An asymptotic test for Quantitative Trait Locus
detection in presence of missing genotypes
Charles-Elie Rabier

To cite this version:
Charles-Elie Rabier. An asymptotic test for Quantitative Trait Locus detection in presence of
missing genotypes. 2013.

HAL Id: hal-00796296
https://hal.archives-ouvertes.fr/hal-00796296
Submitted on 3 Mar 2013

HAL is a multi-disciplinary open access
archive for the deposit and dissemination of sci-
cientific research documents, whether they are pub-
lished or not. The documents may come from
teaching and research institutions in France or
abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est
destinée au dépôt et à la diffusion de documents
scientifiques de niveau recherche, publiés ou non,
émansant des établissements d’enseignement et de
recherche français ou étrangers, des laboratoires
publics ou privés.
An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes

Charles-Elie Rabier

Statistic and Botany departments, University of Wisconsin-Madison
Summary. We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait) on the interval \([0, T]\) representing a chromosome. The originality is in the fact that some genotypes are missing. 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 process is asymptotically the square of a “ non-linear interpolated and normalized Gaussian process “. We have an easy formula in order to compute the supremum of the square of this interpolated process. We prove that the threshold is exactly the same as in the classical situation without missing genotypes.

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

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 \(Y_j, j = 1, \ldots, n\), 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 (see for instance Wu et al. [2007] or Siegmund and Yakir [2007]). 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 modeling that can be represented as follows: \(X(0)\) is a random sign and \(X(t) = X(0)(−1)^N(t)\) where \(N(\cdot)\) is a standard Poisson process on \([0, T]\). Calculations on the Poisson distribution show that

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

we set in addition

\[
    \tilde{r}(t, t') = 1 - r(t, t').
\]

We assume an “analysis of variance model” for the quantitative trait:

\[
    Y = \mu + X(t^\star) q + \sigma \varepsilon
\]  

(1)

where \(\varepsilon\) is a Gaussian white noise and \(t^\star\) is the true location of the QTL.

Usually, in the classical problem of detecting a QTL on a chromosome, the genome information is available only at fixed locations \(t_1 = 0 < t_2 < \ldots < t_K = T\), 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.
The originality of this paper is that we consider the classical problem, but this time, we consider two real thresholds $S_-$ and $S_+$ with $S_- \leq S_+$ and the genotype of one individual is available if and only if the phenotype $Y$ belongs to the interval $S_- \leq Y \leq S_+$. If we call $\bar{X}(t)$ the random variable such as

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

then, in our problem, one observation will be now

$$(Y, \bar{X}(t_1), \ldots, \bar{X}(t_K)).$$

Note that with our notations :

- when $Y \in [S_-, S_+]$, we have $\bar{X}(t_1) = X(t_1), \ldots, \bar{X}(t_K) = X(t_K)$.
- when $Y \not\in [S_-, S_+]$, we have $\bar{X}(t_1) = 0, \ldots, \bar{X}(t_K) = 0$, which means that the genome information is missing at the marker locations.

Note also that, in this paper, the word "genotype" will refer to the genome information at markers locations.

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

It can be proved (see Section 2) that $(Y, \bar{X}(t_1), \ldots, \bar{X}(t_K))$ obeys to a mixture model with known weights, times a function $g(.)$ which does not depend of the parameters $\mu$, $q$ and $\sigma$ :

$$
\begin{align*}
[p(t^*) f_{(\mu+q,\sigma)}(y) & 1_{y \in [S_-,S_+]} + (1-p(t^*)) f_{(\mu-q,\sigma)}(y) 1_{y \notin [S_-,S_+]}] g(\cdot) + \\
\frac{1}{2} f_{(\mu+q,\sigma)}(y) 1_{y \notin [S_-,S_+]} + \frac{1}{2} f_{(\mu-q,\sigma)}(y) 1_{y \in [S_-,S_+]} \end{align*} g(\cdot)
$$

where $f_{(m,\sigma)}$ is the Gaussian density with parameters $(m,\sigma)$ and where the function $p(t)$ 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 oracle situation where all the individuals are genotyped, the asymptotic distribution of the LRT statistic has been given under some approximations by Rebaï et al. [1995], Rebaï et al. [1994], Cierco [1998], Azaïs and Cierco-Ayrolles [2002], Azaïs and Wschebor [2009], Chang et al. [2009]. Recently, Azaïs et al. [2012] 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".

Until now, in QTL detection, different strategies have been studied in order to reduce costs of genotyping. A very famous one is called selective genotyping. It consists of genotyping only the individuals who present an extreme phenotype (i.e. the individuals for which $Y \leq S_-$ or $Y \geq S_+$). Selective genotyping has been studied theoretically by many authors : for instance Lebowitz and al. [1987], Lander and Botstein [1989], Darvasi and Soller [1992], Muranty and
Goffinet [1997], Rabier [2012a]... However, in all these articles, the focus is only on one fixed location of the genome. Recently, in Rabier [2012b], we focused on the whole chromosome, and we proved that the distribution of the LRT statistic was asymptotically that of the maximum of the square of a “non linear normalized interpolated process”.

In this paper, our goal is not to focus on tools for reducing costs due to genotyping (as for selective genotyping), but to help geneticists to analyze data in the case of missing genotypes. According to Arends et al. [2010], “in an ideal world all datasets would be complete (with the genotype for every individual at every marker determined), however in the real world datasets are often incomplete”. As a consequence, 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 when only the genotypes of the non extreme individuals (i.e. the individuals for which the phenotypes $Y$ belong to the interval $[S_-, S_+]$) are available. 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 “non linear normalized interpolated process”. This is a generalization of the results obtained by Azaïs et al. [2012] only for the oracle situation. Under the null hypothesis, despite the missing genotypes, our process is exactly the same as the one obtained by Azaïs et al. [2012]. 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 and Lemma 3. In Theorem 3, we give the Asymptotic Relative Efficiency (ARE) with respect to the oracle situation. In Lemma 3, we present an easy formula (see also formula 21) to compute the maximum of the square of the non linear interpolated process. This formula is original. Usually when we look for a QTL on a chromosome with missing genotypes, 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 extreme phenotypes (for which the genotypes are missing) don’t bring any extra information for statistical inference. This result is complementary to the one obtained in Rabier [2012b], where I show that, under selective genotyping, the non extreme phenotypes don’t bring any information for statistical inference.

