On stochastic processes for Quantitative Trait Locus mapping under selective genotyping

Charles-Elie Rabier

To cite this version:

Charles-Elie Rabier. On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. 2012. <hal-00675414v3>
RESEARCH ARTICLE

On stochastic processes for Quantitative Trait Locus mapping under selective genotyping

Charles-Elie Rabier a,b*

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

We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait) on the interval [0,T] representing a chromosome. The originality is in the fact that we are under selective genotyping: only the individuals with extreme phenotypes are genotyped. 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∗ 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 we have to genotype symmetrically and that the threshold is exactly the same as in the situation where all the individuals are genotyped.

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

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

1. Introduction

We study a backcross population: A × (A × 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 Yj, j = 1,...,n, the observations, which we will assume to be Gaussian, independent and identically distributed (i.i.d.). The mechanism of genetics, or more precisely of meiosis, implies that among the two chromosomes of each individual, one is purely inherited from A while the other (the “recombined” one), consists of parts originated from A and parts originated from B, due to crossing-overs.

The chromosome will be represented by the segment [0, T]. The distance on [0, T] is called the genetic distance, it is measured in Morgans (see for instance Wu et al. [1] or Siegmund and Yakir [2]). 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)

*Corresponding author. Email: rabier@stat.wisc.edu
where $N(.)$ 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

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

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

$$Y = \mu + X(t^*) q + \sigma \varepsilon$$

where $\varepsilon$ is a Gaussian white noise and $t^*$ 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 $t_1 = 0 < t_2 < ... < t_K = T$, called genetic markers. So, usually an observation is

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

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

The 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 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

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

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 $(Y_j, \overline{X}_j(t_1), ..., \overline{X}_j(t_K))$ i.i.d.

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

$$[p(t^*) f_{(\mu+q,\sigma)}(y) 1_{y \notin [S_-, S_+]} + \{1-p(t^*)\} f_{(\mu-q,\sigma)}(y) 1_{y \in [S_-, S_+]}$$

$$+ \frac{1}{2} f_{(\mu+q,\sigma)}(y) 1_{y \in [S_-, S_+]} + \frac{1}{2} f_{(\mu-q,\sigma)}(y) 1_{y \in [S_-, S_+]}] g(.)$$

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 (i.e. without selective genotyping), the asymptotic distribution of the LRT statistic has been given under some approximations by Rebaï et al. [4], Rebaï et al. [5], Cierco [6], Azaïs and Cierco-Ayrolles [7], Azaïs and Wschebor [8], Chang et al. [9]. Recently, Azaïs et al. [10] 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”.

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 with a selective genotyping. Selective genotyping has been studied theoretically by many authors : for instance Lebowitz and al. [11], Lander and Botstein [12], Darvasi and Soller [13], Muranty and Goffinet [14], Rabier [15]... However, in all these articles, the focus is only on one fixed location of the genome. This way, our study which focus on the whole chromosome is totally new, with a real impact for geneticists. In a more practical point of view, we can find in Rabbee et al. [16], a simulation study in which the authors study different strategies for analyzing data in selective genotyping and give the power associated to each strategy. On the other hand, in Manichaikul et al. [17], the authors focus on permutation tests for selective genotyping... This way, our study is complementary to the work of Rabbee et al. [16] and Manichaikul et al. [17].

The main result of the paper (Theorems 2.5 and 4.1) is that the distribution of the LRT statistic is asymptotically that of the maximum of the square of a “non linear normalized interpolated process”. This is a generalization of the results obtained by Azaïs et al. [10] only for the oracle situation. Under the null hypothesis, despite the selective genotyping, our process is exactly the same as the one obtained by Azaïs et al. [10]. 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 4.2. We give the Asymptotic Relative Efficiency (ARE) with respect to the oracle situation. 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 4.2 is the following : if we want to genotype only a percentage $\gamma$ 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 [15], 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 3.1 and formula 22) 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 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 [15] but for the whole chromosome). In other words, we give theoretical answers to the previous work of Rabbee et al. [16].
To conclude, we will illustrate our theoretical results with the help of simulated data. Note that, according to Theorem 2.5 and 4.1, the threshold (i.e. critical value) in selective genotyping, 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. [10], based on Genz [18] is still suitable here. This is an alternative to the permutation method proposed by Manichaikul et al. [17] and inspired by Churchill and Doerge [19], 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 [20] 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 \( t = t^* \).

