On empirical processes for Quantitative Trait Locus mapping under the presence of a selective genotyping and an interference phenomenon

Charles-Elie Rabier

On empirical processes for Quantitative Trait Locus mapping under the presence of a selective genotyping and an interference phenomenon

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 the interval $[0, T]$ representing a chromosome. The originality is twofold. We consider a selective genotyping and an interference phenomenon. We show that, under the null hypothesis and contiguous alternatives, the LRT process is asymptotically the square of a "linear interpolated and normalized Gaussian process ". We prove that we have to genotype symmetrically and that the threshold is exactly the same as in the situation without selective genotyping and without interference.

Keywords: QTL detection, Likelihood Ratio Test, Gaussian process, Selective Genotyping, 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

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

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 (see for instance Wu et al. [27] or Siegmund and Yakir [24]). 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 [11] 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 $[0, T]$. Due to the independence of increments of Poisson process, this model allows double recombinations between markers. For instance, if we consider 3 markers (i.e. $K = 3$), 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 $t^\star$ 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 first originality of this paper is that, inside the marker interval which contains the QTL, we consider an interference phenomenon : a recombination event inhibits the formation of another recombination event nearby. This phenomenon was noticed by geneticists working on the Drosophila (Sturtevant [25], Muller [16]). In McPeek and Speed [15], the authors study several interference models and also mention the importance of modeling interference. We will focus here on the model proposed by Rebaï et al. [23] for

which double recombination between the QTL and its flanking markers is not allowed (see in particular their Section 2). This model was extended to a whole chromosome in Rebaï et al. [22]. For instance, 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 $X(t_1) = 1$, $U(t^\star) = -1$ 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 $U(t^\star) = -1$, then we have automatically $X(t_2) = -1$. In the same way, if $X(t_2) = 1$ and $U(t^\star) = -1$, then we have automatically $X(t_1) = -1$. We will explain in details this model in Section 2 and present the law of $U(t^\star)$, given its flanking markers.

On the other hand, our model presents another originality. Usually, in the problem of detecting a QTL on a chromosome, the genome information is available only at fixed locations $t_1 = 0 < t_2 < ... < t_K = T$, called genetic markers. So, usually an observation is

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

and the challenge is that the location $t^\star$ of the QTL is unknown.

So, the second originality of this paper is that we consider the classical problem, but this time, in order to reduce the costs of genotyping, a selective genotyping has been performed : we consider two real thresholds S_- and S_+, with $S_- \leq S_+$ and we genotype (i.e. we collect the genome information at markers) if and only if the phenotype Y is extreme, that is to say $Y \leq S_-$ or $Y \geq S_+$. If we call $\overline{X}(t)$ the random variable such as

$$\overline{X}(t) = \begin{cases} X(t) & \text{if } Y \notin [S_-\ ,\ S_+] \\ 0 & \text{otherwise}\ , \end{cases}$$

then, in our problem, one observation will be now

$$\left(Y,\ \overline{X}(t_1),\ ...,\ \overline{X}(t_K)\right).$$

Note that with our notations :

- when $Y \notin [S_-\ ,\ S_+]$, we have $\overline{X}(t_1) = X(t_1),\ ..., \overline{X}(t_K) = X(t_K)$.

- when $Y \in [S_-\ ,\ S_+]$, we have $\overline{X}(t_1) = 0,\ ..., \overline{X}(t_K) = 0$, which means that the genome information is missing at the marker locations.

We will observe n observations $\big(Y_j,\ \overline{X}_j(t_1),\ ...,\ \overline{X}_j(t_K)\big)$ i.i.d.

It can be proved that $\big(Y,\ \overline{X}(t_1),\ ...,\ \overline{X}(t_K)\big)$ obeys to a mixture model with known weights, times a function g(.) (fully given in Section 2) which does not depend of the parameters μ, q and σ :

$$\begin{aligned}\big[p(t^*)\ f_{(\mu+q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\ \ +\ \ \{1-p(t^*)\}\ \ f_{(\mu-q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\\ +\ \ \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)\ 1_{y\in[S_-,S_+]}\ \ +\ \ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)\ \ 1_{y\in[S_-,S_+]}\Big]\ g(.)\end{aligned} \tag{2}$$

where $f_{(m,\sigma)}$ is the Gaussian density with parameters (m, σ) and where the function $p(t^\star)$ is fully given in Section 2.

As said before, the challenge is that t^* is unknown. So, at every location $t \in [0, T]$, we perform a Likelihood Ratio Test (LRT), $\Lambda_n(t)$, of the hypothesis "$q = 0$". It leads to a LRT process $\Lambda_n(.)$ and taking as test statistic the maximum of this process comes down to perform a LRT in a model when the localisation of the QTL is an extra parameter.

In the classical problem of detecting a QTL on a chromosome, that is to say in the situation where all the individuals are genotyped (i.e. without selective genotyping) and without interference, the asymptotic distribution of the LRT statistic has been given under some approximations by Rebaï et al. [23], Rebaï et al. [22], Cierco [7], Azaïs et al. [1], Azaïs and Wschebor [4], Chang et al. [5]. Recently, Azaïs et al. [2] have shown that the distribution of the LRT statistic is asymptotically that of the maximum of the square of a "non linear normalized interpolated process". Under an interference phenomenon (but still without selective genotyping), I have shown in Rabier [20] that the LRT statistic is asymptotically that of the maximum of the square of a "linear normalized interpolated process".

On the other hand, selective genotyping has been studied theoretically by many authors : for instance Lebowitz and al. [13], Lander and Botstein [12], Darvasi and Soller [8], Muranty and Goffinet [17], Rabier [19]... However, in all these articles, the focus is only on one fixed location of the genome. In Rabbee et al. [18], the authors compare different strategies for analyzing data in selective genotyping. This simulation study is very interesting since the focus is on the whole chromosome and not only one given location of the genome. As a consequence, in Rabier [21], I address the problem of detecting

a QTL on a chromosome, under the presence of a selective genotyping. I show that the distribution of the LRT statistic is asymptotically that of the maximum of the square of a "non linear normalized interpolated process". This result has been obtained for a model without interference. This way, the originality of this paper is in the fact that we study a problem which has never been studied theoretically before : the detection of a QTL on a chromosome under the presence of a selective genotyping and an interference phenomenon.

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". This is a generalization of the results obtained in Rabier [20] only under interference. Under the null hypothesis, despite the selective genotyping, our process is exactly the same as the one obtained in Rabier [20]. However, under the alternative, we show that the mean functions of the two processes are not the same anymore.

Some important results are also introduced in Theorem 3. We give the Asymptotic Relative Efficiency (ARE) with respect to the oracle situation (i.e. without selective genotyping). Note that we show that we have exactly the same ARE with respect to the oracle situation, if we look for a QTL on a whole chromosome or if we focus only on one locus (even if the QTL is not located on this locus). Another interesting result of Theorem 3 is the following : if we want to genotype only a percentage γ of the population, we should genotype symmetrically, that is to say the $\gamma/2\%$ individuals with the largest phenotypes and $\gamma/2\%$ individuals with the smallest phenotypes. This is a generalization of Rabier [19], where it is proved that we have to genotype symmetrically, when we focus only on one genetic marker.

Furthermore, we have an easy formula (see Lemma 3 and formula 18) to compute the maximum of the square of the linear interpolated process. This formula is original. Usually when we look for a QTL on a chromosome with a selective genotyping, we have to compute an EM algorithm at each location, so it is quite challenging. With our formula, we don't need to perform any EM algorithm and we only have to focus on given locations on the chromosome. Note that in this paper, we also prove that the non extreme phenotypes (for which the genotypes are missing) don't bring any extra information for statistical inference (same result as in Rabier [19] but for the whole chromosome). In other words, we give theoretical answers to the previous study of Rabbee et al. [18].

To conclude, we will illustrate our theoretical results with the help of

simulated data. Note that, a consequence of Theorems 1 and 2 and Lemma 3, the threshold (i.e. critical value) for our model, is exactly the same as the classical threshold obtained without selective genotyping and without interference. So, in order to obtain our threshold, the Monte Carlo Quasi Monte-Carlo methods of Azaïs et al. [2], based on Genz [10] is still suitable here. This is an alternative to the permutation method proposed by Manichaikul et al. [14] and inspired by Churchill and Doerge [6], which is very time consuming and not easy to compute in selective genotyping because of the missing genotypes.

We refer to the book of Van der Vaart [26] 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 $t^\star \in]t_1, t_2[$. 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 $r(t_1, t_2)$ 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,

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

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

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

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

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

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

In the same way as in Rebaï et al. [23], 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 (3) stands with these expressions of $r_{t_1}(t^\star)$ and $r_{t_2}(t^\star)$.

