Likelihood Ratio Test process for Quantitative Trait Locus detection

Jean-Marc Azaïs, Céline Delmas, Charles-Elie Rabier

Statistics
Vol. 00, No. 00, 00 00, 1–13

RESEARCH ARTICLE

Likelihood Ratio Test process for Quantitative Trait Locus detection

Jean-Marc Azaïs [a] , Céline Delmas [b], Charles-Elie Rabier [ab*]

[a] *Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S, 31062 Toulouse, France*; [b] *INRA UR631, Station d'Amélioration Génétique des Animaux, Chemin de Borde-Rouge, 31326 Castanet-Tolosan, France*

()

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". We give the asymptotic distribution of this LRT process under the null hypothesis that there is no QTL on $[0, T]$ and under local alternatives with a QTL at $t^\star$ on $[0, T]$. We show that the LRT is asymptotically the square of some Gaussian process. We give a description of this process as an " non-linear interpolated and normalized process ". We propose a simple method to calculate the maximum of the LRT process using only statistics on markers and their ratio. This gives a new method to calculate thresholds for QTL detection.

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

AMS Subject Classification: 62M86; 65C05; 62P10

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 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. The genome $X(t)$ of one individual takes the value $+1$ if, for example, the "recombined chromosome" is originated from A at location t and takes the value -1 if it is originated from B . We use the Haldane [1] modeling that can be represented as follows: $X(0)$ is a

*Corresponding author. Email: rabier@stat.wisc.edu

ISSN: 0233-1888 print/ISSN 1029-4910 online

DOI: 10.1080/02331888.YYYY.xxxxxx
http://www.tandfonline.com

random sign and $X(t) = X(0)(-1)^{N(t)}$ where $N(.)$ is a standard Poisson process on $[0, T]$. We assume an "analysis of variance model" for the quantitative trait :

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

where ε is a Gaussian white noise and t^* is the true location of the QTL.

In fact the "genome information" will be available only at certain fixed locations $t_1 = 0 < t_2 < ... < t_K = T$ and the observation will be

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

So, we observe n observations $(Y_j,\ X_j(t_1),\ ...,\ X_j(t_K))$ i.i.d. Calculation on the Poisson distribution show that

$$r(t,t') := \mathbb{P}(X(t)X(t') = -1) = \mathbb{P}(\left|N(t) - N(t')\right| \text{ odd}) = \frac{1}{2}\,(1 - e^{-2|t-t'|}),$$

we set in addition

$$\bar{r}(t,t') = 1 - r(t,t').$$

It can be proved that, conditionally to $X(t_1), \ldots, X(t_K)$, Y obeys to a mixture model with known weights :

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

where $f_{(m,\sigma)}$ is the Gaussian density with parameters (m, σ) and where the function $p(t)$ is the probability of $\mathbb{P}(X(t) = 1)$ conditionally to the observations of the markers. It can be expressed from the functions r and $\overline{r}$, see Sections 2 and 3 .

The challenge is that the true location t^* is not known. If $t^* = t$ were known, the model would be a regular model. If we define $\Lambda_n(t)$ and $S_n(t)$ as the likelihood ratio test (LRT) statistic and the score test statistic (see Section 2 for a precise definition) of the null hypothesis "$q = 0$". It is well known that

$$\Lambda_n(t) = \ S_n^2(t) + o_P(1)$$

and that $S_n(t)$ is asymptotically Gaussian. Note that following Van der Vaart [14] we have use a multiplicative coefficient of 2 in the definition of the likelihood ratio test.

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 by Cierco [2], Azaïs and Cierco-Ayrolles [3], Azaïs and Wschebor [4]. In Rebaï et al. [5], Rebaï et al. [6], Chang et al. [7], the authors focus only on the null hypothesis using some approximations.

The main result of the paper (Theorem 2.1 and 3.1) is that the distribution of the LRT statistic is asymptotically that of the maximum of the square of a "non linear normalized interpolated process". It explains the fact that the paths of the LRT process, $\Lambda_n(.)$, are smooth between markers (cf. Wu et al. [8]). The second important result is that we have a close simple formula for the distribution of the maximum of the square of the "non linear normalized interpolated process" see Lemma 2.2.

Finally, we propose a new method suitable whatever the genetic map is , using Monte-Carlo Quasi Monte-Carlo (Genz [9]), to calculate thresholds for QTL detection. We show that our method gives better performances than Rebaï et al. [6]'s method based on Davies [10, 11], and Feingold and al. [12]'s method based on Siegmund [13]. Our method is available in a Matlab package with graphical user interface : "imapping.zip". It can be downloaded at www.stat.wisc.edu/∼rabier .