To conclude, we will illustrate our theoretical results with the help of simulated data. Note that, according to Theorem 1 and 2, the threshold (i.e. critical value) in our study is exactly the same as the classical threshold used in the oracle situation. So, in order to obtain our threshold, the Monte Carlo Quasi Monte-Carlo methods of Azaïs et al. [2012], based on Genz [1992] is still suitable here. This is an alternative to the permutation method proposed by Churchill and Doerge [1994], which is very time consuming and not easy to compute here because of the missing genotypes.

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

2. Main results: two genetic markers

To begin, we suppose that there are only two markers ($K = 2$) located at 0 and $T : 0 = t_1 < t_2 = T$. We look for a QTL located at $t^* \in [t_1, t_2]$. As said before,
since $t^*$ is unknown, we have to consider every locations $t \in [t_1, t_2]$. So, let’s consider a location $t \in [t_1, t_2]$, and let’s suppose $t = t^*$.

**Notations:** For $(i, i') \in \{-1, 1\}^2$, $Q_{i,i'}^i$ is the quantity such as

$$Q_{i,i'}^i = \mathbb{P}\{X(t) = 1 \mid X(t_1) = i, X(t_2) = i'\}.$$

Using Bayes rules, we have

$$Q_{i}^{1,1} = \frac{\bar{r}(t_1, t) \bar{r}(t, t_2)}{\bar{r}(t_1, t_2)}, \quad Q_{i}^{1,-1} = \frac{\bar{r}(t_1, t) \bar{r}(t_2)}{\bar{r}(t_1, t_2)} \quad (3)$$

$$Q_{i}^{-1,1} = \frac{r(t_1, t) \bar{r}(t, t_2)}{r(t_1, t_2)}, \quad Q_{i}^{-1,-1} = \frac{r(t_1, t) \bar{r}(t_2)}{\bar{r}(t_1, t_2)}.$$

We can remark that we have

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

**Notations:** $\mathbb{P}_i \{l \mid i\}$ is the quantity such as $\forall \ l \in \{-1, 0, 1\}$ and $\forall \ i \in \{-1, 1\}$

$$\mathbb{P}_i \{l \mid i\} = \mathbb{P}(X(t) = l \mid X(t) = i).$$

In order to compute the likelihood, we have to study the different probability distributions. To begin, let’s compute $\mathbb{P}(Y \in \lozenge y, y + dy \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1)$ for instance. We have, according to Bayes rules (we remind that we consider $t = t^*$),

$$\mathbb{P}(Y \in \lozenge y, y + dy \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1)$$

$$= \sum_{i \in \{-1, 1\}} \mathbb{P}(Y \in \lozenge y, y + dy \mid \overline{X}(t) = i) \mathbb{P}(\overline{X}(t) = i \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1).$$

Besides,

$$\mathbb{P}(Y \in \lozenge y, y + dy \mid \overline{X}(t) = i) = \frac{\mathbb{P}(Y \in \lozenge y, y + dy \cap \overline{X} \neq 0 \mid X(t) = i)}{\mathbb{P}(\overline{X}(t) \neq 0 \mid X(t) = i)}$$

and

$$\mathbb{P}(\overline{X}(t) = i \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1)$$

$$= \mathbb{P}(X(t) \neq 0) \cap X(t) = i \cap X(t_1) = 1 \cap X(t_2) = 1)$$

$$= \mathbb{P}_i \{l \mid i\} \mathbb{P}(X(t) = i \cap X(t_1) = 1 \cap X(t_2) = 1)$$

$$= \frac{1}{2} \mathbb{P}_i \{1 \mid 1\} \tau(t_1, t) \tau(t_2) 1_{i=1} + \frac{1}{2} \mathbb{P}_i \{-1 \mid -1\} \tau(t_1, t) \tau(t_2) 1_{i=-1}.$$

It comes, using formula (3),

$$\mathbb{P}(Y \in \lozenge y, y + dy \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1)$$

$$= \frac{1}{2} f_{(\mu+q, \sigma)}(y) 1_{y \in [s-, s_+]} \tau(t_1, t_2) Q_{i}^{1,1} + \frac{1}{2} f_{(\mu-q, \sigma)}(y) 1_{y \in [s-, s_+]} \tau(t_1, t_2) Q_{i}^{-1,-1}.$$
In the same way, after some calculations, we find
\[
\mathbb{P}(Y \in [y, y + dy] \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = -1) = \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \in [S_-, S_+]} r(t_1, t_2) Q_t^{1,1} + \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \in [S_-, S_+]} r(t_1, t_2) Q_t^{1,1},
\]
\[
\mathbb{P}(Y \in [y, y + dy] \cap \overline{X}(t_1) = -1 \cap \overline{X}(t_2) = 1) = \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \in [S_-, S_+]} \overline{r}(t_1, t_2) Q_t^{1,1} + \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \in [S_-, S_+]} \overline{r}(t_1, t_2) Q_t^{1,1},
\]
\[
\mathbb{P}(Y \in [y, y + dy] \cap \overline{X}(t_1) = -1 \cap \overline{X}(t_2) = -1) = \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \notin [S_-, S_+]} + \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \notin [S_-, S_+]}.
\]

Finally, when the genotype is missing (i.e. the phenotype is extreme), we find
\[
\mathbb{P}(Y \in [y, y + dy] \cap \overline{X}(t_1) = 0 \cap \overline{X}(t_2) = 0) = \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \notin [S_-, S_+]} + \frac{1}{2} f_{(\mu, q, \sigma)}(y) 1_{y \notin [S_-, S_+]}.
\]