Let's define now

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

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

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

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

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

Besides, according to Bayes rules

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

In the same way,

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

As a consequence,

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

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

As said before, since we consider a selective genotyping, we don't observe $(Y,\ X(t_1),\ X(t_2))$ but $\big(Y,\ \overline{X}(t_1),\ \overline{X}(t_2)\big)$. As a consequence, in order to compute the likelihood, we have to study the corresponding probability distributions. We will use the following notations :

Notations 1. *$\overline{U}(t^\star)$ is the random variable such as*

$$\overline{U}(t^\star) = \begin{cases} U(t^\star) & \text{if } Y \notin [S_-\ ,\ S_+] \\ 0 & \text{otherwise.} \end{cases}$$

Notations 2. *$\mathbb{P}_{t^\star}\{l \mid i\}$ is the quantity such as $\forall\ i \in \{-1, 1\}$ and $\forall\ l \in \{-1, 0, 1\}$,*

$$\mathbb{P}_{t^\star}\{l \mid i\} = \mathbb{P}(\overline{U}(t^\star) = l \mid U(t^\star) = i)\ .$$

To begin, let's compute $\mathbb{P}(Y \in [y\ ,\ y + dy] \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1)$ for instance. We have, according to Bayes rules,

$$\begin{aligned} &\mathbb{P}(Y \in [y\ ,\ y + dy] \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1) \\ &= \sum_{i \in \{-1,1\}} \mathbb{P}(Y \in [y\ ,\ y + dy] \mid \overline{U}(t^\star) = i)\ \mathbb{P}(\overline{U}(t^\star) = i \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1)\ . \end{aligned}$$

Besides,

$$\begin{aligned} \mathbb{P}(Y \in [y\ ,\ y + dy] \mid \overline{U}(t^\star) = i) &= \frac{\mathbb{P}(Y \in [y\ ,\ y + dy] \cap \overline{U}(t^\star) \neq 0 \mid U(t^\star) = i)}{\mathbb{P}(\overline{U}(t^\star) \neq 0 \mid U(t^\star) = i)} \\ &= \frac{f_{(\mu + iq, \sigma)}(y)\ 1_{y \notin [S_-, S_+]}}{\mathbb{P}_{t^\star}\{i \mid i\}} \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}(\overline{U}(t^\star) = i \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1) \\ &= \mathbb{P}(\overline{U}(t^\star) \neq 0 \cap U(t^\star) = i \cap X(t_1) = 1 \cap X(t_2) = 1) \\ &= \mathbb{P}_{t^\star}\{i \mid i\}\ \mathbb{P}(U(t^\star) = i \cap X(t_1) = 1 \cap X(t_2) = 1) \\ &= \frac{1}{2}\,\mathbb{P}_{t^\star}\{1 \mid 1\}\ \overline{r}(t_1, t_2)\, 1_{i=1}\ . \end{aligned}$$

In the same way,

$$\mathbb{P}(\overline{U}(t^\star) = i \,\cap\, \overline{X}(t_1) = -1 \,\cap\, \overline{X}(t_2) = -1) = \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}\, \overline{r}(t_1,t_2)\, 1_{i=-1}\ .$$

Furthermore,

$$\begin{aligned}
&\mathbb{P}(\overline{U}(t^\star) = i \,\cap\, \overline{X}(t_1) = 1 \,\cap\, \overline{X}(t_2) = -1)\\
&= \mathbb{P}_{t^\star}\left\{i \mid i\right\}\, \mathbb{P}(U(t^\star) = i \,\cap\, X(t_1) = 1 \,\cap\, X(t_2) = -1)\\
&= \frac{t_2 - t^\star}{2(t_2 - t_1)}\,\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}\, r(t_1,t_2)\, 1_{i=1} \;+\; \frac{t^\star - t_1}{2(t_2 - t_1)}\,\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}\, r(t_1,t_2)\, 1_{i=-1}\ ,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(\overline{U}(t^\star) = i \,\cap\, \overline{X}(t_1) = -1 \,\cap\, \overline{X}(t_2) = 1)\\
&= \mathbb{P}_{t^\star}\left\{i \mid i\right\}\, \mathbb{P}(U(t^\star) = i \,\cap\, X(t_1) = -1 \,\cap\, X(t_2) = 1)\\
&= \frac{t^\star - t_1}{2(t_2 - t_1)}\,\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}\, r(t_1,t_2)\, 1_{i=1} \;+\; \frac{t_2 - t^\star}{2(t_2 - t_1)}\,\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}\, r(t_1,t_2)\, 1_{i=-1}\ .
\end{aligned}$$

It comes:

$$\begin{aligned}
&\mathbb{P}(Y \in [y\ ,\ y+dy] \,\cap\, \overline{X}(t_1) = 1 \,\cap\, \overline{X}(t_2) = 1) = \frac{1}{2}\, f_{(\mu+q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\, \overline{r}(t_1,t_2)\ ,\\
&\mathbb{P}(Y \in [y\ ,\ y+dy] \,\cap\, \overline{X}(t_1) = -1 \,\cap\, \overline{X}(t_2) = -1) = \frac{1}{2}\, f_{(\mu-q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\, \overline{r}(t_1,t_2)\ ,\\
&\mathbb{P}(Y \in [y\ ,\ y+dy] \,\cap\, \overline{X}(t_1) = 1 \,\cap\, \overline{X}(t_2) = -1)\\
&= \frac{t_2 - t^\star}{2(t_2 - t_1)} r(t_1,t_2)\, f_{(\mu+q,\sigma)}(y)\, 1_{y\notin[S_-,S_+]} \;+\; \frac{t^\star - t_1}{2(t_2 - t_1)} r(t_1,t_2)\, f_{(\mu-q,\sigma)}(y)\, 1_{y\notin[S_-,S_+]},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(Y \in [y\ ,\ y+dy] \,\cap\, \overline{X}(t_1) = -1 \,\cap\, \overline{X}(t_2) = 1)\\
&= \frac{t^\star - t_1}{2(t_2 - t_1)} r(t_1,t_2)\, f_{(\mu+q,\sigma)}(y)\, 1_{y\notin[S_-,S_+]} \;+\; \frac{t_2 - t^\star}{2(t_2 - t_1)} r(t_1,t_2)\, f_{(\mu-q,\sigma)}(y)\, 1_{y\notin[S_-,S_+]}.
\end{aligned}$$

Finally, when the genome information is missing at marker locations (i.e. the phenotype is not extreme), we have

$$\begin{aligned}
&\mathbb{P}(Y \in [y\ ,\ y+dy] \,\cap\, \overline{X}(t_1) = 0 \,\cap\, \overline{X}(t_2) = 0)\\
&= \frac{1}{2}\, f_{(\mu+q,\sigma)}(y)\, 1_{y\in[S_-,S_+]} \;+\; \frac{1}{2}\, f_{(\mu-q,\sigma)}(y)\, 1_{y\in[S_-,S_+]}\ .
\end{aligned}$$

As explained previouly, 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. Note that we 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. As a consequence, the likelihood of the triplet $\big(Y,\ \overline{X}(t_1),\ \overline{X}(t_2)\big)$ with respect to the measure $\lambda \otimes N \otimes N$, λ being the Lebesgue measure, N the counting measure on $\mathbb{N}$, is :

$$\begin{aligned} L_t(\theta) = \big[\, p(t)\ f_{(\mu+q,\sigma)}(y) 1_{y\notin[S_-,S_+]} + \{1-p(t)\}\ f_{(\mu-q,\sigma)}(y) 1_{y\notin[S_-,S_+]} \\ +\ \frac{1}{2}\, f_{(\mu+q,\sigma)}(y) 1_{y\in[S_-,S_+]}\ +\ \frac{1}{2}\, f_{(\mu-q,\sigma)}(y) 1_{y\in[S_-,S_+]} \Big]\ g(t) \end{aligned} \tag{5}$$

where the function

$$\begin{aligned} g(t) = &\frac{1}{2}\left\{ \bar{r}(t_1,t_2)\ 1_{\overline{X}(t_1)=1} 1_{\overline{X}(t_2)=1}\ +\ r(t_1,t_2)\ 1_{\overline{X}(t_1)=1} 1_{\overline{X}(t_2)=-1} \right\} \\ &+ \frac{1}{2}\left\{ r(t_1,t_2)\ 1_{\overline{X}(t_1)=-1} 1_{\overline{X}(t_2)=1}\ +\ \bar{r}(t_1,t_2)\ 1_{\overline{X}(t_1)=-1} 1_{\overline{X}(t_2)=-1} \right\} \\ &+\ 1_{\overline{X}(t_1)=0} 1_{\overline{X}(t_2)=0} \end{aligned}$$

can be removed because it does not depend on the parameters. Note that, for $t = t^\star$ we find our formula (2) of the introduction where $p(t^\star)$ is described in formula (4).

Notations 3. *γ, γ_+ and γ_- are respectively the quantities $\mathbb{P}_{H_0}(Y \notin [S_-,\ S_+])$, $\mathbb{P}_{H_0}(Y > S_+)$ and $\mathbb{P}_{H_0}(Y < S_-)$.*

Notations 4. *$\mathcal{A}$ is the quantity such as $\mathcal{A} = \sigma^2\ \big\{\gamma\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\big\}$, where $\varphi(x)$ and z_α denote respectively the density of a standard normal distribution taken at the point x, and the quantile of order $1-\alpha$ of a standard normal distribution.*

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,

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

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*
- $D(.)$ *is the Gaussian process with unit variance such as :*

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

where

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

and with expectation :

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

$$m_{t^\star}(t) = \frac{\alpha(t)\, m_{t^\star}(t_1) + \beta(t)\, m_{t^\star}(t_2)}{\sqrt{\mathbb{V}\left\{\alpha(t)D(t_1) + \beta(t)D(t_2)\right\}}}$$

where

$$m_{t^\star}(t_1) = \frac{a\sqrt{\mathcal{A}}}{\sigma^2}\left\{\alpha(t^\star) + \beta(t^\star)\rho(t_1, t_2)\right\} \ , \ m_{t^\star}(t_2) = \frac{a\sqrt{\mathcal{A}}}{\sigma^2}\left\{\alpha(t^\star)\rho(t_1, t_2) + \beta(t^\star)\right\} \ .$$

In the sense of this equation, $D(.)$ will be called a "linear normalized interpolated process". We can see that under the null hypothesis, despite the selective genotyping, $D(.)$ is exactly the same process as the process $W(.)$ of Theorem 2.1 of Rabier [20] obtained for the oracle situation (i.e. without selective genotyping). However, under the alternative, the mean functions

of the two processes are not the same anymore : the mean functions are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$. Note also that $D(.)$ is the generalization of $W(.)$. Indeed, if we choose $S_- = S_+$, that is to say we genotype all the individuals, the factor $\sqrt{\mathcal{A}}/\sigma$ is equal to 1, and $D(.)$ is the same process as $W(.)$.

