On stochastic processes for Quantitative Trait Locus mapping under selective genotyping

Charles-Elie Rabier

Hal
Vol. 1, No. 2, 1–18
10.1080/0233-1888.YYYY.xxxxxx 1029-4910 0233-1888 00 00 2013 October

ISSN: print/ISSN online

DOI:
http://www.tandfonline.com

Hal
Vol. , No. , , 2–18

RESEARCH ARTICLE

On stochastic processes for Quantitative Trait Locus mapping under selective genotyping

Charles-Elie Rabier [abc*]

[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*; [c] *University of Wisconsin-Madison, Department of Statistics, 1300 University Avenue, Madison WI 53706, USA*

(*v3 October 2013*)

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 of this study is 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 $t^\star$ 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 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: $A \times (A \times B)$, where A and B are purely homozygous lines and we address the problem of detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured) on a given chromosome. The trait is observed on n individuals (progenies) and we denote by $Y_j,\ j = 1, ..., n$, these observations. 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. The Haldane modeling, which assumes no crossover interference, can be represented as follows: $X(0)$ is a random

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

sign and $X(t) = X(0)(-1)^{N(t)}$ where $N(.)$ is a standard Poisson process on $[0, T]$. Calculations on the Poisson distribution show that

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

we set in addition

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

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

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

where ε is standard Gaussian 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 $t^\star$ of the QTL is unknown.

The originality of this paper is that we consider the classical problem, but in order to reduce the costs of genotyping, a selective genotyping has been performed : we consider two real thresholds S_- and S_+, with $S_- \leqslant S_+$ and we genotype if and only if the phenotype Y is extreme, that is to say $Y \leqslant S_-$ or $Y \geqslant S_+$. Note that in practice, the cutoffs for genotyping are based on quantiles. However, in most of the theoretical studies about selective genotyping (e.g. Darvasi and Soller [13], Muranty and Goffinet [14]), authors consider fixed thresholds. This approximation is reasonable when we deal with a large number of observations.

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

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

Note that with our notations :

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

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

It can be proved that $\big(Y,\ \overline{X}(t_1),\ ...,\ \overline{X}(t_K)\big)$ obeys to a mixture model with known weights, times a function g(.) which does not depend on the parameters μ,

q and σ :

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

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

As mentioned 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 complete data 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".

In this paper, 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 focuses 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 complete data 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 complete data situation. Recall that the Asymptotic Relative Efficiency (ARE) determines the relative sample size required to obtain the same local asymptotic power as the one of the test under the complete data situation where all the genotypes are known. Note that we show that we have exactly the same ARE, 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 γ of the population, we should genotype symmetrically, that is to say the $\gamma/2\%$ individuals with the largest phenotypes and $\gamma/2\%$ individuals with the smallest phenotypes. This is a generalization of Rabier [15], where it is proved that

we have to genotype symmetrically, when we focus only on one genetic marker.

Furthermore, we propose a test statistic (see Lemma 3.1 and formula 15) asymptotically distributed as the LRT, but which presents computational advantages. Indeed, usually, in order to perform a LRT, we have to compute an EM algorithm at each location of the genome, which is quite challenging. In contrast, our test statistic does not require the use of any EM algorithm. 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 relevant to the previous work of Rabbee et al. [16]. However, we have to mention that the non-extreme phenotypes are useful for the estimation of the QTL effect.

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 complete data 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 stratified permutation method proposed by Manichaikul et al. [17] and inspired by Churchill and Doerge [19], which requires to permute the genotypes within the extremes. Our method is very fast since it relies on very powerful algorithms developed by Genz [18]. In contrast, permutation methods are usually time consuming since a large number of permutations has to be performed in order to obtain an accurate threshold.

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 $t^\star \in [t_1, t_2]$. As said before, since $t^\star$ 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 $t = t^\star$.