Notation 2.1: For \((i,i') \in \{-1,1\}^2 \), \( Q_{i,i'}^t \) is the quantity such as

\[
Q_{i,i'}^t = P \{ X(t) = 1 | X(t_1) = i, X(t_2) = i' \}.
\]

Using Bayes rules, we have

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

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

We can remark that we have

\[
Q_{1,-1}^t = 1 - Q_{1,1}^t \quad \text{and} \quad Q_{-1,1}^t = 1 - Q_{-1,-1}^t.
\]

Notation 2.2: \( P_t \{ l | i \} \) is the quantity such as \( \forall \ i \in \{-1,1\} \) and \( \forall \ l \in \{-1,0,1\} \)

\[
P_t \{ l | i \} = P(\overline{X}(t) = l | X(t) = i) .
\]

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

\[
P(Y \in [y, y + dy] \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1) = \sum_{i \in \{-1,1\}} P(Y \in [y, y + dy] \mid \overline{X}(t) = i) \ P(\overline{X}(t) = i \cap \overline{X}(t_1) = 1 \cap \overline{X}(t_2) = 1).
\]
Let’s define the quantity $p(t)$ such as

\[
p(t) = Q_t^{1,1} 1_{X(t)=1} 1_{X(t_1)=1} + Q_t^{1,-1} 1_{X(t)=1} 1_{X(t_2)=1} + Q_t^{1,1} 1_{X(t)=1} 1_{X(t_2)=-1} + Q_t^{1,-1} 1_{X(t)=1} 1_{X(t_2)=-1}
\]

(4)
and let $\theta = (q, \mu, \sigma)$ be the parameter of the model at $t$ fixed. It comes, the likelihood of the triplet $(Y, \X(t_1), \X(t_2))$ with respect to the measure $\lambda \otimes N \otimes N$, $\lambda$ being the Lebesgue measure, $N$ the counting measure on $\mathbb{N}$, is $\forall t \in [t_1, t_2]$

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

where the function

\[
g(t) = \frac{1}{2} \left\{ \bar{r}(t_1, t_2) 1_{\X(t_1)=1} 1_{\X(t_2)=1} + r(t_1, t_2) 1_{\X(t_1)=1} 1_{\X(t_2)=0} + \bar{r}(t_1, t_2) 1_{\X(t_1)=-1} 1_{\X(t_2)=-1} \right\}
\]

\[
+ \frac{1}{2} \left\{ r(t_1, t_2) 1_{\X(t_1)=-1} 1_{\X(t_2)=1} + \bar{r}(t_1, t_2) 1_{\X(t_1)=-1} 1_{\X(t_2)=0} \right\}
\]

\[
+ 1_{\X(t_1)=0} 1_{\X(t_2)=0}
\]

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).

\textbf{Notation 2.3}: $\gamma, \gamma_+$ and $\gamma_-$ are respectively the quantities $\mathbb{P}_{H_0}(Y \notin [S_- S_+]), \mathbb{P}_{H_0}(Y > S_+)$ and $\mathbb{P}_{H_0}(Y < S_-)$.

\textbf{Notation 2.4}: $A$ is the quantity such as $A = \sigma^2 \left\{ \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) \right\}$, where $\varphi(x)$ and $z_\alpha$ 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

\textbf{Theorem 2.5}: 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

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

With the previous defined notations,

\[
S_n(.) \Rightarrow V(.) , \quad \Lambda_n(.) \xrightarrow{F.d} V^2(.) , \quad \sup \Lambda_n(.) \xrightarrow{L} \sup V^2(.)\]

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

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

\[
V(t) = \frac{\alpha(t) V(t_1) + \beta(t) V(t_2)}{\sqrt{\{\alpha(t) V(t_1) + \beta(t) V(t_2)\}}}
\]
where

\[
\text{Cov} \{V(t_1), V(t_2)\} = \rho(t_1, t_2) = \exp(-2|t_1 - t_2|)
\]

\[
\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}, \quad \beta(t) = Q_t^{1,1} - Q_t^{1,-1}
\]

and with expectation:

- under \(H_0\), \(m(t) = 0\),
- under \(H_{alt}\),

\[
m_{t^*}(t) = \frac{\alpha(t) m_{t^*}(t_1) + \beta(t) m_{t^*}(t_2)}{\sqrt{V\{\alpha(t)V(t_1) + \beta(t)V(t_2)\}}}.
\]

In the sense of this equation, \(V(.)\) will be called a "non linear normalized interpolated process". We can see that under the null hypothesis, despite the selective genotyping, \(V(.)\) is exactly the same process as the process \(Z(.)\) of Theorem 2.1 of Azaïs et al. [10] 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 \(\sqrt{A}/\sigma\). Note also that \(V(.)\) is the generalization of \(Z(.)\). Indeed, if we choose \(S_- = S_+\), that is to say we genotype all the individuals, the factor \(\sqrt{A}/\sigma\) is equal to 1, and \(V(.)\) is the same process as \(Z(.)\).