Proof of Theorem 1

Fisher Information Matrix

Let $t \in]t_1, t_2[$ and let $l_t(\theta)$ be the loglikelihood. We first compute the Fisher information at a point θ_0 that belongs to H_0. We have

$$\frac{\partial l_t}{\partial q} \mid_{\theta_0} = \frac{y-\mu}{\sigma^2} \ \{2p(t)-1\} \ 1_{y \notin [S_-, S_+]} \tag{6}$$

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

Then,

$$\mathbb{E}_{H_0}\left\{\left(\frac{\partial l_t}{\partial q} \mid_{\theta_0}\right)^2\right\} = \mathbb{E}_{H_0}\left\{\left(\frac{y-\mu}{\sigma^2}\right)^2 \{2p(t)-1\}^2 \, 1_{y \notin [S_-, S_+]}\right\} .$$

Let's introduce two key lemmas.

Lemma 1. *We have the following relationship :*

$$\{2p(t)-1\} \, 1_{y \notin [S_-, S_+]} = \alpha(t)\overline{X}(t_1) + \beta(t)\overline{X}(t_2)$$

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

To prove this lemma, use formula (4) and check that both sides coincide when $y \notin [S_-, S_+]$.

Lemma 2. *Let* $V \sim N(\mu, \, \sigma^2)$*, then :*

i) $\mathbb{E}\left(V^2 1_{V \notin [S_-, \, S_+]}\right) = \ (\mu^2 + \sigma^2) \, \mathbb{P}(V \notin [S_-, \ S_+]) \ + \ \sigma \, (S_+ + \mu) \, \varphi\left(\frac{S_+ - \mu}{\sigma}\right)$

$-\ \sigma\,(S_- + \mu)\,\varphi\left(\frac{S_- - \mu}{\sigma}\right)$

ii) $\mathbb{E}\left(V 1_{V \notin [S_-,\, S_+]}\right) = \mu\,\mathbb{P}(V \notin [S_-,\, S_+]) \ + \ \sigma\,\varphi\left(\frac{S_+ - \mu}{\sigma}\right) \ - \ \sigma\,\varphi\left(\frac{S_- - \mu}{\sigma}\right)$

iii) $\mathbb{E}\left\{(V-\mu)^2 1_{V \notin [S_-,\, S_+]}\right\} = \sigma^2\,\mathbb{P}(V \notin [S_-,\, S_+]) \ + \ \sigma\,(S_+ - \mu)\,\varphi\left(\frac{S_+ - \mu}{\sigma}\right)$
$-\ \sigma\,(S_- - \mu)\,\varphi\left(\frac{S_- - \mu}{\sigma}\right)$

iv) $\mathbb{E}\left\{(V-\mu) 1_{V \notin [S_-,\, S_+]}\right\} = \sigma\,\varphi\left(\frac{S_+ - \mu}{\sigma}\right) \ - \ \sigma\,\varphi\left(\frac{S_- - \mu}{\sigma}\right)$

v) $\mathbb{E}\left\{(V-\mu)^2 1_{V \in [S_-,\, S_+]}\right\} = \sigma^2 - \sigma^2\mathbb{P}(V \notin [S_-,\, S_+]) - \sigma(S_+ - \mu)\,\varphi\left(\frac{S_+ - \mu}{\sigma}\right)$
$+\,\sigma\,(S_- - \mu)\,\varphi\left(\frac{S_- - \mu}{\sigma}\right)$.

To prove this lemma, use integration by parts.

According to iii) of Lemma 2, we have $\mathcal{A} = \mathbb{E}_{H_0}\left\{(y-\mu)^2 1_{y \notin [S_-, S_+]}\right\}$. It comes, according to Lemma 1:

$$
\begin{aligned}
&\mathbb{E}_{H_0}\left\{\left(\frac{\partial l_t}{\partial q}\mid_{\theta_0}\right)^2\right\} \\
&= \mathbb{E}_{H_0}\left[\left(\frac{y-\mu}{\sigma^2}\right)^2 \left\{\alpha(t)\overline{X}(t_1) + \beta(t)\overline{X}(t_2)\right\}^2\right] \\
&= \mathbb{E}_{H_0}\left[\left(\frac{y-\mu}{\sigma^2}\right)^2 \left\{\alpha(t)X(t_1) + \beta(t)X(t_2)\right\}^2 1_{y \notin [S_-, S_+]}\right] \\
&= \mathbb{E}_{H_0}\left\{\left(\frac{y-\mu}{\sigma^2}\right)^2 1_{y \notin [S_-, S_+]}\right\}\ \mathbb{E}_{H_0}\left[\left\{\alpha(t)X(t_1) + \beta(t)X(t_2)\right\}^2\right] \\
&= \mathcal{A}\left\{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)e^{-2(t_2-t_1)}\right\}/\sigma^4 \,.
\end{aligned}
$$

To conclude, after some calculations, we find

$$
I_{\theta_0} = Diag\left[\mathcal{A}\left\{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)e^{-2(t_2-t_1)}\right\}/\sigma^4\,,\ \frac{1}{\sigma^2}\,,\ \frac{2}{\sigma^2}\right] \quad . \qquad (7)
$$

Only the computation of $\mathbb{E}_{H_0}\left\{-\frac{\partial l_t}{\partial q \partial \mu}\mid_{\theta_0}\right\}$ and $\mathbb{E}_{H_0}\left\{-\frac{\partial l_t}{\partial q \partial \sigma}\mid_{\theta_0}\right\}$, were not easy. Let's prove now why these two terms are equal to zero. We have

$$
\frac{\partial l_t}{\partial q \partial \mu}\mid_{\theta_0} = -\frac{2p(t)-1}{\sigma^2} 1_{y \notin [S_-, S_+]} \ .
$$

It comes, using Lemma 1,

$$\begin{aligned}\mathbb{E}_{H_0}\left\{-\frac{\partial l_t}{\partial q\partial\mu}\mid_{\theta_0}\right\} &= -\frac{1}{\sigma^2}\,\mathbb{E}_{H_0}\left[\alpha(t)\overline{X}(t_1)+\beta(t)\overline{X}(t_2)\,\right]\\ &= -\frac{1}{\sigma^2}\,\mathbb{E}_{H_0}\left[\alpha(t)X(t_1)+\beta(t)X(t_2)\mid y\notin[S_-,S_+]\,\right]\,\mathbb{P}_{H_0}(y\notin[S_-,S_+])\\ &= -\frac{1}{\sigma^2}\,\mathbb{E}_{H_0}\left\{\alpha(t)X(t_1)+\beta(t)X(t_2)\right\}\,\mathbb{P}_{H_0}(y\notin[S_-,S_+])=0\,.\end{aligned}$$

Besides,

$$\frac{\partial l_t}{\partial q\partial\sigma}\mid_{\theta_0}=-\frac{2}{\sigma^3}(y-\mu)\left\{2p(t)-1\right\}1_{y\notin[S_-,S_+]}\,.$$

It comes

$$\begin{aligned}&\mathbb{E}_{H_0}\left(\frac{\partial l_t}{\partial q\partial\sigma}\mid_{\theta_0}\right)\\ &=-\frac{2}{\sigma^3}\,\mathbb{E}_{H_0}\left\{(y-\mu)1_{y\notin[S_-,S_+]}\right\}\,\mathbb{E}_{H_0}\left\{\alpha(t)X(t_1)+\beta(t)X(t_2)\right\}=0\,.\end{aligned}$$

It concludes the proof for the Fisher Information matrix.

Study of the score process under H_0

Let $l_t^n(\theta)$ be the log likelihood for n observations. Since the Fisher Information matrix is diagonal, the score statistic of the hypothesis "$q=0$" will be defined as

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

Now using formula (6) and using Lemma 1, it is clear that

$$\begin{aligned}\frac{\partial l_t^n}{\partial q}\mid_{\theta_0} &= \sum_{j=1}^{n}\frac{y_j-\mu}{\sigma^2}\,\left\{2p_j(t)-1\right\}\,1_{y_j\notin[S_-,S_+]}\\ &= \frac{\alpha(t)}{\sigma}\sum_{j=1}^{n}\varepsilon_j\,\overline{X}_j(t_1)\,+\,\frac{\beta(t)}{\sigma}\sum_{j=1}^{n}\,\varepsilon_j\,\overline{X}_j(t_2) \qquad (8)\end{aligned}$$

this proves that $D(.)$ is a linear interpolated process.