Let’s define the quantity \( p(t) \) such as
\[
p(t) = Q_t^{1,1} 1_{\overline{X}(t_1) = 1} 1_{\overline{X}(t_2) = 1} + Q_t^{1,1} 1_{\overline{X}(t_1) = 1} 1_{\overline{X}(t_2) = -1} + Q_t^{1,1} 1_{\overline{X}(t_1) = -1} 1_{\overline{X}(t_2) = 1} + Q_t^{1,1} 1_{\overline{X}(t_1) = -1} 1_{\overline{X}(t_2) = -1}
\]
\[
(4)
\]
and let \( \theta = (q, \mu, \sigma) \) be the parameter of the model at \( t \) fixed. As a consequence, the likelihood of the triplet \((Y, \overline{X}(t_1), \overline{X}(t_2))\) with respect to the measure \( \lambda \otimes N \otimes N \), \( \lambda \) being the Lebesgue measure, \( N \) the counting measure on \( \mathbb{N} \), is \( \forall t \in [t_1, t_2] : \)
\[
L_t(\theta) = \bigg[ p(t) f_{(\mu, q, \sigma)}(y) 1_{y \in [S_-, S_+]} + (1 - p(t)) f_{(\mu, q, \sigma)}(y) 1_{y \in [S_-, S_+]} \bigg] g(t)
\]
\[
(5)
\]
where the function
\[
g(t) = \frac{1}{2} \left\{ r(t_1, t_2) 1_{\overline{X}(t_1) = 1} 1_{\overline{X}(t_2) = 1} + r(t_1, t_2) 1_{\overline{X}(t_1) = 1} 1_{\overline{X}(t_2) = -1} \right\}
\]
\[
+ \frac{1}{2} \left\{ r(t_1, t_2) 1_{\overline{X}(t_1) = -1} 1_{\overline{X}(t_2) = 1} + r(t_1, t_2) 1_{\overline{X}(t_1) = -1} 1_{\overline{X}(t_2) = -1} \right\}
\]
\[
+ \frac{1}{2} 1_{\overline{X}(t_1) = 0} 1_{\overline{X}(t_2) = 0}
\]
can be removed because it does not depend on the parameters. Note that for \( t = t^* \), we find our formula (2) of the introduction where \( p(t^*) \) is described in formula (4).

**Notations**: \( \gamma, \gamma_+ \) and \( \gamma_- \) 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**: \( \mathcal{B} \) is the quantity such as
\[ B = \sigma^2 \{ 1 - \gamma - z_\gamma \varphi(z_\gamma) + z_{1-\gamma} \varphi(z_{1-\gamma}) \}, \text{ where } \varphi(x) \text{ and } z_\alpha \text{ denote}
\]
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 \( (q, \mu, \sigma^2) \) vary in a compact and
that \( \sigma^2 \) is bounded away from zero. Let \( H_0 \) be the null hypothesis \( q = 0 \) and
define the following local alternative
\[ H_{at*}: \text{"the QTL is located at the position } t^* \text{ with effect } q = a/\sqrt{n} \text{ where } a \neq 0." \]

With the previous defined notations,
\[ S_n(.) \Rightarrow U(.) \ , \ \Lambda_n(.) \overset{F.d.}{\to} U^2(.) \ , \ \sup \Lambda_n(.) \overset{\mathcal{L}}{\to} \sup U^2(.)\]
as \( n \) tends to infinity, under \( H_0 \) and \( H_{at*} \), where:
- \( S_n(.) \) is the score process
- \( \Rightarrow \) is the weak convergence, \( \overset{F.d.}{\to} \) is the convergence of finite-dimensional
  distributions and \( \overset{\mathcal{L}}{\to} \) is the convergence in distribution
- \( U(.) \) is the Gaussian process with unit variance such as :
\[ U(t) = \frac{\alpha(t)U(t_1) + \beta(t)U(t_2)}{\sqrt{\{ \alpha(t)U(t_1) + \beta(t)U(t_2) \}}} \tag{6} \]
where
\[ \text{Cov} \{ U(t_1), U(t_2) \} = \rho(t_1, t_2) = \exp(-2|t_1 - t_2|) \]
\[ \alpha(t) = Q_{t}^{1,1} - Q_{t}^{-1,1} \]
\[ \beta(t) = Q_{t}^{1,1} - Q_{t}^{-1,1} \]

and with expectation :
- under \( H_0 \), \( m(t) = 0 \),
- under \( H_{at*} \),
\[ m_{t^*}(t) = \frac{\alpha(t) m_{t^*}(t_1) + \beta(t) m_{t^*}(t_2)}{\sqrt{\{ \alpha(t)U(t_1) + \beta(t)U(t_2) \}}} \]

where
\[ m_{t^*}(t_1) = \frac{a \sqrt{B} \rho(t_1, t^*)}{\sigma^2} \]
\[ m_{t^*}(t_2) = \frac{a \sqrt{B} \rho(t^*, t_2)}{\sigma^2} \]

In the sense of this equation, \( U(.) \) will be called a "non linear normalized interpolated process". We can see that under the null hypothesis, despite the missing genotypes, \( U(.) \) is exactly the same process as the process \( Z(.) \) of Theorem 2.1 of Azaïs et al. [2012] obtained for the oracle situation. However, under the alternative, the mean functions of the two processes are not the same anymore : the mean functions are proportional of a factor \( \sqrt{B}/\sigma \). Note also that \( U(.) \) is the generalization of \( Z(.) \). Indeed, if we choose \( S_- = -\infty \) and \( S_+ = +\infty \), that is to say the genotypes of all the individuals are available, the factor \( \sqrt{B}/\sigma \) is equal to 1, and \( U(.) \) is the same process as \( Z(.) \).

**Proof.** Theorem 1
Fisher Information Matrix

Let $l_t(\theta)$ be the loglikelihood. We first compute the Fisher information at a point $\theta_0$ that belongs to $H_0$. We have

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

(7)

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

Then,

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

Let’s introduce two key lemmas :

**Lemma 1.** We have the following relationship :

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

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 (4) and check that both sides coincide when $y \in [S_-, S_+]$.

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

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

ii) $\mathbb{E} \left\{ (W - \mu)^2 1_{W \in [S_-, S_+]} \right\} = -\sigma \varphi \left( \frac{s_+ - \mu}{\sigma} \right) + \sigma \varphi \left( \frac{s_- - \mu}{\sigma} \right)$

To prove this lemma, use integration by parts.

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

$$
\mathbb{E}_{H_0} \left\{ \left( \frac{\partial l_t}{\partial q} \bigg|_{\theta_0} \right)^2 \right\} = \mathbb{E}_{H_0} \left[ \left( \frac{y - \mu}{\sigma^2} \right)^2 \{\alpha(t) \overline{X}(t_1) + \beta(t) \overline{X}(t_2)\}^2 \right]
$$

$$
= \mathbb{E}_{H_0} \left[ \left( \frac{y - \mu}{\sigma^2} \right)^2 \{\alpha(t) X(t_1) + \beta(t) X(t_2)\}^2 1_{y \in [S_-, S_+]} \right]
$$

$$
= \mathbb{E}_{H_0} \left[ \left( \frac{y - \mu}{\sigma^2} \right)^2 1_{y \in [S_-, S_+]} \right] \mathbb{E}_{H_0} \left[ \{\alpha(t) X(t_1) + \beta(t) X(t_2)\}^2 \right]
$$

$$
= \mathbb{B} \left\{ \alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)e^{-2(t_2 - t_1)} \right\} / \sigma^4 .
$$
To conclude, after some calculations, we find
\[ I_{\theta_0} = \text{Diag} \left[ B \left\{ \alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)e^{-2(t_2-t_1)} \right\} / \sigma^4, \frac{1}{\sigma^2}, \frac{2}{\sigma^2} \right] . \] (9)

Only the computation of \( \mathbb{E}_{H_0} \left\{ -\frac{\partial l_t}{\partial q \partial \mu} \bigg|_{\theta_0} \right\} \) and \( \mathbb{E}_{H_0} \left\{ -\frac{\partial l_t}{\partial q \partial \sigma} \bigg|_{\theta_0} \right\} \), were not easy.