**Proof: Theorem 2.5**

**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

\[
\frac{\partial l_t}{\partial q} \bigg|_{\theta_0} = \frac{y - \mu}{\sigma^2} \{2p(t) - 1\} \mathbf{1}_{y \notin [S_-, S_+]} \tag{7}
\]

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

Then,

\[
E_{H_0} \left\{ \left( \frac{\partial l_t}{\partial q} \bigg|_{\theta_0} \right)^2 \right\} = E_{H_0} \left\{ \left( \frac{y - \mu}{\sigma^2} \right)^2 \{2p(t) - 1\}^2 \mathbf{1}_{y \notin [S_-, S_+]} \right\} \tag{8}
\]

Let’s introduce two key lemmas:

**Lemma 2.6:** We have the following relationship:

\[
\{2p(t) - 1\} \mathbf{1}_{y \notin [S_-, S_+]} = \alpha(t) \overline{X}(t_1) + \beta(t) \overline{X}(t_2)
\]
\[ \alpha(t) = Q_t^{1,1} - Q_t^{-1,1} \text{ and } \beta(t) = Q_t^{1,1} - Q_t^{-1,1} \].

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

**Lemma 2.7:** Let \( W \sim N(\mu, \sigma^2) \), then:

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

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

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

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

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

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

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

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

To prove this lemma, use integration by parts.

According to iii) of Lemma 2.7, we have \( \mathcal{A} = \mathbb{E}_{H_0} \{ (y - \mu)^2 1_{y \notin [S_-, S_+]} \} \). It comes, according to Lemma 2.6:

\[
\mathbb{E}_{H_0} \left\{ \left( \frac{\partial l_t}{\partial q} \right)_{\theta_0} \right\}^2 \quad (9)
\]

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

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

\[
= \mathbb{E}_{H_0} \left\{ \left( \frac{y - \mu}{\sigma^2} \right)^2 1_{y \notin [S_-, S_+]} \right\} \mathbb{E}_{H_0} \left\{ \{ \alpha(t)X(t_1) + \beta(t)X(t_2) \}^2 \right\} / \sigma^4.
\]

To conclude, after some calculations, we find

\[
I_{\theta_0} = \text{Diag} \left[ \mathcal{A} \left\{ \alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)e^{-2(t_2-t_1)} \right\} / \sigma^4, \frac{1}{\sigma^2}, \frac{2}{\sigma^2} \right]. \quad (10)
\]

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

\[
\frac{\partial l_t}{\partial q\partial \mu} \bigg|_{\theta_0} = \frac{2\mu(t) - 1}{\sigma^2} 1_{y \notin [S_-, S_+]}.
\]
It comes, using Lemma 2.6,

$$
\mathbb{E}_{H_0} \left\{ -\frac{\partial l_t}{\partial q \partial \mu} \mid \theta_0 \right\} = -\frac{1}{\sigma^2} \mathbb{E}_{H_0} \left[ \alpha(t) \overline{X}(t_1) + \beta(t) \overline{X}(t_2) \right]
$$

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

$$
= -\frac{1}{\sigma^2} \mathbb{E}_{H_0} \left\{ \alpha(t) X(t_1) + \beta(t) X(t_2) \right\} \mathbb{P}_{H_0}(y \notin [S_-, S_+]) = 0.
$$

Besides,

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

It comes

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

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 of the hypothesis \( q = 0 \) will be defined as

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

Now using formula (7) and using Lemma 2.6, it is clear that

$$
\frac{\partial l^n}{\partial q} \mid \theta_0 = \sum_{j=1}^{n} \frac{y_j - \mu}{\sigma^2} \left\{ 2p_j(t) - 1 \right\} 1_{y_j \in [S_-, S_+]} - \frac{\alpha(t)}{\sigma} \sum_{j=1}^{n} \frac{\varepsilon_j \overline{X}_j(t_1)}{\sigma} + \frac{\beta(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j \overline{X}_j(t_2)
$$

(12)

this proves that \( V(.) \) is a non linear interpolated process.

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

$$
S_n(t_k) = \sum_{j=1}^{n} \frac{\sigma \varepsilon_j \overline{X}_j(t_k)}{\sqrt{nA}}.
$$
We have:

\[
E \left\{ \sigma \varepsilon X(t_k) \right\} = E \left( \sigma \varepsilon 1_{y \in [S-, S_+]} \mid X(t_k) = 1 \right) P \left\{ X(t_k) = 1 \right\}
- E \left( \sigma \varepsilon 1_{y \in [S-, S_+]} \mid X(t_k) = -1 \right) P \left\{ X(t_k) = -1 \right\}
= E \left( \sigma \varepsilon 1_{y \in [S-, S_+]} \right) / 2 - E \left( \sigma \varepsilon 1_{y \in [S-, S_+]} \right) / 2
= 0.
\]