On the other hand, we have $\forall k = 1, 2$:

$$S_n(t_k) = \sum_{j=1}^{n} \frac{\sigma \varepsilon_j \, \overline{X}_j(t_k)}{\sqrt{n \, \mathcal{A}}} \, .$$

We have :

$$\begin{aligned}
\mathbb{E}\left\{\sigma\varepsilon \, \overline{X}(t_k)\right\} &= \mathbb{E}\left(\sigma \, \varepsilon \, 1_{y \notin [S_-, S_+]} \mid X(t_k) = 1\right) \mathbb{P}\left\{X(t_k) = 1\right\} \\
&\quad - \mathbb{E}\left(\sigma \, \varepsilon \, 1_{y \notin [S_-, S_+]} \mid X(t_k) = -1\right) \mathbb{P}\left\{X(t_k) = -1\right\} \\
&= \mathbb{E}\left(\sigma \, \varepsilon \, 1_{y \notin [S_-, S_+]}\right)/2 \; - \; \mathbb{E}\left(\sigma \, \varepsilon \, 1_{y \notin [S_-, S_+]}\right)/2 \\
&= 0 \, .
\end{aligned}$$

Besides :

$$\mathbb{E}\left[\sigma^2 \, \varepsilon^2 \left\{\overline{X}(t_k)\right\}^2\right] = \mathbb{E}(\sigma^2 \, \varepsilon^2 \, 1_{y \notin [S_-, S_+]}) = \mathcal{A} \, .$$

According to the Central Limit Theorem, it comes

$$S_n(t_k) \xrightarrow{\mathcal{L}} N(0, \, 1) \, .$$

Let's compute the covariance of the score statistics on markers, i.e. $\mathrm{Cov}\left\{S_n(t_1), S_n(t_2)\right\}$. Since $\mathbb{E}\left\{(y-\mu)^2 1_{y \notin [S_-, S_+]}\right\} = \mathcal{A}$, we have :

$$\begin{aligned}
\mathbb{E}\left\{S_n(t_1) S_n(t_2)\right\} &= \frac{1}{\mathcal{A}} \, \mathbb{E}\left\{(y-\mu)^2 \, X(t_1) \, X(t_2) \, 1_{y \notin [S_-, S_+]}\right\} \\
&= \frac{1}{\mathcal{A}} \, \mathbb{E}\left\{(y-\mu)^2 1_{y \notin [S_-, S_+]}\right\} \mathbb{E}\left\{X(t_1) X(t_2)\right\} = e^{-2(t_2 - t_1)} \quad .
\end{aligned}$$

As a consequence, $\mathrm{Cov}\left\{S_n(t_1), S_n(t_2)\right\} = \rho(t_1, t_2)$. The weak convergence of the score process, $S_n(.)$, is then a direct consequence of (8), the convergence of $(S_n(t_1), S_n(t_2))$ and the Continuous Mapping Theorem.

Study under the local alternative

Let's consider a local alternative defined by t^* and $q = a/\sqrt{n}$.

It remains to compute the asymptotic distribution of $S_n(.)$ under this alternative. Since we have already proved that $S_n(.)$ is a linear interpolated process (see formula 8), we only need to compute the distribution of $S_n(t_1)$ and $S_n(t_2)$ under the alternative. The mean function of the process is obviously a linear interpolated function (same interpolation as previously).

So, let's consider the score statistic at location t_k $\forall k = 1, 2$. We have

$$\begin{aligned} S_n(t_k) &= \sum_{j=1}^{n} \frac{(y_j - \mu)\, \overline{X}_j(t_k)}{\sqrt{n\, \mathcal{A}}} \\ &= \sum_{j=1}^{n} \frac{q\overline{U}_j(t^\star)\, \overline{X}_j(t_k)}{\sqrt{n\, \mathcal{A}}} \; + \; \sum_{j=1}^{n} \frac{\sigma\varepsilon_j\, \overline{X}_j(t_k)}{\sqrt{n\, \mathcal{A}}} \; . \end{aligned}$$

We will see, that we can apply the Law of Large Numbers for the first term and the Central Limit Theorem for the second term. To begin, let's focus on the second term. So, first we compute

$$\begin{aligned} &\mathbb{E}\left\{\sigma\varepsilon\, \overline{X}(t_k)\right\} \\ &= \frac{1}{2}\, \mathbb{E}\left\{\sigma\varepsilon\, \overline{X}(t_k) \mid U(t^\star) = 1\right\} \; + \; \frac{1}{2}\, \mathbb{E}\left\{\sigma\varepsilon\, \overline{X}(t_k) \mid U(t^\star) = -1\right\} \; . \end{aligned} \tag{9}$$

We have

$$\begin{aligned} &\mathbb{E}\left\{\sigma\varepsilon\, \overline{X}(t_k) \mid U(t^\star) = 1\right\} \\ &= \mathbb{E}\left\{\sigma\varepsilon\, 1_{y \notin [S_-, S_+]} \mid U(t^\star) = 1\right\} \mathbb{P}(X(t_k) = 1 \mid U(t^\star) = 1) \\ &- \mathbb{E}\left\{\sigma\varepsilon\, 1_{y \notin [S_-, S_+]} \mid U(t^\star) = 1\right\} \mathbb{P}(X(t_k) = -1 \mid U(t^\star) = 1) \end{aligned}$$

Let's compute $\mathbb{P}\left\{X(t_k) = 1 \mid U(t^\star) = 1\right\}$ and $\mathbb{P}\left\{X(t_k) = -1 \mid U(t^\star) = 1\right\}$. We have :

$$\begin{aligned} &\mathbb{P}\left\{X(t_k) = 1 \mid U(t^\star) = 1\right\} = \frac{\mathbb{P}\left\{X(t_k) = 1 \cap U(t^\star) = 1\right\}}{\mathbb{P}\left\{U(t^\star)\right\}} \\ &= 2\mathbb{P}\left\{X(t_k) = 1 \cap U(t^\star) = 1 \cap X(t_{k+1}) = 1\right\} + 2\mathbb{P}\left\{X(t_k) = 1 \cap U(t^\star) = 1 \cap X(t_{k+1}) = -1\right\} \\ &= \left\{1 - r(t_1, t_2)\right\}\; \mathbb{P}\left\{U(t^\star) = 1 \mid X(t_k) = 1 \cap X(t_{k+1}) = 1\right\} \\ &+ r(t_1, t_2)\, \mathbb{P}\left\{U(t^\star) = 1 \mid X(t_k) = 1 \cap X(t_{k+1}) = -1\right\} \\ &= 1 - r(t_1, t_2) \; + \; r(t_1, t_2)\, \frac{t_2 - t^\star}{t_2 - t_1} = 1 - r(t_1, t_2)\, \frac{t^\star - t_1}{t_2 - t_1} \\ &= 1 - \beta(t^\star) r(t_1, t_2)\, . \end{aligned} \tag{10}$$

It comes :

$$\mathbb{P}\left\{X(t_k) = -1 \mid U(t^\star) = 1\right\} = \beta(t^\star) r(t_1, t_2)\, .$$

Besides, according to iv) of Lemma 2,

$$\mathbb{E}\left\{\sigma\varepsilon\ 1_{y\notin[S_-,S_+]}\mid U(t^\star)=1\right\}=\sigma\varphi\left(\frac{S_+-\mu-q}{\sigma}\right)-\sigma\varphi\left(\frac{S_--\mu-q}{\sigma}\right)\ .$$

Using the relationship

$$1-2\beta(t^\star)r(t_1,t_2)=\alpha(t^\star)+\beta(t^\star)\ \rho(t_1,t_2)\ ,$$

it comes

$$\begin{aligned}
&\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\mid U(t^\star)=1\right\}\\
&=\{1-\beta(t^\star)r(t_1,t_2)\}\ \sigma\ \left\{\varphi\left(\frac{S_+-\mu-q}{\sigma}\right)-\varphi\left(\frac{S_--\mu-q}{\sigma}\right)\right\}\\
&-\beta(t^\star)\ r(t_1,t_2)\ \sigma\ \left\{\varphi\left(\frac{S_+-\mu-q}{\sigma}\right)-\varphi\left(\frac{S_--\mu-q}{\sigma}\right)\right\}\\
&=\{\alpha(t^\star)+\beta(t^\star)\ \rho(t_1,t_2)\}\ \sigma\ \left\{\varphi\left(\frac{S_+-\mu-q}{\sigma}\right)-\varphi\left(\frac{S_--\mu-q}{\sigma}\right)\right\}\ .
\end{aligned}\tag{11}$$

In the same way, after some calculations, we obtain :

$$\begin{aligned}
&\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\mid U(t^\star)=-1\right\}\\
&=-\{\alpha(t^\star)+\beta(t^\star)\ \rho(t_1,t_2)\}\ \sigma\ \left\{\varphi\left(\frac{S_+-\mu+q}{\sigma}\right)-\varphi\left(\frac{S_--\mu+q}{\sigma}\right)\right\}\ .
\end{aligned}\tag{12}$$

Since we consider q small, using a Taylor expansion at first order, we obtain for instance :

$$\varphi\left(\frac{S_--\mu+q}{\sigma}\right)=\frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{S_--\mu}{\sigma}\right)^2}\ \left\{1-\frac{(S_--\mu)\ q}{\sigma^2}+o(q)\right\}\ .$$

Finally, using Taylor expansions in formulae (11) and (12), we have :

$$\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\right\}=\{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)\}\ q\ \{z_{\gamma^+}\ \varphi(z_{\gamma^+})-z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\}\ +\ o(q)\ .$$

It comes

$$\mathbb{E}\left\{\sum_{j=1}^{n}\frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\right\}\to\frac{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)}{\sqrt{\mathcal{A}}}\ a\ \{z_{\gamma^+}\ \varphi(z_{\gamma^+})-z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\}\ .$$