Let’s prove now why these two terms are equal to zero. We have
\[ \frac{\partial l_t}{\partial q \partial \mu} \bigg|_{\theta_0} = -\frac{2p(t) - 1}{\sigma^2} 1_{y \in [S_-,S_+]} . \]

It comes, using Lemma 1,
\[ \mathbb{E}_{H_0} \left\{ -\frac{\partial l_t}{\partial q \partial \mu} \bigg|_{\theta_0} \right\} = -\frac{1}{\sigma^2} \mathbb{E}_{H_0} \left[ \alpha(t)\overline{X}(t_1) + \beta(t)\overline{X}(t_2) \right] \]
\[ = -\frac{1}{\sigma^2} \mathbb{E}_{H_0} \left[ \alpha(t)X(t_1) + \beta(t)X(t_2) \bigg| y \in [S_-,S_+] \right] \mathbb{P}_{H_0}(y \in [S_-,S_+]) \]
\[ = -\frac{1}{\sigma^2} \mathbb{E}_{H_0} \left\{ \alpha(t)X(t_1) + \beta(t)X(t_2) \right\} \mathbb{P}_{H_0}(y \in [S_-,S_+]) = 0 . \]

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

It comes
\[ \mathbb{E}_{H_0} \left( \frac{\partial l_t}{\partial q \partial \sigma} \bigg|_{\theta_0} \right) \]
\[ = -\frac{2}{\sigma^3} \mathbb{E}_{H_0} \left\{ (y - \mu)1_{y \in [S_-,S_+]} \right\} \mathbb{E}_{H_0} \left\{ \alpha(t)X(t_1) + \beta(t)X(t_2) \right\} = 0 . \]

It concludes the proof for the Fisher Information matrix.

**Study of the score process under \( H_0 \)**

Since the Fisher Information matrix is diagonal, the score statistic (for \( n \) observations) of the hypothesis “\( q = 0 \)” will be defined as
\[ S_n(t) = \frac{\partial l_n}{\partial q} \bigg|_{\theta_0} \sqrt{\mathbb{V} \left( \frac{\partial l_n}{\partial q} \bigg|_{\theta_0} \right)} . \]

Now using formula (7) and using Lemma 1, it is clear that
\[ \frac{\partial l_n}{\partial q} \bigg|_{\theta_0} = \sum_{j=1}^{n} \frac{y_j - \mu}{\sigma^2} \left\{ 2p_j(t) - 1 \right\} 1_{y_j \in [S_-,S_+]} \]
\[ = \frac{\alpha(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j X_j(t_1) + \frac{\beta(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j X_j(t_2) \] (11)
this proves that $U(.)$ is a non linear interpolated process.

On the other hand, according to formula (7) and Lemma 1, we have $\forall k = 1, 2$:

$$S_n(t_k) = \sum_{j=1}^{n} \frac{\sigma \varepsilon_j X_j(t_k)}{\sqrt{n} B}.$$  

We have:

$$E\{\sigma \varepsilon X(t_k)\} = E(\sigma 1_{y \in [S_- , S_+] | X(t_k) = 1} \right) P\{X(t_k) = 1\}$$

$$- E(\sigma 1_{y \in [S_- , S_+] | X(t_k) = -1} \right) P\{X(t_k) = -1\}$$

$$= E(\sigma 1_{y \in [S_- , S_+]}) / 2 - E(\sigma 1_{y \in [S_- , S_+]}) / 2$$

$$= 0.$$  

Besides:

$$E\left[\sigma^2 \varepsilon^2 \{X(t_k)\}^2\right] = E(\sigma^2 \varepsilon^2 1_{y \in [S_- , S_+]}) = B.$$  

According to the Central Limit Theorem, it comes

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

Let’s compute the covariance of the score statistics on markers, i.e. $\text{Cov}\{S_n(t_1), S_n(t_2)\}$. Since $E\{ (y - \mu)^2 1_{y \in [S_- , S_+]} \} = B$, we have:

$$E\{S_n(t_1)S_n(t_2)\} = \frac{1}{B} E\{(y - \mu)^2 X(t_1) X(t_2) 1_{y \in [S_- , S_+]}\}$$

$$= \frac{1}{B} E\{(y - \mu)^2 1_{y \in [S_- , S_+]}\} E\{X(t_1)X(t_2)\} = e^{-2(t_2 - t_1)}.$$  

As a consequence, $\text{Cov}\{S_n(t_1), S_n(t_2)\} = e^{-2(t_2 - t_1)}$. The weak convergence of the score process, $S_n(.)$, is then a direct consequence of (11), 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 non linear interpolated process (see formula 11), 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 non linear interpolated function (same interpolation as previously).

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

$$S_n(t_k) = \sum_{j=1}^{n} \frac{(y_j - \mu) X_j(t_k)}{\sqrt{n} B}$$

$$= \sum_{j=1}^{n} q X_j(t^*) X_j(t_k) + \sum_{j=1}^{n} \frac{\sigma \varepsilon_j X_j(t_k)}{\sqrt{n} B}.$$
An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes

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

\[ E\{\sigma \varepsilon \overline{X}(t_k)\} = \frac{1}{2} E\{\sigma \varepsilon \overline{X}(t_k) | X(t^*) = 1\} + \frac{1}{2} E\{\sigma \varepsilon \overline{X}(t_k) | X(t^*) = -1\}. \] (12)

We have

\[ E\{\sigma \varepsilon \overline{X}(t_k) | X(t^*) = 1\} = E\{\sigma \varepsilon 1_{y \in [s_- , s_+]} | X(t^*) = 1\} \approx_r (t_k, t^*) - E\{\sigma \varepsilon 1_{y \notin [s_- , s_+]} | X(t^*) = 1\} r(t_k, t^*) = E\{\sigma \varepsilon 1_{y \notin [s_- , s_+]} | X(t^*) = 1\} e^{-2|t_k-t^*|}. \]

