ANNALES
DE LA FACULTÉ
DES SCIENCES
DE
TOULOUSE
Mathématiques

CHARLES-ELIE RABIER

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

Tome XXIII, n° 4 (2014), p. 755-778.

<http://afst.cedram.org/item?id=AFST_2014_6_23_4_755_0>

Annales de la Faculté des Sciences de Toulouse Vol. XXIII, n° 4, 2014
pp. 755-778

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

Charles-Elie Rabier[(1)]

Abstract. — We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait) on the interval $[0, T]$ representing a chromosome. The originality is in the fact that some genotypes are missing. We give the asymptotic distribution of this LRT process under the null hypothesis that there is no QTL on $[0, T]$ and under local alternatives with a QTL at $t^\star$ on $[0, T]$. We show that the LRT process is asymptotically the square of a "non-linear interpolated and normalized Gaussian process". We have an easy formula in order to compute the supremum of the square of this interpolated process. We prove that the threshold is exactly the same as in the classical situation without missing genotypes.

Résumé. — Nous considérons le processus de test de rapport de vraisemblance (LRT) relatif au test d'absence de QTL (un QTL désigne un gène à effet quantitatif sur un trait) sur un intervalle $[0, T]$ représentant un chromosome. L'originalité de cette étude vient du fait que certains génotypes s'avèrent manquants. Nous donnons la distribution asymptotique du processus de LRT, sous l'hypothèse nulle d'absence de QTL sur $[0, T]$, et sous des alternatives locales où le QTL se situe en $t^\star$ sur $[0, T]$. Nous montrons que le processus de LRT est asymptotiquement le carré d'un "processus Gaussien d'interpolation non linéaire et renormalisé". Nous présentons une formule simple permettant le calcul du maximum du carré du processus interpolé. Pour finir, nous prouvons que la valeur critique est exactement la même que dans la configuration classique sans génotypes manquants.

(*) Reçu le 05/10/2012, accepté le 02/06/2014

(1) University of Wisconsin-Madison, Statistic department, Medical Science Center, 1300 University Ave., Madison, WI 53706-1532, USA.
rabier@stat.wisc.edu

Article proposé par Jean-Pierre Otal.

1. Introduction

We study a backcross population: $A \times (A \times B)$, where A and B are purely homozygous lines and we address the problem of detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured) on a given chromosome. The trait is observed on n individuals (progenies) and we denote by $Y_j,\ j = 1, ..., n$, the observations, which we will assume to be Gaussian, independent and identically distributed (i.i.d.). The mechanism of genetics, or more precisely of meiosis, implies that among the two chromosomes of each individual, one is purely inherited from A while the other (the "recombined" one), consists of parts originated from A and parts originated from B, due to crossing-overs.

The chromosome will be represented by the segment $[0, T]$. The distance on $[0, T]$ is called the genetic distance, it is measured in Morgans (see for instance [29] or [27]). 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 admitted model for the stochastic structure of $X(.)$ is due to Haldane which states that:

$$X(0) \sim \frac{1}{2}(\delta_{+1} + \delta_{-1}), \quad X(t) = X(0)(-1)^{N(t)}$$

where for any $b \in \mathbb{R}$, δ_b denotes the point mass at b and $N(.)$ is a standard Poisson process on $[0, T]$. In a more practical point of view, the Haldane [17] model assumes no crossover interference and the Poisson process represents the number of crossovers on $[0, T]$ which happen during meiosis. Calculations on the Poisson distribution show that

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

we set in addition

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

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

$$Y = \mu \ + \ X(t^\star)\, q \ + \ \sigma\varepsilon \tag{1.1}$$

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

Indeed, it is well known that there is a finite number of loci underlying the variation in quantitative traits (e.g. in aquaculture and livestock, see [18]). In this study, we will focus only on one locus (so-called QTL) and on only one quantitative trait. We will study the concept of QTL mapping: we will look for associations between allele variation at the QTL and variation in the quantitative trait of interest.

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. Note that in the following, the word "genotype" will refer to the genome information at all the markers locations. 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. An important aspect of QTL mapping is the presence of missing genotypes. Genotype data is rarely complete due to failures in the genotyping assays (see [7], [1]) or in order to reduce genotyping costs. For instance, under selective genotyping, only the individuals with extreme phenotypes are genotyped (e.g. [20], [19], [11], [21], [23]). The originality of this paper is that we consider the classical problem (i.e. without missing genotypes), but this time, we consider two real thresholds S_- and S_+ with $S_- \leqslant S_+$ and the genotype of one individual is available if and only if the phenotype Y belongs to the interval $S_- \leqslant Y \leqslant S_+$. If we call $\overline{X}(t)$ the random variable such as

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

then, in our problem, one observation will be now

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

Note that with our notations:

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

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

$$\begin{aligned} \big(p(t^\star)\ f_{(\mu+q,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]} \ &+\ (1-p(t^\star))\ f_{(\mu-q,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]} \\ +\ \tfrac{1}{2}\ f_{(\mu+q,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]} \ &+\ \tfrac{1}{2}\ f_{(\mu-q,\sigma)}(Y)\ \ 1_{Y\notin[S_-,S_+]}\big)\ g(.) \end{aligned} \quad (1.2)$$

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

We consider that we have n observations $\left(Y_j,\ \overline{X}_j(t_1),\ ...,\ \overline{X}_j(t_K)\right)$, $j = 1, ..., n$ which are i.i.d., with the same distribution as described previously, and we want to test the presence of a QTL. Since its true location is

unknown, we have to consider the location $t^\star$ as an unknown parameter t, and the likelihood process will also depend on the parameter t. The absence of a QTL is given by the null hypothesis H_0:"q=0," and the likelihood ratio test (LRT) of H_0 against its general alternative, has test statistic $\sup_t \Lambda_n(t)$, where $\Lambda_n(t)$ is the LRT statistic at location t. This paper gives the exact asymptotic distribution of this LRT statistic under the null hypothesis and under contiguous alternatives. Note that $\arg\sup_t \Lambda_n(t)$ is a natural estimator of the QTL location.