We have now just to remark that

$$\begin{aligned}
&\mathbb{E}\left(\left\{\sigma\varepsilon\,\overline{X}(t_k)\right\}^2\right) = \mathbb{E}\left\{\sigma^2\,\varepsilon^2\,1_{y\notin[S_-,S_+]}\right\}\\
&= \mathbb{E}\left\{\sigma^2\,\varepsilon^2\,1_{y\notin[S_-,S_+]} \mid U(t^\star)=1\right\}/2 + \mathbb{E}\left\{\sigma^2\,\varepsilon^2\,1_{y\notin[S_-,S_+]} \mid U(t^\star)=-1\right\}/2\\
&\to \mathcal{A}/2 \;+\; \mathcal{A}/2 \to \mathcal{A}\,.
\end{aligned}$$

It comes

$$\mathbb{V}\left\{\sum_{j=1}^{n}\frac{\sigma\varepsilon_j\,\overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}}\right\}\to 1\,,$$

and according to the Central Limit Theorem

$$\sum_{j=1}^{n}\frac{\sigma\varepsilon_j\,\overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}}\xrightarrow{\mathcal{L}} N\left[\frac{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)}{\sqrt{\mathcal{A}}}\;a\;\left\{z_{\gamma^+}\,\varphi(z_{\gamma^+})-z_{1-\gamma^-}\,\varphi(z_{1-\gamma^-})\right\},\;1\right]\,. \tag{13}$$

Besides,

$$\begin{aligned}
&\mathbb{E}\left\{\overline{U}(t^\star)\,\overline{X}(t_k)\right\}\\
&= \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{1\mid 1\right\}\;\mathbb{P}(X(t_k)=1\mid U(t^\star)=1)\;-\;\frac{1}{2}\,\mathbb{P}_{t^\star}\left\{1\mid 1\right\}\;\mathbb{P}(X(t_k)=-1\mid U(t^\star)=1)\\
&-\frac{1}{2}\mathbb{P}_{t^\star}\left\{-1\mid -1\right\}\;\mathbb{P}(X(t_k)=1\mid U(t^\star)=-1)+\frac{1}{2}\mathbb{P}_{t^\star}\left\{-1\mid -1\right\}\;\mathbb{P}(X(t_k)=-1\mid U(t^\star)=-1)\\
&= \frac{1}{2}\mathbb{P}_{t^\star}\left\{1\mid 1\right\}\left\{1-2\beta(t^\star)r(t_1,t_2)\right\}-\frac{1}{2}\mathbb{P}_{t^\star}\left\{-1\mid -1\right\}\left\{2\beta(t^\star)r(t_1,t_2)-1\right\}\\
&= \frac{1}{2}\;\left\{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)\right\}\;\left(\mathbb{P}_{t^\star}\left\{1\mid 1\right\}\;+\;\mathbb{P}_{t^\star}\left\{-1\mid -1\right\}\right)\,.
\end{aligned}$$

Using Taylor expansion and after some work on integrals, we have :

$$P_{t^\star}\left\{1\mid 1\right\}=\Phi\left(\frac{S_--\mu}{\sigma}\right)\;-\;\frac{q}{\sigma}\varphi\left(\frac{S_--\mu}{\sigma}\right)\;+\;1\;-\;\Phi\left(\frac{S_+-\mu}{\sigma}\right)\;+\;\frac{q}{\sigma}\varphi\left(\frac{S_+-\mu}{\sigma}\right)\;+\;o(q) \tag{14}$$

where $\Phi(.)$ is the cumulative distribution of the standard normal distribution.

Note that we can replace q by $-q$ in order to obtain the expression of $P_{t^\star}\{-1 \mid -1\}$. It comes

$$\begin{aligned}\mathbb{E}\left\{\overline{U}(t^\star)\,\overline{X}(t_k)\right\} &= \{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)\}\left\{1+\Phi\left(\frac{S_- -\mu}{\sigma}\right)+\Phi\left(\frac{S_+ -\mu}{\sigma}\right)\right\}+o(q)\\ &= \{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)\}\ \gamma\ +\ o(q)\ .\end{aligned}$$

As a consequence, according to the Law of Large Numbers,

$$\sum_{j=1}^{n}\frac{q\overline{U}_j(t^\star)\,\overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}} \to \frac{a\ \{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)\}\ \gamma}{\sqrt{\mathcal{A}}}\ . \tag{15}$$

Finally, using formulae (13) and (15), we obtain

$$S_n(t_k) \xrightarrow{\mathcal{L}} N\left[\frac{a\sqrt{\mathcal{A}}}{\sigma^2}\left\{\alpha(t^\star)+\beta(t^\star)\rho(t_1,t_2)\right\},\ 1\right]\ . \tag{16}$$

Study of the supremum of the LRT process

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 $\widehat{\theta}_{|H_0} = \left(0, \overline{Y} = \sum Y_j/n, 1/n\sum(Y_j-\overline{Y})^2\right)$ (the genetic markers are useless under H_0). The likelihood ratio statistics will be defined as

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

on n independent observations.

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. Our goal is now to prove that the rest above is uniform in t.

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 for θ belonging to H_0. It is easy to check that at H_0 the Fisher information relative to t is zero so that the model is not regular.

It can be proved that assumptions 1, 2 and 3 of Azaïs et al. [3] holds. So, we can apply Theorem 1 of Azaïs et al. [3] and we have

$$\sup_{(t,\theta)} l_t(\theta) - l_{t^*}(\theta^*) = \sup_{d\in\mathcal{D}}\left[\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} d(X_j)\right\}^2 1_{d(X_j)\geq 0}\right] + o_P(1) \tag{17}$$

where the observation X_j stands for $Y_j, \overline{X}_j(t_1), \overline{X}_j(t_2)$ and where $\mathcal{D}$ is the set of scores defined in Azaïs et al. [3], see also Gassiat [9]. 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.

It is easy to see that

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

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 formula (17). Then, since the Fisher information matrix is diagonal (see formula (7)) , 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. Note that 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 (17) remains true under the alternative. It concludes the proof.

Remark 1. *According to the Law of Large Numbers, under the null hypothesis H_0 and under the local alternative $H_{at^\star}$, $\frac{1}{n}\sum 1_{y_j \notin [S_+, S_-]} \to \gamma$. So, γ corresponds asymptotically to the percentage of individuals genotyped. In the same way, γ_+ (resp. γ_-) corresponds asymptotically to the percentage of individuals genotyped with the largest (resp. the smallest) phenotypes.*

3. An easy way to perform the statistical test

Since $D(.)$ is a "linear normalized interpolated process", we can use Lemma 2.2 of Azaïs et al. [2] in order to compute easily the supremum of $D^2(.)$. Indeed, if we replace $\gamma_1(.)$ by $\alpha(.)$, $\gamma_2(.)$ by $\beta(.)$, and $\tilde{\rho}$ by $\rho(t_1, t_2)$, we can remark that all the conditions for applying Lemma 2.2 are fulfilled.

It comes

$$\max_{t\in[t_1,t_2]} \frac{\{\alpha(t)D(t_1)+\beta(t)D(t_2)\}^2}{\alpha^2(t)+\beta^2(t)+2\rho(t_1,t_2)\alpha(t)\beta(t)}$$
$$= \max\left(D^2(t_1)\ ,\ D^2(t_2)\ ,\ \frac{D^2(t_1)+D^2(t_2)-2\rho(t_1,t_2)D(t_1)D(t_2)}{1-\rho^2(t_1,t_2)} 1_{\frac{D(t_2)}{D(t_1)} \in\ ]\ \rho(t_1,t_2)\ ,\ \frac{1}{\rho(t_1,t_2)}\ [} \right) . \tag{18}$$

Before interpreting this formula, we have to remind that in Azaïs et al. [2], the authors prove that, under a model without interference and without selective genotyping, the LRT process converges to the square of a "non linear interpolated process", called $Z(.)$. Here, we can remark that the formula above is exactly the same if we want to compute the supremum of $D^2(.)$ or the supremum of $Z^2(.)$. Besides, under H_0, the processes $D(.)$ and $Z(.)$ discretized at markers locations, are both the squeleton of an Ornstein Uhlenbeck process. As a consequence, we will have exactly the same threshold if we consider a selective genotyping and an interference phenomenon, or if we deal with a model without selective genotyping and without interference. So, the Monte-Carlo Quasi Monte-Carlo method of Azaïs et al. [2] and based on Genz [10], is still suitable here.

Let's focus now on the data analysis. Which test statistic should we use in order to make the data analysis easy ? It is well known that under selective genotyping, when we focus only on one location of the genome which is a marker location, performing a LRT or a Wald test is time consuming : an EM algorithm is required to obtain the maximum likelihood estimators. In Rabier [19], I focus on only one location of the genome, and I propose a very easy test which is almost a comparison of means and which has the same asymptotic properties as LRT and Wald tests. So, the idea now is to adapt this comparison of means to our problem which focus on the whole chromosome.

As a consequence, $\forall k = 1, 2$, let's define now the test statistic $T_n(t_k)$ such

as

$$T_n(t_k) = \frac{\sum_{j=1}^n (Y_j - \overline{Y})\, \overline{X}_j(t_k)}{\sqrt{\sum_{j=1}^n (Y_j - \overline{Y})^2\, 1_{Y_j \notin [S_-, S_+]}}} \ .$$

We introduce the following lemma.