We refer to the book of Van der Vaart [14] 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 : $0 = t_1 < t_2 = T$. For $t \in [t_1, t_2]$ we define

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

and

$$x(t) = \mathbb{E}\left\{X(t) \middle| X(t_1), X(t_2)\right\} = 2p(t) - 1.$$

It is clear that $p(t^*)$ is effectively the probability appearing in (2). An application of Bayes rule leads to

$$\begin{aligned} p(t) = {} & Q_t^{1,1}\, 1_{X(t_1)=1} 1_{X(t_2)=1} \;+\; Q_t^{1,-1}\, 1_{X(t_1)=1} 1_{X(t_2)=-1} \\ & + Q_t^{-1,1}\, 1_{X(t_1)=-1} 1_{X(t_2)=1} \;+\; Q_t^{-1,-1}\, 1_{X(t_1)=-1} 1_{X(t_2)=-1} \end{aligned} \tag{3}$$

where

$$\begin{aligned} Q_t^{1,1} &= \frac{\bar{r}(t_1,t)\,\bar{r}(t,t_2)}{\bar{r}(t_1,t_2)} \quad , \quad Q_t^{1,-1} = \frac{\bar{r}(t_1,t)\,r(t,t_2)}{r(t_1,t_2)} \\ Q_t^{-1,1} &= \frac{r(t_1,t)\,\bar{r}(t,t_2)}{r(t_1,t_2)} \quad , \quad Q_t^{-1,-1} = \frac{r(t_1,t)\,r(t,t_2)}{\bar{r}(t_1,t_2)}. \end{aligned}$$

We can remark that we have

$$Q_t^{-1,-1} = 1 - Q_t^{1,1} \quad \text{and} \quad Q_t^{-1,1} = 1 - Q_t^{1,-1}.$$

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 $\forall t \in [t_1, t_2]$:

$$L_t(\theta) = \left[p(t) f_{(\mu+q,\sigma)}(y) + \{1 - p(t)\}\, f_{(\mu-q,\sigma)}(y)\right]\ g(t) \tag{4}$$

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. By a small abuse of notation we still denote $L_t(\theta)$ for the likelihood without this function. Thus we set

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

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

$$\frac{\partial l_t}{\partial q} \mid_{\theta_0} = \frac{y-\mu}{\sigma^2}\, x(t) \tag{5}$$

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

After some calculations, we find

$$I_{\theta_0} = Diag\left[\frac{\mathbb{E}\left\{x^2(t)\right\}}{\sigma^2}\,,\, \frac{1}{\sigma^2}\,,\, \frac{2}{\sigma^2}\right] \tag{6}$$

Our main result is the following

Theorem 2.1 : *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,

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

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*
- *$Z(.)$ is the Gaussian process with unit variance and -covariance function*

$$\begin{aligned}\Gamma(t,t') &= \frac{\mathbb{E}\left\{x(t)x(t')\right\}}{\sqrt{\mathbb{E}\left\{x^2(t)\right\}}\sqrt{\mathbb{E}\left\{x^2(t')\right\}}} \\ &= \frac{\alpha(t)\alpha(t') + \beta(t)\beta(t') + \left\{\alpha(t)\beta(t') + \alpha(t')\beta(t)\right\}\rho(t_1,t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1,t_2)}\,\sqrt{\alpha^2(t') + \beta^2(t') + 2\alpha(t')\beta(t')\rho(t_1,t_2)}}\end{aligned}$$

where

$$\begin{aligned}\rho(t_1,t_2) &= \exp(-2|t_1-t_2|) \\ \alpha(t) &= Q_t^{1,1} - Q_t^{-1,1} \\ \beta(t) &= Q_t^{1,1} - Q_t^{1,-1}\end{aligned}$$

-expectation $\forall (t, t^\star) \in [t_1, t_2]^2$ *:*

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

$$m_{t^\star}(t) = \frac{a\,\mathbb{E}\left\{X(t^\star)x(t)\right\}}{\sigma\,\sqrt{\mathbb{E}\left\{x^2(t)\right\}}} = a/\sigma\sqrt{\mathbb{E}\left\{x^2(t^\star)\right\}}\,\Gamma(t, t^*).$$