Besides:

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

According to the Central Limit Theorem, it comes

\[ S_n(t_k) \overset{L}{\rightarrow} N(0, 1). \]

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

\[
E \left\{ S_n(t_1)S_n(t_2) \right\} = \frac{1}{A} E \left\{ (y - \mu)^2 X(t_1) X(t_2) 1_{y \in [S-, S_+]} \right\}
= \frac{1}{A} E \left\{ (y - \mu)^2 1_{y \in [S-, S_+]} \right\} E \{ X(t_1)X(t_2) \} = e^{-2(t_2 - t_1)}.
\]

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, \( S_n(\cdot) \), is then a direct consequence of (12), 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 \( S_n(\cdot) \) under this alternative. Since we have already proved that \( S_n(\cdot) \) is a non linear interpolated process (see Lemma 2.6), 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

\[
S_n(t_k) = \sum_{j=1}^{n} \frac{(y_j - \mu) X_j(t_k)}{\sqrt{nA}}
= \sum_{j=1}^{n} \frac{qX_j(t^*) X_j(t_k)}{\sqrt{nA}} + \sum_{j=1}^{n} \frac{\sigma\varepsilon_j X_j(t_k)}{\sqrt{nA}}.
\]

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
\[
\mathbb{E} \left\{ \sigma \varepsilon \bar{X}(t_k) \right\} = \frac{1}{2} \mathbb{E} \left\{ \sigma \varepsilon \bar{X}(t_k) \mid X(t^*) = 1 \right\} + \frac{1}{2} \mathbb{E} \left\{ \sigma \varepsilon \bar{X}(t_k) \mid X(t^*) = -1 \right\}.
\]
We have
\[
\mathbb{E} \left\{ \sigma \varepsilon \bar{X}(t_k) \mid X(t^*) = 1 \right\} = \mathbb{E} \left\{ \sigma \varepsilon 1_{y \notin [S_{-},S_{+}]} \mid X(t^*) = 1 \right\} \bar{r}(t_k,t^*) - \mathbb{E} \left\{ \sigma \varepsilon 1_{y \notin [S_{-},S_{+}]} \mid X(t^*) = 1 \right\} r(t_k,t^*)
\]

Besides, according to iv) of Lemma 2.7,
\[
\mathbb{E} \left\{ \sigma \varepsilon 1_{y \notin [S_{-},S_{+}]} \mid X(t^*) = 1 \right\} = \sigma \varphi \left( \frac{S_{+} - \mu - q}{\sigma} \right) - \sigma \varphi \left( \frac{S_{-} - \mu - q}{\sigma} \right).
\]

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

In the same way, after some calculations, we obtain
\[
\mathbb{E} \left\{ \sigma \varepsilon \bar{X}(t_k) \mid X(t^*) = -1 \right\} = -e^{-2|t_k-t^*|} \sigma \left\{ \varphi \left( \frac{S_{+} - \mu + q}{\sigma} \right) - \varphi \left( \frac{S_{-} - \mu + q}{\sigma} \right) \right\}.
\]

Since we consider \( q \) small, using a Taylor expansion at first order, we obtain for instance :
\[
\varphi \left( \frac{S_{-} - \mu + q}{\sigma} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{S_{-} - \mu - q}{\sigma} \right)^2} \left\{ 1 - \frac{(S_{-} - \mu)q}{\sigma^2} + o(q) \right\}.
\]

Finally, using Taylor expansions in formulae (14) and (15), we have :
\[
\mathbb{E} \left\{ \sigma \varepsilon \bar{X}(t_k) \right\} = e^{-2|t_k-t^*|} q \left\{ z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma} \varphi(z_{1-\gamma}) \right\} + o(q).
\]

It comes
\[
\mathbb{E} \left\{ \sum_{j=1}^{n} \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n} A} \right\} \to \frac{e^{-2|t_k-t^*|}}{\sqrt{A}} a \left\{ z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma} \varphi(z_{1-\gamma}) \right\}.
\]