Notation 2.1: *For $(i, i') \in \{-1, 1\}^2$, $Q_t^{i,i'}$ is the quantity such as*

$$Q_t^{i,i'} = \mathrm{P}\left\{X(t) = 1 \middle| X(t_1) = i, X(t_2) = i'\right\} .$$

Using Bayes rules, we have

$$Q_t^{1,1} = \frac{\bar{r}(t_1,t)\,\bar{r}(t,t_2)}{\bar{r}(t_1,t_2)} \quad , \quad Q_t^{1,-1} = \frac{\bar{r}(t_1,t)\,r(t,t_2)}{r(t_1,t_2)} \tag{3}$$
$$Q_t^{-1,1} = \frac{r(t_1,t)\,\bar{r}(t,t_2)}{r(t_1,t_2)} \quad , \quad Q_t^{-1,-1} = \frac{r(t_1,t)\,r(t,t_2)}{\bar{r}(t_1,t_2)} .$$

We can remark that we have

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

Notation 2.2: $\mathrm{P}_t\{l \mid i\}$ *is the quantity such as* $\forall\, i \in \{-1, 1\}$ *and* $\forall\, l \in \{-1, 0, 1\}$

$$\mathrm{P}_t\{l \mid i\} = \mathrm{P}(\overline{X}(t) = l \mid X(t) = i) .$$

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

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

Besides,

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

and

$$\begin{aligned}&\mathrm{P}(\overline{X}(t)=i\ \cap\ \overline{X}(t_1)=1\ \cap\ \overline{X}(t_2)=1)\\ &=\mathrm{P}(\overline{X}(t)\neq 0\ \cap\ X(t)=i\ \cap\ X(t_1)=1\ \cap\ X(t_2)=1)\\ &=\mathrm{P}_t\{i \mid i\}\ \mathrm{P}(X(t)=i\ \cap\ X(t_1)=1\ \cap\ X(t_2)=1)\\ &=\frac{1}{2}\,\mathrm{P}_t\{1 \mid 1\}\ \overline{r}(t_1,t)\,\overline{r}(t,t_2)1_{i=1}\ +\frac{1}{2}\,\mathrm{P}_t\{-1 \mid -1\}\ r(t_1,t)\,r(t,t_2)1_{i=-1}\ .\end{aligned}$$

As a result, using formula (3),

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

In the same way, after some calculations, we find

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

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

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

Finally, when the genome information is missing at marker locations (i.e. the phe-

notype is not extreme), we find

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

Let's define the quantity $p(t)$ such as

$$\begin{aligned}p(t) = Q_t^{1,1}\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=1}\ +\ Q_t^{1,-1}\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=-1}\\ +\ Q_t^{-1,1}\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=1}\ +\ Q_t^{-1,-1}\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=-1}\end{aligned} \tag{4}$$

and let $\theta=(q,\ \mu,\ \sigma)$ be the parameter of the model at t fixed. As a consequence, the likelihood of the triplet $\left(Y,\ \overline{X}(t_1),\ \overline{X}(t_2)\right)$ with respect to the measure $\lambda\otimes N\otimes N$, λ being the Lebesgue measure, N the counting measure on N, is $\forall t\in[t_1,t_2]$:

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

where the function

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

can be removed because it does not depend on the parameters. Recall that $\forall k$, $1_{\overline{X}(t_k)\neq 0}=1_{Y\notin[S_-,S_+]}$ and note also that for $t=t^\star$, we find our formula (2) of the introduction where $p(t^\star)$ is described in formula (4).

Notation 2.3: *γ, γ_+ and γ_- are respectively the quantities* $\mathrm{P}_{H_0}\left(Y\notin[S_-,\ S_+]\right)$, $\mathrm{P}_{H_0}\left(Y>S_+\right)$ *and* $\mathrm{P}_{H_0}\left(Y<S_-\right)$.

Notation 2.4: *$\mathcal{A}$ is the quantity such as* $\mathcal{A}=\sigma^2\ \left\{\gamma\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\right\}$, *where $\varphi(x)$ and z_α denote respectively the density of a standard normal distribution taken at the point x, and the quantile of order $1-\alpha$ of a standard normal distribution.*

Before introducing our main theorem, let us define the score statistic and the LRT statistic at t. Since the Fisher Information matrix is diagonal (cf. proof of Theorem 2.5 below), the score statistic of the hypothesis "$q=0$" at t, for n independent observations, will be defined as

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

where $l_t^n(\theta)$ denotes the log likelihood at t, associated to n observations.