In the classical problem of detecting a QTL on a chromosome, that is to say in the oracle situation where all the individuals are genotyped, the asymptotic distribution of the LRT statistic has been given under some approximations by [25], [24], [9], [2], [5], [8]. Recently, [6] 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 when only the genotypes of the non extreme individuals (i.e. the individuals for which the phenotypes Y belong to the interval $[S_-, S_+]$) are available. The main result of the paper (Theorems 2.1 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 [6] only for the oracle situation. Under the null hypothesis, despite the missing genotypes, our process is exactly the same as the one obtained by [6]. 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 and Lemma 3.1. In Theorem 4.2, we give the Asymptotic Relative Efficiency (ARE) with respect to the oracle situation. Furthermore, we propose a test statistic (see Lemma 3.1 and formula (3.1)) 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 extreme phenotypes (for which the genotypes are missing) don't bring any extra information for statistical inference. This result is complementary to the one obtained in [22], where I show that, under selective genotyping, the non extreme phenotypes don't bring any information for statistical inference.

To conclude, we will illustrate our theoretical results with the help of simulated data. Note that, according to Theorems 2.1 and 4.1, the threshold

(i.e. critical value) in our study is exactly the same as the classical threshold used in the oracle situation. So, in order to obtain our threshold, the Monte-Carlo Quasi Monte-Carlo methods of [6], based on [16] is still suitable here. This is an alternative to the permutation method proposed by [10]. Our method is very fast since it relies on very powerful algorithms developed by [16]. 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 [28] 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 $t \in [t_1, t_2]$.

Notations. — For $(i, i') \in \{-1, 1\}^2$, $Q_t^{i,i'}$ is the quantity such as

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

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

Notations. — $\mathcal{B}$ is the quantity such as $\mathcal{B} = \sigma^2\ \left(1 - \gamma\ -\ z_{\gamma_+}\ \varphi\left(z_{\gamma_+}\right)\ +\ z_{1-\gamma_-}\ \varphi\left(z_{1-\gamma_-}\right)\right)$, 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.

Using Bayes rules, we have

$$\begin{aligned} Q_t^{1,1} &= \frac{\bar{r}(t_1, t)\ \bar{r}(t, t_2)}{\bar{r}(t_1, t_2)}, \quad Q_t^{1,-1} = \frac{\bar{r}(t_1, t)\ r(t, t_2)}{r(t_1, t_2)} \\ Q_t^{-1,1} &= \frac{r(t_1, t)\ \bar{r}(t, t_2)}{r(t_1, t_2)}, \quad Q_t^{-1,-1} = \frac{r(t_1, t)\ r(t, t_2)}{\bar{r}(t_1, t_2)}. \end{aligned} \tag{2.1}$$

We can remark that we have

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

Let us define the quantity $p(t)$ such as

$$p(t) = Q_t^{1,1}\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=1}\ +\ Q_t^{1,-1}\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=-1} \\ +Q_t^{-1,1}\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=1}\ +\ Q_t^{-1,-1}\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=-1} \quad (2.2)$$

and let $\theta = (q,\ \mu,\ \sigma)$ be the parameter of the model at t fixed. According to straightforward calculations present in Appendix, 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 $\mathbb{N}$, is $\forall t \in [t_1, t_2]$:

$$L_t(\theta) = \big(\ p(t) f_{(\mu+q,\sigma)}\,(Y)\,1_{Y\in[S_-,S_+]} + (1-p(t))\,f_{(\mu-q,\sigma)}\,(Y)\,1_{Y\in[S_-,S_+]} \\ +\ \tfrac{1}{2}\ f_{(\mu+q,\sigma)}\,(Y)\,1_{Y\notin[S_-,S_+]}\ +\ \tfrac{1}{2}\ f_{(\mu-q,\sigma)}\,(Y)\,1_{Y\notin[S_-,S_+]}\big)\ g\,(t) \quad (2.3)$$

where the function

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

can be removed because it does not depend on the parameters. Note that for $t = t^\star$, we find our formula (1.2) of the introduction where $p(t^\star)$ is described in formula (2.2).

Before introducing our main theorem, let us define the different hypotheses, the score statistic and the LRT statistic at t. 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 $a \neq 0$".

Since the Fisher Information matrix is diagonal (cf. proof of Theorem 2.1 below), the score statistic of the hypothesis "$q = 0$" at t, for n independent observations, will be defined as

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q}\ |_{\theta_0}}{\sqrt{\mathrm{Var}_{H_0}\left(\frac{\partial l_t^n}{\partial q}\ |_{\theta_0}\right)}},$$

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

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

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

Notations. — $\Rightarrow$ is the weak convergence, $\stackrel{F.d.}{\rightarrow}$ is the convergence of finite-dimensional distributions and $\stackrel{\mathcal{L}}{\longrightarrow}$ is the convergence in distribution.

Our main result is the following:

THEOREM 2.1. — *Suppose that the parameters (q, μ, σ^2) vary in a compact and that σ^2 is bounded away from zero. With the previous defined notations,*

$$S_n(.) \Rightarrow U(.), \quad \Lambda_n(.) \stackrel{F.d.}{\longrightarrow} U^2(.), \quad \sup \Lambda_n(.) \stackrel{\mathcal{L}}{\longrightarrow} \sup U^2(.)$$

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

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

$$U(t) = \frac{\alpha(t)\, U(t_1) + \beta(t)\, U(t_2)}{\sqrt{Var\left(\alpha(t)\, U(t_1) + \beta(t)\, U(t_2)\right)}} \tag{2.4}$$

where

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

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

and with expectation:

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

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

where

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

In the sense of this equation, $U(.)$ will be called a "non linear normalized interpolated process". We can see that under the null hypothesis, despite the missing genotypes, $U(.)$ is exactly the same process as the process $Z(.)$ of Theorem 2.1 of [6] 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 $\sqrt{\mathcal{B}}/\sigma$. Note also

that $U(.)$ is the generalization of $Z(.)$. Indeed, if we choose $S_- = -\infty$ and $S_+ = +\infty$, that is to say the genotypes of all the individuals are available, the factor $\sqrt{\mathcal{B}}/\sigma$ is equal to 1, and $U(.)$ is the same process as $Z(.)$.