Lemma 3. *Let $T_n(.)$ be the process such as*

$$T_n(t) = \frac{\alpha(t)T_n(t_1) + \beta(t)T_n(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\rho(t_1,t_2)\alpha(t)\beta(t)}} \quad \text{, then } T_n(.) \Rightarrow D(.) \quad \text{and} \quad T_n^2(.) \Rightarrow D^2(.) \ .$$

Then, for the data analysis, we just have to consider as a test statistic $\sup T_n^2(.)$, which can be obtained easily using formula (18) and replacing $D(t_1)$ and $D(t_2)$ by respectively $T_n(t_1)$ and $T_n(t_2)$. Note that, according to Lemma 3, this test has the same asymptotic properties as the test based on the test statistic $\sup \Lambda_n(.)$, which corresponds to a LRT on the whole chromosome. So, Lemma 3 is an answer to the work of Rabbee et al. [18] where the authors study different strategies for analyzing data in selective genotyping.

On the other hand, a consequence of Lemma 3 is that the non extreme phenotypes (for which the genotypes are missing) don't bring any information for statistical inference. Indeed, our test statistics $T_n(t)$ are based only on the extreme phenotypes, as soon as we replace the empirical mean $\overline{Y}$ by $\hat{\mu}$, an estimator $\sqrt{n}$ consistent based only on the extreme phenotypes ($\hat{\mu}$ can be obtained by the method of moments for instance). This is a generalization of Rabier [19], where I proved that the non extreme phenotypes don't bring any information for statistical inference, when we look for a QTL only on one genetic marker.

Proof of Lemma 3

For $k = 1, 2$, we define $\tilde{T}(t_k)$ such as

$$\tilde{T}_n(t_k) = \frac{\sum_{j=1}^n (Y_j - \overline{Y})\, \overline{X}_j(t_k)}{\sqrt{n\, \mathcal{A}}} \ .$$

To begin, in order to make the proof easier, let's consider that we are under H_0. Since $\overline{Y} = \mu + O_P\left(1/\sqrt{n}\right)$, we have

$$\tilde{T}_n(t_k) = \frac{\sum_{j=1}^n (Y_j - \mu)\, \overline{X}_j(t_k)}{\sqrt{n\, \mathcal{A}}} \ + \ O_P\left(\frac{1}{\sqrt{n}}\right) \frac{\sum_{j=1}^n \overline{X}_j(t_k)}{\sqrt{n\, \mathcal{A}}} \ .$$

Let's focus on the second term under H_0. We have

$$\begin{aligned}\mathbb{E}\left[\overline{X}(t_k)\right] &= \mathbb{E}\left[X(t_k) \mid Y \notin [S_-, S_+]\right] \mathbb{P}(Y \notin [S_-, S_+]) \\ &= \mathbb{E}\left[X(t_k)\right] \gamma \;=\; 0\,.\end{aligned}$$

By Prohorov, it comes $\sum_{j=1}^{n} \overline{X}_j(t_k) = O_P\left(1/\sqrt{n}\right)$.
It comes $\tilde{T}_n(t_k) = S_n(t_k) + O_P\left(1/\sqrt{n}\right)$ and as a consequence $\tilde{T}_n(t_k) = S_n(t_k) + o_P(1)$. As said before, 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, the remainder converges also to 0 in probability under the alternative.
So, if we apply the Multivariate Central Limit Theorem, we have now $\left(\tilde{T}_n(t_1), \tilde{T}_n(t_2)\right) \xrightarrow{\mathcal{L}}$ $(D(t_1), D(t_2))$ whatever the hypothesis. We set in addition

$$\hat{\mathcal{A}} = \frac{1}{n}\sum_{j=1}^{n} (Y_j - \overline{Y})^2 \, 1_{Y_j \notin [S_-, S_+]}\,.$$

We have the relationship $(T_n(t_1), T_n(t_2)) = \sqrt{\frac{\mathcal{A}}{\hat{\mathcal{A}}}}\left(\tilde{T}_n(t_1), \tilde{T}_n(t_2)\right)$. Since $\hat{\mathcal{A}} \xrightarrow{\mathcal{L}} \mathcal{A}$ whatever the hypothesis, according to Slutsly and then Continuous Mapping theorem, we have $\sqrt{\frac{\mathcal{A}}{\hat{\mathcal{A}}}} \xrightarrow{\mathcal{L}} 1$. Using Slutsky, it comes $(T_n(t_1), T_n(t_2)) \xrightarrow{\mathcal{L}} (D(t_1), D(t_2))$. To conclude the proof, we just have to use the Continuous Mapping Theorem : $T_n(.) \Rightarrow D(.)$ and obviously $T_n^2(.) \Rightarrow D^2(.)$. It concludes the proof.

4. Several markers : the "genome scan"

We suppose now that there are K markers $0 = t_1 < t_2 < ... < t_K = T$. A QTL is lying at a position $t^\star$. So, in order to find the QTL, we will perform tests at every positions t on the chromosome. Note that we use the terminology "genome scan" instead of "interval mapping", since the "interval mapping" of Lander and Botstein [12] is usually computed by geneticists with a model without interference. 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 :

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

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

Since we use the Haldane modeling for the genome information at marker locations (cf. Section 1), in order to infer the value of $U(t^\star)$, we just need to keep the flanking markers. In others words, the information brought by the other markers is useless. So, we have

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

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 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* $\alpha(t^\star)$ *and* $\beta(t^\star)$*, where* t_1 *becomes* $t^{\star\ell}$ *and* t_2 *becomes* $t^{\star r}$
- $m_{t^\star}(t^\ell) = a\,\sqrt{\mathcal{A}}\,\rho(t^\ell, t^{\star\ell})\left\{\alpha(t^\star) + \beta(t^\star)\rho(t^{\star\ell}, t^{\star r})\right\}/\sigma^2$ *if* $t^\star > t^\ell$
- $m_{t^\star}(t^\ell) = a\ \sqrt{\mathcal{A}}\,\rho(t^\ell, t^{\star r})\left\{\alpha(t^\star)\rho(t^{\star r}, t^{\star\ell}) + \beta(t^\star)\right\}/\sigma^2$ *if* $t^\star < t^\ell$
- $m_{t^\star}(t^r) = a\ \sqrt{\mathcal{A}}\,\rho(t^r, t^{\star\ell})\left\{\alpha(t^\star) + \beta(t^\star)\rho(t^{\star\ell}, t^{\star r})\right\}/\sigma^2$ *if* $t^\star > t^r$
- $m_{t^\star}(t^r) = a\ \sqrt{\mathcal{A}}\,\rho(t^r, t^{\star r})\left\{\alpha(t^\star)\rho(t^{\star r}, t^{\star\ell}) + \beta(t^\star)\right\}/\sigma^2$ *if* $t^\star < t^r$.

Proof of Theorem 2
The proof of the theorem is the same as 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 and our process $D(.)$ is still a linear interpolated process. However, under the local alternative, the proof is more complicated than the proof of Theorem 1. Indeed, the location $t^\star$ of the QTL and the location t, can belong to a different marker interval.

In order to make the proof easier, instead of considering one location t inside a marker interval, we will focus on one genetic marker located at t_k. Indeed, since we deal with an interpolated process, we can obtain by interpolation the value of our process at t as soon as we know the values at the flanking markers.

According to the proof of Theorem 1, under the alternative

$$S_n(t_k) = q \sum_{j=1}^{n} \frac{\overline{U}_j(t^*)\overline{X}_j(t_k)}{\sqrt{n\mathcal{A}}} + \sum_{j=1}^{n} \frac{\sigma\varepsilon_j \overline{X}_j(t_k)}{\sqrt{n\mathcal{A}}} \,. \tag{19}$$

Note that in this proof, we will only consider the case $t_k < t^\star$ and the result corresponding to $t^\star < t_k$ will be deduced by symmetry.

Let's focus on the first term. We have :

$$\begin{aligned}\mathbb{E}\left\{\overline{U}(t^\star)\overline{X}(t_k)\right\} &= \mathbb{P}\left\{\overline{U}(t^\star) = 1 \cap \overline{X}(t_k) = 1\right\} + \mathbb{P}\left\{\overline{U}(t^\star) = -1 \cap \overline{X}(t_k) = -1\right\} \\ &\quad - \mathbb{P}\left\{\overline{U}(t^\star) = 1 \cap \overline{X}(t_k) = -1\right\} - \mathbb{P}\left\{\overline{U}(t^\star) = -1 \cap \overline{X}(t_k) = 1\right\} \,.\end{aligned}$$

Besides, since $\mathbb{P}\left\{X(t^{\star\ell}) = 1 \mid U(t^\star = 1)\right\} = 1 - \beta(t^\star) r(t^{\star\ell}, t^{\star r})$ (cf. formula 10), we have

$$\begin{aligned}&\mathbb{P}\left\{\overline{U}(t^\star) = 1 \cap \overline{X}(t_k) = 1\right\} = \mathbb{P}_{t^\star}\left\{1 \mid 1\right\} \mathbb{P}\left\{U(t^\star) = 1 \cap X(t_k) = 1 \cap X(t^{\star\ell}) = 1\right\} \\ &+ \mathbb{P}_{t^\star}\left\{1 \mid 1\right\} \mathbb{P}\left\{U(t^\star) = 1 \cap X(t_k) = 1 \cap X(t^{\star\ell}) = -1\right\} \\ &= \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}\, \left\{1 - r(t_k, t^{\star\ell})\right\}\, \left\{1 - \beta(t^\star) r(t^{\star\ell}, t^{\star r})\right\} + \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}\, r(t_k, t^{\star\ell})\, \beta(t^\star)\, r(t^{\star\ell}, t^{\star r}) \,.\end{aligned}$$