The LRT at t, for n independent observations, will be defined as

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

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

Our main result is the following :

Theorem 2.5 : *Suppose that the parameters (q, μ, σ^2) vary in a compact and that σ^2 is bounded away from zero. Let H_0 be the null hypothesis $q = 0$ and define the following local alternative*

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

With the previous defined notations,

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

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

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

$$V(t) = \frac{\alpha(t)V(t_1) + \beta(t)V(t_2)}{\sqrt{\mathrm{V}\left\{\alpha(t)V(t_1) + \beta(t)V(t_2)\right\}}} \tag{6}$$

where

$$\mathrm{Cov}\left\{\mathrm{V}(\mathrm{t}_1), \mathrm{V}(\mathrm{t}_2)\right\} = \rho(\mathrm{t}_1, \mathrm{t}_2) = \exp(-2|\mathrm{t}_1 - \mathrm{t}_2|)$$

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

and with expectation :

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

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

where

$$m_{t^\star}(t_1) = \frac{a\,\sqrt{\mathcal{A}}\,\rho(t_1, t^\star)}{\sigma^2} \quad , \quad m_{t^\star}(t_2) = \frac{a\,\sqrt{\mathcal{A}}\,\rho(t^\star, t_2)}{\sigma^2} \ .$$

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 complete data 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 $\sqrt{\mathcal{A}}/\sigma$. 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 $\sqrt{\mathcal{A}}/\sigma$ 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 θ_0 that belongs to H_0. The proof relies on two key lemmas.

Lemma 2.6: *We have the following relationship :*

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

with $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$ *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 $Y \notin [S_-, S_+]$.

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

$$\mathrm{E}\left\{(W-\mu)^2 1_{W \notin [S_-,\, S_+]}\right\} = \sigma^2\, \mathrm{P}(W \notin [S_-,\, S_+]) \;+\; \sigma\,(S_+ - \mu)\,\varphi\left(\tfrac{S_+ - \mu}{\sigma}\right) - \sigma\,(S_- - \mu)\,\varphi\left(\tfrac{S_- - \mu}{\sigma}\right) \;.$$

To prove this lemma, use integration by parts. A consequence of Lemma 2.7 is that we have the relationship $\mathcal{A} = \mathrm{E}_{H_0}\left\{(Y-\mu)^2 1_{Y \notin [S_-, S_+]}\right\}$. To conclude, after some easy calculations, we find that the Fisher information is diagonal :

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

Study of the score process under H_0

Using Lemma 2.6, it is clear that

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

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

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

$$S_n(t_k) = \frac{\frac{\partial l_{t_k}^n}{\partial q}\,|_{\theta_0}}{\sqrt{\mathrm{V}\left(\frac{\partial l_{t_k}^n}{\partial q}\,|_{\theta_0}\right)}} = \sum_{j=1}^{n} \frac{\sigma\varepsilon_j\, \overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}} \;.$$

Since $\frac{\partial l_t^n}{\partial q}\,|_{\theta_0}$ is centered under H_0, a direct application of the central limit theorem

implies that

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

Let's compute the covariance of the score statistics on markers, i.e. $\mathrm{Cov_{H_0}}\left\{\mathrm{S_n(t_1)}, \mathrm{S_n(t_2)}\right\}$. Since $\mathrm{E}_{H_0}\left\{(Y-\mu)^2 1_{Y\notin[S_-,S_+]}\right\} = \mathcal{A}$, we have :

$$\begin{aligned}\mathrm{E}_{H_0}\left\{S_n(t_1)S_n(t_2)\right\} &= \frac{1}{\mathcal{A}}\,\mathrm{E}_{H_0}\left\{(Y-\mu)^2\ X(t_1)\ X(t_2)\ 1_{Y\notin[S_-,S_+]}\right\}\\ &= \frac{1}{\mathcal{A}}\,\mathrm{E}_{H_0}\left\{(Y-\mu)^2 1_{Y\notin[S_-,S_+]}\right\}\mathrm{E}\left\{X(t_1)X(t_2)\right\} = \rho(t_1,t_2)\ .\end{aligned}$$

As a consequence, $\mathrm{Cov_{H_0}}\left\{\mathrm{S_n(t_1)}, \mathrm{S_n(t_2)}\right\} = \rho(\mathrm{t_1}, \mathrm{t_2})$. The weak convergence of the score process, $S_n(.)$, is then a direct consequence of (8), the convergence of $(S_n(t_1), S_n(t_2))$ and the Continuous Mapping Theorem.