It is clear that we have

$$Z(t) = \frac{\alpha(t)Z(t_1) + \beta(t)Z(t_2)}{\sqrt{\mathbb{V}\left\{\alpha(t)Z(t_1) + \beta(t)Z(t_2)\right\}}}. \tag{7}$$

In the sense of this equation, $Z(.)$ will be called a "non linear normalized interpolated process". As a consequence, the mean function, $m_{t^\star}(t)$, is also an interpolated function. In particular, we have :

$$m_{t^\star}(t) = \frac{\left\{\alpha(t)\; m_{t^\star}(t_1) + \beta(t)\; m_{t^\star}(t_2)\right\}}{\sqrt{\mathbb{V}\left\{\alpha(t)Z(t_1) + \beta(t)Z(t_2)\right\}}} \tag{8}$$

where

$$m_{t^\star}(t_1) = \frac{a}{\sigma}\left\{\alpha(t^\star) + \beta(t^\star)\rho(t_1, t_2)\right\} = a\;\rho(t_1, t^\star)/\sigma$$
$$m_{t^\star}(t_2) = \frac{a}{\sigma}\left\{\alpha(t^\star)\rho(t_1, t_2) + \beta(t^\star)\right\} = a\;\rho(t^\star, t_2)/\sigma.$$

The computation of the maximum of the process $Z^2(.)$ can be performed by using the following lemma.

Lemma 2.2: *Let* $\gamma_1(t)$ *and* $\gamma_2(t)$ *be two functions such that* $\frac{\gamma_i(t)}{\gamma_1(t)+\gamma_2(t)}$ *takes every value in* $[0, 1]$*,* $i = 1, 2$*. Let* C_1 *and* C_2 *be two real numbers and* $0 < \tilde{\rho} < 1$ *then*

$$\max_{t\in[t_1,t_2]} \frac{\left\{\gamma_1(t)C_1 + \gamma_2(t)C_2\right\}^2}{\gamma_1^2(t) + \gamma_2^2(t) + 2\tilde{\rho}\gamma_1(t)\gamma_2(t)} = \max\left(C_1^2\;,\; C_2^2\;,\; \frac{C_1^2 + C_2^2 - 2\tilde{\rho}C_1C_2}{1-\tilde{\rho}^2} 1_{\frac{C_2}{C_1}\in\,]\,\tilde{\rho}\,,\,\frac{1}{\tilde{\rho}}\,[}\right).$$

In particular, if C_1 and C_2 are two random variables defined on the same probability space with $\mathbb{V}(C_i) = 1$, $i = 1, 2$, $\mathrm{Cov}(\mathrm{C_1}, \mathrm{C_2}) = \tilde{\rho}$ with $0 < \tilde{\rho} < 1$ and if $\gamma_1(t)$ and $\gamma_2(t)$ are two functions as above, the lemma gives the distribution of the maximum on $[t_1, t_2]$ of the square of the following normalized interpolated process $D(.)$:

$$\forall t \in [t_1, t_2], \quad D(t) = \frac{\gamma_1(t)C_1 + \gamma_2(t)C_2}{\sqrt{\gamma_1^2(t) + \gamma_2^2(t) + 2\tilde{\rho}\gamma_1(t)\gamma_2(t)}}.$$

So, the lemma can be applied to the process $Z(.)$ by taking $\gamma_1(t) = \alpha(t)$, $\gamma_2(t) = \beta(t)$, $\tilde{\rho} = \rho(t_1, t_2)$, $C_1 = Z(t_1)$, $C_2 = Z(t_2)$, as soon as we prove that $\gamma_3(t) = \frac{\beta(t)}{\alpha(t)+\beta(t)}$ takes every value in $[0, 1]$. Let's now prove this.
Since $\alpha(t_1) = 1$ and $\beta(t_1) = 0$, $\gamma_3(t_1) = 0$. Since $\alpha(t_2) = 0$ and $\beta(t_2) = 1$, $\gamma_3(t_2) = 1$. So, the bounds 0 and 1 are reached. Besides,

$$\beta(t) = \frac{\overline{r}(t_1, t)\overline{r}(t, t_2)r(t_1, t_2) - \overline{r}(t_1, t)r(t, t_2)\overline{r}(t_1, t_2)}{r(t_1, t_2)\overline{r}(t_1, t_2)}$$

has the same sign as

$$\overline{r}(t,t_2)r(t_1,t_2) - r(t,t_2)\overline{r}(t_1,t_2) = r(t_1,t_2) - r(t,t_2) \geqslant 0.$$

Furthermore, $\alpha(t) + \beta(t) = 2Q_t^{1,1} - 1 > 0$ since t is bounded. So, $\gamma_3(t)$ which is the ratio of two positive and continuous functions, takes every value in $[0,1]$.