Proof of Theorem 2.1. — The proof is divided into four parts:

(a) Preliminaries (i.e. computation of the Fisher Information Matrix)

(b) Study of the score process under H_0

(c) Study of the score process under the local alternative $H_{at^\star}$

(d) Study of the supremum of the LRT process.

Preliminaries

Recall that $l_t(\theta)$ denotes the log-likelihood. We first compute the Fisher information at a point θ_0 that belongs to H_0. The proof relies on two key lemmas.

LEMMA 2.2. — *We have the following relationship:*

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

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

To prove this lemma, use formula (2.2) and check that both sides coincide when $Y \in [S_-, S_+]$.

LEMMA 2.3. — *Let* $W \sim N(\mu,\ \sigma^2)$, *then*

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

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

$$I_{\theta_0} = Diag\left(\mathcal{B}\left(\alpha^2(t) + \beta^2(t) + 2\alpha(t)\,\beta(t)\,\rho(t_1,t_2)\right)/\sigma^4\ ,\ \frac{1}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right). \quad (2.5)$$

Study of the score process under H_0

Using Lemma 2.2, it is clear that

$$\begin{aligned}\frac{\partial l_t^n}{\partial q} \mid_{\theta_0} &= \sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma^2} \ (2p_j(t) - 1) \ 1_{Y_j \in [S_-, S_+]} \\ &= \frac{\alpha(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j \ \overline{X}_j(t_1) \ + \ \frac{\beta(t)}{\sigma} \sum_{j=1}^{n} \ \varepsilon_j \ \overline{X}_j(t_2) \end{aligned} \tag{2.6}$$

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

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

$$S_n(t_k) = \frac{\frac{\partial l_{t_k}^n}{\partial q} \mid_{\theta_0}}{\sqrt{\mathrm{Var}_{H_0}\left(\frac{\partial l_{t_k}^n}{\partial q} \mid_{\theta_0}\right)}} = \sum_{j=1}^{n} \frac{\sigma \varepsilon_j \ \overline{X}_j(t_k)}{\sqrt{n \ \mathcal{B}}}.$$

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

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

Then, since we have the relationship (cf. formula (2.6))

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

the continous mapping theorem implies that

$$S_n(t) \xrightarrow{\mathcal{L}} U(t) \ \ \forall t \in [0, T].$$

It proves the convergence of finite-dimensional.

Let us now compute the covariance of the score statistics on markers, that is to say the covariance between $S_n(t_1)$ and $S_n(t_2)$. Since $\mathbb{E}_{H_0}\left((Y - \mu)^2 \, 1_{Y \in [S_-, S_+]}\right) = \mathcal{B}$, we have

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

As a consequence,

$$\mathrm{Cov}_{\mathrm{H_0}}\left(S_{\mathrm{n}}(\mathrm{t}_1), S_{\mathrm{n}}(\mathrm{t}_2)\right) = \rho(\mathrm{t}_1, \mathrm{t}_2).$$

Let us now prove the weak convergence of the score process. Recall that the tightness and the convergence of finite-dimensional imply the weak convergence of the score process (see for instance Theorem 4.9 of [5]). Since we have already proved the convergence of finite-dimensional, let us focus on the tightness of the score process. Since $p(t)$ and $\alpha^2(t)+\beta^2(t)+2\alpha(t)\beta(t)\rho(t_1,t_2)$ are continuous functions, each path of the process $S_n(.)$ is a continuous function on $[t_1,t_2]$. Recall the modulus of continuity of a continous function $x(t)$ on $[t_1,t_2]$:

$$w_x(\delta) = \sup_{|t'-t|<\delta} |x(t') - x(t)| \quad \text{where} \quad t_1 < \delta \leqslant t_2.$$

According to Theorem 8.2 of Billingsley (1999), the score process is tight if and only if the two following conditions hold:

(a) the sequence $S_n(t_1)$ is tight.

(b) for each positive ε and η, there exists a δ, with $t_1 < \delta < t_2$, and an integer n_0 such that $\mathbb{P}\left(w_{S_n}(\delta) \geqslant \eta\right) \leqslant \varepsilon \quad \forall n \geqslant n_0$.

According to Prohorov, the sequence $S_n(t_1)$ is tight. Then, (a) is verified. Let us define the functions $\tilde{\alpha}(t)$ and $\tilde{\beta}(t)$ in the following way:

$$\tilde{\alpha}(t) = \alpha(t)/\sqrt{\alpha^2(t)+\beta^2(t)+2\alpha(t)\beta(t)\rho(t_1,t_2)},$$
$$\tilde{\beta}(t) = \beta(t)/\sqrt{\alpha^2(t)+\beta^2(t)+2\alpha(t)\beta(t)\rho(t_1,t_2)}.$$

First, we can notice that $\forall\delta$ such as $t_1 < \delta \leqslant t_2$,

$$\begin{aligned} w_{S_n}(\delta) &= \sup_{|t'-t|<\delta} |S_n(t') - S_n(t)| \\ &= \sup_{|t'-t|<\delta} \left| \left(\tilde{\alpha}(t') - \tilde{\alpha}(t)\right) S_n(t_1) + \left(\tilde{\beta}(t') - \tilde{\beta}(t)\right) S_n(t_2) \right| \\ &\leqslant \max\left(|S_n(t_1)|, |S_n(t_2)|\right) \left(w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta)\right). \end{aligned} \tag{2.7}$$