Study under the local alternative

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

It remains to compute the asymptotic distribution of $S_n(.)$ under this alternative. Since we have already proved that $S_n(.)$ is a 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 recall that under H_0,

$$S_n(t_k) = \sum_{j=1}^{n} \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}} \quad , \quad S_n(t_k) \xrightarrow{\mathcal{L}} N(0,\ 1)\ . \tag{9}$$

Since our model is differentiable in quadratic mean, according to Theorem 7.2 of Van der Vaart [20], under H_0, the log likelihood ratio verifies

$$l^n_{t^\star}(\theta)\ -\ l^n_{t^\star}(\theta_0) = \frac{a}{\sqrt{n}}\frac{\partial l^n_{t^\star}}{\partial q}\ |_{\theta_0}\ -\ \frac{a^2}{2}\,\mathrm{E}_{H_0}\left\{\left(\frac{\partial l_{t^\star}}{\partial q}\ |_{\theta_0}\right)^2\right\}\ +\ o_P(1) \tag{10}$$

where $o_P(1)$ denotes a sequence which converges in probability to zero.

According to the central limit theorem and formula (7), under H_0

$$l^n_{t^\star}(\theta)\ -\ l^n_{t^\star}(\theta_0) \xrightarrow{\mathcal{L}} N(-\frac{1}{2}\vartheta^2,\ \vartheta^2)\ \text{ with } \vartheta^2 = a^2\mathcal{A}\ \left\{\alpha^2(t^\star) + \beta^2(t^\star) + 2\alpha(t^\star)\beta(t^\star)\rho(t_1,t_2)\right\}/\sigma^4\ . \tag{11}$$

As a consequence, conditions required to apply Lecam's third lemma are fulfilled (cf. formulae 9 and 11). Recall that Lecam's third lemma allows to obtain the asymptotic distribution of $S_n(t_k)$ under the local alternative, by computing the covariance beween the log likelihood ratio and $S_n(t_k)$ under the null hypothesis.

In order to compute this covariance easily, we need an explicit expression of the

log likelihood ratio. According to formulae (7), (8) and (10), under H_0,

$$\begin{aligned} & l^n_{t^\star}(\theta) \ - \ l^n_{t^\star}(\theta_0) \\ & = \frac{a}{\sigma\sqrt{n}} \left\{ \alpha(t^\star) \sum_{j=1}^{n} \varepsilon_j \, \overline{X}_j(t_1) + \beta(t^\star) \sum_{j=1}^{n} \varepsilon_j \, \overline{X}_j(t_2) \right\} \\ & - \frac{a^2}{2\sigma^4} \mathcal{A} \left\{ \alpha^2(t^\star) + \beta^2(t^\star) + 2\alpha(t^\star)\beta(t^\star)\rho(t_1,t_2) \right\} + o_P(1). \end{aligned} \tag{12}$$

First, let us focus on the score statistic at location t_1. Then, we have

$$\begin{aligned} \mathrm{Cov}_{\mathrm{H_0}} \left\{ \mathrm{S_n(t_1)}, \frac{\mathrm{a}\,\alpha(\mathrm{t}^\star)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}} \, \overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t}_1) \right\} &= \mathrm{Cov}_{\mathrm{H_0}} \left\{ \sum_{\mathrm{j}=1}^{\mathrm{n}} \frac{\sigma\varepsilon_{\mathrm{j}} \, \overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t}_1)}{\sqrt{\mathrm{n}\,\mathcal{A}}}, \frac{\mathrm{a}\,\alpha(\mathrm{t}^\star)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}} \, \overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t}_1) \right\} \\ &= \frac{a\,\alpha(t^\star)}{\sqrt{\mathcal{A}}} \mathrm{V}_{H_0} \left\{ \varepsilon \, \overline{X}(t_1) \right\} = \frac{a\,\alpha(t^\star)\,\sqrt{\mathcal{A}}}{\sigma^2} \, . \end{aligned}$$

In the same way,

$$\begin{aligned} & \mathrm{Cov}_{\mathrm{H_0}} \left\{ \mathrm{S_n(t_1)}, \frac{\mathrm{a}\,\beta(\mathrm{t}^\star)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}} \, \overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t}_2) \right\} = \frac{\mathrm{a}\,\beta(\mathrm{t}^\star)}{\sqrt{\mathcal{A}}} \, \mathrm{Cov}_{\mathrm{H_0}} \left\{ \varepsilon \, \overline{\mathrm{X}}(\mathrm{t}_1), \varepsilon \, \overline{\mathrm{X}}(\mathrm{t}_2) \right\} \\ & = \frac{a\,\beta(t^\star)}{\sigma^2\,\sqrt{\mathcal{A}}} \, \mathrm{E}_{H_0} \left\{ (Y-\mu)^2 \, X(t_1) \, X(t_2) \, 1_{Y \notin [S_-, S_+]} \right\} = \frac{a\,\beta(t^\star)}{\sigma^2\,\sqrt{\mathcal{A}}} \, \mathrm{E}_{H_0} \left\{ (Y-\mu)^2 \, 1_{Y \notin [S_-, S_+]} \right\} \mathrm{E} \left\{ X(t_1) \, X(t_2) \right\} \\ & = \frac{a\,\beta(t^\star)\,\sqrt{\mathcal{A}}\,\rho(t_1,t_2)}{\sigma^2} \, . \end{aligned} \tag{13}$$