Proof: Theorem 2.1

Preliminaries

We define some additional notation. For every t, the statistical model is regular with an invertible Fisher information matrix given by (5) under H_0. Its likelihood $L_t(\theta)$ is given by (4) with $\theta = (q, \mu, \sigma^2)$. The log likelihood, associated to n observations will be denoted by $l_t^n(\theta)$.
Let $l_t^n(\widehat{\theta})$ be the maximized log likelihood and let $l_t^n(\widehat{\theta}_{|H_0})$ be the maximized log likelihood under H_0, with $\widehat{\theta}_{|H_0} = \big(0, \overline{Y} = \sum Y_j/n, 1/n\sum(Y_j - \overline{Y})^2\big)$.
The likelihood ratio statistics will be defined as

$$\Lambda_n(t) = 2\big[l_t^n(\widehat{\theta}) - l_t^n(\widehat{\theta}_{|H_0}\big],$$

on n independent observations. Since the Fisher Information matrix is diagonal, the score statistics of the hypothesis "$q = 0$" will be defined as

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

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

$$\Lambda_n(t) = S_n^2(t) + o_P(1)$$

under the null hypothesis. Note that no coefficient $1/2$ is present since we have introduced a coefficient 2 in the definition of the likelihood ratio. Our goal is now to prove that the rest above is uniform in t.

Study of the supremum of the LRT process

Let us consider now t as an extra parameter. Let t^*, θ^* be the true parameter that will be assumed to belong to H_0. Note that t^* makes no sense. It is easy to check that at H_0 the Fisher information relative to t is zero so that the model is not regular.

Conditionally to $X(t_1)$ and $X(t_2)$, the model is a mixture of Gaussian distributions with different means, common unknown variance and a probability that varies between two bounds as a consequence of Equation (2). This is a sub-model of the general mixture of Gaussian distributions (with a probability that varies freely between 0 and 1) as studied, for example in Section 4.3 of Azaïs et al. [15]. In particular it proves that Theorem 1 of Azaïs et al. [15] applies in the sense that

$$\sup_{t,\theta)} l_t(\theta) - l_{t^*}(\theta^*) = \sup_{d\in\mathcal{D}} \left[\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^n d(X_j)\right\}^2 1_{d(X_j)\geqslant 0}\right] + o_P(1) \qquad (9)$$

where the observation X_j stands for $Y_j, X_j(t_1), X_j(t_2)$ and where $\mathcal{D}$ is the set of scores defined in Azaïs et al. [15], see also Gassiat [16]. A similar result is true under H_0 with a set $\mathcal{D}_0$. Let us precise the sets of scores $\mathcal{D}$ and $\mathcal{D}_0$. This sets are defined at the sets of scores of one parameter families that converge to the true model p_{t^*,θ^*} and that are differentiable in quadratic mean.

These sets are subset of the subsets obtained in the general model ($p \in [0,1]$) so it is easy to see that when we sum the four cases for $X(t_1)$ and $X(t_2)$

$$\mathcal{D} = \Big\{ \frac{\langle V, l'_t(\theta^*)\rangle}{\sqrt{\mathbb{V}(\langle V, l'_t(\theta^*)\rangle)}}, V \in \mathbb{R}^3, t \in [t_1, t_2] \Big\}$$

where l' is the gradient with respect to θ. In the same manner

$$\mathcal{D}_0 = \Big\{ \frac{\langle V, l'_t(\theta^*)\rangle}{\sqrt{\mathbb{V}(\langle V, l'_t(\theta^*)\rangle)}}, V \in \mathbb{R}^2 \Big\},$$

where now the gradient is taken with respect to μ and σ only. Of course this gradient does not depend on t.

Using the transform $V \to -V$ in the expressions of the sets of score, we see that the indicator function can be removed in (9). Then, since the Fisher information matrix is diagonal (see formula (6)) , it is easy to see that

$$\sup_{d\in\mathcal{D}} \left[\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right\}^2 \right] - \sup_{d\in\mathcal{D}_0} \left[\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right\}^2 \right]$$
$$= \sup_{t\in[t_1,t_2]} \left(\left[\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\frac{\partial l_t}{\partial q}(X_j) \mid_{\theta_0}}{\sqrt{\mathbb{V}\left\{ \frac{\partial l_t}{\partial q}(X_j) \mid_{\theta_0} \right\}}} \right]^2 \right).$$

This is exactly the desired result.