Furthermore, the sequence $\max\left(|S_n(t_1)|, |S_n(t_2)|\right)$ is uniformly tight. This way,

$$\forall\varepsilon > 0 \;\; \exists M > 0 \;\; \forall n \geqslant 1 \;\; \mathbb{P}\left(\max\left(|S_n(t_1)|, |S_n(t_2)|\right) \geqslant M\right) \leqslant \varepsilon. \tag{2.8}$$

According to Heine's theorem, since $\tilde{\alpha}(t)$ and $\tilde{\beta}(t)$ are continuous on the compact $[t_1,t_2]$, these functions are uniformly continuous. So,

$$\forall\upsilon > 0 \;\; \exists\delta \;\text{ such as }\; t_1 < \delta < t_2, \;\; w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta) < \upsilon. \tag{2.9}$$

Let η be a positive quantity. Using formulae (2.8) and (2.9) and imposing $\upsilon = \eta/M$, we have

$$\mathbb{P}\left(\max\left(|S_n(t_1)|, |S_n(t_2)|\right)\left(w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta)\right) \geqslant \eta\right) \leqslant \varepsilon.$$

As a consequence, according to formula (2.7), we have

$$\forall n \geqslant 1 \quad \mathbb{P}\left(w_{S_n}(\delta) \geqslant \eta\right) \leqslant \varepsilon.$$

It proves (b) of Theorem 8.2 of Billingsley (1999). As a result, the tightness of the score process is proved. To conclude, the tightness and the convergence of finite-dimensional imply the weak convergence of the score process.

Study of the score process 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 2.2), 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 $\forall k = 1, 2$. Recall that under H_0,

$$S_n(t_k) = \sum_{j=1}^{n} \frac{\sigma\varepsilon_j \ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{B}}}, \quad S_n(t_k) \xrightarrow{\mathcal{L}} N(0,\ 1). \tag{2.10}$$

Since our model is differentiable in quadratic mean, according to Theorem 7.2 of [28], under H_0, the log likelihood ratio verifies

$$l^n_{t^\star}(\theta) \ - \ l^n_{t^\star}(\theta_0) = \frac{a}{\sqrt{n}} \frac{\partial l^n_{t^\star}}{\partial q} \mid_{\theta_0} \ - \ \frac{a^2}{2} \mathbb{E}_{H_0}\left(\left(\frac{\partial l_{t^\star}}{\partial q} \mid_{\theta_0}\right)^2\right) + o_P(1) \tag{2.11}$$

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

According to the central limit theorem and formula (2.5), under H_0

$$l^n_{t^\star}(\theta) \ - \ l^n_{t^\star}(\theta_0) \xrightarrow{\mathcal{L}} N\left(-\frac{1}{2}\vartheta^2,\ \vartheta^2\right) \tag{2.12}$$

with $\vartheta^2 = a^2\mathcal{B}\left(\alpha^2(t^\star) + \beta^2(t^\star) + 2\alpha(t^\star)\beta(t^\star)\rho(t_1, t_2)\right)/\sigma^4$.

As a consequence, conditions required to apply Le Cam's third lemma are fulfilled (cf. formulae (2.10) and (2.12)). Recall that Le Cam'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 (2.5), (2.6) and (2.11), under H_0,

$$\begin{aligned} l^n_{t^\star}(\theta) \ - \ & l^n_{t^\star}(\theta_0) \\ & = \frac{a}{\sigma\sqrt{n}} \ \left(\alpha\left(t^\star\right) \sum_{j=1}^{n} \varepsilon_j \ \overline{X}_j(t_1) + \beta\left(t^\star\right) \sum_{j=1}^{n} \varepsilon_j \ \overline{X}_j(t_2) \right) \\ & -\frac{a^2\mathcal{B}}{2\sigma^4} \left(\alpha^2\left(t^\star\right) + \beta^2\left(t^\star\right) + 2\alpha\left(t^\star\right)\beta\left(t^\star\right)\rho\left(t_1,t_2\right) \right) + o_P(1). \end{aligned} \tag{2.13}$$

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

$$\begin{aligned} \mathrm{Cov}_{\mathrm{H}_0} & \left(S_{\mathrm{n}}\left(\mathrm{t}_1\right), \frac{\mathrm{a}\ \alpha\left(\mathrm{t}^\star\right)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}} \ \overline{\mathrm{X}}_{\mathrm{j}}\left(\mathrm{t}_1\right) \right) \\ & = \ \mathrm{Cov}_{\mathrm{H}_0} \left(\sum_{\mathrm{j}=1}^{\mathrm{n}} \frac{\sigma\varepsilon_{\mathrm{j}} \ \overline{\mathrm{X}}_{\mathrm{j}}\left(\mathrm{t}_1\right)}{\sqrt{\mathrm{n}\ \mathcal{B}}}, \frac{\mathrm{a}\ \alpha\left(\mathrm{t}^\star\right)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}} \ \overline{\mathrm{X}}_{\mathrm{j}}\left(\mathrm{t}_1\right) \right) \\ & = \ \frac{a\ \alpha\left(t^\star\right)}{\sqrt{\mathcal{B}}} \mathrm{Var}_{H_0}\left(\varepsilon\ \overline{X}\left(t_1\right)\right) = \frac{a\ \alpha\left(t^\star\right)\ \sqrt{\mathcal{B}}}{\sigma^2}. \end{aligned}$$