As consequence, since $\alpha(t^\star) + \beta(t^\star)\rho(t_1,t_2) = \rho(t_1,t^\star)$,

$$\mathrm{Cov}_{\mathrm{H_0}} \left\{ \mathrm{S_n(t_1)}, \mathrm{l^n_{t^\star}}(\theta) \ - \ \mathrm{l^n_{t^\star}}(\theta_0) \right\} = \frac{\mathrm{a}\,\sqrt{\mathcal{A}}\,\rho(\mathrm{t}_1,\mathrm{t}^\star)}{\sigma^2} \, .$$

Using the same kind of proof, and the fact that $\alpha(t^\star)\rho(t_1,t_2) + \beta(t^\star) = \rho(t^\star,t_2)$, we obtain

$$\mathrm{Cov}_{\mathrm{H_0}} \left\{ \mathrm{S_n(t_2)}, \mathrm{l^n_{t^\star}}(\theta) \ - \ \mathrm{l^n_{t^\star}}(\theta_0) \right\} = \frac{\mathrm{a}\,\sqrt{\mathcal{A}}\,\rho(\mathrm{t}^\star,\mathrm{t}_2)}{\sigma^2} \, .$$

As a result, under the local alternative, according to Lecam's third lemma,

$$S_n(t_1) \xrightarrow{\mathcal{L}} N(\frac{a\,\sqrt{\mathcal{A}}\,\rho(t_1,t^\star)}{\sigma^2}, \, 1) \quad \text{and} \quad S_n(t_2) \xrightarrow{\mathcal{L}} N(\frac{a\,\sqrt{\mathcal{A}}\,\rho(t^\star,t_2)}{\sigma^2}, \, 1)$$

which concludes the proof.

Study of the supremum of the LRT process

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

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

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

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

It can be proved that assumptions 1, 2 and 3 of Azaïs et al. [21] holds. So, we can apply Theorem 1 of Azaïs et al. [21] and we have

$$\sup_{(t,\theta)} l_t^n(\theta) - l_{t^*}^n(\theta^*) = \sup_{d\in\mathcal{D}} \left[\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right\}^2 1_{d(X_j)\geqslant 0} \right] + o_P(1) \tag{14}$$

where the observation X_j stands for $Y_j, \overline{X}_j(t_1), \overline{X}_j(t_2)$ and where $\mathcal{D}$ is the set of scores defined in Azaïs et al. [21], see also Gassiat [22]. A similar result is true under H_0 with a set $\mathcal{D}_0$. Let us precise the sets of scores $\mathcal{D}$ and $\mathcal{D}_0$. This sets are defined at the sets of scores of one parameter families that converge to the true model p_{t^*,θ^*} and that are differentiable in quadratic mean.

It is easy to see that

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

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

$$\mathcal{D}_0 = \left\{ \frac{\langle U, l_t'(\theta^*)\rangle}{\sqrt{\mathrm{V}(\langle U, l_t'(\theta^*)\rangle)}}, U \in \mathrm{R}^2 \right\},$$

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

Using the transform $U \to -U$ in the expressions of the sets of score, we see that the indicator function can be removed in formula (14). Then, since the Fisher information matrix is diagonal (see formula (7)) , it is easy to see that

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

This is exactly the desired result. Since the model with t^* fixed is differentiable in quadratic mean, the alternative defines a contiguous sequence of alternatives. By Le Cam's first lemma, relation (14) remains true under the alternative. □

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

3. An easy way to perform the statistical test

Since $V(.)$ 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]. As a result

$$\max_{t\in[t_1,t_2]} \frac{\{\alpha(t)V(t_1)+\beta(t)V(t_2)\}^2}{\alpha^2(t)+\beta^2(t)+2\rho(t_1,t_2)\alpha(t)\beta(t)}$$
$$= \max\left(V^2(t_1)\ ,\ V^2(t_2)\ ,\ \frac{V^2(t_1)+V^2(t_2)-2\rho(t_1,t_2)V(t_1)V(t_2)}{1-\rho^2(t_1,t_2)} 1_{\frac{V(t_2)}{V(t_1)}\in\ ]\ \rho(t_1,t_2)\ ,\ \frac{1}{\rho(t_1,t_2)}\ [}\right)\ . \tag{15}$$

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

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

We introduce the following lemma.