Besides, according to ii) of Lemma 2,

\[ E\{\sigma \varepsilon 1_{y \in [s_- , s_+]} | X(t^*) = 1\} = -\sigma \varphi \left( \frac{S_+ - \mu - q}{\sigma} \right) + \sigma \varphi \left( \frac{S_- - \mu - q}{\sigma} \right). \]

It comes

\[ E\{\sigma \varepsilon \overline{X}(t_k) | X(t^*) = 1\} = e^{-2|t_k-t^*|} \sigma \left\{ -\varphi \left( \frac{S_+ - \mu - q}{\sigma} \right) + \varphi \left( \frac{S_- - \mu - q}{\sigma} \right) \right\}. \] (13)

In the same way, after some calculations, we obtain

\[ E\{\sigma \varepsilon \overline{X}(t_k) | X(t^*) = -1\} = e^{-2|t_k-t^*|} \sigma \left\{ \varphi \left( \frac{S_+ - \mu + q}{\sigma} \right) - \varphi \left( \frac{S_- - \mu + q}{\sigma} \right) \right\}. \] (14)

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 (13) and (14), we have :

\[ E\{\sigma \varepsilon \overline{X}(t_k)\} = e^{-2|t_k-t^*|} q \left\{ -z_{\gamma^+} \varphi(z_{\gamma^+}) + z_{1-\gamma^-} \varphi(z_{1-\gamma^-}) \right\} + o(q). \]

It comes

\[ E\left\{ \sum_{j=1}^{n} \frac{\sigma \varepsilon_j \overline{X}_j(t_k)}{\sqrt{n B}} \right\} \rightarrow \frac{e^{-2|t_k-t^*|}}{\sqrt{B}} a \left\{ -z_{\gamma^+} \varphi(z_{\gamma^+}) + z_{1-\gamma^-} \varphi(z_{1-\gamma^-}) \right\}. \] (15)

We have now just to remark that

\[ E\left\{ \sigma \varepsilon \overline{X}(t_k)\right\}^2 = E\{\sigma^2 \varepsilon^2 1_{y \in [s_- , s_+]}\} = E\{\sigma^2 \varepsilon^2 1_{y \in [s_- , s_+]} | X(t^*) = 1\}/2 + E\{\sigma^2 \varepsilon^2 1_{y \in [s_- , s_+]} | X(t^*) = -1\}/2 \rightarrow B/2 + B/2 \rightarrow B. \]
It comes
\[ \forall \left\{ \sum_{j=1}^{n} \frac{\sigma \varepsilon_j X_j(t_k)}{\sqrt{n} \overline{B}} \right\} \to 1, \]
and according to the Central Limit Theorem
\[ \sum_{j=1}^{n} \frac{\sigma \varepsilon_j X_j(t_k)}{\sqrt{n} \overline{B}} \overset{\mathcal{L}}{\to} N\left[ \frac{e^{-2|t_k-t^*|}}{\sqrt{B}} a \left\{ -z_{\gamma^+} \varphi(z_{\gamma^+}) + z_{1-\gamma^-} \varphi(z_{1-\gamma^-}) \right\}, 1 \right]. \] (16)

Let us focus now on the first term of the score statistic. We have
\[ \mathbb{E}\{X(t^*) X(t_k)\} \]
\[ = \frac{1}{2} \mathbb{P}_{t^*} \{1 \mid 1\} \mathbb{E}\{X(t_k) \mid X(t^*) = 1\} - \frac{1}{2} \mathbb{P}_{t^*} \{-1 \mid -1\} \mathbb{E}\{X(t_k) \mid X(t^*) = -1\} \]
\[ = \frac{1}{2} e^{-2|t_k-t^*|} \left( \mathbb{P}_{t^*} \{1 \mid 1\} + \mathbb{P}_{t^*} \{-1 \mid -1\} \right). \]

Using Taylor expansion and after some work on integrals, we have:
\[ P_{t^*} \{1 \mid 1\} = -\Phi\left( \frac{S_- - \mu}{\sigma} \right) + \frac{q}{\sigma} \varphi\left( \frac{S_- - \mu}{\sigma} \right) + \Phi\left( \frac{S_+ - \mu}{\sigma} \right) - \frac{q}{\sigma} \varphi\left( \frac{S_+ - \mu}{\sigma} \right) + o(q) \]
where \( \Phi(\cdot) \) is the cumulative distribution of the standard normal distribution.

Note that we can replace \( q \) by \(-q\) in order to obtain the expression of \( P_{t^*} \{-1 \mid -1\} \).

It comes
\[ \mathbb{E}\{X(t^*) \overline{X}(t_k)\} = e^{-2|t_k-t^*|} \left\{ -\Phi\left( \frac{S_- - \mu}{\sigma} \right) + \Phi\left( \frac{S_+ - \mu}{\sigma} \right) \right\} + o(q) \]
\[ = e^{-2|t_k-t^*|} (1 - \gamma) + o(q). \]

As a consequence, according to the Law of Large Numbers,
\[ \sum_{j=1}^{n} q\overline{X}_j(t^*) X_j(t_k) \overset{\mathcal{L}}{\to} a e^{-2|t_k-t^*|} (1 - \gamma). \] (18)

Finally, using formulae (16) and (18), we obtain
\[ S_n(t_k) \overset{\mathcal{L}}{\to} N\left( \frac{a \sqrt{B}}{\sigma^2} e^{-2|t_k-t^*|}, 1 \right), \] (19)
which concludes the proof.

**Study of the supremum of the LRT process**
Let \( l^n_t(\hat{\theta}) \) be the log likelihood for \( n \) observations log likelihood. Let \( l^n_t(\hat{\theta}_{|H_0}) \) be the maximized log likelihood and let \( l^n_t(\hat{\theta}_{|H_0}) \) be the maximized log likelihood under \( H_0 \), with \( \hat{\theta}_{|H_0} = (0, \overline{Y} = \sum Y_j/n, 1/n \sum (Y_j - \overline{Y})^2) \) (the genetic markers are useless under \( H_0 \)). The likelihood ratio statistics will be defined as
\[ \Lambda_n(t) = 2\left[ l^n_t(\hat{\theta}) - l^n_t(\hat{\theta}_{|H_0}) \right], \]
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 \( t^*, \theta^* \) be the true parameter
that will be assumed to belong to \( H_0 \). Note that \( t^* \) makes no sense for \( \theta \) 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. [2009] holds. So,
we can apply Theorem 1 of Azaïs et al. [2009] and we have
\[
\sup_{(t, \theta)} l_t(\theta) - l_{t^*}(\theta^*) = \sup_{d \in D} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right\}^2 1_{d(X_j) \geq 0} + o_P(1) \tag{20}
\]
where the observation \( X_j \) stands for \( Y_j, X_{\bar{j}}(t_1), X_{\bar{j}}(t_2) \) and where \( D \) is the set
of scores defined in Azaïs et al. [2009], see also Gassiat [2002] and Azaïs et al.
[2006]. A similar result is true under \( H_0 \) with a set \( D_0 \). Let us precise the sets
of scores \( D \) and \( D_0 \). This sets are defined at the sets of scores of one parameter
families that converge to the true model \( p_{t^*, \theta^*} \) and that are differentiable in
quadratic mean.

