}$,*

$$\max_{t\in[0,T]} \tilde{Z}^2(t) = \max_{t\in[0,T]} Z^2(t) \ .$$

where $Z(.)$ is the "non linear normalized interpolated process" of Theorem 3.1 of Azaïs et al. [2].

To prove this theorem, just consider that the maximum on $[0,T]$ is the maximum of the maximums obtained for the different marker intervals. In the same way, Lemma 1 can be generalized to the case of several markers.

As a consequence, we have the same conclusion as in the part of this paper dealing with only two genetic markers : if we choose the false recombination model in order to construct our LRT statistic, we will keep asymptotically the optimal power, but the location of the QTL can be estimated differently than if we used the true recombination model.

4. The reverse configuration

In order to make the analysis developed in this paper more general, we propose in this section to focus on the reverse configuration : now the true recombination model is the model with interference and the false model is the one without interference. Note that as previously, Haldane modeling is used for the genetic information at marker locations. The main question is : do we still have an "asymptotic robustness of the LRT" ?

To begin with, let us consider only two genetic markers located at $t_1 = 0$ and $t_2 = T$. We remind that $\sup \Lambda_n(.)$ (resp. $\sup \tilde{\Lambda}_n(.)$) denotes the LRT statistic for the model without interference (resp. with interference), that is to say based on formula (3) (resp. based on formula 4). In Rabier [13], under the interference model, I have proved that the distribution of $\sup \tilde{\Lambda}_n(.)$ is asymptotically that of the maximum of the square of a "linear normalized interpolated process". This "linear normalized interpolated process" is the same process as our process $\tilde{Z}(.)$ except that the mean functions are totally different under the alternative. In both cases, the mean functions are linear interpolated functions. However, since the expectation at marker locations are different, the mean functions are totally different. More precisely, according to our Theorem 1, the expectation at t_1 is $a\rho(t_1,t^\star)/\sigma$, whereas in Rabier [13], the expectation at t_1 is $a\left[\tilde{\alpha}(t^\star) + \tilde{\beta}(t^\star)\rho(t_1,t_2)\right]/\sigma$. In the same way, here, the expectation at t_2 is $a\rho(t^\star,t_2)/\sigma$, whereas in Rabier [13], the

expectation at t_2 is $a\left[\tilde{\alpha}(t^\star)\rho(t_1,t_2)+\tilde{\beta}(t^\star)\right]/\sigma$. It is due to the fact that the model for the quantitative trait Y is not the same if the true model is without interference (cf. formula 1) or with interference (cf. formula 2). Note also that in both cases, we have a "linear normalized interpolated process" since we consider the weights $\tilde{p}(t)$ of the mixture model of formula (4), which verify (cf. Lemma 2):

$$2\tilde{p}(t)-1=\tilde{\alpha}(t)X(t_1)+\tilde{\beta}(t)X(t_2)\ .$$

Let us now focus on the statistic of interest : $\sup\Lambda_n(.)$. In Azaïs et al. [2], under a model without interference, the authors have proved that the distribution of $\sup\Lambda_n(.)$ is that of the maximum of the square of a "non linear normalized interpolated process". If we consider a model with interference, the distribution of $\sup\Lambda_n(.)$ will still be that of the maximum of the square of a "non linear normalized interpolated process" since the weights $p(t)$ of the mixture model of formula (3) verify (cf. Lemma 2.3 of Azaïs et al. [2]):

$$2p(t)-1=\alpha(t)X(t_1)+\beta(t)X(t_2)\ .$$

In the same way as before, the mean function will still be a non linear interpolated function but the values at marker locations will be obtained from the true model, i.e. with interference (cf. formula 2). In other words, the expectation at t_1 will be $a\left[\tilde{\alpha}(t^\star)+\tilde{\beta}(t^\star)\rho(t_1,t_2)\right]/\sigma$ and the expectation at t_2 will be $a\left[\tilde{\alpha}(t^\star)\rho(t_1,t_2)+\tilde{\beta}(t^\star)\right]/\sigma$.

Finally, under the interference model, $\sup\Lambda_n(.)$ will converge to the square of a "non linear normalized interpolated process" whereas $\sup\tilde{\Lambda}_n(.)$ to the square of a "linear normalized interpolated process". The mean functions of these two Gaussian processes are exactly the same at marker locations. Using the same kind of proof as the one of Theorem 2, the maximum of the square of the two processes will be the same, and $\sup\Lambda_n(.)=\sup\tilde{\Lambda}_n(.)+o_P(1)$ under the interference model. Note that we will still have the analogue of Lemma 1.

The result can easily be generalized to several markers. Note that the expectation at marker locations is given in Theorem 3.1 of Rabier [13].

5. Conclusion

To conclude, in this paper, we have considered two different recombination models : a model with interference and a model without interference.

We have proved that even if we choose the false recombination model in order to construct our statistical test, we will keep asymptotically the optimal power. However, the location of the QTL can be estimated differently than if we had chosen the true recombination model. This is a result which could be of interest for geneticists.

6. Acknowledgements

I thank Professor Jean-Marc Azaïs from university Paul-Sabatier Toulouse (FR) and Céline Delmas, Researcher at "Institut National de la Recherche Agronomique" Toulouse (FR) for fruitful discussions.

References

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

[2] J.M. Azaïs, C. Delmas, C.E. Rabier, Likelihood ratio test process for Quantitative Trait Locus detection. to appear in Statistics (2012).

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

[4] J.M. Azaïs and M. Wschebor, Level sets and extrema of random processes and fields, Wiley, New-York, 2009.

[5] M.N. Chang, R. Wu, S.S. Wu, G. Casella, Score statistics for mapping quantitative trait loci. Stat. Appl. Genet. Mol. Biol., 8 (2009), 1, 16.

[6] C. Cierco, Asymptotic distribution of the maximum likelihood ratio test for gene detection. Statistics, 31 (1998), 261-285.

[7] E. Gassiat, Likelihood ratio inequalities with applications to various mixtures. Ann. Inst. Henri Poincaré (B), 6 (2002), 897-906.

[8] A. Genz, Numerical computation of multivariate normal probabilities. J. Comp. Graph. Stat., 1 (1992), 141-149.

[9] J.B.S. Haldane, The combination of linkage values and the calculation of distance between the loci of linked factors. Journal of Genetics, 8 (1919), 299-309.

[10] E.S. Lander and D. Botstein, Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 138 (1989), 235-240.

[11] M. S. McPeek, T. P. Speed, Modeling interference in genetic recombination. Genetics, 139 (1995), 1031-1044.

[12] H.J. Muller, The mechanism of crossing-over. Am. Nat., 50 (1916), 193-221, 284-305, 350-366, 421-434.

[13] C.E. Rabier, On Quantitative Trait Locus mapping with an interference phenomenon. Unpublished result, hal-00658586 (2012).

[14] A. Rebaï, B. Goffinet, B. Mangin, Approximate thresholds of interval mapping tests for QTL detection. Genetics, 138 (1994), 235-240.

[15] A. Rebaï, B. Goffinet, B. Mangin, Comparing power of different methods for QTL detection. Biometrics, 51 (1995), 87-99.

[16] D. Siegmund, B. Yakir, The statistics of gene mapping, Springer, 2007.

[17] A.H. Sturtevant, The behavior of the chromosomes as studied through linkage. Z. Indukt. Abstammungs. Vererbungsl., 13 (1915), 234-287.

[18] A.W. Van der Vaart, Asymptotic statistics, Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[19] R. Wu, C.X. MA, G. Casella, Statistical Genetics of Quantitative Traits, Springer, 2007.