We have now just to remark that
\[
\mathbb{E} \left( \left\{ \sigma \varepsilon \bar{X}(t_k) \right\}^2 \right) = \mathbb{E} \left\{ \sigma^2 \varepsilon^2 1_{y \notin [S_{-},S_{+}]} \right\}
\]

\[
= \mathbb{E} \left\{ \sigma^2 \varepsilon^2 1_{y \notin [S_{-},S_{+}]} \mid X(t^*) = 1 \right\} / 2 + \mathbb{E} \left\{ \sigma^2 \varepsilon^2 1_{y \notin [S_{-},S_{+}]} \mid X(t^*) = -1 \right\} / 2
\]

\[
\to A/2 + A/2 \to A.
\]
It comes
\[ \forall \left\{ \sum_{j=1}^{n} \frac{\sigma \varepsilon_j X_j(t_k)}{\sqrt{nA}} \right\} \to 1 , \]
and according to the Central Limit Theorem
\[ \sum_{j=1}^{n} \frac{\sigma \varepsilon_j X_j(t_k)}{\sqrt{nA}} \overset{\mathcal{L}}{\to} N \left[ \frac{e^{-2|\mu_k - t'|}}{\sqrt{A}} a \left\{ z_{1-\gamma} \varphi(z_{1-\gamma}) - z_{1-\gamma} \varphi(z_{1-\gamma}) \right\} , 1 \right] . \] (17)

Besides,
\[ \mathbb{E} \{ X(t^*) X(t_k) \} \]
\[ = \frac{1}{2} \mathbb{P}_{t^*} \{ 1 | 1 \} \mathbb{E} \{ X(t_k) | X(t^*) = 1 \} - \frac{1}{2} \mathbb{P}_{t^*} \{ -1 | -1 \} \mathbb{E} \{ X(t_k) | X(t^*) = -1 \} \]
\[ = \frac{1}{2} e^{-2|\mu_k - t'|} \left( \mathbb{P}_{t^*} \{ 1 | 1 \} + \mathbb{P}_{t^*} \{ -1 | -1 \} \right) . \]

Using Taylor expansion and after some work on integrals, we have:
\[ P_{t^*} \{ 1 | 1 \} = \Phi \left( \frac{S_- - \mu}{\sigma} \right) - \frac{q}{\sigma^2} \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) + \frac{q}{\sigma^2} \left( \frac{S_+ - \mu}{\sigma} \right) + o(q) \]
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 \( P_{t^*} \{ -1 | -1 \} \).

It comes
\[ \mathbb{E} \{ X(t^*) X(t_k) \} = e^{-2|\mu_k - t'|} \left\{ 1 + \Phi \left( \frac{S_- - \mu}{\sigma} \right) + \Phi \left( \frac{S_+ - \mu}{\sigma} \right) \right\} + o(q) \]
\[ = e^{-2|\mu_k - t'|} \gamma + o(q) . \]

As a consequence, according to the Law of Large Numbers,
\[ \sum_{j=1}^{n} \frac{qX_j(t^*) X_j(t_k)}{\sqrt{nA}} \overset{\mathcal{L}}{\to} a \frac{e^{-2|\mu_k - t'|}}{\sqrt{A}} \gamma . \] (19)

Finally, using formulae (17) and (19), we obtain
\[ S_n(t_k) \overset{\mathcal{L}}{\to} N \left( \frac{a \sqrt{A}}{\sigma^2} e^{-2|\mu_k - t'|} , 1 \right) , \] (20)
which concludes the proof.

**Study of the supremum of the LRT process**

Let \( l_n(\theta) \) be the log likelihood for \( n \) observations log likelihood. Let \( l_n^*(\hat{\theta}) \) be the maximized log likelihood and let \( l_n^*(\hat{\theta}_{H_0}) \) be the maximized log likelihood under \( H_0 \), with \( \hat{\theta}_{H_0} = (0, \bar{Y} = \sum Y_j/n, 1/n \sum (Y_j - \bar{Y})^2) \) (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. 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. [21] holds. So, we can apply Theorem 1 of Azaïs et al. [21] and we have

$$
\sup_{(t, \theta)} l_t(\theta) - l^*_t(\theta^*) = \sup_{d \in D} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right\}^2 1_{d(X_j) \geq 0} + o_P(1)
$$
(21)

where the observation $X_j$ stands for $Y_j, X_j(t_1), X_j(t_2)$ and where $D$ is the set of scores defined in Azaïs et al. [21], see also Gassiat [22]. 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

$$
D = \left\{ \frac{\langle U, l'_t(\theta^*) \rangle}{\sqrt{V(\langle U, l'_t(\theta^*) \rangle)}}, U \in \mathbb{R}^3, t \in [t_1, t_2] \right\}
$$
where $l'$ is the gradient with respect to $\theta$. In the same manner

$$
D_0 = \left\{ \frac{\langle U, l'_t(\theta^*) \rangle}{\sqrt{V(\langle U, l'_t(\theta^*) \rangle)}}, U \in \mathbb{R}^2 \right\},
$$
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 $U \rightarrow -U$ in the expressions of the sets of score, we see that the indicator function can be removed in formula (21). Then, since the Fisher information matrix is diagonal (see formula (10)) , it is easy to see that

$$
= \sup_{t \in [t_1, t_2]} \left( \frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{\partial l_t}{\partial \theta}(X_j) \bigg|_{\theta_0} \right)^2.
$$
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 (21) remains true under the alternative. \( \square \)