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

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

Note that this lemma can easily be proved by contiguity and using Slutsky's lemma.

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 (15) 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 $\sup \Lambda_n(.)$, 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.

4. Several markers : the "Interval Mapping'' of Lander and Botstein [12] under selective genotyping

In that case suppose that there are K markers $0 = t_1 < t_2 < ... < t_K = T$. We consider values t, t' or $t^\star$ of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions. For $t \in [t_1, t_K] \backslash \mathrm{T}_K$ where $\mathrm{T}_K = \{t_1, ..., t_K\}$, we define t^ℓ and t^r as :

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

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 $V(.)$ 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 $V(.)$ considered at marker positions is the "squeleton" of an Ornstein-Uhlenbeck process: the stationary Gaussian process with covariance $\rho(t_k, t_{k'}) = \exp(-2|t_k - t_{k'}|)$*
- *at the other positions, $V(.)$ is obtained from $V(t^\ell)$ 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 $m_{t^\star}(t_k) = \frac{a\sqrt{\mathcal{A}}\,\rho(t_k,t^\star)}{\sigma^2}$*
- *at other positions, the expection is obtained from $m_{t^\star}(t^\ell)$ and $m_{t^\star}(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$.*

Proof :

Due to Haldane model with Poisson increments, for a position t, we can limit our attention to the interval (t^ℓ, t^r). As a result when $t^\star$ does belong to the marker interval (t^ℓ, t^r), the proof is the same as the proof of Theorem 2.5. On the other hand, when $t^\star$ does not belong to the marker interval (t^ℓ, t^r), some adjustments have to be done for computing the distribution of the test statistic under the local alternative. In particular, in order to obtain an explicit expression of the log likelihood ratio, we can still use formula (12) provided that we replace t_1 and t_2 by respectively $t^{\star\ell}$ and $t^{\star r}$. As a consequence, if we consider $t_k = t^\ell$, we have

$$\mathrm{Cov}_{\mathrm{H}_0}\left\{\mathrm{S_n(t_k)}, \frac{\mathrm{a}\,\alpha(\mathrm{t}^\star)}{\sigma\sqrt{\mathrm{n}}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\varepsilon_{\mathrm{j}}\,\overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t}^{\star\ell})\right\} = \mathrm{Cov}_{\mathrm{H}_0}\left\{\sum_{\mathrm{j}=1}^{\mathrm{n}}\frac{\sigma\varepsilon_{\mathrm{j}}\,\overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t_k})}{\sqrt{\mathrm{n}\,\mathcal{A}}}, \frac{\mathrm{a}\,\alpha(\mathrm{t}^\star)}{\sigma\sqrt{\mathrm{n}}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\varepsilon_{\mathrm{j}}\,\overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t}^{\star\ell})\right\}$$

$$= \frac{a\,\alpha(t^\star)}{\sigma^2\sqrt{\mathcal{A}}}\,\mathrm{E}_{H_0}\left\{(Y-\mu)^2\,X(t_k)\,X(t^{\star\ell})\,1_{Y\notin[S_-,S_+]}\right\} = \frac{a\,\alpha(t^\star)\,\sqrt{\mathcal{A}}\,\rho(t_k,t^{\star\ell})}{\sigma^2} .$$

In the same way,

$$\mathrm{Cov}_{\mathrm{H}_0}\left\{\mathrm{S_n(t_k)}, \frac{\mathrm{a}\,\beta(\mathrm{t}^\star)}{\sigma\sqrt{\mathrm{n}}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\varepsilon_{\mathrm{j}}\,\overline{\mathrm{X}}_{\mathrm{j}}(\mathrm{t}^{\star\mathrm{r}})\right\} = \frac{\mathrm{a}\,\beta(\mathrm{t}^\star)\,\sqrt{\mathcal{A}}\,\rho(\mathrm{t_k},\mathrm{t}^{\star\mathrm{r}})}{\sigma^2} .$$