In the same way, after some calculations, we obtain

$$\begin{aligned}&\mathbb{P}\left\{\overline{U}(t^\star) = -1 \cap \overline{X}(t_k) = -1\right\} = \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}\, r(t_k, t^{\star\ell})\, \beta(t^\star) r(t^{\star\ell}, t^{\star r}) \\ &+ \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}\, \left\{1 - r(t_k, t^{\star\ell})\right\}\, \left\{1 - \beta(t^\star)\, r(t^{\star\ell}, t^{\star r})\right\} \,,\end{aligned}$$

$$\begin{aligned}&\mathbb{P}\left\{\overline{U}(t^\star) = 1 \cap \overline{X}(t_k) = -1\right\} = \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}\, r(t_k, t^{\star\ell})\, \left\{1 - \beta(t^\star) r(t^{\star\ell}, t^{\star r})\right\} \\ &+ \frac{1}{2}\,\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}\, \left\{1 - r(t_k, t^{\star\ell})\right\}\, \beta(t^\star)\, r(t^{\star\ell}, t^{\star r}) \,,\end{aligned}$$

$$\begin{aligned}&\mathbb{P}\left\{\overline{U}(t^\star) = -1 \cap \overline{X}(t_k) = 1\right\} = \frac{1}{2}\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\} \left\{1 - r(t_k, t^{\star\ell})\right\} \beta(t^\star) r(t^{\star\ell}, t^{\star r}) \\ &+ \frac{1}{2}\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\} r(t_k, t^{\star\ell}) \left\{1 - \beta(t^\star) r(t^{\star\ell}, t^{\star r})\right\} \,.\end{aligned}$$

As a consequence,

$$\begin{aligned}\mathbb{E}\left\{\overline{U}(t^\star)\overline{X}(t_k)\right\} &= -\frac{1}{2}\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}\rho(t_k,t^{\star\ell})\left\{2\beta(t^\star)r(t^{\star\ell},t^{\star r})-1\right\}\\ &\quad+\frac{1}{2}\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}\rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\\ &= \frac{1}{2}\rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\left(\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}+\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}\right)\ .\end{aligned}$$

Using a Taylor expansion of $\mathbb{P}_{t^\star}\left\{1 \mid 1\right\}$ and $\mathbb{P}_{t^\star}\left\{-1 \mid -1\right\}$ (cf. formula 14), it comes

$$\begin{aligned}\mathbb{E}\left\{\overline{U}(t^\star)\overline{X}(t_k)\right\} &= \rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\left\{1+\Phi\left(\frac{S_- -\mu}{\sigma}\right)+\Phi\left(\frac{S_+ -\mu}{\sigma}\right)\right\}+o(q)\\ &= \rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\gamma+o(q)\ ,\end{aligned}$$

and finally, according to the Law of Large Numbers

$$q\sum_{j=1}^{n}\frac{\overline{U}_j(t^*)\overline{X}_j(t_k)}{\sqrt{n\mathcal{A}}}\to\frac{a\rho(t_k,t^{\star\ell})}{\sqrt{\mathcal{A}}}\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\gamma\ . \tag{20}$$

Let's focus now on the second term of formula (19). First, according to iv) of Lemma 2

$$\mathbb{E}\left\{\sigma\varepsilon\ 1_{y\notin[S_-,S_+]} \mid U(t^\star)=1\right\}=\sigma\varphi\left(\frac{S_+-\mu-q}{\sigma}\right)-\sigma\varphi\left(\frac{S_--\mu-q}{\sigma}\right)\ , \tag{21}$$

$$\mathbb{E}\left\{\sigma\varepsilon\ 1_{y\notin[S_-,S_+]} \mid U(t^\star)=-1\right\}=\sigma\varphi\left(\frac{S_+-\mu+q}{\sigma}\right)-\sigma\varphi\left(\frac{S_--\mu+q}{\sigma}\right)\ . \tag{22}$$

Besides,

$$\begin{aligned}&\mathbb{P}\left\{X(t_k)=1 \mid U(t^\star)=1\right\}=\frac{\mathbb{P}\left\{X(t_k)=1\cap U(t^\star)=1\right\}}{\mathbb{P}\left\{U(t^\star)\right\}}\\ &=2\mathbb{P}\left\{X(t_k)=1\cap U(t^\star)=1\cap X(t^{\star\ell})=1\right\}+2\mathbb{P}\left\{X(t_k)=1\cap U(t^\star)=1\cap X(t^{\star\ell})=-1\right\}\\ &=\left\{1-r(t_k,t^{\star\ell})\right\}\left\{1-\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}+r(t_k,t^{\star\ell})\beta(t^\star)r(t^{\star\ell},t^{\star r})\ .\end{aligned}$$

It comes

$$2\mathbb{P}\left\{X(t_k)=1 \mid U(t^\star)=1\right\}-1=\rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\ .$$

In the same way, after some calculations,

$$2\mathbb{P}\left\{X(t_k)=1 \mid U(t^\star)=-1\right\}-1=-\rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\ .$$

It comes

$$\begin{aligned}&\mathbb{E}\left[\sigma\varepsilon\overline{X}(t_k)\mid U(t^\star)=1\right]\\&=\left[\rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\right]\ \mathbb{E}\left[\sigma\varepsilon 1_{y\notin[S_-,S_+]}\mid U(t^\star)=1\right]\ .\end{aligned}$$

Besides,

$$\begin{aligned}&\mathbb{E}\left[\sigma\varepsilon\overline{X}(t_k)\mid U(t^\star)=-1\right]\\&=\left[-\rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\right]\ \mathbb{E}\left[\sigma\varepsilon 1_{y\notin[S_-,S_+]}\mid U(t^\star)=-1\right]\ .\end{aligned}$$

Using a Taylor expansion in formulae (21) and (22), and using the relationship $1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})=\alpha(t^\star)+\beta(t^\star)\rho(t^{\star\ell},t^{\star r})$, it comes

$$\begin{aligned}&\mathbb{E}\left[\sigma\varepsilon\overline{X}(t_k)\right]\\&=q\ \left[\rho(t_k,t^{\star\ell})\left\{1-2\beta(t^\star)r(t^{\star\ell},t^{\star r})\right\}\right]\ \left\{z_{\gamma_+}\varphi(z_{\gamma_+})-z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\}+o(q)\\&=q\ \left[\rho(t_k,t^{\star\ell})\left\{\alpha(t^\star)+\beta(t^\star)\rho(t^{\star\ell},t^{\star r})\right\}\right]\ \left\{z_{\gamma_+}\varphi(z_{\gamma_+})-z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\}+o(q)\ .\end{aligned}$$

As in the proof of Theorem 1

$$\mathbb{E}\left(\left\{\sigma\varepsilon\ \overline{X}(t_k)\right\}^2\right)\to\mathcal{A}\ \text{ and }\ \mathbb{V}\left\{\sum_{j=1}^n\frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}}\right\}\to 1\ .$$

So, according to the Central Limit Theorem

$$\sum_{j=1}^n\frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}}\xrightarrow{\mathcal{L}}N\left[\frac{\rho(t_k,t^{\star\ell})\left\{\alpha(t^\star)+\beta(t^\star)\rho(t^{\star\ell},t^{\star r})\right\}}{\sqrt{\mathcal{A}}}\ a\ \left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})-z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\},\ 1\right]\ . \tag{23}$$

Finally, according to formulae (20) and (23) :

$$S_n(t_k)\xrightarrow{\mathcal{L}}N\left[\rho(t_k,t^{\star\ell})\left\{\alpha(t^\star)+\beta(t^\star)\rho(t^{\star\ell},t^{\star r})\right\}\ a\sqrt{\mathcal{A}}/\sigma^2,\ 1\right]\ .$$

As said before, the result for $t^\star<t_k$ is deduced by symmetry. It concludes the proof.

We introduce now our Theorem 3.

Theorem 3. *Let κ be the Asymptotic Relative Efficiency (ARE) with respect to the oracle situation where all the genotypes are known. Then, we have*

$$\begin{array}{l} \textit{i)}\ \kappa = \gamma\ +\ z_{\gamma_+}\,\varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-}) \\ \textit{ii)}\ \kappa \textit{ reaches its maximum for } \gamma_+ = \gamma_- = \gamma/2\ . \end{array}$$

According to i) of Theorem 3, the ARE with respect to the oracle situation (i.e. without selective genotyping), does not depend on the constant a linked to the QTL effect, and does not depend on the location of the QTL $t^\star$. Besides, we can remark that we have exactly the same ARE with respect to the oracle situation, if we scan the chromosome or if we focus only on one locus (even if the QTL is not on this locus). Indeed, since the mean functions (oracle situation and selective genotyping) are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$, it is obvious that the ARE will be the same if we scan the chromosome or if we focus only on one locus. On the other hand, according to ii) of Theorem 3, if we want to genotype only a percentage γ of the population, we should genotype the $\gamma/2\%$ individuals with the largest phenotypes and $\gamma/2\%$ individuals with the smallest phenotypes. It confirms by the theory what geneticists do in practice. It is also a generalization of Rabier [19] where I prove that we have to genotype symmetrically when we look for a QTL on only one genetic marker.