**Remark 1:** According to the Law of Large Numbers, under the null hypothesis \( H_0 \) and under the local alternative \( H_{\alpha^*} \), \( \frac{1}{n} \sum_{j=1}^{n} 1_{y_j \in [S_+, S_-]} \to \gamma \). So, \( \gamma \) corresponds asymptotically to the percentage of individuals genotyped. In the same way, \( \gamma_+ \) (resp. \( \gamma_- \)) 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 \( V(.) \) is a “non linear normalized interpolated process”, we can use Lemma 2.2 of Azaïs et al. [10] in order to compute easily the supremum of \( V^2(.) \). Note that this lemma is suitable here because we have exactly the same interpolation as in Theorem 2.1 of Azaïs et al. [10]. It comes

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

Note that since under \( H_0 \), the process \( V(.) \) is exactly the same process as the process \( Z(.) \) obtained by Azaïs et al. [10], we will have exactly the same threshold if we are under selective genotyping or not. So, the Monte-Carlo Quasi Monte-Carlo method of Azaïs et al. [10] and based on Genz [18], 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 [15], 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

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

We introduce the following lemma.

**Lemma 3.1:** Let \( T_n(.) \) be the process such as

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

Then, for the data analysis, we just have to consider as a test statistic \( \sup T^2_n(.) \), which can be obtained easily using formula (22) and replacing \( V(t_1) \) and \( V(t_2) \) by respectively \( T_n(t_1) \) and \( T_n(t_2) \). Note that, according to Lemma 3.1, this test has
the same asymptotic properties as the test based on the test statistic sup \( \Lambda_n(\cdot) \), which corresponds to a LRT on the whole chromosome. So, Lemma 3.1 is an answer to the work of Rabbee et al. [16] where the authors study different strategies for analyzing data in selective genotyping.

On the other hand, a consequence of Lemma 3.1 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 [15], where I have 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: Lemma 3.1**

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

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

To begin, in order to make the proof easier, let’s consider that we are under \( H_0 \).

Since \( \overline{Y} = \mu + O_P(1/\sqrt{n}) \), we have

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

Let’s focus on the second term under \( H_0 \). We have

\[
\mathbb{E}[\overline{X}(t_k)] = \mathbb{E}[X(t_k) | Y \notin [S_-, S_+]] \mathbb{P}(Y \notin [S_-, S_+]) = \mathbb{E}[X(t_k)] \gamma = 0.
\]

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

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

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

We have the relationship \( (T_n(t_1), T_n(t_2)) = \sqrt{\mathbf{A} / \mathbf{A}} (\tilde{T}_n(t_1), \tilde{T}_n(t_2)) \). Since \( \mathbf{A} \xrightarrow{\mathcal{L}} \mathbf{A} \) whatever the hypothesis, according to Slutsky and then Continuous Mapping theorem, we have \( \sqrt{\mathbf{A} / \mathbf{A}} \xrightarrow{\mathcal{L}} 1 \). Using Slutsky, it comes \( (T_n(t_1), T_n(t_2)) \xrightarrow{\mathcal{L}} (V(t_1), V(t_2)) \). To conclude the proof, we just have to use the Continuous Mapping Theorem: \( T_n(.) \Rightarrow V(.) \) and obviously \( T_n^2(.) \Rightarrow V^2(.) \). \( \square \)
4. Several markers: the “Interval Mapping” of Lander and Botstein [12]
under selective genotyping

In that case suppose that there are $K$ markers $0 = t_1 < t_2 < \ldots < t_K = T$.
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 $t \in [t_1, t_K] \setminus \mathbb{T}_K$ where $\mathbb{T}_K = \{t_1, \ldots, t_K\}$, we define $t^l$ and $t^r$ as:

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

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

**Theorem 4.1:** We have the same result as in Theorem 2.5, provided that we
make some adjustments and that we redefine $V(\cdot)$ in the following way:

- in the definition of $\alpha(t)$ and $\beta(t)$, $t_1$ becomes $t^l$ and $t_2$ becomes $t^r$.
- under the null hypothesis, the process $V(\cdot)$ considered at marker positions is the
  “skeleton” of an Ornstein-Uhlenbeck process: the stationary Gaussian process
  with covariance $r(t_k, t_{k'}) = \exp(-2|t_k - t_{k'}|)$.
- at the other positions, $V(\cdot)$ is obtained from $V(t^l)$ and $V(t^r)$ by interpolation
  and normalization using the functions $\alpha(t)$ and $\beta(t)$.
- at the marker positions, the expectation is such as $m_{t^l}(t_k) = \frac{\sqrt{A} r(t_k, t^*)}{\sigma}$
- at other positions, the expectation is obtained from $m_{t^l}(t^l)$ and $m_{t^r}(t^r)$ by inter-
  polation and normalization using the functions $\alpha(t)$ and $\beta(t)$.