Since $\alpha(t^\star)\rho(t_k,t^{\star\ell}) + \beta(t^\star)\rho(t_k,t^{\star r}) = \rho(t_k,t^\star)$ and according to Lecam's third lemma, we have under the local alternative

$$S_n(t_k) \xrightarrow{\mathcal{L}} N(\frac{a\sqrt{\mathcal{A}}\,\rho(t_k,t^\star)}{\sigma^2},\ 1)\ .$$

□

An 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 (15) 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 $\sup T_n^2(.)$.

Before introducing our Theorem 4.2, let us recall that the Asymptotic Relative Efficiency (ARE) determines the relative sample size required to obtain the same local asymptotic power as the one of the test under the complete data situation where all the genotypes are known.

Theorem 4.2 : *Let κ denote the ARE, then we have*

i) $\kappa = \gamma\ +\ z_{\gamma_+}\,\varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-})$

ii) κ *reaches its maximum for* $\gamma_+ = \gamma_- = \gamma/2$.

According to i) of Theorem 4.2, the ARE with respect to the complete data situation, does not depend on the constant a linked to the QTL effect, and does not depend on the location of the QTL $t^\star$. Besides, we can remark that we have exactly the same ARE with respect to the complete data 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 (complete data situation and selective genotyping) are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$, it is obvious that the ARE will be the same if we scan the chromosome or if we focus only on one locus. On the other hand, according to ii) of Theorem 4.2, if we want to genotype only a percentage γ of the population, we should genotype the $\gamma/2\%$ individuals with the largest phenotypes and $\gamma/2\%$ individuals with the smallest phenotypes. This confirms theoretically 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 complete data situation, are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$. Let's now prove that the maximum is reached for $\gamma_+ = \gamma_- = \gamma/2$. We have to answer the following question : how must we choose γ_+ and γ_- to maximize the efficiency ? We remind that $\gamma_+ + \gamma_- = \gamma$ and that $\varphi(.)$ and $\Phi(.)$ denote respectively the density and the cumulative distribution of the standard normal distribution. Let $u(.)$ be the function such as : $u(z_{\gamma_+}) = \Phi^{-1}\left\{\gamma - 1 + \Phi(z_{\gamma_+})\right\}$. Then, $z_{1-\gamma_-} = u(z_{\gamma_+})$.

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

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

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

Then, $k_1'(z_{\gamma/2}) = \varphi(z_{\gamma/2}) - \left\{z_{\gamma/2}\right\}^2 \varphi(z_{\gamma/2}) - \varphi(z_{1-\gamma/2}) + \left\{z_{1-\gamma/2}\right\}^2 \varphi(z_{1-\gamma/2}) = 0$. As a result, the efficiency κ 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

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

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

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

Our first result is that the threshold (i.e. critical value) is the same if we are under selective genotyping or in the complete data situation. So, the Monte-Carlo Quasi Monte-Carlo method, proposed by Azaïs et al. [10] (based on Genz [18]) for the complete data 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 $T = 1$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 γ_+, 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 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 γ as previously. For the QTL effect q, we consider $a = 4$: we remind that $q = a/\sqrt{n}$. We focus on different locations $t^\star$ of the QTL and different values of γ_+. As expected (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 γ : $\gamma = 0.6$. 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 with, we have proved in this study that the LRT process is asymptotically the same under the null hypothesis, whether selective genotyping was performed or not. However, under the alternative, the mean functions are not the same anymore. Finally, we have introduced a test statistic asymptotically similarly distributed as the LRT, and which presents a computational advantage.

Table 1. Percentage of False Positives as a function of n and the percentage γ_+ of individuals genotyped in the right tail. The chromosome is of length $T = 1$M and two markers are located at each extremity ($\gamma = 0.3$, $a = 0$, $\mu = 0$, $\sigma = 1$, 10000 samples of size n).

γ_+ \ n	1000	200	50
γ	4.98%	4.96%	4.50%
$\gamma/2$	5.27%	4.89%	4.65%
$\gamma/4$	4.79%	4.91%	4.43%
$\gamma/8$	5.21%	4.99%	4.58%

Table 2. Theoretical Power and Empirical Power (in brackets) as a function of the location of the QTL $t^\star$ and the percentage γ_+ of individuals genotyped in the right tail . The chromosome is of length $T = 1$M 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).