Study of the score process under the null hypothesis

The study is based on the following key lemma :

Lemma 2.3: *The conditional expectation $x(t)$ of $X(t)$ is linear in $X(t_1), X(t_2)$:*

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

with $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$ and $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$.

To prove this lemma use formula (3) and check that both sides coincide whatever the value of $X(t_1), X(t_2)$ is.
Now using (5) it is clear that

$$\frac{\partial l_t^n}{\partial q} \mid_{\theta_0} = \sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma^2} x_j(t) = 1/\sigma \sum_{j=1}^{n} \varepsilon_j x_j(t) = \frac{\alpha(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j X_j(t_1) + \frac{\beta(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j X_j(t_2) \tag{10}$$

this proves (7).
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. The weak convergence of the score process, $S_n(.)$, is then a direct consequence of (10), the convergence of $(S_n(t_1), S_n(t_2))$ and the Continuous Mapping Theorem.

Study under the local alternative

Let us consider a local alternative defined by t^* and $q = a/\sqrt{n}$. 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. It remains to compute the asymptotic distribution of $S_n(t)$ under this alternative. Indeed, under the alternative

$$S_n(t) = \frac{a}{n\sigma} \sum_{j=1}^{n} \frac{X_j(t^*)x_j(t)}{\sqrt{\mathbb{V}\{x(t)\}}} + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \frac{x_j(t)}{\sqrt{\mathbb{V}\{x(t)\}}}$$

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

$$\mathbb{E}\{S_n(t)\} = \frac{a\,\mathbb{E}\{X(t^*)x(t)\}}{\sigma\sqrt{\mathbb{V}\{x(t)\}}}.$$

By the properties of conditional expectation

$$\mathbb{E}\{X(t^*)x(t)\} = \mathbb{E}\{x(t^*)x(t)\}.$$

This gives the result. □

Proof: Lemma 2.2

Without loss of generality, we can consider $t \in [0, 1]$, $\gamma_1(t) = 1 - t$ and $\gamma_2(t) = t$. So, the focus is on the function on $[0, 1]$

$$\psi(t) = \frac{(1-t)\,C_1 \;+\; t\,C_2}{\sqrt{(1-t)^2 + t^2 + 2\,\tilde{\rho}\,t\,(1-t)}} \;\text{ where }\; 0 < \tilde{\rho} < 1.$$

We find that

$$\begin{aligned}
&\frac{\partial \psi^2(t)}{\partial t} = 0 \\
&\Leftrightarrow \{(1-t)C_1 + tC_2\} \\
&\times \ [\{C_2 - C_1\}\{1 - 2(1-\tilde{\rho})t(1-t)\} + \ (1-\tilde{\rho})(1-2t)\{(1-t)C_1 + tC_2\}] = 0.
\end{aligned}$$

Since $\{(1-t)C_1 + tC_2\}$ corresponds to a minimum, the focus is on the second term. After some calculations, we find that this second term is equal to zero for

$$\xi = \frac{\tilde{\rho}\, C_1 - C_2}{(\tilde{\rho} - 1)(C_2 + C_1)}.$$

So, we just have to consider the cases $\xi \in [0,1]$ and $\xi \notin [0,1]$. Note that

$$\psi^2(\xi) = \frac{C_1^2 + C_2^2 - 2\tilde{\rho} C_1 C_2}{1 - \tilde{\rho}^2}.$$

This gives the result. □

3. Several markers : the "Interval Mapping" of Lander and Botstein [17]

In that case suppose that there are K markers $0 = t_1 < t_2 < ... < t_K = T$. We consider values t, 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 :

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

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

Theorem 3.1 : *We have the same result as in Theorem 2.1, provided that we make some adjustments and that we redefine $Z(.)$ in the following way :*

- *in the definition of $\alpha(t)$ and $\beta(t)$, t_1 becomes t^ℓ and t_2 becomes t^r*
- *under the null hypothesis, the process $Z(.)$ 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, $Z(.)$ is obtained from $Z(t^\ell)$ and $Z(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$*
- *at the marker positions, the expectation is such as $m_{t^\star}(t_k) = a\rho(t_k, t^\star)/\sigma$*
- *at other positions, the expection is obtained from $m_{t^\star}(t^\ell)$ and $m_{t^\star}(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$.*