In the same way,

$$\begin{aligned} \mathrm{Cov}_{\mathrm{H}_0} & \left(S_{\mathrm{n}}\left(\mathrm{t}_1\right), \frac{\mathrm{a}\ \beta\left(\mathrm{t}^\star\right)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}} \ \overline{\mathrm{X}}_{\mathrm{j}}\left(\mathrm{t}_2\right) \right) \\ & = \frac{a\ \beta\left(t^\star\right)}{\sqrt{\mathcal{B}}} \ \mathrm{Cov}_{\mathrm{H}_0}\left(\varepsilon\ \overline{\mathrm{X}}\left(\mathrm{t}_1\right), \varepsilon\ \overline{\mathrm{X}}\left(\mathrm{t}_2\right)\right) \\ & = \frac{a\ \beta\left(t^\star\right)}{\sigma^2\ \sqrt{\mathcal{B}}} \ \mathbb{E}_{H_0}\left((Y-\mu)^2 \ X\left(t_1\right) \ X\left(t_2\right) \ 1_{Y\in[S_-,S_+]} \right) \\ & = \frac{a\ \beta\left(t^\star\right)}{\sigma^2\ \sqrt{\mathcal{B}}} \ \mathbb{E}_{H_0}\left((Y-\mu)^2 \ 1_{Y\in[S_-,S_+]} \right) \mathbb{E}\left(X\left(t_1\right) \ X\left(t_2\right) \right) \\ & = \frac{a\ \beta\left(t^\star\right)\ \sqrt{\mathcal{B}}\ \rho\left(t_1,t_2\right)}{\sigma^2}. \end{aligned} \tag{2.14}$$

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

$$\mathrm{Cov}_{\mathrm{H}_0}\left(S_{\mathrm{n}}\left(\mathrm{t}_1\right), l^{\mathrm{n}}_{\mathrm{t}^\star}\left(\theta\right) \ - \ l^{\mathrm{n}}_{\mathrm{t}^\star}\left(\theta_0\right)\right) = \frac{\mathrm{a}\,\sqrt{\mathcal{B}}\,\rho(\mathrm{t}_1, \mathrm{t}^\star)}{\sigma^2}.$$

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

$$\mathrm{Cov}_{\mathrm{H}_0}\left(S_{\mathrm{n}}(\mathrm{t}_2), l^{\mathrm{n}}_{\mathrm{t}^\star}\left(\theta\right) \ - \ l^{\mathrm{n}}_{\mathrm{t}^\star}\left(\theta_0\right)\right) = \frac{\mathrm{a}\,\sqrt{\mathcal{B}}\,\rho(\mathrm{t}^\star, \mathrm{t}_2)}{\sigma^2}.$$

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

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

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 remainder is uniform in t.

Let us consider now t as an extra parameter. Let $t^\star$, $\theta^\star$ be the true parameter that will be assumed to belong to H_0. Note that $t^\star$ 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 [4] hold. So, we can apply Theorem 1 of [4] and we have

$$\sup_{(t,\theta)} l_t\left(\theta\right) - l_{t^\star}\left(\theta^\star\right) = \sup_{d\in\mathcal{D}} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right)^2 1_{\sum_{j=1}^n d(X_j)\geqslant 0} \right) + o_P(1) \tag{2.15}$$

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 [4], see also [15] and [3]. A similar result is true under H_0 with a set $\mathcal{D}_0$. Let us precise the sets of scores $\mathcal{D}$ and $\mathcal{D}_0$. These sets are defined at the sets of scores of one parameter families that converge to the true model $p_{t^\star,\theta^\star}$ and that are differentiable in quadratic mean.

It is easy to see that

$$\mathcal{D} = \left\{ \frac{\langle W, l'_t\left(\theta^\star\right)\rangle}{\sqrt{\mathrm{Var}_{H_0}\left(\langle W, l'_t(\theta^\star)\rangle\right)}}, W \in \mathbb{R}^3, t \in [t_1, t_2] \right\}$$

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

$$\mathcal{D}_0 = \Big\{ \frac{\langle W, l'_t(\theta^\star)\rangle}{\sqrt{\mathrm{Var}_{H_0}\left(\langle W, l'_t(\theta^\star)\rangle\right)}}, W \in \mathbb{R}^2 \Big\},$$

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

Using the transform $W \to -W$ in the expressions of the sets of score, we see that the indicator function can be removed in formula (2.15). Then, since the Fisher information matrix is diagonal (see formula (2.5)) , it is easy to see that

$$\sup_{d\in\mathcal{D}} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right)^2 \right) - \sup_{d\in\mathcal{D}_0} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right)^2 \right)$$
$$= \sup_{t\in[t_1,t_2]} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\frac{\partial l_t}{\partial q}(X_j) \mid_{\theta_0}}{\sqrt{\mathrm{Var}_{H_0}\left(\frac{\partial l_t}{\partial q}(X_j) \mid_{\theta_0}\right)}} \right)^2 \right).$$

This is exactly the desired result. Note that the model with $t^\star$ fixed is differentiable in quadratic mean, this implies that the alternative defines a contiguous sequence of alternatives. By Le Cam's first lemma, relation (2.15) remains true under the alternative. □

Remark. — According to the Law of Large Numbers, under the null hypothesis H_0 and under the local alternative $H_{at^\star}$, $\frac{1}{n}\sum 1_{Y_j\in[S_+,S_-]} \to 1-\gamma$. So, $1-\gamma$ corresponds asymptotically to the percentage of individuals genotyped. In the same way, γ_+ (resp. γ_-) corresponds asymptotically to the percentage of non-genotyped individuals in the right tail (resp. the left tail) of the distribution.