The proof of the theorem is the same as the proof of Theorem 2.5 since for a
position $t$, we can limit our attention to the interval $(t^l, t^r)$. 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 $(t^l, t^r)$, we can still use the section
“Study under the alternative” of the proof of Theorem 2.5.

Another important point is that since for a position $t$ we can limit our attention
to the interval $(t^l, t^r)$, Lemma 3.1 and formula (22) are still true here. We just
have to replace $t_1$ and $t_2$ by $t^l$ and $t^r$ in order to have the good expressions. As a
consequence, we can easily compute $\sup T^2_{\alpha}(\cdot)$.

We introduce now our Theorem 4.2.

**Theorem 4.2:** Let $\kappa$ be the Asymptotic Relative Efficiency (ARE) with respect
to the oracle situation where all the genotypes are known. Then, we have

$$
\text{i) } \kappa = \gamma + z_{\gamma+} \varphi(z_{\gamma+}) - z_{1-\gamma} \varphi(z_{1-\gamma}) \\
\text{ii) } \kappa \text{ reaches its maximum for } \gamma_+ = \gamma_- = \gamma/2.
$$

According to i) of Theorem 4.2, the ARE with respect to the oracle situation,
does not depend on the constant $a$ linked to the QTL effect, and does not depend on
the location of the QTL $t^*$. 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
$\sqrt{A}/\sigma$, 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 4.2,
if we want to genotype only a percentage $\gamma$ 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 [15] where I prove that we have to genotype symmetrically when we look for a QTL on only one genetic marker.

**Proof:** The proof of i) is obvious since the mean functions of the selective genotyping and the oracle situation, are proportional of a factor $\sqrt{\lambda}/\sigma$. 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 $\gamma_+$ and $\gamma_-$ 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 : $u(z_{\gamma_+}) = \Phi^{-1}\{\gamma - 1 + \Phi(z_{\gamma_+})\}$. 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 maximize the function $k_1(.)$. Let $k_1'(.), u'(.)$ and $\varphi'(.)$ be respectively the derivative of $k_1(.)$, $u(.)$ and $\varphi(.)$. We have :

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

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

Then, $k_1'(z_{\gamma/2}) = \varphi(z_{\gamma/2}) - \{z_{\gamma/2}\}^2 \varphi(z_{\gamma/2}) - \varphi(z_{1-\gamma/2}) + \{z_{1-\gamma/2}\}^2 \varphi(z_{1-\gamma/2}) = 0$. As a result, the efficiency $\kappa$ reaches its maximum when $\gamma_+ = \gamma_- = \frac{\gamma}{2}$. □

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

$$h_n(t_k, t_{k+1}) = \frac{T_n^2(t_k) + T_n^2(t_{k+1}) - 2\rho(t_k, t_{k+1})T_n(t_k)T_n(t_{k+1})}{1 - \rho^2(t_k, t_{k+1})} \frac{1}{\text{max}\{\rho(t_k, t_{k+1}) \cap |\rho(t_k, t_{k+1})|\}},$$

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

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

Our first result is that the threshold (i.e. critical value) is the same if we are under selective genotyping or in the oracle situation. So, the Monte-Carlo Quasi Monte-Carlo method, proposed by Azaïs et al. [10] (based on Genz [18]) for the oracle situation, is still suitable here to obtain our threshold. This way, in Figure 1, we propose to check these asymptotic results on simulated data. We consider a chromosome of length $T = 1\text{M}$, with two genetic markers located at each extremity. For such a configuration, if we choose a level 5%, the corresponding threshold is 5.40. We consider here $\gamma = 0.3$, 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 $\gamma_+$, the Percentage of False Positives is close to the true level of the test (i.e. 5%) even for small values of $n$ (see $n = 50$). Note that our method to compute thresholds in an alternative to the permutation method proposed by Manichaikul et al. [17] and inspired by Churchill and Doerge [19]. The permutation method
Table 1. Percentage of False Positives as a function of $n$ and the percentage $\gamma_+$ of individuals genotyped in the right tail. The chromosome is of length $T = 1M$ and two markers are located at each extremity ($\gamma = 0.3$, $a = 0$, $\mu = 0$, $\sigma = 1$, 10000 samples of size $n$).