It is easy to see that
\[
D = \left\{ \frac{\langle W, l_t'(\theta^*) \rangle}{\sqrt{\text{V}((W, l_t'(\theta^*)))}, W \in \mathbb{R}^3, t \in [t_1, t_2]} \right\}
\]
where \( l' \) is the gradient with respect to \( \theta \). In the same manner
\[
D_0 = \left\{ \frac{\langle W, l_t'(\theta^*) \rangle}{\sqrt{\text{V}((W, l_t'(\theta^*)))}, W \in \mathbb{R}^2} \right\},
\]
where now the gradient is taken with respect to \( \mu \) and \( \sigma \) only. Of course this
gradient does not depend on \( t \).

Using the transform \( W \rightarrow -W \) in the expressions of the sets of score, we see
that the indicator function can be removed in formula (20). Then, since the
Fisher information matrix is diagonal (see formula (9)) , it is easy to see that
\[
\sup_{d \in D} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right\}^2 - \sup_{d \in D_0} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right\}^2
= \sup_{t \in [t_1, t_2]} \left( \frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{\partial l_t}{\partial \theta}(X_j | \theta_0) \right)^2.
\]
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 (20) remains true
under the alternative.
Remark: According to the Law of Large Numbers, under the null hypothesis $H_0$ and under the local alternative $H_{alt}$, $\frac{1}{n} \sum_{i=1}^{n} 1_{y_i \in [s_+, s_-]} \to 1 - \gamma$. So, $1 - \gamma$ corresponds asymptotically to the percentage of individuals genotyped. In the same way, $\gamma_+$ (resp. $\gamma_-$) corresponds asymptotically to the percentage of non-genotyped individuals in the right tail (resp. the left tail) of the distribution.

3. An easy way to perform the statistical test

Since $U(.)$ is a "non linear normalized interpolated process", we can use Lemma 2.2 of Azaïs et al. [2012] in order to compute easily the supremum of $U^2(.)$. Note that this lemma is suitable here because we have exactly the same interpolation as in Theorem 2.1 of Azaïs et al. [2012]. It comes

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

(21)

Note that since under $H_0$, the process $U(.)$ is exactly the same process as the process $Z(.)$ obtained by Azaïs et al. [2012], we will have exactly the same threshold as the one under the oracle situation (i.e. all the individuals genotyped). So, the Monte-Carlo Quasi Monte Carlo method of Azaïs et al. [2012] and based on Genz [1992], 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? Indeed, 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. So, since we focus here on the whole chromosome, we have to propose the easiest statistical test for geneticists.

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

$$T_n(t_k) = \frac{\sum_{j=1}^{n} (Y_j - \bar{Y}) \overline{X_j(t_k)}}{\sqrt{\sum_{j=1}^{n} (Y_j - \bar{Y})^2} \cdot 1_{y_j \in [s_-, s_+]}}.$$  

We introduce the following lemma.

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

$$T_n(t) = \frac{\alpha(t)T_n(t_1) + \beta(t)T_n(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\rho(t_1, t_2)\alpha(t)\beta(t)}}$$  

, then $T_n(.) \Rightarrow U(.)$ and $T_n^2(.) \Rightarrow U^2(.)$.

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 (21) and replacing $U(t_1)$ and $U(t_2)$ 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 $\Lambda_n(.)$, which corresponds to a LRT on the whole chromosome.

On the other hand, a consequence of Lemma 3 is that the 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 non extreme phenotypes, as soon as we replace the empirical mean $\bar{Y}$ by $\hat{\mu}$, an estimator $\sqrt{n}$ consistent based only on the non extreme phenotypes ($\hat{\mu}$ can be obtained by the method of moments for instance). This result is complementary to the one obtained in Rabier [2012b], where I show that, under selective genotyping, the non extreme phenotypes (i.e. $Y \in [S_-, S_+]$ in the case of the selective genotyping) don’t bring any information for statistical inference.

**Proof. Lemma 3**

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

$$\hat{T}_n(t_k) = \frac{\sum_{j=1}^{n} (Y_j - \bar{Y}) X_j(t_k)}{\sqrt{n} \hat{\mathcal{B}}}.$$

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

$$\hat{T}_n(t_k) = \frac{\sum_{j=1}^{n} (Y_j - \mu) X_j(t_k)}{\sqrt{n} \hat{\mathcal{B}}} + O_P\left(\frac{1}{\sqrt{n}}\right) \frac{\sum_{j=1}^{n} X_j(t_k)}{\sqrt{n} \hat{\mathcal{B}}}.$$

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

$$\mathbb{E}\left[\hat{X}(t_k)\right] = \mathbb{E}[X(t_k) \mid Y \in [S_-, S_+)] \mathbb{P}(Y \in [S_-, S_+]) = \mathbb{E}[X(t_k)] (1 - \gamma) = 0.$$

By Prohorov, it comes $\sum_{j=1}^{n} \hat{X}_j(t_k) = O_P(1/\sqrt{n})$.

It comes $\hat{T}_n(t_k) = S_n(t_k) + O_P(1/\sqrt{n})$ and as a consequence $\hat{T}_n(t_k) = S_n(t_k) + o_P(1)$. 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 $(\hat{T}_n(t_1), \hat{T}_n(t_2)) \xrightarrow{\mathcal{L}} (U(t_1), U(t_2))$ whatever the hypothesis. We set in addition

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

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


In that case suppose that there are $K$ markers $0 = t_1 < t_2 < ... < t_K = T$. We consider values $t$, $t'$ or $t^*$ of the parameters that are distinct of the markers
positions, and the result will be prolonged by continuity at the markers positions.