3. An easy way to perform the statistical test

Since $U(.)$ is a "non linear normalized interpolated process", we can use Lemma 2.2 of [6] in order to compute easily the supremum of $U^2(.)$. Note that this lemma is suitable here because we have exactly the same interpolation as in Theorem 2.1 of [6]. It comes

$$\max_{t\in[t_1,t_2]} \frac{\left(\alpha(t)U(t_1)+\beta(t)U(t_2)\right)^2}{\alpha^2(t)+\beta^2(t)+2\rho(t_1,t_2)\alpha(t)\beta(t)} \tag{3.1}$$
$$= \max\left(U^2(t_1), U^2(t_2), \frac{U^2(t_1)+U^2(t_2)-2\rho(t_1,t_2)U(t_1)U(t_2)}{1-\rho^2(t_1,t_2)} 1_{\frac{U(t_2)}{U(t_1)}\in\left]\rho(t_1,t_2),\frac{1}{\rho(t_1,t_2)}\right[} \right).$$

Note that since under H_0, the process $U(.)$ is exactly the same process as the process $Z(.)$ obtained by [6], we will have exactly the same threshold as the one under the oracle situation (i.e. all the individuals genotyped). So, the Monte-Carlo Quasi Monte-Carlo method of [6] and based on [16], is still suitable here.

Let's focus now on the data analysis. Which test statistic should we use in order to make the data analysis easy? Indeed, when we focus only on one location of the genome which is a marker location, performing a LRT or a Wald test is time consuming: an EM algorithm is required to obtain the maximum likelihood estimators. So, since we focus here on the whole chromosome, we have to propose the easiest statistical test for geneticists.

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

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

We introduce the following lemma.

LEMMA 3.1. — *Let $T_n(.)$ be the process such as*

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

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

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 (3.1) and replacing $U(t_1)$ and $U(t_2)$ by respectively $T_n(t_1)$ and $T_n(t_2)$. Note that, according to Lemma 3.1, 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.

On the other hand, a consequence of Lemma 3.1 is that the extreme phenotypes (for which the genotypes are missing) don't bring any information for statistical inference. Indeed, our test statistics $T_n(t)$ are based only on the non extreme phenotypes, as soon as we replace the empirical mean $\overline{Y}$ by $\hat{\mu}$, an estimator $\sqrt{n}$ consistent based only on the non extreme phenotypes ($\hat{\mu}$ can be obtained by the method of moments for instance). This result is complementary to the one obtained in [22], where it is shown that, under selective genotyping, the non extreme phenotypes (i.e. $Y \in [S_-, S_+]$ in the case of the selective genotyping) don't bring any information for statistical inference.

4. Several markers : the "Interval Mapping" of [19] in presence of missing genotypes

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

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

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

THEOREM 4.1. — *We have the same result as in Theorem 2.1, provided that we make some adjustments and that we redefine $U(.)$ in the following way:*

- *in the definition of $\alpha(t)$ and $\beta(t)$, t_1 becomes t^ℓ and t_2 becomes t^r*
- *under the null hypothesis, the process $U(.)$ 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, $U(.)$ is obtained from $U(t^\ell)$ and $U(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$*
- *at the marker positions, the expectation is such as $m_{t^\star}(t_k) = \frac{a\sqrt{\mathcal{B}}\rho(t_k, t^\star)}{\sigma^2}$*
- *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 of Theorem 4.1. — 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 $t^\star$ belongs to the marker interval (t^ℓ, t^r), the proof is the same as the proof of Theorem 2.1. 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 (2.13) 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

$$\begin{aligned}
&\mathrm{Cov}_{\mathrm{H}_0}\left(S_{\mathrm{n}}\left(\mathrm{t_k}\right), \frac{\mathrm{a}\ \alpha\left(\mathrm{t}^{\star}\right)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}}\ \overline{\mathrm{X}}_{\mathrm{j}}\left(\mathrm{t}^{\star\ell}\right)\right) \\
&\quad = \mathrm{Cov}_{\mathrm{H}_0}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}} \frac{\sigma\varepsilon_{\mathrm{j}}\ \overline{\mathrm{X}}_{\mathrm{j}}\left(\mathrm{t_k}\right)}{\sqrt{\mathrm{n}\ \mathcal{B}}}, \frac{\mathrm{a}\ \alpha\left(\mathrm{t}^{\star}\right)}{\sigma\sqrt{\mathrm{n}}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \varepsilon_{\mathrm{j}}\ \overline{\mathrm{X}}_{\mathrm{j}}\left(\mathrm{t}^{\star\ell}\right)\right) \\
&\quad = \frac{a\ \alpha\left(t^{\star}\right)}{\sigma^2\ \sqrt{\mathcal{B}}}\ \mathbb{E}_{H_0}\left(\left(Y-\mu\right)^2\ X\left(t_k\right)\ X\left(t^{\star\ell}\right)\ 1_{Y\notin[S_-,S_+]}\right) \\
&\quad = \frac{a\ \alpha\left(t^{\star}\right)\ \sqrt{\mathcal{B}}\ \rho\left(t_k, t^{\star\ell}\right)}{\sigma^2}.
\end{aligned}$$

In the same way,

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

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 Le Cam's third lemma, we have under the local alternative

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

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 (3.1) are still true here. We just have to replace t_1 and t_2 by t^{ℓ} and t^{r} in order to obtain the good expressions. As a consequence, we can easily compute $\sup T_n^2(.)$.
We introduce now our Theorem 4.2.

THEOREM 4.2. — *Let κ be the Asymptotic Relative Efficiency (ARE) with respect to the oracle situation where all the genotypes are known. Then, we have*

$$\begin{aligned}
&\textit{i)}\ \kappa = 1 - \gamma\ -\ z_{\gamma_+}\ \varphi\left(z_{\gamma_+}\right)\ +\ z_{1-\gamma_-}\ \varphi\left(z_{1-\gamma_-}\right) \\
&\textit{ii)}\ \kappa \textit{ reaches its maximum for } \gamma_+ = \gamma \textit{ or } \gamma_- = \gamma \\
&\textit{iii)}\ \kappa > 1-\gamma \Leftrightarrow z_{1-\gamma_-}\ \varphi\left(z_{1-\gamma_-}\right) > z_{\gamma_+}\ \varphi\left(z_{\gamma_+}\right).
\end{aligned}$$

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 $t^\star$. On the other hand, according to ii) of Theorem 4.2, if only a percentage $1-\gamma$ of genotypes is available in the population considered, the efficiency of our test is maximum when all the missing genotypes are located in the right tail of the distribution (i.e. $\gamma_+ = \gamma$). Obviously, by symmetry, the efficiency of our test is also maximum when all the missing genotypes are located in the left tail of the distribution (i.e. $\gamma_- = \gamma$). Note also, that according to iii), our test can reduce costs due to genotyping when $z_{1-\gamma_-}\ \varphi\left(z_{1-\gamma_-}\right) > z_{\gamma_+}\ \varphi\left(z_{\gamma_+}\right)$. However, this condition is very restrictive due to the properties of the Gaussian distribution.