The proof of the theorem is the same the proof of Theorem 2.1 as soon as we can limit our attention to the interval (t^ℓ, t^r) when considering a unique instant t and to the intervals $(t^\ell, t^r)(t'^\ell, t'^r)$ when considering two instants t and t'. For that we need to prove that

$$x(t) = \mathbb{E}\{X(t)|X(t_1), \ldots, X(t_K)\} = \mathbb{E}\left\{X(t)|X(t^\ell), X(t^r)\right\}$$

which is a direct consequence of the independance of the increments of Poisson process.

3.1. *Application to the calculation of thresholds*

The theoretical results presented in this article allow us to propose a new method to obtain the $\alpha\%$ quantile of the maximum of the process $Z^2(.)$ under H_0. This method is a direct application of Lemma 2.2. If we call

$$h(t_k,t_{k+1})=\frac{Z^2(t_k)+Z^2(t_{k+1})-2\rho(t_k,t_{k+1})Z(t_k)Z(t_{k+1})}{1-\rho^2(t_k,t_{k+1})}1_{\frac{Z(t_{k+1})}{Z(t_k)}\in]\rho(t_k,t_{k+1}),\frac{1}{\rho(t_k,t_{k+1})}[}$$

we have to compute the distribution of

$$M=\max\left\{Z^2(t_1),Z^2(t_2),h(t_1,t_2),...,Z^2(t_{K-1}),Z^2(t_K),h(t_{K-1},t_K)\right\}.$$

According to Bayes rules, we have $\forall c\in\mathbb{R}$

$$\mathbb{P}(0\leqslant M\leqslant c^2)=\mathbb{P}\left\{-c\leqslant Z(t_1)\leqslant c,...,-c\leqslant Z(t_K)\leqslant c\right\}\ \times$$
$$\mathbb{P}\left\{0\leqslant h(t_1,t_2)\leqslant c^2,...,0\leqslant h(t_{K-1},t_K)\leqslant c^2\mid -c\leqslant Z(t_1)\leqslant c,...,-c\leqslant Z(t_K)\leqslant c\right\}.$$

The first term is an integral over the density of a dimension K Gaussian vector. It can be performed for large K using the function QSIMVNEF of Genz which is a MCQMC program. QSIMVNEF also allows to calculate the second term. Monte-Carlo Quasi Monte-Carlo (MCQMC) methods of Genz [9] are very fast. As the numerical computation of a multivariate normal distribution is often a difficult problem, Genz described in his paper, a transformation that simplifies the problem and places it on $[0,1]^K$. A form that allows efficient calculations using standard numerical multiple integration algorithms. He suggests to use in particular MCQMC algorithms. Indeed, a simple Monte-Carlo method (MC) using N points has errors that are typically $O(1/\sqrt{N})$ whereas Quasi Monte-Carlo methods (QMC) have errors which can be aproximatively $O(1/N)$. In order to have a confidence interval an extra Monte-Carlo step is added, this is MCQMC. We refer to Genz [9] for more details.
Note that here the function QSIMVNEF has been adapted and a Newton method has been used in order to find the threshold c_α^2 such as $\mathbb{P}(0\leqslant M\leqslant c_\alpha^2)=\alpha$.

Our method is available in a Matlab package with graphical user interface : "imapping.zip". It can be downloaded at www.stat.wisc.edu/∼rabier .

In this section, we propose to compare the performances of our method with other methods usually used in QTL detection. Note that all the methods are asymptotic in terms of the number of individuals n.

In Rebaï et al. [6], the authors focus on another recombination model. They propose an upper bound for the threshold corresponding to their model. This bound is the quantity $\tilde{c}_\alpha^2$ such as :

$$1-\alpha=2\,\Phi(-\tilde{c}_\alpha)\ +\ \frac{2\,e^{-\tilde{c}_\alpha^2/2}}{\pi}\ \sum_{k=1}^{K-1}\ \arctan\left(\ \sqrt{\frac{1-\rho(t_k,t_{k+1})}{1+\rho(t_k,t_{k+1})}}\ \right)$$

where Φ is the cumulative distribution of the standardized normal distribution.
This method is based on Davies [10]. However, it is sensitive to the number of genetic markers. Indeed, the derivative of the process has a jump at each markers

location, and Davies [10] upper bound is suitable when the derivative of the process has a finite number of jumps.