γ_+ \ $t^\star$	10cM	30cM	60cM	80cM
γ	53.72% (53.84%)	30.70% (30.02%)	25.59% (25.88%)	39.88% (39.10%)
$\gamma/2$	76.22% (75.98%)	46.64% (46.15%)	38.91% (38.30%)	59.82% (59.07%)
$\gamma/4$	72.71% (72.41%)	43.80% (43.51%)	36.42% (35.72%)	56.53% (56.02%)
$\gamma/8$	67.95% (67.56%)	40.15% (39.65%)	33.22% (33.77%)	52.16% (51.44%)

Table 3. Theoretical Power and Empirical Power (in brackets) as a function of the location of the QTL $t^\star$ and the percentage γ_+ of individuals genotyped in the right tail . The chromosome is of length $T = 1$M 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).

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

Acknowledgements

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

References

[1] R. Wu, C.X. Ma, G. Casella, *Statistical Genetics of Quantitative Traits*, Springer (2007).
[2] D. Siegmund, B. Yakir, *The statistics of gene mapping*, Springer (2007).

[3] J.B.S. Haldane. *The combination of linkage values and the calculation of distance between the loci of linked factors*, Journal of Genetics 8 (1919), pp. 299–309.
[4] A. Rebaï, B. Goffinet, B. Mangin, *Comparing power of different methods for QTL detection*, Biometrics 51 (1995), pp. 87–99.
[5] A. Rebaï, B. Goffinet, B. Mangin, *Approximate thresholds of interval mapping tests for QTL detection*, Genetics 138 (1994), pp. 235–240.
[6] C. Cierco, *Asymptotic distribution of the maximum likelihood ratio test for gene detection*, Statistics 31 (1998), pp. 261–285.
[7] J.M. Azaïs and C. Cierco-Ayrolles, *An asymptotic test for quantitative gene detection*, Ann. Inst. Henri Poincaré (B) 38(6) (2002), pp. 1087–1092.
[8] J.M. Azaïs and M. Wschebor, *Level sets and extrema of random processes and fields*, Wiley, New-York (2009).
[9] M.N. Chang, R. Wu, S.S. Wu, G. Casella, *Score statistics for mapping quantitative trait loci*, Statistical Application in Genetics and Molecular Biology 8(1) 16 (2009).
[10] J.M. Azaïs, C. Delmas, C.E. Rabier, *Likelihood ratio test process for Quantitative Trait Locus detection*, to appear in Statistics (2012).
[11] R.J. Lebowitz, M. Soller, J.S. Beckmann, *Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines*, Theoretical and Applied Genetics, 73 (1987), pp. 556–562.
[12] E.S. Lander and D. Botstein, *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps*, Genetics 138 (1989), pp. 235–240.
[13] D. Darvasi and M. Soller, *Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus*, Theoretical and Applied Genetics 85 (1992), pp. 353–359.
[14] H. Muranty and B. Goffinet, *Selective genotyping for location and estimation of the effect of the effect of a quantitative trait locus*, Biometrics 53 (1997), pp. 629–643.
[15] C.E. Rabier, *On statistical inference for selective genotyping*, hal-00658583 (2012).
[16] N. Rabbee, D. Speca, N. Armstrong, T. Speed, *Power calculations for selective genotyping in QTL mapping in backcross mice*, Genet. Res. Camb. 84 (2004), pp. 103–108.
[17] A. Manichaikul, A. Palmer, S. Sen, K. Broman, *Significance thresholds for Quantitative Trait Locus mapping under selective genotyping*, Genetics 177 (2007), pp. 1963–1966.
[18] A. Genz, *Numerical computation of multivariate normal probabilities*, J. Comp. Graph. Stat. 1 (1992), pp. 141–149.
[19] G.A. Churchill, R.W. Doerge, *Empirical threshold values for quantitative trait mapping*, Genetics 138 (1994), pp. 963–971.
[20] A.W. Van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (1998).
[21] J.M. Azaïs, E. Gassiat, C. Mercadier, *Asymptotic distribution and local power of the likelihood ratio test for mixtures*, Bernoulli 12(5) (2006), pp. 775–799.
[22] E. Gassiat, *Likelihood ratio inequalities with applications to various mixtures*, Ann. Inst. Henri Poincaré (B) 6 (2002), pp. 897–906.