Proof of Theorem 4.2. — The proof of i) is obvious since the mean function of the process $U(.)$ and the one of the process $Z(.)$ corresponding to the oracle situation, are proportional of a factor $\sqrt{\mathcal{B}}/\sigma$. Let's now prove that the maximum is reached for $\gamma_- = \gamma$, that is to say $\gamma_+ = 0$, since $\gamma = \gamma_+ + \gamma_-$. Note that without loss of generality, it will also prove that the maximum is reached for $\gamma_+ = \gamma$ and $\gamma_- = 0$. We have to answer the following question : how must we choose 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 : $u\left(z_{\gamma_+}\right) = \Phi^{-1}\left(\gamma - 1 + \Phi\left(z_{\gamma_+}\right)\right)$. Then, $z_{1-\gamma_-} = u\left(z_{\gamma_+}\right)$.

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

$$\begin{aligned} k_1'\left(z_{\gamma_+}\right) &= \varphi\left(z_{\gamma_+}\right) + z_{\gamma_+}\varphi'\left(z_{\gamma_+}\right) - u'\left(z_{\gamma_+}\right)\varphi\left(u\left(z_{\gamma_+}\right)\right) \\ &\quad -u\left(z_{\gamma_+}\right)u'\left(z_{\gamma_+}\right)\varphi'\left(u\left(z_{\gamma_+}\right)\right), \\ u'(z_{\gamma_+}) &= \frac{\varphi(z_{\gamma_+})}{\varphi(z_{1-\gamma_-})}\ . \end{aligned}$$

As a consequence,

$$k_1'\left(z_{\gamma_+}\right) = \varphi\left(z_{\gamma_+}\right)\left(z_{\gamma_-}^2 - z_{\gamma_+}^2\right).$$

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

5. Applications

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

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

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 only the genotypes of the non extreme individuals (i.e. the individuals for which $Y \in [S_-, S_+]$) are available or if all the genotypes are available (i.e. the oracle situation). So, the Monte-Carlo Quasi Monte-Carlo method, proposed by [6] (based on [16]) for the oracle situation, is still suitable here to obtain our threshold. Note that in [6], the authors show that their method gives better results than the method of [14] based on [26], and the method of [24] based on [12] and [13]. This way, in Tables 1 and 2, we propose to check on simulated data, the fact that the threshold is the same as in the oracle situation. In the following, 1M will denote 1 Morgan whereas 1cM will stand for 1 centiMorgan. First, in Table 1, we consider a sparse map: a chromosome of length $T = 1$M, with two genetic markers located at each extremity. For such a configuration, if we choose a 5% level, the corresponding threshold is 4.89. We consider $\gamma = 0.2$. In other words we have 20% of missing genotypes. Besides, we consider different values for the percentage γ_+ of indivuals not genotyped in the right tail of the distribution. 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 $n = 50$). Then, in Table 2, we consider a more dense genetic map. We still consider a chromosome of length $T = 1$M, but 6 genetic markers are now equally spaced every 20cM. We can notice that, as previously, the Percentage of False Positives is close to 5%.

Let's now focus on the alternative hypothesis. To begin, in Table 3, we consider the sparse map and the same value of 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_+. We present the Theoretical Power based on 100000 paths of the asymptotic process, and also the Empirical Power (in brackets) obtained for $n = 1000$. We can see that the Theoretical Power and the Empirical Power are very close whatever

the values of $t^\star$ and γ_+ are. Besides, as expected (cf. Theorem 4.2), we can see that the Theoretical Power is maximum when the missing genotypes are all located in the right tail of the distribution (i.e. $\gamma_+ = \gamma$). Note that we would have obtained the same result for $\gamma_- = \gamma$ (i.e. left tail). Finally, in Table 4, we consider the dense map, and we change the value of a : $a = 6$. We obtain the same kind of conclusions as before. This result was expected since all the theoretical results obtained in this paper, are suitable for any kind of genetic map.

To conclude, we present in this paper easy ways to analyze data in presence of missing genotypes. That's why it must be interesting for geneticists.

6. Appendix

Notations. — $\mathbb{P}_t\{l \mid i\}$ is the quantity such as $\forall\, l \in \{-1, 0, 1\}$ and $\forall\, i \in \{-1, 1\}$

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

In order to compute the likelihood, we have to study the different probability distributions.
To begin with, let us compute $\mathbb{P}\left(Y \in [y\ ,\ y+dy] \ \cap\ \overline{X}(t_1) = 1 \ \cap\ \overline{X}(t_2) = 1\right)$, and let consider that the location tested t is equal to the true location $t^\star$ of the QTL. We have, according to Bayes rules,

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

Besides,

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

and

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

It comes, using formula (2.1),

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

In the same way, after some calculations, we find

$$\begin{aligned}\mathbb{P}(Y \in [y\ ,\ y+dy] \ \cap\ \overline{X}(t_1) = 1\ \cap\ \overline{X}(t_2) = -1)\\ = \frac{1}{2}\, f_{(\mu+q,\sigma)}(y)\ 1_{y\in[S_-,S_+]}\ r(t_1,t_2)\ Q_t^{1,-1}\\ + \frac{1}{2}\, f_{(\mu-q,\sigma)}(y)\ 1_{y\in[S_-,S_+]}\ r(t_1,t_2)\ Q_t^{-1,1}, \quad (6.1)\end{aligned}$$

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

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

Finally, when the genotype is missing (i.e. the phenotype is extreme), we find

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

As a result, we obtain the expression of the likelihood described in formula (2.3) of Section 2.