Proof of Theorem 3

The proof of i) is obvious since the mean functions of the selective genotyping and the oracle situation, are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$. Let's now prove that the maximum is reached for $\gamma_+ = \gamma_- = \gamma/2$. We have to answer the following question : how must we choose γ_+ and γ_- to maximize the efficiency ? We remind that $\gamma_+ + \gamma_- = \gamma$ and that $\varphi(.)$ and $\Phi(.)$ denote respectively the density and the cumulative distribution of the standard normal distribution. Let $u(.)$ be the function such as : $u(z_{\gamma_+}) = \Phi^{-1}\left\{\gamma - 1 + \Phi(z_{\gamma_+})\right\}$. Then, $z_{1-\gamma_-} = u(z_{\gamma_+})$.

Let $k_1(.)$ be the following function : $k_1(z_{\gamma_+}) = z_{\gamma_+}\varphi(z_{\gamma_+}) - u(z_{\gamma_+})\,\varphi\left\{u(z_{\gamma_+})\right\}$. In order to maximize κ, we have to maximize the function $k_1(.)$. Let $k_1'(.)$, $u'(.)$ and $\varphi'(.)$ be respectively the derivative of $k_1(.)$, $u(.)$ and $\varphi(.)$. We have

$$k_1'(z_{\gamma_+}) = \varphi(z_{\gamma_+})\ +\ z_{\gamma_+}\varphi'(z_{\gamma_+})\ -\ u'(z_{\gamma_+})\,\varphi\left\{u(z_{\gamma_+})\right\}\ -\ u(z_{\gamma_+})\,u'(z_{\gamma_+})\,\varphi'\left\{u(z_{\gamma_+})\right\}\ ,$$

$u'(z_{\gamma_+}) = \dfrac{\varphi(z_{\gamma_+})}{\varphi(z_{1-\gamma_-})}$ .Then

$$k_1'(z_{\gamma/2}) = \varphi(z_{\gamma/2}) - \left\{z_{\gamma/2}\right\}^2 \varphi(z_{\gamma/2}) - \varphi(z_{1-\gamma/2}) + \left\{z_{1-\gamma/2}\right\}^2 \varphi(z_{1-\gamma/2}) = 0$$

. It can be proved that it corresponds to a maximum. As a result, the efficiency κ reaches its maximum when $\gamma_+ = \gamma_- = \frac{\gamma}{2}$. It concludes the proof.

5. Applications

In this Section, we propose to illustrate the theoretical results obtained in this paper. For all the following applications, we will consider statistical tests at the 5% level. If we call

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

as explained before, an easy way to perform our statistical test is to use the test statistic

$$M_n = \max\left\{T_n^2(t_1), T_n^2(t_2), h_n(t_1, t_2), ..., T_n^2(t_{K-1}), T_n^2(t_K), h_n(t_{K-1}, t_K)\right\}.$$

Our first result is that the threshold (i.e. critical value) is exactly the same as the classical threshold obtained without selective genotyping and without interference. So, in order to obtain our threshold, the Monte Carlo Quasi Monte-Carlo methods of Azaïs et al. [2], based on Genz [10] is still suitable here. The advantage of this method is that it is very fast and it can be performed very easily (just download the Matlab package with graphical user interface, called "imapping.zip", on www.stat.wisc.edu/∼rabier). This is an alternative to the permutation method proposed by Manichaikul et al. [14] and inspired by Churchill and Doerge [6], which is very time consuming and not easy to compute in selective genotyping because of the missing genotypes. This way, in Figure 1, we propose to check these asymptotic results on simulated data. We consider a chromosome of length $T = 3$M and a sparse map : seven genetic markers are equally spaced every 50cM. For such a configuration, if we choose a level 5%, the corresponding threshold of Azaïs et al. [2] is 7.75. We consider here $\gamma = 0.4$, and different ways of performing the selective genotyping : genotyping symmetrically (i.e. $\gamma_+ = \gamma/2$), genotyping only the individuals with the largest phenotypes (i.e. $\gamma_+ = \gamma$) .... We can see that, whatever the value of γ_+, the Percentage of False Positives is close to the true level of the test (i.e. 5%) as soon as the number of individuals n is at least 100. Note that for small values of n (see $n = 50$), the threshold

seems to be too conservative. In Figure 2, we consider a smaller chromosome $T = 1$M and a denser map : 11 genetic markers are equally spaced every 10cM. Besides, we consider now $\gamma = 0.3$. We obtain the same kind of conclusions as previously.

In Figures 3 and 4, we focus on the alternative hypothesis. In Figure 3, we consider the sparse map. For the QTL effect q, we consider $a = 4$: we remind that $q = a/\sqrt{n}$. We focus on different locations $t^\star$ of the QTL and different values of γ_+. As expected (cf. Theorem 3), we can see that the Theoretical Power is maximum when we genotype symmetrically (i.e. $\gamma_+ = \gamma/2$). Note that, we also give in brackets the Empirical Power obtained for $n = 1000$, just to confirm our asymptotic results. Finally, in Figure 4, we focus on our dense genetic map. We obtain the same kind of conclusions as before. This result was expected since all the theoretical results obtained in this paper, are suitable for any kind of genetic map.

To conclude, we present in this paper easy ways to analyze data under selective genotyping and interference. That's why it must be interesting for geneticists.

6. Acknowledgements

I thank Jean-Marc Azaïs and Céline Delmas for fruitful discussions.

γ_+ \ n	1000	200	100	50
γ	4.89%	4.71%	4.48%	3.51%
$\gamma/2$	5.18%	5.14%	4.70%	3.95%
$\gamma/4$	5.07%	4.89%	4.61%	3.63%
$\gamma/8$	4.69%	4.98%	4.46%	3.71%

Figure 1: Percentage of False Positives as a function of n and the percentage γ_+ of individuals genotyped in the right tail. The chromosome is of length $T = 3$M and 7 markers are equally spaced every 50cM ($\gamma = 0.4$, $a = 0$, $\mu = 0$, $\sigma = 1$, 10000 samples of size n).

γ_+ \ n	1000	200	100	50
γ	5.05%	4.58%	4.20%	3.47%
$\gamma/2$	4.94%	4.73%	4.59%	4.21%
$\gamma/4$	4.82%	4.56%	4.65%	3.83%
$\gamma/8$	5.02%	4.87%	4.31%	3.40%

Figure 2: Percentage of False Positives as a function of n and the percentage γ_+ of individuals genotyped in the right tail. The chromosome is of length $T = 1$M and 11 markers are equally spaced every 10cM ($\gamma = 0.3$, $a = 0$, $\mu = 0$, $\sigma = 1$, 10000 samples of size n).

γ_+ \ $t^\star$	80cM	130cM	205cM	265cM
γ	45.08% (45.20%)	45.26% (44.71%)	54.66% (53.92%)	46.34% (46%)
$\gamma/2$	72.12% (71.50%)	72.05% (71.41%)	82.13% (82.21%)	74.27% (73.78%)
$\gamma/4$	68.06% (67.35%)	68.08% (67.98%)	78.79% (78.88%)	70.49% (69.73%)
$\gamma/8$	61.78% (61.51%)	61.75% (60.98%)	72.74% (72.20%)	64.07% (63.61%)

Figure 3: Theoretical power and Empirical Power (in brackets) as a function of the location of the QTL $t^\star$ and the percentage γ_+ of individuals non genotyped in the right tail. The chromosome is of length $T = 3$M and 7 markers are equally spaced every 50cM ($\gamma = 0.4$, $a = 4$, $\sigma = 1$, $\mu = 0$, 10000 samples of $n = 1000$ individuals, 100000 paths for the Theoretical Power).

γ_+ \ $t^\star$	12cM	36cM	52cM	75cM
γ	61.37% (60.74%)	62.03% (61.36%)	62.82% (62.23%)	61.68% (61.10%)
$\gamma/2$	83.83% (83.54%)	83.70% (83.15%)	84.61% (84.49%)	83.28% (83.79%)
$\gamma/4$	80.97% (80.95%)	80.86% (80.18%)	81.85% (81.10%)	80.67% (80.61%)
$\gamma/8$	76.15% (75.70%)	75.98% (75.36%)	76.75% (76.75%)	75.63% (75.14%)

Figure 4: Theoretical power and Empirical Power (in brackets) as a function of the location of the QTL $t^\star$ and the percentage γ_+ of individuals non genotyped in the right tail. The chromosome is of length $T = 1$M and 11 markers are equally spaced every 10cM ($\gamma = 0.3$, $a = 4$, $\sigma = 1$, $\mu = 0$, 10000 samples of $n = 1000$ individuals, 100000 paths for the Theoretical Power).

References

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] G.A. Churchill, R.W. Doerge, Empirical threshold values for quantitative trait mapping. Genetics, 138 (1994), 963-971.

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

[8] D. Darvasi and M. Soller, Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet., 85 (1992), 353-359.

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

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

[11] 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.

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

[13] R.J. Lebowitz, M. Soller, J.S. Beckmann, Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet., 73 (1987), 556-562.

[14] A. Manichaikul, A. Palmer, S. Sen, K. Broman, Significance thresholds for Quantitative Trait Locus mapping under selective genotyping. Genetics, 177 (2007), 1963-1966.

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

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

[17] H. Muranty and B. Goffinet, Selective genotyping for location and estimation of the effect of the effect of a quantitative trait locus. Biometrics, 53 (1997), 629-643.

[18] N. Rabbee, D. Speca, N. Armstrong, T. Speed, Power calculations for selective genotyping in QTL mapping in backcross mice. Genet. Res. Camb., 84 (2004), 103-108.

[19] C.E. Rabier, On statistical inference for selective genotyping. Unpublished result, hal-00658583 (2012).

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

[21] C.E. Rabier, On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. Unpublished result, hal-00675414 (2012).

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

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

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

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

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

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