In Feingold and al. [12], the authors propose a threshold based on the discrete process resulting from tests only on markers. Besides, they suppose constant the distance between genetic markers. The threshold c_α^2 is such as :

$$1 - \alpha = 1 \; - \; \Phi(c_\alpha) \; + \; 2\, T \, c_\alpha \, \varphi(c_\alpha) \, \nu(2c_\alpha\sqrt{\Delta})$$

where φ is the density of a normal standardized, Δ is distance between two consecutive markers. This method is inspired from Siegmund [13] where the function ν is fully described. The method is also largely described in Siegmund and Yakir [18].

In Tables 1 and 2, we propose to compare the different methods. The different thresholds computed correspond to $\alpha = 95\%$, and since the three methods are based on asymptotic results, we propose to check the percentage of false positives on simulated data, in order to obtain the true level corresponding to each method. We simulated under the null hypothesis, 10000 samples of $n = 200$ individuals. We analyzed data using our Lemma 1, 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 fulfills the given condition.

To begin, in Table 1, we consider a chromosome of length $T = 1$ Morgan. We consider the markers equally spaced and we consider different densities of markers. We can see that Feingold's method and our method, give reasonable results : the percentage of false positives is close to 5%. On the other hand, as expected, Rebaï's method is very sensitive to the number of genetic markers. We can remark that the more markers there are, the more conservative Rebai's method is.

Let's focus now more in details on the differences between our method and Feingold's method. As said before, in Feingold and al. [12], the authors focus only on the discrete process which results from tests only at marker locations. Besides, in order to obtain a theoretical result, they consider that the markers are equally spaced. In our study, we consider the true "Interval Mapping" of Lander and Botstein [17] : we consider the stochastic process which results from tests on the whole chromosome (ie. on markers and between markers). Furthermore, we allow the markers not to be equally spaced, which is generally the case in a biological experiment.

In Table 2, we compare the performances of the two methods. We consider different genetic maps for which markers are not equally spaced : the maps are described in Table 3. Note that in order to compute Feingold's method, since the markers are not equally spaced anymore, we use for Δ (ie. the distance between markers), the mean distance between markers. We can see on the different examples, that our method generally respects the 5% level, which is not the case of Feingold's method. The performances of Feingold's method could have maybe been improoved, by testing different values of Δ. However, there is not any rule in order to choose an appropriate Δ. This way, our method which is suitable for any genetic map, must be the most interesting for geneticists.

To conclude, in Table 4, we focus on the alternative hypothesis. Using our threshold, we compare the Theoretical Power (cf. our Theorem 2) and the Empirical Power for different values of n. We can see that for $n = 1000$, the asymptotic is reached. It validates the asymptotic results of this paper.

Table 1. Threshold and Percentage of False Positives (10000 samples and $n = 200$) as a function of the number of markers and the method considered.The chromosome is of length $T = 1$ Morgan and the markers are equally spaced.

number of markers	101	51	41	26	21	11	6	3	2
Rebai	9.74 2.55%	9.09 3.23%	8.88 3.25%	8.43 3.82%	8.20 4.05%	7.58 4.48%	6.92 4.53%	6.07 4.55%	5.22 5.13%
Feingold	8.45 4.67%	8.17 4.91%	8.06 4.74%	7.81 5.37%	7.67 5.19%	7.18 5.57%	6.59 5.25%	5.71 5.46%	5.08 5.55%
this paper	8.41 4.76%	8.27 4.71%	8.16 4.63%	7.91 5.17%	7.75 4.99%	7.29 5.24%	6.76 4.76%	6.04 4.64%	5.41 4.61%

Table 2. Threshold and Percentage of False Positives (10000 samples and $n = 200$) as a function of the genetic map and the method considered.

Genetic Map	map 1*	map 2*	map 3*	map 4*
this paper	6.14 5.31%	6.25 4.87%	6.85 4.92%	7.07 4.90%
Feingold	5.79 6.25%	6.59 4.03%	7.38 3.74%	7.76 3.63%

*The different maps are described in Table 3.

Table 3. The different genetic maps considered (K is the number of markers, T is the length of the chromosome in Morgan, t_k is the location of marker k in Morgan).