For \( t \in [t_1, t_K] \setminus T_K \) where \( T_K = \{ t_1, \ldots, t_K \} \), we define \( t^\ell \) and \( t^r \) as:

\[
t^\ell = \sup \{ t_k \in T_K : t_k < t \}, \quad t^r = \inf \{ t_k \in T_K : t < t_k \}.
\]

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

**Theorem 2.** We have the same result as in Theorem 1, provided that we make some adjustments and that we redefine \( U(.) \) in the following way:

- in the definition of \( \alpha(t) \) and \( \beta(t) \), \( t_1 \) becomes \( t^\ell \) and \( t_2 \) becomes \( t^r \)
- under the null hypothesis, the process \( U(.) \) considered at marker positions is the "skeleton" 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, \( U(.) \) is obtained from \( U(t^\ell) \) and \( U(t^r) \) by interpolation and normalization using the functions \( \alpha(t) \) and \( \beta(t) \)
- at the marker positions, the expectation is such as \( m_{t^\ell}(t_k) = \frac{\alpha \sqrt{3} \rho(t_k, t^r)}{\sigma^2} \)
- at other positions, the expection is obtained from \( m_{t^\ell}(t_k) \) and \( m_{t^r}(t_k) \) by interpolation and normalization using the functions \( \alpha(t) \) and \( \beta(t) \).

The proof of the theorem is the same the proof of Theorem 1 since for a position \( t \), we can limit our attention to the interval \((t^\ell, t^r)\). Note that it is due to Haldane model with Poisson increments. Another key point for the proof, is that when \( t^* \) does not belong to the marker interval \((t^\ell, t^r)\), we can still use the section "Study under the alternative" of the proof of Theorem 1.

Another important point is that since for a position \( t \) we can limit our attention to the interval \((t^\ell, t^r)\), Lemma 3 and formula (21) are still true here. We just have to replace \( t_1 \) and \( t_2 \) by \( t^\ell \) and \( t^r \) in order to have the good expressions. As a consequence, we can easily compute \( \sup T^2_n(.) \).

We introduce now our Theorem 3.

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

\[
i) \quad \kappa = 1 - \gamma - z_{\gamma_+} \phi(z_{\gamma_+}) + z_{1-\gamma_-} \phi(z_{1-\gamma_-})
\]

\[
ii) \quad \kappa \text{ reaches its maximum for } \gamma_+ = \gamma \text{ or } \gamma_- = \gamma
\]

\[
iii) \quad \kappa > 1 - \gamma \iff z_{\gamma_+} \phi(z_{\gamma_+}) > z_{1-\gamma_-} \phi(z_{1-\gamma_-}).
\]

According to i) of Theorem 3, the ARE with respect to the oracle situation, does not depend on the constant \( \alpha \) linked to the QTL effect, and does not depend on the location of the QTL \( t^* \). On the other hand, according to ii) of Theorem 3, if only a percentage \( 1 - \gamma \) of genotypes is available in the population considered, the efficiency of our test is maximum when all the missing genotypes are located in the right tail of the distribution (i.e. \( \gamma_+ = \gamma \)). Obviously, by symmetry, the efficiency of our test is also maximum when all the missing genotypes are located in the left tail of the distribution (i.e. \( \gamma_- = \gamma \)). Note also, that according to iii),
our test can reduce costs due to genotyping when \( z_{\gamma} \varphi(z_{\gamma}) > z_{1-\gamma} \varphi(z_{1-\gamma}) \). However, this condition is very restrictive due to the properties of the Gaussian distribution.

**Proof.** The proof of i) is obvious since the mean function of the process \( U(.) \) and the one of the process \( Z(.) \) corresponding to the oracle situation, are proportional of a factor \( \sqrt{B}/\sigma \). Let’s now prove that the maximum is reached for \( \gamma_\rightarrow = \gamma \), that is to say \( \gamma_\rightarrow = 0 \), since \( \gamma = \gamma_+ + \gamma_- \). Note that without loss generality, it will also prove that the maximum is reached for \( \gamma_+ = \gamma \) and \( \gamma_- = 0 \). We have to answer the following question : how must we choose \( \gamma_+ \) and \( \gamma_- \) to maximize the efficiency ? We remind that \( \gamma_+ + \gamma_- = \gamma \) and \( \varphi(.) \) and \( \Phi(.) \) denote respectively the density and the cumulative distribution of the standard normal distribution. Let \( u(.) \) be the function such as : \( u(z_{\gamma_+}) = \Phi^{-1}\{\gamma - 1 + \Phi(z_{\gamma_+})\} \). Then, \( z_{1-\gamma_-} = u(z_{\gamma_+}) \).

Let \( k_1(.) \) be the following function : \( k_1(z_{\gamma}) = z_{\gamma} \varphi(z_{\gamma}) - u(z_{\gamma}) \varphi\{u(z_{\gamma})\} \).

In order to maximize \( \kappa \), we have to minimize the function \( k_1(.) \). Let \( k_1'(.), u'(.) \) and \( \varphi'(.) \) be respectively the derivative of \( k_1(.) \), \( u(.) \) and \( \varphi(.) \). We have :

\[
k_1'(z_{\gamma}) = \varphi(z_{\gamma}) + z_{\gamma} \varphi'(z_{\gamma}) - u'(z_{\gamma}) \varphi\{u(z_{\gamma})\} - u(z_{\gamma}) u'(z_{\gamma}) \varphi'\{u(z_{\gamma})\},
\]
\[
u'(z_{\gamma}) = \frac{\varphi(z_{\gamma})}{\varphi(z_{1-\gamma})}.
\]

As a consequence,

\[
k_1'(z_{\gamma}) = \varphi(z_{\gamma})(z_{\gamma}^2 - z_{\gamma_+}^2).
\]

If \( z_{\gamma} = +\infty \), then \( k_1'(z_{\gamma}) = 0 \). It can been proved that \( \gamma_+ = 0 \) corresponds to a minimum for \( k_1(.) \). As a result, the efficiency \( \kappa \) reaches its maximum when \( \gamma_\rightarrow = \gamma \).