<table>
<thead>
<tr>
<th>$\gamma_+$</th>
<th>$n$</th>
<th>1000</th>
<th>200</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td></td>
<td>4.98%</td>
<td>4.96%</td>
<td>4.50%</td>
</tr>
<tr>
<td>$\gamma/2$</td>
<td></td>
<td>5.27%</td>
<td>4.89%</td>
<td>4.65%</td>
</tr>
<tr>
<td>$\gamma/4$</td>
<td></td>
<td>4.79%</td>
<td>4.91%</td>
<td>4.43%</td>
</tr>
<tr>
<td>$\gamma/8$</td>
<td></td>
<td>5.21%</td>
<td>4.99%</td>
<td>4.58%</td>
</tr>
</tbody>
</table>

Table 2. Theoretical Power and Empirical Power (in brackets) as a function of the location of the QTL $t^*$ and the percentage $\gamma_+$ of individuals genotyped in the right tail. The chromosome is of length $T = 1M$ and two markers are located at each extremity ($\gamma = 0.3$, $a = 4$, $\mu = 0$, $\sigma = 1$, 10000 samples of size $n = 1000$, 100000 paths for the Theoretical Power).

<table>
<thead>
<tr>
<th>$\gamma_+$</th>
<th>$t^*$</th>
<th>10cM</th>
<th>30cM</th>
<th>60cM</th>
<th>80cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td></td>
<td>53.72%</td>
<td>30.70%</td>
<td>25.59%</td>
<td>39.88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(53.84%)</td>
<td>(30.02%)</td>
<td>(25.88%)</td>
<td>(39.10%)</td>
</tr>
<tr>
<td>$\gamma/2$</td>
<td></td>
<td>76.22%</td>
<td>46.64%</td>
<td>38.91%</td>
<td>59.82%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(75.98%)</td>
<td>(46.15%)</td>
<td>(38.30%)</td>
<td>(59.07%)</td>
</tr>
<tr>
<td>$\gamma/4$</td>
<td></td>
<td>72.71%</td>
<td>43.80%</td>
<td>36.42%</td>
<td>56.53%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(72.41%)</td>
<td>(43.51%)</td>
<td>(35.72%)</td>
<td>(56.02%)</td>
</tr>
<tr>
<td>$\gamma/8$</td>
<td></td>
<td>67.95%</td>
<td>40.15%</td>
<td>33.22%</td>
<td>52.16%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(67.56%)</td>
<td>(39.65%)</td>
<td>(33.77%)</td>
<td>(51.44%)</td>
</tr>
</tbody>
</table>

is very time consuming and not easy to compute because of the missing genotypes. The advantage of our 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).

In Figures 2 and 3, we focus on the alternative hypothesis. In Figure 2, we consider the same 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 $\gamma_+$. As expected (c.f. Theorem 4.2), 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 3, we focus on a more dense genetic map (6 genetic markers), and we change the value of $\gamma$: $\gamma = 0.6$. 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 under selective genotyping. That’s why it must be interesting for geneticists.

Acknowledgements

I thank Jean-Marc Azaïs and Céline Delmas for fruitful discussions.
Table 3. Theoretical Power and Empirical Power (in brackets) as a function of the location of the QTL $t^\star$ and the percentage $\gamma_+$ of individuals genotyped in the right tail. The chromosome is of length $T = 1M$ and 6 markers are equally spaced every 20cM ($\gamma = 0.6, a = 4, \mu = 0, \sigma = 1, 10000$ samples of size $n = 1000, 100000$ paths for the Theoretical Power).

<table>
<thead>
<tr>
<th>$\gamma_+$</th>
<th>$t^\star$</th>
<th>18cM</th>
<th>44cM</th>
<th>70cM</th>
<th>90cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>64.16%</td>
<td>62.45%</td>
<td>59.43%</td>
<td>58.05%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(63.34%)</td>
<td>(62.48%)</td>
<td>(58.86%)</td>
<td>(57.67%)</td>
<td></td>
</tr>
<tr>
<td>$\gamma/2$</td>
<td>91.57%</td>
<td>90.45%</td>
<td>87.87%</td>
<td>87.42%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(91.71%)</td>
<td>(89.47%)</td>
<td>(88.05%)</td>
<td>(87.25%)</td>
<td></td>
</tr>
<tr>
<td>$\gamma/4$</td>
<td>89.22%</td>
<td>87.84%</td>
<td>85.06%</td>
<td>84.35%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(88.82%)</td>
<td>(88.17%)</td>
<td>(84.92%)</td>
<td>(83.77%)</td>
<td></td>
</tr>
<tr>
<td>$\gamma/8$</td>
<td>84.19%</td>
<td>82.55%</td>
<td>79.66%</td>
<td>78.43%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(84.84%)</td>
<td>(82.09%)</td>
<td>(79.51%)</td>
<td>(78.55%)</td>
<td></td>
</tr>
</tbody>
</table>

References