	T	K	marker locations
map 1	1.50	3	$t_1 = 0$, $t_2 = 0.50$, $t_3 = 1.50$
map 2	1	6	$t_1 = 0$, $t_2 = 0.80$, $t_3 = 0.85$, $t_4 = 0.90$, $t_5 = 0.95$, $t_6 = 1$
map 3	1	14	$\forall k = 1, ..., 11 \quad t_k = 0.01(k-1)$, $t_{12} = 0.40$, $t_{13} = 0.70$, $t_{14} = 1$
map 4	1	23	$\forall k = 1, ..., 11 \quad t_k = 0.01(k-1)$, $t_{12} = 0.40$, $t_k = 0.90 + 0.01(k-13)$ $\forall k = 13, ..., 23$

Table 4. Theoretical Power and Empirical Power (EP) as a function of the location of the QTL $t^\star$ in Morgan ($a = 4$, 100000 paths for the Theoretical Power, 10000 samples for EP). The chromosome is of length $T = 1$ Morgan, 6 markers are equally spaced every 0.2 Morgan.

$t^\star$	0.10	0.43	0.75	0.88
EP for $n = 50$	82.32%	87.76%	84.68%	82.10%
EP for $n = 100$	85.62%	90.12%	88.56%	85.84%
EP for $n = 200$	87.14%	91.17%	89.73%	87.73%
EP for $n = 1000$	88.51%	92.20%	90.33%	89.20%
Theoretical Power	88.61%	92.01%	90.56%	88.94%

Acknowledgements

The authors thank Jean-Michel Elsen for having proposed this subject of research and 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.

References

[1] 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), pp. 299–309.

[2] C. Cierco, *Asymptotic distribution of the maximum likelihood ratio test for gene detection*, Statistics 31 (1998), pp. 261–285.
[3] J.M. Azaïs and C. Cierco-Ayrolles, *An asymptotic test for quantitative gene detection*, Ann. Inst. Henri Poincaré (B) 38(6) (2002), pp. 1087–1092.
[4] J.M. Azaïs and M. Wschebor, *Level sets and extrema of random processes and fields*, Wiley, New-York (2009).
[5] A. Rebaï, B. Goffinet, B. Mangin, *Comparing power of different methods for QTL detection*, Biometrics 51 (1995), pp. 87–99.
[6] ———, *Approximate thresholds of interval mapping tests for QTL detection*, Genetics 138 (1994), pp. 235–240.
[7] M.N. Chang, R. Wu, S.S. Wu, G. Casella, *Score statistics for mapping quantitative trait loci*, Statistical Application in Genetics and Molecular Biology 8(1) 16 (2009).
[8] R. Wu, C.X. Ma, G. Casella, *Statistical Genetics of Quantitative Traits*, Springer (2007).
[9] A. Genz, *Numerical computation of multivariate normal probabilities*, J. Comp. Graph. Stat. 1 (1992), pp. 141–149.
[10] R.B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, Biometrika 64 (1977), pp. 247–254.
[11] ———, *Hypothesis testing when a nuisance parameter is present only under the alternative*, Biometrika 74 (1987), pp. 33–43.
[12] E. Feingold, P.O. Brown, D. Siegmund, *Gaussian models for genetic linkage analysis using complete high-resolution maps of idendity by descent*, Am. J. Human. Genet. 53 (1993), pp. 234–251.
[13] D. Siegmund, *Sequential analysis : tests and confidence intervals*, Springer, New York (1985).
[14] A.W. Van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (1998).
[15] J.M. Azaïs, E. Gassiat, C. Mercadier, *The likelihood ratio test for general mixture models with possibly structural parameter*, ESAIM 13 (2009), pp. 301–327.
[16] E. Gassiat, *Likelihood ratio inequalities with applications to various mixtures* , Ann. Inst. Henri Poincaré (B) 6 (2002), pp. 897–906.
[17] E.S. Lander and D. Botstein, *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps*, Genetics 138 (1989), pp. 235–240.
[18] D. Siegmund, B. Yakir, *The statistics of gene mapping*, Springer, New York (2007).
[19] J.M. Azaïs, E. Gassiat, C. Mercadier, *Asymptotic distribution and local power of the likelihood ratio test for mixtures*, Bernoulli 12(5) (2006), pp. 775–799.
[20] P. Billingsley, *Convergence of probability measures*, Wiley, New-York (1999).
[21] J.K. Ghosh and P.K. Sen, *On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results*, Inst. Statistics Mimeo Series 1467 (1984).
[22] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Springer (1986).
[23] C.E. Rabier, *PhD thesis*, Université Toulouse 3 Paul Sabatier (2010).