Acknowledgements. — I thank Professor Laurent Bordes from University of Pau (FR) for having proposed this subject of research. I also thank Professor Jean-Marc Azaïs from University Paul-Sabatier Toulouse (FR) for fruitful discussions, and the reviewers who helped me to improve the paper.

Table 1. — Percentage of False Positives as a function of the number of individuals n and the percentage of individuals γ_+ not genotyped in the right tail. The chromosome is of length $T = 1$M and 2 markers are located at each extremity ($\gamma = 0.2$, $1 - \gamma = 0.8$, $a = 0$, $\sigma = 1$, $\mu = 0$, 10000 samples of n individuals).

γ_+ \ n	1000	200	100	50
$\gamma/4$	4.98%	5.43%	4.60%	4.79%
$\gamma/2$	5.10%	4.87%	5.17%	4.94%
γ	5.18%	5.31%	4.49%	4.61%

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

γ_+ \ n	1000	200	100	50
$\gamma/4$	5.08%	4.65%	4.59%	4.48%
$\gamma/2$	5.01%	4.70%	4.56%	4.44%
γ	5.06%	4.67%	4.28%	4.20%

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 non genotyped in the right tail. The chromosome is of length $T = 20$cM and 2 markers are located at each extremity ($\gamma = 0.2$, $1-\gamma = 0.8$, $a = 4$, $\sigma = 1$, $\mu = 0$, 10000 samples of $n = 1000$ individuals, 100000 paths for the Theoretical Power).

γ_+ \ $t^\star$	5cM	12cM	18cM
$\gamma/4$	60.46% (60.75%)	58.88% (58.43%)	62.88% (62.18%)
$\gamma/2$	56.17% (56.68%)	54.57% (54.82%)	58.44% (58.16%)
γ	76.61% (76.34%)	74.79% (74.71%)	79.10% (78.60%)

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

$t^\star$ / γ_+	12cM	35cM	48cM	77cM
$\gamma/4$	83.99% (83.47%)	86.39% (85.91%)	84.91% (84.20%)	87.87% (87.19%)
$\gamma/2$	80.14% (79.62%)	82.75% (81.88%)	80.96% (80.37%)	84.00% (83.25%)
γ	95.36% (94.87%)	96.23% (96.07%)	95.48% (94.83%)	96.85% (96.89%)

Bibliography

[1] Arends (D.), Prins (P.), Broman (K.), Jansen (R.). — Tutorial Multiple QTL Mapping (MQM) Analysis, Technical Report (2010).

[2] Azaïs (J.M.), Cierco-Ayrolles (C.). — An asymptotic test for quantitative gene detection, Ann. Inst. Henri Poincaré (B), 38, 6, p. 1087-1092 (2002).

[3] Azaïs (J.M.), Gassiat (E.), Mercadier (C.). — Asymptotic distribution and local power of the likelihood ratio test for mixtures, Bernoulli, 12(5), p. 775-799 (2006).

[4] Azaïs (J.M.), Gassiat (E.), Mercadier (C.). — The likelihood ratio test for general mixture models with possibly structural parameter, ESAIM, 13, p. 301-327 (2009).

[5] Azaïs (J.M.), Wschebor (M.). — Level sets and extrema of random processes and fields, Wiley, New-York (2009).

[6] Azaïs (J.M.), Delmas (C.), Rabier (C-E.). — 2012. Likelihood Ratio Test process for QTL detection, Statistics, DOI:10.1080/02331888.2012.760093.

[7] Broman (K.). — Use of hidden Markov models for QTL mapping, Technical Report (2006).

[8] Chang (M. N.), Wu (R.), Wu (S. S.), Casella (G.). — Score statistics for mapping quantitative trait loci, Statistical Application in Genetics and Molecular Biology, 8(1), 16 (2009).

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

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

[11] Darvasi (D.,) Soller (M.). — Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus, Theoretical and Applied Genetics, 85, p. 353-359 (1992).

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

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

[14] Feingold (E.), Brown (P.O.), Siegmund (D.). — Gaussian models for genetic linkage analysis using complete high-resolution maps of idendity by descent, Am. J. Human. Genet., 53, p. 234-251 (1993).

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

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

[17] Haldane (J.B.S.). — The combination of linkage values and the calculation of distance between the loci of linked factors, Journal of Genetics, 8, p. 299-309 (1919).

[18] Hayes (B.). — QTL Mapping, MAS, and Genomic Selection, Short course organized by Iowa State University (2007).

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

[20] Lebowitz (R.J.), Soller (M.), Beckmann (J.S.). — 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, p. 556-562 (1987).

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

[22] Rabier (C.-E.). — On stochastic processes for Quantitative Trait Locus mapping under selective genotyping, Statistics, DOI:10.1080/02331888.2013.858720 (2013).

[23] Rabier (C.-E.). — On statistical inference for selective genotyping, Journal of Statistical Planning and Inference, 147, p. 24-52 (2014).

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

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

[26] Siegmund (D.). — Sequential analysis: tests and confidence intervals, Springer, New York (1985).

[27] Siegmund (D.), Yakir (B.). — The statistics of gene mapping, Springer, New York (2007).

[28] Van der Vaart (A.W.). — Asymptotic statistics, Cambridge Series in Statistical and Probabilistic Mathematics (1998).

[29] Wu (R.), MA (C.X.), Casella (G.). — Statistical Genetics of Quantitative Traits, Springer, New York (2007).