**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

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

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

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

Our first result is that the threshold (i.e. critical value) is the same if only the genotypes of the non extreme individuals (i.e. the individuals for which \( Y \in [S_-,S_+] \)) are available or if all the genotypes are available (i.e. the oracle situation). So, the Monte-Carlo Quasi Monte-Carlo method, proposed by Azaïs et al. [2012] (based on Genz [1992]) for the oracle situation, is still suitable here to obtain our threshold. Note that in Azaïs et al. [2012], the authors show
that their method gives better results than the method of Feingold and al. [1993]
based on Siegmund [1985], and the method of Rebaï et al. [1994] based on Davies
[1977] and Davies [1987]. This way, in Tables 1 and 2, we propose to check on
simulated data, the fact that the threshold is the same as in the oracle situation.
First, in Table 1, we consider a sparse map: a chromosome of length $T = 1M$,
with two genetic markers located at each extremity. For such a configuration, if
we choose a 5% level, the corresponding threshold is 4.89. We consider $\gamma = 0.2$.
In other words we have 20% of missing genotypes. Besides, we consider different
values for the percentage $\gamma_+$ of individuals not genotyped in the right tail of the
distribution. We can see that, whatever the value of $\gamma_+$, the Percentage of False
Positives is close to the true level of the test (i.e. 5%) even for small values of $n$ (see $n = 50$).
Then, in Table 2, we consider a more dense genetic map. We still consider a chromosome of length $T = 1M$, but 6 genetic markers are now equally spaced every 20cM. We can remark that, as previously, the Percentage of False Positives is close to 5%.

Let’s now focus on the alternative hypothesis. To begin, in Table 3, we
consider the sparse map and the same value of $\gamma$ as previously. For the QTL
effect $q$, we consider $a = 4$: we remind that $q = a/\sqrt{n}$. We focus on different
locations $t^*$ of the QTL and different values of $\gamma_+$. We present the Theoretical
Power based on 100000 paths of the asymptotic process, and also the Empirical
Power (in brackets) obtained for $n = 1000$. We can see that the Theoretical
Power and the Empirical Power are very close whatever the values of $t^*$ and
$\gamma_+$ are. Besides, as expected (cf. Theorem 3), we can see that the Theoretical
Power is maximum when the missing genotypes are all located in the right tail
of the distribution (i.e. $\gamma_+ = \gamma$). Note that we would have obtained the same
result for $\gamma_+ = \gamma$ (i.e. left tail). Finally, in Table 4, we consider the dense map,
and we change the value of $a$: $a = 6$. 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 in presence
of missing genotypes. That’s why it must be interesting for geneticists.

<table>
<thead>
<tr>
<th>$\gamma_+$</th>
<th>$n$</th>
<th>1000</th>
<th>200</th>
<th>100</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma/4$</td>
<td></td>
<td>4.98%</td>
<td>5.43%</td>
<td>4.60%</td>
<td>4.79%</td>
</tr>
<tr>
<td>$\gamma/2$</td>
<td></td>
<td>5.10%</td>
<td>4.87%</td>
<td>5.17%</td>
<td>4.94%</td>
</tr>
<tr>
<td>$\gamma$</td>
<td></td>
<td>5.18%</td>
<td>5.31%</td>
<td>4.49%</td>
<td>4.61%</td>
</tr>
</tbody>
</table>

Fig. 1. Percentage of False Positives as a function of the number of individuals $n$ and
the percentage of individuals $\gamma_+$ not genotyped in the right tail. The chromosome is of
length $T = 1M$ and 2 markers are located at each extremity ($\gamma = 0.2$, $1 - \gamma = 0.8$, $a = 0$,
$\sigma = 1$, $\mu = 0$, 10000 samples of $n$ individuals).
An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$n$</th>
<th>1000</th>
<th>200</th>
<th>100</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma/4$</td>
<td>5.08%</td>
<td>4.65%</td>
<td>4.59%</td>
<td>4.48%</td>
<td></td>
</tr>
<tr>
<td>$\gamma/2$</td>
<td>5.01%</td>
<td>4.70%</td>
<td>4.56%</td>
<td>4.44%</td>
<td></td>
</tr>
<tr>
<td>$\gamma$</td>
<td>5.06%</td>
<td>4.67%</td>
<td>4.28%</td>
<td>4.20%</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Percentage of False Positives as a function of the number of individuals $n$ and the percentage of individuals $\gamma_+$ not genotyped in the right tail. The chromosome is of length $T = 1M$ and 6 markers are equally spaced every 20cM ($\gamma = 0.2$, $1 - \gamma = 0.8$, $a = 0$, $\sigma = 1$, $\mu = 0$, 10000 samples of $n$ individuals).

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$t^*$</th>
<th>5cM</th>
<th>12cM</th>
<th>18cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma/4$</td>
<td>60.46%</td>
<td>(60.75%)</td>
<td>58.88%</td>
<td>(62.18%)</td>
</tr>
<tr>
<td>$\gamma/2$</td>
<td>56.17%</td>
<td>(56.68%)</td>
<td>54.57%</td>
<td>(58.16%)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>76.61%</td>
<td>(76.34%)</td>
<td>74.79%</td>
<td>(78.60%)</td>
</tr>
</tbody>
</table>

Fig. 3. Theoretical power and Empirical Power (in brackets) as a function of the location of the QTL $t^*$ and the percentage $\gamma_+$ of individuals non genotyped in the right tail. The chromosome is of length $T = 20cM$ and 2 markers are located at each extremity ($\gamma = 0.2$, $1 - \gamma = 0.8$, $a = 4$, $\sigma = 1$, $\mu = 0$, 10000 samples of $n = 1000$ individuals, 100000 paths for the Theoretical Power).

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$t^*$</th>
<th>12cM</th>
<th>35cM</th>
<th>48cM</th>
<th>77cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma/4$</td>
<td>83.99%</td>
<td>(83.47%)</td>
<td>86.39%</td>
<td>(85.91%)</td>
<td>84.91%</td>
</tr>
<tr>
<td>$\gamma/2$</td>
<td>80.14%</td>
<td>(79.62%)</td>
<td>82.75%</td>
<td>(81.88%)</td>
<td>80.96%</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>95.36%</td>
<td>(94.87%)</td>
<td>96.23%</td>
<td>(96.07%)</td>
<td>95.48%</td>
</tr>
</tbody>
</table>

Fig. 4. Theoretical power and Empirical Power (in brackets) as a function of the location of the QTL $t^*$ and the percentage $\gamma_+$ of individuals non genotyped in the right tail. The chromosome is of length $T = 1M$ and 6 markers are equally spaced every 20cM ($\gamma = 0.2$, $1 - \gamma = 0.8$, $a = 6$, $\sigma = 1$, $\mu = 0$, 10000 samples of $n = 1000$ individuals, 100000 paths for the Theoretical Power).
References


Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 64, 247-254.

Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 74, 33-43.


An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes


