

1.

## ON GENE MAPPING WITH THE MIXTURE MODEL AND THE EXTREMES

BY CHARLES-ELIE RABIER<sup>\*,†,‡</sup>, CÉLINE DELMAS<sup>‡</sup>

*ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France* <sup>\*</sup>

*LIRMM, UMR 5506, CNRS, Université de Montpellier, France* <sup>†</sup>

*MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France* <sup>‡</sup>

We introduce a new variable selection method, suitable when the correlation between regressors is known. It is appropriate in genomics since once the genetic map has been built, the correlation is perfectly known. Our method, based on the LASSO, is original since the number of selected variables is bounded by the number of predictors, instead of being bounded by the number of observations as in the classical LASSO. It is made possible by the construction of a specific statistical test, a transformation of the data and by the knowledge of the correlation between regressors. We prove that the signal to noise ratio is largely increased by considering the extremes. This new technique is inspired by stochastic processes arising from statistical genetics. Our approach and existing methods are compared for simulated and real data, and the results points to the validity of our approach.

### 1. Introduction and background.

1.1. *Preliminaries.* There are many issues related to high dimensional data. As mentioned in Fan and Lv [17], one of the challenge is that “important predictors can be highly correlated with some unimportant ones”. In genomics, correlation between predictors is highly linked to recombination between genetic markers. Then, once the genetic map is built (see [50] for instance), the correlation between predictors is perfectly known and we do not have to estimate these correlations. In this context, we propose in this study, to exploit this extra information and to introduce a new variable selection method.

The number of selected variables by our method, is bounded by the number of predictors, instead of being bounded by the number of observations

---

*AMS 2000 subject classifications:* Primary 60G15, 62F03, 62F05

*Keywords and phrases:* Gaussian process, Mixture model, Hypothesis testing, Extreme values, Selective genotyping, Quantitative Trait Locus detection

as in the classical LASSO ([46]). It is made possible by the construction of a specific statistical test, a transformation of the data and by this knowledge of the correlation between regressors. In other words, we transform the original problem of  $n$  observations and  $K$  predictors ( $K \geq n$ ), to a situation where  $K$  is now the number of observations and  $L$  the number of predictors ( $L \geq K$ ). The use of a mixture model allows us to look for variables that can be viewed as unobserved predictors. The quantity  $L$  denotes the total number of predictors (observed and unobserved).

Moreover, in high dimensional problem, it is well known that the sparse coefficient should be large enough (see for instance [10]) in order to recover the true model ( $\beta$ -min condition). We prove that the performances of our method can be largely increased by considering the extremes. Indeed, the signal to noise ratio can be largely improved with the help of the selective genotyping concept proposed by Lebowitz and al. [25], and heavily used in agronomy.

Our study, inspired by stochastic processes arising from biology, focuses on the backcross design (see below): the mathematical theory behind this concept has been largely studied for many years (e.g. [14]). Note that we could have focused on an evolutionary process such as the Wright Fisher model. It could be investigated in future research.

**1.2. A statistical genetic context.** 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 Quantitative Trait Loci, so-called QTL (genes 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  $Y_j$ ,  $j = 1, \dots, n$ , the observations, which we will assume to be 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. 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  $X(\cdot)$  is due to Haldane [20] which states that:

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

where for any  $b \in \mathbb{R}$ ,  $\delta_b$  denotes the point mass at  $b$  and  $N(\cdot)$  is a standard Poisson process on  $[0, T]$ . In a more practical point of view, the Haldane [20] model assumes no crossover interference and the Poisson process represents the number of crossovers on  $[0, T]$  which happen during meiosis.

**1.3. Analysis of variance model.** The quantitative trait  $Y$  is affected by  $m$  additive QTLs located on the chromosome. 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 [22]). Let  $q_s$  and  $t_s^*$  denote respectively the QTL effect and the location of the  $s$ th QTL. Besides, we will consider  $0 < t_1^* < \dots < t_m^* < T$ . We assume an “analysis of variance model” for the quantitative trait:

$$(1.1) \quad Y = \mu + \sum_{s=1}^m X(t_s^*) q_s + \sigma \varepsilon$$

where  $\varepsilon$  is a Gaussian white noise.

**1.4. Mixture model.** In fact the “genome information” will be available only at marker locations, that is to say at certain fixed locations  $t_1 = 0 < t_2 < \dots < t_K = T$ , and the observation will be

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

So, we observe  $n$  observations  $(Y_j, X_j(t_1), \dots, X_j(t_K))$  i.i.d.

The aim of this study is to estimate the number  $m$  of QTLs, their locations  $t_1^*, \dots, t_m^*$  and their effects  $q_1, \dots, q_m$ . If the QTLs were located exactly on marker locations, a classical way to solve this problem would be to perform the “least absolute shrinkage and selection operator”, so called LASSO ([46]) using as regressors the genome information at marker locations. However, since QTLs lie on the chromosome at unknown locations, we should not look for QTLs only at marker locations but rather focus on the whole chromosome. As a result, our problem can not be solved using classical variable selection tools: QTL mapping requires the use of mixture models in a way we will explain below. In what follows,  $r(t, t')$  will denote the probability of recombination between two loci (i.e. positions) located at  $t$  and  $t'$ . Calculations on the Poisson distribution show that

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

we set in addition

$$\bar{r}(t, t') = 1 - r(t, t'), \quad \rho(t, t') = e^{-2|t-t'|}.$$

When there is only one QTL (i.e.  $m = 1$ ), conditionally to  $X(t_1), \dots, X(t_K)$ ,  $Y$  obeys to a mixture model with known weights

$$(1.2) \quad p(t_1^*)f_{(\mu+q_1, \sigma)}(\cdot) + \{1 - p(t_1^*)\} f_{(\mu-q_1, \sigma)}(\cdot),$$

where  $f_{(\mu, \sigma)}$  is the Gaussian density with parameters  $(\mu, \sigma)$  and where  $p(t_1^*)$  is the probability  $\mathbb{P}(X(t_1^*) = 1)$  conditionally to the observations of the markers. It can be expressed from the functions  $r$  and  $\bar{r}$  (see formula 3.1 for  $K = 2$ , and formula 4.1 for  $K > 2$ ).

1.5. *The “Interval Mapping” of Lander and Botstein [24].* In a famous article, Lander and Botstein [24] proposed to test the presence of the QTL (i.e.  $m = 0$  vs  $m = 1$ ), performing a likelihood ratio test (LRT) of the null hypothesis “ $q_1 = 0$ ” in equation (1.2). Since  $t_1^*$  is unknown, the authors suggested to scan the chromosome and to perform a LRT,  $\Lambda_n(t)$ , at each location  $t$  of the interval  $[0, T]$ . It leads to a “LRT process”,  $\Lambda_n(\cdot)$ , and considering the supremum of  $\Lambda_n(\cdot)$  gives the LRT of “ $q_1 = 0$ ” on the whole chromosome. Note that when the null hypothesis of the absence of QTL on  $[0, T]$  is rejected,  $\arg \sup \Lambda_n(t)$  is a natural estimator of the QTL location. This method, very popular in genetics, is called the “Interval Mapping”. The distribution of the LRT statistic,  $\sup \Lambda_n(\cdot)$ , has been given using some approximations by [14, 1, 5]. In [42, 41, 11], the authors focus only on the null hypothesis and are still using some approximations. Theoretical results are also present in [12] under non contiguous hypotheses. However, geneticists are usually interested in detecting QTLs with small effects (see [22]). Then, in Azaïs et al. [2], we have recently given the exact asymptotic distribution of the LRT statistic under the null and contiguous hypotheses. We showed that the LRT process,  $\Lambda_n(\cdot)$ , is asymptotically the square of a “non linear interpolated process” centered under  $H_0$  (i.e. no QTL on the chromosome) and uncentered of a mean function under the alternative which depends on the QTL effect  $q_1$  and its location  $t_1^*$ . Then, we presented a formula (due to the interpolation) to compute the supremum of  $\Lambda_n(\cdot)$ .

1.6. *First contribution: Asymptotic results on max test and LRT process, and a new gene mapping method.* The problem is that the use of the test statistic  $\sup \Lambda_n(\cdot)$  is appropriate for testing and localizing one QTL on  $[0, T]$ , but it is not so rewarding when more than one QTL (i.e.  $m > 1$ ) lie on  $[0, T]$ . When there are  $m$  QTLs, conditionally on  $X(t_1), \dots, X(t_K)$ ,  $Y$  obeys to a mixture of  $2^m$  components

$$(1.3) \quad \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} w_{\tilde{t}^*}(u_1, \dots, u_m) f_{(\mu+u_1 q_1 + \dots + u_m q_m, \sigma)}(Y)$$

where  $w_{\vec{t}^*}(u_1, \dots, u_m)$  is the probability  $\mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\}$  conditionally on the observations of the markers (see Sections 3 and 4).

In this paper, we propose to generalize the results of Azaïs et al. [2] to the general alternative that there exist  $m$  QTLs on  $[0, T]$  at  $t_1^*, \dots, t_m^*$  with additive effects  $q_1, \dots, q_m$ . We will show that under the general alternative, the LRT process is still asymptotically the square of a “non linear interpolated process”. However, the mean function depends this time on the number of QTLs, their positions and their effects. This theoretical result allows us to propose a new method to estimate the number of QTLs, their positions and their effects using the LASSO.

Note that we will also give the asymptotic distribution of the statistic  $\sup \Lambda_n(\cdot)$  when  $m > 1$ , since this test can be viewed as a global test or max test (see for instance [6]). In this context,  $\sup \Lambda_n(\cdot)$  matches the test statistic corresponding to the statistical test with the smallest pvalue in a multiple testing framework. It could be used before performing our new gene mapping method, in order to look for “some signal” on the chromosome.

Section 8.1 illustrates on simulated data, our theoretical result regarding the max test. We will show that the empirical power matches the theoretical power for moderate values of  $n$ . A way to increase the signal to noise ratio is to focus on the extremes (see below).

*1.7. Some background about the use of the extremes in genomics: the selective genotyping.* In the past, collecting the genome information at one marker for all the individuals was very expensive. In such a context, Lebowitz and al. [25] proposed to genotype only the individuals who present an extreme phenotype (i.e. the smallest and the largest  $Y$ ), since they noticed that most of the information about the QTL is present in the extreme phenotypes. This way, at a given power, a large increase of the number of individuals leads to a decrease of the number of individuals genotyped. Later, Lander and Botstein [24] formalized this approach and called it “selective genotyping”. This design has been studied theoretically by many authors considering only one fixed location of the genome (e.g. [25], [24], [15], [30], [33]). More recently, in Rabier [37], we investigated the asymptotic properties of the LRT statistic on the chromosome: it can be viewed as an answer to the simulation study presented by Rabbee et al. [32].

The model corresponding to selective genotyping is the following: we consider two real thresholds  $S_-$  and  $S_+$ , with  $S_- \leq S_+$  and we genotype if and only if the phenotype  $Y$  is extreme, that is to say  $Y \leq S_-$  or  $Y \geq S_+$ . Note that in practice, the cutoffs for genotyping are based on quantiles. However, in most of the theoretical studies about selective genotyping, authors con-

sider 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

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

Note that with our notations :

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

When there is only one QTL (i.e.  $m = 1$ ), we have proved (see [37]) that the probability distribution of  $(Y, \overline{X}(t_1), \dots, \overline{X}(t_K))$  is proportional to the mixture

$$(1.4) \quad p(t_1^*) f_{(\mu+q_1, \sigma)}(Y) 1_{Y \notin [S_-, S_+]} + \{1 - p(t_1^*)\} f_{(\mu-q_1, \sigma)}(Y) 1_{Y \notin [S_-, S_+]} \\ + \frac{1}{2} f_{(\mu+q_1, \sigma)}(Y) 1_{Y \in [S_-, S_+]} + \frac{1}{2} f_{(\mu-q_1, \sigma)}(Y) 1_{Y \in [S_-, S_+]}.$$

Recall that the function  $p(t_1^*)$  is the probability  $\mathbb{P}(X(t_1^*) = 1 \mid X(t_1), \dots, X(t_K))$ . Note that although  $p(t_1^*)$  is a function of  $X(t_1), \dots, X(t_K)$ , the quantity  $p(t_1^*) 1_{Y \notin [S_-, S_+]}$  present in (1.4) is a function of  $\overline{X}(t_1), \dots, \overline{X}(t_K)$ . In this context, we have proved ([37]) that the LRT process,  $\Lambda_n(\cdot)$ , converges to the square of a non linear interpolated process. This limiting process is the same as the one of the complete data situation (as above) except that the mean functions are proportional of a factor linked to the selective genotyping.

*1.7.1. Second contribution: Asymptotic results regarding the extremes (selective genotyping).* As explained before, we propose to tackle in this study, the problem of recovering several genes lying on the genome. So, under selective genotyping, when there are  $m$  QTLs, the probability distribution of  $(Y, \overline{X}(t_1), \dots, \overline{X}(t_K))$  is proportional to the mixture of  $2^m$  components

$$\sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} w_{\vec{t}^*}(u_1, \dots, u_m) f_{(\mu+u_1 q_1 + \dots + u_m q_m, \sigma)}(Y) 1_{Y \notin [S_-, S_+]} \\ + v_{\vec{t}^*}(u_1, \dots, u_m) f_{(\mu+u_1 q_1 + \dots + u_m q_m, \sigma)}(Y) 1_{Y \in [S_-, S_+]}$$

where  $v_{t^*}(u_1, \dots, u_m)$  is the probability  $\mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m)$  and where  $w_{t^*}(u_1, \dots, u_m)$  is the same quantity as previously. Explanations about this mixture model are given in Section 6.

We will show that the mean function of the LRT process is still proportional to the one of the complete data situation. Besides, we will compare theoretically the case where the  $n$  genotyped individuals are extreme or not.

In Section 8.2, we will present a simulation study to illustrate our new gene mapping method under selective genotyping. As expected, the signal to noise ratio is largely increased by considering extreme individuals. Recall that our proposed method takes into account explicitly the fact that the individuals are extreme, since it relies on the mean function of the LRT process under selective genotyping. Section 8.3 compares our new method and existing methods. According to a simulation study, our method outperforms existing methods specially when the selective genotyping is performed unilaterally. For instance, in dairy cattle, only the superior animals are genotyped ([8]). Besides, our L1 penalization method dedicated to the extremes, is easy to use and does not require a threshold as “Interval Mapping” methods. Last, Section 8.4 is devoted to a rice data analysis.

*1.7.2. Extra models studied in this paper.* In this article, we will also investigate the asymptotic properties of the LRT process regarding two other models present in the statistical genetics literature: the epistatic model and the interference model. First, it is well known that interactions between QTLs (so-called epistasis phenomenon) can be responsible for a non-negligible part of the genetic variability of a quantitative trait (see for instance [50]). On the other hand, biologists have observed that a recombination event can inhibit the formation of another recombination event nearby (e.g. [45, 27]). As a result, we will also deal with a model incorporating this interference phenomenon. Recall that Haldane modeling assumes that crossovers occur independently along the genome.

**2. Roadmap.** In this article, we will present the following theorems and lemmas corresponding to the different models studied:

- Theorem 3.1: Two genetic markers
- Theorem 4.1: Several markers
- Theorem 5.1: Epistasis
- Theorem 6.1: Selective genotyping
- Lemma 6.1: Asymptotic Relative Efficiency for selective genotyping
- Theorem 6.2: Selective genotyping and epistasis

Section 7 introduces our new gene mapping method, and Section 8 illustrates the different results on simulated and real data. Note that in [Supplement C](#), we will introduce Lemma 6.3 based on the reverse configuration of selective genotyping, and we will also present Theorem 7.1 and Theorem 7.2 dealing with interference, and interference combined with epistasis, respectively.

**3. Two genetic markers.** To begin, we suppose that there are only two markers ( $K = 2$ ) located at 0 and  $T : 0 = t_1 < t_2 = T$ . Besides, let us consider the case  $m = 1$  (i.e one QTL located at  $t_1^*$ ). For  $t \in [t_1, t_2]$  we define

$$p(t) = \mathbb{P} \{X(t) = 1 | X(t_1), X(t_2)\}$$

and

$$x(t) = \mathbb{E} \{X(t) | X(t_1), X(t_2)\} = 2p(t) - 1.$$

It is clear that  $p(t_1^*)$  is effectively the probability appearing in (1.2). An application of the rule of total probabilities leads to

$$(3.1) \quad \begin{aligned} p(t) = & Q_t^{1,1} 1_{X(t_1)=1} 1_{X(t_2)=1} + Q_t^{1,-1} 1_{X(t_1)=1} 1_{X(t_2)=-1} \\ & + Q_t^{-1,1} 1_{X(t_1)=-1} 1_{X(t_2)=1} + Q_t^{-1,-1} 1_{X(t_1)=-1} 1_{X(t_2)=-1} \end{aligned}$$

where

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

We can notice 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  $\theta^1 = (q_1, \mu, \sigma)$  be the parameter of the model at  $t$  fixed. The likelihood of the triplet  $(Y, X(t_1), X(t_2))$  with respect to the measure  $\lambda \otimes N \otimes N$ ,  $\lambda$  being the Lebesgue measure,  $N$  the counting measure on  $\mathbb{N}$ , is  $\forall t \in [t_1, t_2]$ :

$$(3.2) \quad L_t(\theta^1) = [p(t) f_{(\mu+q_1, \sigma)}(Y) + \{1 - p(t)\} f_{(\mu-q_1, \sigma)}(Y)] g(t)$$

where the function

$$(3.3) \quad \begin{aligned} g(t) = & \frac{1}{2} \{ \bar{r}(t_1, t_2) 1_{X(t_1)=1} 1_{X(t_2)=1} + r(t_1, t_2) 1_{X(t_1)=1} 1_{X(t_2)=-1} \} \\ & + \frac{1}{2} \{ r(t_1, t_2) 1_{X(t_1)=-1} 1_{X(t_2)=1} + \bar{r}(t_1, t_2) 1_{X(t_1)=-1} 1_{X(t_2)=-1} \} \end{aligned}$$



can be removed because it does not depend on the parameters. By a small abuse of notation we still denote  $L_t(\theta^1)$  for the likelihood without this function. Thus we set

$$L_t(\theta^1) = [p(t)f_{(\mu+q_1, \sigma)}(Y) + \{1 - p(t)\}f_{(\mu-q_1, \sigma)}(Y)]$$

and  $l_t(\theta^1)$  will be the loglikelihood.

Before defining the score statistic and the LRT statistic at  $t$ , let us introduce the notation  $\theta_0^1 = (0, \mu, \sigma)$  which will refer to the parameter  $\theta^1$  under  $H_0$ . Since the Fisher Information matrix is diagonal (cf. Section 1.1 of [Supplement A](#)), the score statistic of the hypothesis “ $q_1 = 0$ ” at  $t$ , for  $n$  independent observations, will be defined as

$$(3.4) \quad S_n(t) = \frac{\frac{\partial l_t^n}{\partial q_1} | \theta_0^1}{\sqrt{\mathbb{V}_{H_0} \left( \frac{\partial l_t^n}{\partial q_1} | \theta_0^1 \right)}},$$

where  $l_t^n(\theta^1)$  denotes the log-likelihood at  $t$ , associated to  $n$  observations.

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

$$\Lambda_n(t) = 2 \left\{ l_t^n(\hat{\theta}^1) - l_t^n(\hat{\theta}^1|_{H_0}) \right\},$$

where  $\hat{\theta}^1$  is the maximum likelihood estimator (MLE), and  $\hat{\theta}^1|_{H_0}$  the MLE under  $H_0$ . As previously said,  $\sup \Lambda_n(\cdot)$  is the LRT statistic of  $q_1 = 0$  on the whole interval  $[0, T]$ . We refer to Azaïs et al. [\[2\]](#) for the asymptotic distributions of the LRT statistic.

Let us now suppose that more than one QTL (i.e.  $m > 1$ ) lie on  $[0, T]$ . In what follows,  $t_1^*, \dots, t_m^*$  will denote the QTL locations, and we define the parameter  $\theta^m$  and  $\theta_0$  in the following way :  $\theta^m = (q_1, \dots, q_m, \mu, \sigma)$  and  $\theta_0^m = (0, \dots, 0, \mu, \sigma)$ . Then, the full likelihood of the triplet  $(Y, X(t_1), X(t_2))$ , with respect to the measure  $\lambda \otimes N \otimes N$ , is

$$(3.5) \quad L_{t^*}^m(\theta^m) = \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} w_{t^*}(u_1, \dots, u_m) f_{(\mu+u_1q_1+\dots+u_mq_m, \sigma)}(Y) \\ \times g^m(t_1^*, \dots, t_m^*)$$

where the function  $g^m(\cdot)$  is equal to the function  $g(\cdot)$  given in formula [\(3.3\)](#) and where

$$w_{t^*}(u_1, \dots, u_m) = \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \mid X(t_1), X(t_2) \}.$$

Note that at this time, it is clear that  $g^m(\cdot)$  does not depend on  $t_1^*, \dots, t_m^*$ . However, these parameters will be useful in the generalization in the next

Section. Note that calculations on the Poisson process lead to (proof included in Section 1 of [Supplement A](#)).

$$\begin{aligned}
w_{\vec{t}^\star}(u_1, \dots, u_m) &= \{r(t_1, t_1^\star) 1_{X(t_1)u_1=-1} + \bar{r}(t_1, t_1^\star) 1_{X(t_1)u_1=1}\} \\
&\times \{r(t_1^\star, t_2^\star) 1_{u_1u_2=-1} + \bar{r}(t_1^\star, t_2^\star) 1_{u_1u_2=1}\} \\
&\times \dots \times \{r(t_{m-1}^\star, t_m^\star) 1_{u_{m-1}u_m=-1} + \bar{r}(t_{m-1}^\star, t_m^\star) 1_{u_{m-1}u_m=1}\} \\
&\times \{r(t_m^\star, t_2) 1_{u_mX(t_2)=-1} + \bar{r}(t_m^\star, t_2) 1_{u_mX(t_2)=1}\} / \{r(t_1, t_2) 1_{X(t_1)X(t_2)=-1} + \\
&\bar{r}(t_1, t_2) 1_{X(t_1)X(t_2)=1}\}.
\end{aligned}$$

As previously, we set  $L_{\vec{t}^\star}^m(\theta^m)$  the likelihood without the function  $g(\cdot)$ . Note that since  $p(t^\star)$  is equal to

$$\sum_{(u_2, \dots, u_m) \in \{-1, 1\}^{m-1}} \mathbb{P}\{X(t_1^\star) = 1, X(t_2^\star) = u_2, \dots, X(t_m^\star) = u_m \mid X(t_1), X(t_2)\},$$

we have the relationship

$$L_{\vec{t}^\star}^m(\theta_{10\dots 0}^m) = L_{t_1^\star}(\theta^1) \text{ where } \theta_{10\dots 0}^m = (q_1, 0, \dots, 0, \mu, \sigma).$$

In the same way, under  $H_0$ ,

$$L_{\vec{t}^\star}^m(\theta_0^m) = L_{t_1^\star}(\theta_0^1).$$

NOTATIONS 3.1.  $\Rightarrow$  is the weak convergence,  $\xrightarrow{F.d.}$  is the convergence of finite-dimensional distributions and  $\xrightarrow{\mathcal{L}}$  is the convergence in distribution.

Our main result is the following:

THEOREM 3.1. Suppose that the parameters  $(q_1, \dots, q_m, \mu, \sigma^2)$  vary in a compact and that  $\sigma^2$  is bounded away from zero, and also that  $m$  is finite. Let  $H_0$  be the null hypothesis of no QTL on  $[0, T]$ , and let define the following local alternatives  $H_{at^\star}$ : “there are  $m$  QTLs located respectively at  $t_1^\star, \dots, t_m^\star$  with effect  $q_1 = a_1/\sqrt{n}, \dots, q_m = a_m/\sqrt{n}$  where  $a_1 \neq 0, \dots, a_m \neq 0$ ”. Then,

$$S_n(\cdot) \Rightarrow Z(\cdot) \text{ , } \Lambda_n(\cdot) \xrightarrow{F.d.} Z^2(\cdot) \text{ , } \sup \Lambda_n(\cdot) \xrightarrow{\mathcal{L}} \sup Z^2(\cdot)$$

as  $n$  tends to infinity, under  $H_0$  and  $H_{at^\star}$  where  $Z(\cdot)$  is the Gaussian process with unit variance such as

$$\begin{aligned}
Z(t) &= \frac{\alpha(t) Z(t_1) + \beta(t) Z(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1, t_2)}} \text{ ,} \\
\text{Cov}\{Z(t_1), Z(t_2)\} &= \rho(t_1, t_2) = e^{-2|t_1 - t_2|}
\end{aligned}$$

where  $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$ ,  $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$  and with mean function

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

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

where

$$m_{t^*}(t_1) = \sum_{s=1}^m a_s \rho(t_1, t_s^*) / \sigma, \quad m_{t^*}(t_2) = \sum_{s=1}^m a_s \rho(t_2, t_s^*) / \sigma.$$

A proof is given in Section 1 of [Supplement A](#). This theorem is a generalization of Theorem 2.1 of Azaïs et al. [2], that considers only one QTL on the chromosome. According to Theorem 3.1, under the general alternative, the LRT process is still asymptotically the square of a non linear interpolated process. However, the mean function depends this time on the number of QTLs, their positions and their effects.

**4. Several markers.** In that case suppose that there are  $K$  markers  $0 = t_1 < t_2 < \dots < t_K = T$ . We consider values  $t, t'$  or  $t^*$  of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions. For  $t \in [t_1, t_K] \setminus \mathbb{T}_K$  where  $\mathbb{T}_K = \{t_1, \dots, t_K\}$ , we define  $t^\ell$  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”  $(t^\ell, t^r)$ .

Let us briefly describe the changes with the previous section. Due to the independent increments of the Poisson process,

$$\mathbb{P} \{X(t) \mid X(t_1), \dots, X(t_K)\} = \mathbb{P} \{X(t) \mid X(t^\ell), X(t^r)\}.$$

As a consequence, the likelihood ratio test  $\Lambda_n(t)$  is now built on the likelihood of the triplet  $(Y, X(t^\ell), X(t^r))$  and the quantities  $p(t)$  and  $g(t)$ , introduced in formulae (3.1) and (3.3), become

$$(4.1) \quad p(t) = \mathbb{P} \{X(t) = 1 \mid X(t^\ell), X(t^r)\}, \quad g(t) = \mathbb{P} \{X(t^\ell), X(t^r)\}.$$

Recall that our test statistic  $\Lambda_n(t)$ , is the LRT corresponding to the test of the presence of only one QTL at  $t$ . Let us now consider the true probability distribution. Since  $\mathbb{P} \{X(t_1^*), \dots, X(t_m^*) \mid X(t_1), \dots, X(t_K)\}$  is equal to

$$\mathbb{P} \{X(t_1^*), \dots, X(t_m^*) \mid X(t_1^{\ell}), X(t_1^r), \dots, X(t_m^{\ell}), X(t_m^r)\}$$

all the information is contained in the markers flanking the QTL locations. As a result, the focus is on the probability distribution of  $(Y, X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r}))$  and the quantities  $w_{\vec{t}^\star}(u_1, \dots, u_m)$  and  $g^m(\cdot)$  present in formula (3.5), verify now

(4.2)

$$\begin{aligned} & w_{\vec{t}^\star}(u_1, \dots, u_m) \\ &= \mathbb{P} \left\{ X(t_1^\star) = u_1, \dots, X(t_m^\star) = u_m \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r}) \right\} \end{aligned}$$

and

$$(4.3) \quad g^m(t_1^\star, \dots, t_m^\star) = \mathbb{P} \left\{ X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r}) \right\}.$$

Let us now introduce Theorem 4.1.

**THEOREM 4.1.** *We have the same result as in Theorem 3.1, provided that we make some adjustments and that we redefine  $Z(\cdot)$  in the following way :*

- *in the definition of  $\alpha(t)$  and  $\beta(t)$ ,  $t_1$  becomes  $t^\ell$  and  $t_2$  becomes  $t^r$*
- *under the null hypothesis, the process  $Z(\cdot)$  considered at marker positions is the "skeleton" of an Ornstein-Uhlenbeck process: the stationary Gaussian process with covariance  $\rho(t_k, t_{k'}) = \exp(-2|t_k - t_{k'}|)$*
- *at the other positions,  $Z(\cdot)$  is obtained from  $Z(t^\ell)$  and  $Z(t^r)$  by interpolation and normalization using the functions  $\alpha(t)$  and  $\beta(t)$*
- *at the marker positions, the expectation is such as*  

$$m_{\vec{t}^\star}(t_k) = \sum_{s=1}^m a_s \rho(t_k, t_s^\star) / \sigma$$
- *at other positions, the expectation is obtained from  $m_{\vec{t}^\star}(t^\ell)$  and  $m_{\vec{t}^\star}(t^r)$  by interpolation and normalization using the functions  $\alpha(t)$  and  $\beta(t)$ .*

The proof of the theorem is the same as the proof of Theorem 3.1 since we can limit our attention to the interval  $(t^\ell, t^r)$  when considering a unique instant  $t$ .

**5. Epistasis.** It is well known that interactions between QTLs (so-called epistasis phenomenon) can be responsible for a non-negligible part of the genetic variability of a quantitative trait (see for instance [50]). Then, we propose now to include interactions between QTLs into our model. We will assume that only loci with additive effects on the trait, are involved in interactions. The "analysis of variance model" of formula (1.1) for the

quantitative trait becomes

$$(5.1) \quad Y = \mu + \sum_{s=1}^m X(t_s^*) q_s + \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m X(t_s^*) X(t_{\tilde{s}}^*) q_{s,\tilde{s}} + \sigma \varepsilon$$

where  $\varepsilon$  is a Gaussian white noise, and  $q_{s,\tilde{s}}$  is the interaction effect between loci  $t_s^*$  and  $t_{\tilde{s}}^*$ .

The probability distribution of  $(Y, X(t_1^*), X(t_1^*), \dots, X(t_m^*), X(t_m^*))$  is

$$(5.2) \quad \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} w_{\vec{t}^*}(u_1, \dots, u_m) f_{(\mu + \sum_{s=1}^m u_s q_s + \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m u_s u_{\tilde{s}} q_{s,\tilde{s}}, \sigma)}(Y) \times g^m(t_1^*, \dots, t_m^*)$$

where  $w_{\vec{t}^*}(u_1, \dots, u_m)$  and  $g^m(\cdot)$  are given in formulae (4.2) and (4.3).

**THEOREM 5.1.** *Suppose that the parameters  $(q_1, \dots, q_m, q_{12}, \dots, q_{m-1,m}, \mu, \sigma^2)$  vary in a compact and that  $\sigma^2$  is bounded away from zero, and also that  $m$  is finite. Let define the local alternative*

- $H_{a\vec{t}^*, b\vec{t}^*}$ : “There are  $m$  additive QTLs located respectively at  $t_1^*, \dots, t_m^*$  with effects respectively  $q_1 = a_1/\sqrt{n}, \dots, q_m = a_m/\sqrt{n}$  where  $a_1 \neq 0, \dots, a_m \neq 0$ . Besides, all these QTLs interact with each other: the interaction effects are respectively  $q_{1,2} = b_{1,2}/\sqrt{n}$  for loci  $t_1^*$  and  $t_2^*$ ,  $\dots, q_{m-1,m} = b_{m-1,m}/\sqrt{n}$  for loci  $t_{m-1}^*$  and  $t_m^*$  where  $b_{1,2} \neq 0, \dots, b_{m-1,m} \neq 0$ ”.

then, with the previous notations, under  $H_{a\vec{t}^*, b\vec{t}^*}$ ,

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

where  $Z(\cdot)$  is the Gaussian process of Theorem 4.1 uncentered with mean function  $m_{\vec{t}^*}(\cdot)$  defined in Theorem 4.1.

A proof is given in Section 2 of [Supplement A](#). Note that the interaction effects are not included in the mean function. In other words, those effects are unidentifiable when the classical LRT is used. It is due to independent increments of the Poisson process.

**6. Selective genotyping.** We propose to consider here the classical problem (as in Sections 3 and 4), but incorporating now a selective genotyping in order to reduce the costs of genotyping. To begin with, in order to

make the reading easier, we won't consider interactions in our model. However, the epistasis will be investigated later in this section. As mentioned in Section 1.7, the selective genotyping model is the following: we consider two real thresholds  $S_-$  and  $S_+$ , with  $S_- \leq S_+$  and we genotype if and only if the phenotype  $Y$  is extreme, that is to say  $Y \leq S_-$  or  $Y \geq S_+$ .

If we call  $\overline{X}(t)$  the random variable such as

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

then, in our problem, one observation will be now

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

Note that with our notations :

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

To begin with, let us consider the case  $m = 1$ . According to Rabier [37], the likelihood of the triplet  $(Y, \overline{X}(t^\ell), \overline{X}(t^r))$  with respect to the measure  $\lambda \otimes N \otimes N$ ,  $\lambda$  is  $\forall t \in [t_1, t_K] \setminus \mathbb{T}_K$  :

(6.1)

$$\begin{aligned} \overline{L}_t(\theta^1) = & \left[ p(t) f_{(\mu+q_1, \sigma)}(Y) 1_{Y \notin [S_-, S_+]} + \{1 - p(t)\} f_{(\mu-q_1, \sigma)}(Y) 1_{Y \notin [S_-, S_+]} \right. \\ & \left. + \frac{1}{2} f_{(\mu+q_1, \sigma)}(Y) 1_{Y \in [S_-, S_+]} + \frac{1}{2} f_{(\mu-q_1, \sigma)}(Y) 1_{Y \in [S_-, S_+]} \right] \overline{g}(t) \end{aligned}$$

with

$$(6.2) \quad \overline{g}(t) = \mathbb{P} \left\{ X(t^\ell), X(t^r) \right\} 1_{Y \notin [S_-, S_+]} + 1_{Y \in [S_-, S_+]}.$$

Note that we use the same notations  $p(t)$  for the weights since they are exactly the same as in Section 4. Recall that

$$\begin{aligned} p(t) = & \mathbb{P} \left\{ X(t) = 1 \mid X(t^\ell), X(t^r) \right\} \\ = & Q_t^{1,1} 1_{X(t^\ell)=1} 1_{X(t^r)=1} + Q_t^{1,-1} 1_{X(t^\ell)=1} 1_{X(t^r)=-1} \\ & + Q_t^{-1,1} 1_{X(t^\ell)=-1} 1_{X(t^r)=1} + Q_t^{-1,-1} 1_{X(t^\ell)=-1} 1_{X(t^r)=-1}. \end{aligned}$$

Although  $p(t)$  is not a function of  $X(t^\ell)$  and  $X(t^r)$ ,  $p(t)1_{Y \notin [S_-, S_+]}$  is the following function of  $\bar{X}(t^\ell)$  and  $\bar{X}(t^r)$ :

$$\begin{aligned} p(t)1_{Y \notin [S_-, S_+]} &= Q_t^{1,1} 1_{\bar{X}(t^\ell)=1} 1_{\bar{X}(t^r)=1} + Q_t^{1,-1} 1_{\bar{X}(t^\ell)=1} 1_{\bar{X}(t^r)=-1} \\ &\quad + Q_t^{-1,1} 1_{\bar{X}(t^\ell)=-1} 1_{\bar{X}(t^r)=1} + Q_t^{-1,-1} 1_{\bar{X}(t^\ell)=-1} 1_{\bar{X}(t^r)=-1} . \end{aligned}$$

In the same way, the quantity  $\mathbb{P}\{X(t^\ell), X(t^r)\} 1_{Y \notin [S_-, S_+]}$  present in the definition of  $\bar{g}(t)$  verifies

$$\begin{aligned} \mathbb{P}\{X(t^\ell), X(t^r)\} 1_{Y \notin [S_-, S_+]} &= \frac{1}{2} \left\{ \bar{r}(t^\ell, t^r) 1_{\bar{X}(t^\ell)\bar{X}(t^r)=1} \right. \\ &\quad \left. + r(t^\ell, t^r) 1_{\bar{X}(t^\ell)\bar{X}(t^r)=-1} \right\} . \end{aligned}$$

As a result, as expected, the likelihood is a function of  $Y, \bar{X}(t^\ell), \bar{X}(t^r)$ , which was not obvious at first reading. However, the expression given in formula (6.1) will be very convenient for the generalization to several QTLs.

The score statistic of the hypothesis “ $q_1 = 0$ ” at  $t$ , for  $n$  independent observations, will be defined as

$$\bar{S}_n(t) = \frac{\frac{\partial \bar{l}_t^n}{\partial q_1} |_{\theta_0^1}}{\sqrt{\mathbb{V}\left(\frac{\partial \bar{l}_t^n}{\partial q_1} |_{\theta_0^1}\right)}} ,$$

where  $\bar{l}_t^n(\theta^1)$  denotes the log likelihood at  $t$ , associated to  $n$  observations.

In the same way, the LRT at  $t$ , for  $n$  independent observations, will be defined as

$$\bar{\Lambda}_n(t) = 2 \left\{ \bar{l}_t^n(\hat{\theta}^1) - \bar{l}_t^n(\hat{\theta}^1|_{H_0}) \right\} ,$$

where  $\hat{\theta}^1$  is the maximum likelihood estimator (MLE), and  $\hat{\theta}^1|_{H_0}$  the MLE under  $H_0$ .

Let us now suppose that more than one QTL (i.e.  $m > 1$ ) lie on  $[0, T]$ . Using the same notations as in Sections 3 and 4,  $t_1^*, \dots, t_m^*$  denote the QTL locations, and the parameter  $\theta^m$  and  $\theta_0$  are defined in the following way :  $\theta^m = (q_1, \dots, q_m, \mu, \sigma)$  and  $\theta_0^m = (0, \dots, 0, \mu, \sigma)$ . Besides, recall that all the information is contained in the markers flanking the QTL locations. Then, the probability distribution of  $(Y, \bar{X}(t_1^{*\ell}), \bar{X}(t_1^{*r}), \dots, \bar{X}(t_m^{*\ell}), \bar{X}(t_m^{*r}))$ , with respect to the measure  $\lambda \otimes N \otimes \dots \otimes N$ , is

$$\begin{aligned} \bar{L}_{t^*}^m(\theta^m) &= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \left[ w_{t^*}(u_1, \dots, u_m) f_{(\mu+u_1 q_1 + \dots + u_m q_m, \sigma)}(Y) 1_{Y \notin [S_-, S_+]} \right. \\ (6.3) \quad &\quad \left. + v_{t^*}(u_1, \dots, u_m) f_{(\mu+u_1 q_1 + \dots + u_m q_m, \sigma)}(Y) 1_{Y \in [S_-, S_+]} \right] \bar{g}^m(t_1^*, \dots, t_m^*) \end{aligned}$$

with

$$\begin{aligned} v_{\vec{t}^*}(u_1, \dots, u_m) &= \mathbb{P}\{X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m\} , \\ \bar{g}^m(t_1^*, \dots, t_m^*) &= \mathbb{P}\left\{X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\} 1_{Y \notin [S_-, S_+]} \\ &\quad + 1_{Y \in [S_-, S_+]} . \end{aligned}$$

Recall also that the definition of  $w_{\vec{t}^*}(u_1, \dots, u_m)$  is given in formula (4.2). As before, formula (6.3) is a function of  $Y, \bar{X}(t_1^{\star\ell}), \dots, \bar{X}(t_m^{\star r})$ . Note that the proof of formula (6.3) is included in the proof of the following Theorem 6.1.

NOTATIONS 6.1.  $\gamma, \gamma_+$  and  $\gamma_-$  are respectively the quantities  $\mathbb{P}_{H_0}(Y \notin [S_-, S_+]), \mathbb{P}_{H_0}(Y > S_+)$  and  $\mathbb{P}_{H_0}(Y < S_-)$ .

NOTATIONS 6.2.  $\mathcal{A}$  is the quantity such as  $\mathcal{A} = \sigma^2 \{ \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) \}$ , where  $\varphi(x)$  and  $z_\alpha$  denote respectively the density of a standard normal distribution taken at the point  $x$ , and the quantile of order  $1 - \alpha$  of a standard normal distribution.

THEOREM 6.1. Suppose that the parameters  $(q_1, \dots, q_m, \mu, \sigma^2)$  vary in a compact and that  $\sigma^2$  is bounded away from zero, and also that  $m$  is finite. Then,

$$\bar{S}_n(\cdot) \Rightarrow V(\cdot) , \quad \bar{\Lambda}_n(\cdot) \xrightarrow{F.d.} V^2(\cdot) , \quad \sup \bar{\Lambda}_n(\cdot) \xrightarrow{\mathcal{L}} \sup V^2(\cdot)$$

as  $n$  tends to infinity, under  $H_0$  and  $H_{a\vec{t}^*}$  where  $V(\cdot)$  is the Gaussian process with unit variance such as

$$\begin{aligned} V(t) &= \frac{\alpha(t) V(t^\ell) + \beta(t) V(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} , \\ \text{Cov}\{V(t^\ell), V(t^r)\} &= \rho(t^\ell, t^r) = e^{-2|t^\ell - t^r|} \end{aligned}$$

where the functions  $\alpha(\cdot)$  and  $\beta(\cdot)$  are given in Theorem 3.1, and with mean function

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

$$\bar{m}_{\vec{t}^*}(t) = \frac{\alpha(t) \bar{m}_{\vec{t}^*}(t^\ell) + \beta(t) \bar{m}_{\vec{t}^*}(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}}$$

where

$$\bar{m}_{\vec{t}^*}(t^\ell) = \sum_{s=1}^m a_s \sqrt{\mathcal{A}} \rho(t^\ell, t_s^*) / \sigma^2 , \quad \bar{m}_{\vec{t}^*}(t^r) = \sum_{s=1}^m a_s \sqrt{\mathcal{A}} \rho(t^r, t_s^*) / \sigma^2 .$$



The proof is given in Section 3 of [Supplement A](#). Under the null hypothesis, despite the selective genotyping,  $V(\cdot)$  is exactly the same process as the process  $Z(\cdot)$  of Theorem 4.1 obtained for the complete data situation. However, under the general alternative, the mean functions of the two processes are not the same anymore : the mean functions are proportional of a factor  $\sqrt{\mathcal{A}}/\sigma$ .

Before introducing our Lemma 6.1, 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 the genome information at markers is known for all the individuals.

LEMMA 6.1. *Let  $\kappa$  denote the ARE, then we have*

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

This lemma is a generalization of Theorem 4.2 of Rabier [37] where the focus was only on the case  $m = 1$ . To prove Lemma 6.1, just use the same proof as the one of Theorem 4.2 of [37].

According to i) of Lemma 6.1, the ARE with respect to the complete data situation, does not depend on the number of QTLs  $m$ , the constants  $a_1, \dots, a_m$  linked to the QTL effects, and the QTLs locations  $t_1^*, \dots, t_m^*$ . 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 does not depend on those parameters. On the other hand, according to ii) of Lemma 6.1, if we want to genotype only a percentage  $\gamma$  of the population, we should genotype the  $\gamma/2\%$  individuals with the largest phenotypes and  $\gamma/2\%$  individuals with the smallest phenotypes.

Let us consider now  $n^*$  individuals for a selective genotyping experiment, and let us assume that we have the relationship  $n = n^*\gamma$ . In other words, we focus on the case where, for economical reasons, we are allowed to genotype only  $n$  individuals. By considering  $n = n^*\gamma$ , we are allowed to genotype  $n$  extreme individuals, provided that the overall population size has been increased to  $n^*$ . In this context, we have

$$\bar{S}_{n^*}(t_k) \xrightarrow{H_0} N \left( \sqrt{\frac{\mathcal{A}}{\gamma \sigma^4}} \sum_{s=1}^m a_s \rho(t_k, t_s^*), 1 \right)$$

and the mean function of the process is still interpolated. As a result, the ratio between the signal corresponding to selective genotyping and the one

matching the complete data situation is equal to  $\sqrt{\frac{\mathcal{A}}{\gamma \sigma^2}}$ . This quantity verifies the following relationship

$$\sqrt{\frac{\mathcal{A}}{\gamma \sigma^2}} = \sqrt{z_{\gamma_+} \varphi(z_{\gamma_+})/\gamma - z_{1-\gamma_-} \varphi(z_{1-\gamma_-})/\gamma + 1}$$

and if we are willing to genotype symmetrically (i.e.  $\gamma_+ = \gamma_-$ ), it becomes

$$\sqrt{\frac{\mathcal{A}}{\gamma \sigma^2}} = \sqrt{2z_{\gamma/2} \varphi(z_{\gamma/2})/\gamma + 1}.$$

In other words, provided that the phenotyping is free, the signal can be largely increased, by genotyping extreme individuals (i.e. selective genotyping) instead of genotyping random individuals (i.e. complete data situation). According to Figure 1, when the selective genotyping is performed symmetrically, the signal corresponding respectively to the cases  $\gamma = 0.1$ ,  $\gamma = 0.2$  and  $\gamma = 0.3$ , is respectively 2.09, 1.80 and 1.61 times larger under selective genotyping than under random genotyping. The worst case is obtained when genotyping only the largest phenotypes (see  $\gamma_+/\gamma = 1$ ) or genotyping only the smallest phenotypes (same curve as the one for  $\gamma_+/\gamma = 1$ ). In that case, the selective genotyping can be less rewarding than the random genotyping (cf.  $\gamma_+ = \gamma = 0.8$ ).

Obviously, when all the individuals are genotyped ( $\gamma = 1$ ), all the efficiencies are equal to one.

Let us now move on to the case where interactions are present into our model (see formula 5.1). Then, under selective genotyping, we have the following result:

**THEOREM 6.2.** *Suppose that the parameters  $(q_1, \dots, q_m, q_{1,2}, \dots, q_{m-1,m}, \mu, \sigma^2)$  vary in a compact and that  $\sigma^2$  is bounded away from zero, and also that  $m$  is finite. Then, with the previous notations, under  $H_{a\vec{t}^*, b\vec{t}^*}$ ,*

$$\bar{S}_n(\cdot) \Rightarrow V(\cdot) \quad , \quad \bar{\Lambda}_n(\cdot) \xrightarrow{F.d.} V^2(\cdot) \quad , \quad \sup \bar{\Lambda}_n(\cdot) \xrightarrow{\mathcal{L}} \sup V^2(\cdot)$$

where  $V(\cdot)$  is the Gaussian process of Theorem 6.1 uncentered with mean function  $m_{t^*}(\cdot)$  defined in Theorem 6.1.

The proof is given in Section 4 of [Supplement A](#). As under the complete data situation (Theorem 5.1), the interaction effects are not included in the mean function.

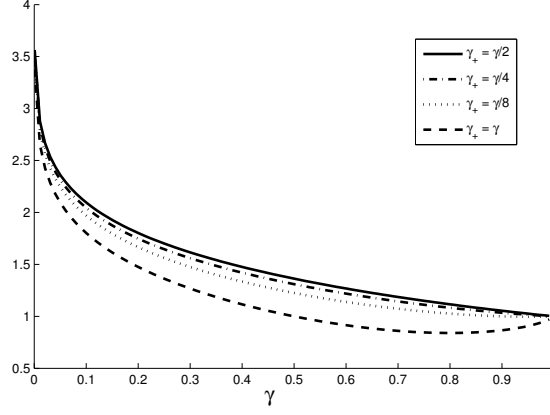


FIG 1. Function  $\sqrt{z_{\gamma+} \varphi(z_{\gamma+})/\gamma - z_{1-\gamma-} \varphi(z_{1-\gamma-})/\gamma + 1}$  as a function of the percentage  $\gamma$  of individuals genotyped and as a function of the ratio  $\gamma_+/\gamma$ .

**7. A new method for gene mapping.** In this section, the goal is to propose a method to estimate the number of QTLs, their effects and their positions combining results of the theorem and a penalized likelihood method. Note that in order to make the reading easier, we will introduce the method in the context of Sections 3 and 4 (Haldane mapping, no selective genotyping and no epistasis). However, it can easily be adapted to the different models studied in this paper.

**7.1. Introducing our method.** According to Theorem 4.1, as soon as we discretize the score process at markers positions, we have the following relationship when  $n$  is large:

$$\vec{S}_n = \vec{m}_{\vec{t}^*} + \vec{\varepsilon} + o_P(1)$$

where  $\vec{S}_n = (S_n(t_1), S_n(t_2), \dots, S_n(t_K))'$ ,  $\vec{m}_{\vec{t}^*} = (m_{\vec{t}^*}(t_1), m_{\vec{t}^*}(t_2), \dots, m_{\vec{t}^*}(t_K))'$  and  $\vec{\varepsilon} \sim N(0, \Sigma)$  with  $\Sigma_{kk'} = \rho(t_k, t_{k'})$ .

Since most of the penalized likelihood methods rely on i.i.d. observations, we will decorrelate the components of  $\vec{S}_n$  keeping only points of the process taken at marker positions. Recall that  $S_n(\cdot)$  is an “interpolated process”. Let us consider the Cholesky decomposition  $\Sigma = AA'$ . We have

$$A^{-1}\vec{S}_n = A^{-1}B \left( \frac{a_1}{\sigma}, \dots, \frac{a_m}{\sigma} \right)' + A^{-1}\vec{\varepsilon} + o_P(1)$$

where  $B$  is a matrix of size  $K \times m$  such as  $B_{ks} = e^{-2|t_k - t_s^*|}$ .

Since the number  $m$  of QTLs and their positions  $t_1^*, \dots, t_m^*$  are unknown, we

propose to focus on a new discretization of  $[0, T]$  corresponding to all the putative QTL locations:  $0 \leq t'_1 < t'_2 < \dots < t'_L \leq T$ .  $\Delta_1, \dots, \Delta_L$  will be the corresponding effects divided by  $\sigma$ . Note that although we focus only on the discretized process at markers locations, we look for QTL not only on markers. The model can be rewritten in the following way:

$$(7.1) \quad A^{-1}\vec{S}_n = A^{-1}C(\Delta_1, \dots, \Delta_L)' + A^{-1}\vec{\varepsilon} + o_P(1)$$

where  $C$  is a matrix of size  $K \times L$  such as  $C_{kl} = e^{-2|t_k - t'_l|}$ .

Last, in order to find the non zero  $\Delta_l$ , a natural approach is to use the L1 penalized regression, so-called LASSO ([46]):

$$\arg \min_{(\Delta_1, \dots, \Delta_L)} \left\| A^{-1}\vec{S}_n - A^{-1}C\Delta \right\|_2^2 + \zeta \|\Delta\|_1$$

where  $\|\cdot\|_2$  is the L2 norm,  $\|\cdot\|_1$  is the L1 norm,  $\Delta = (\Delta_1, \dots, \Delta_L)'$  and  $\zeta$  denotes the tuning parameter.  $\zeta$  can be estimated, for instance, by cross validation as described in Chapter 7 of Hastie and al. [21].

## 8. Illustrations.

8.1. *About the max test.* To begin with, in [Supplement B](#), we briefly illustrate our theoretical results regarding the max test. Recall that it relies on the test statistic,  $\sup \Lambda_n(\cdot)$ . The focus is on a sparse map: a chromosome of length 1M ( $T = 1$ ), with 21 markers ( $K = 21$ ) equally spaced every 5cM. In this context, we show (see Table 1 of [Supplement B](#)) that there is a good agreement between the empirical power and the theoretical power under different configurations ( $m$  is equal to either 1, 2 or 3, and  $\gamma$  is equal to 0.3 or 1). This validates our theoretical results presented in Theorems 4.1 and 6.1.

8.2. *Selective genotyping improves the detection process.* Figure 2, based on one simulated data set, illustrates the performances of our new gene mapping method (see Section 7) under selective genotyping. The considered genome is of length 10M ( $T = 10$ ), with 201 markers ( $K = 201$ ) equally spaced every 5cM. 16 QTLs ( $m = 16$ ) lie on the interval  $[0M, 4M]$  whereas no QTLs are present on the rest of the genome (i.e.  $[6M, 10M]$ ). The QTL effects are equal to either +0.2 or -0.2, each QTL having its own random sign. The presence of QTL is tracked every 2.5cM. As a consequence, 401 regressors ( $L = 401$ ) are present in the linear model (formula 7.1). In other words, we use the discretization  $t'_l = 0.025(l - 1)$ ,  $l = 1, \dots, 401$ . Recall that this grid is different from the one corresponding to marker locations:

$t_k = 0.05(k - 1)$ ,  $k = 1, \dots, 201$ . Figure 2A refers to the case  $n = 200$  whereas Figure 2B focuses on  $n = 100$ .

Assuming that, for economical reasons, the geneticist is allowed to genotype only  $n$  individuals, we compare here the case where those  $n$  individuals are extreme or not. In particular, when a selective genotyping was performed, the total number of individuals was increased to  $n^*$ , with the relationship  $n^* = n/\gamma$ . This way, on average,  $n$  individuals are genotyped under selective genotyping (cf. Section 6). For instance, when the sample size  $n$  was equal to 100 under the complete data situation ( $\gamma = 1$ ), we considered respectively 1000, 500 and 333 individuals to handle the cases  $\gamma = 0.1, 0.2$ , and  $0.3$  respectively. According to Figure 2A, the largest estimated effects are the ones corresponding to the case  $\gamma = 0.1$ : a few QTL effects are estimated at approximately 5 (see around 1M and 4M), and at  $-6$  around 2M. It was expected since under such selective genotyping (i.e with  $n^* = n/\gamma$ ), the quantities  $\Delta_l$ , present in formula (7.1), are increased by a factor  $\sqrt{A}/\sqrt{\gamma}$  at each gene location. Then, under the configuration studied, the quantities  $|a|\sqrt{A}/\sqrt{\gamma}$  are equal respectively to 5.92, 4.56 and 2.50 when  $\gamma$  takes respectively the values 0.1, 0.3, and 1. Note that the number of selected regressors was between 15 and 17 in all studied cases.

In what follows, the L1 ratio will denote the ratio L1 norm of estimated effects on  $[0M, 4M]$  to L1 norm of estimated effects on  $[0M, 10M]$ . This L1 ratio is an indicator of whether or not the detected QTLs belong to the “signal area”. Recall that on our example, all the simulated QTLs belong to the interval  $[0M, 4M]$ . Table 1 reports in a general framework, the mean L1 ratio over 100 samples of size  $n = 100$  or  $n = 200$ . Different QTL effects are taken into consideration :  $|q_s|$  is either equal to 0.2, 0.1, or 0.05. Since a large number of markers are now available in genomic studies, we also considered a dense map consisting in  $K = 10,001$  markers equally spaced every 0.1cM. Due to this high marker density, the presence of QTL was only investigated on markers ( $K = L$ ). For both maps (sparse and dense), we can notice that whatever the parameter values, the more extremes the genotyped individuals are, the larger the L1 ratio is. In other words, by considering extreme individuals, we largely improve the detection process. Besides, we can notice that the more markers there are, the more powerful the method is. Another interesting aspect (not shown here) is in the choice of the tuning parameter value. The MSE curve obtained by cross validation is flat under the complete data situation ( $\gamma = 1$ ), suggesting the absence of signal. In contrast, under selective genotyping, we can clearly distinguished the minimum of the curve, due to the increase of signal.

Last, Table 2 of [Supplement B](#) focuses on different ways of performing the

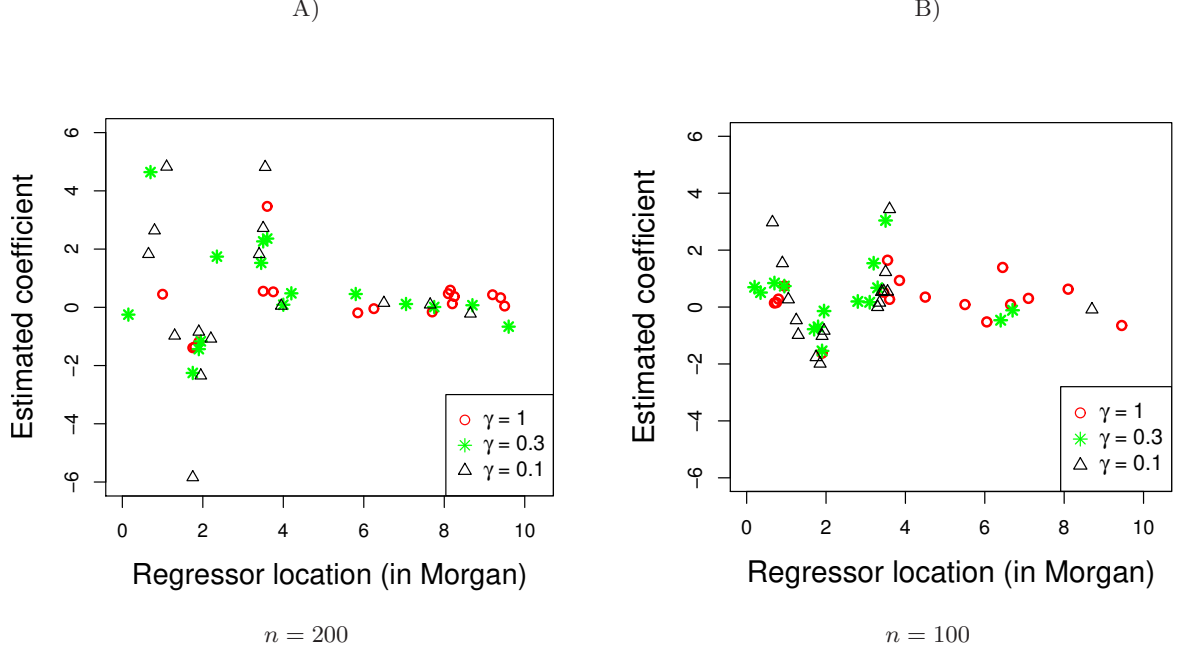


FIG 2. *Estimated coefficients according to our new method as a function of the percentage  $\gamma$  of genotyped individuals (1 sample,  $m = 16$ ,  $T = 10$ ,  $|q_1| = \dots = |q_{16}| = 0.2$ ,  $QTLs$  randomly located only on  $[0M, 4M]$ ,  $K = 201$ ,  $t_k = 0.05(k-1)$ ,  $L = 401$ ,  $t'_l = 0.025(k-1)$ ,  $\gamma_+/\gamma = 1/2$ , on average  $n$  individuals genotyped).*

selective genotyping: different ratios  $\gamma_+/\gamma$  are investigated under both maps. As expected, when only the largest (or the smallest) individuals are genotyped ( $\gamma_+/\gamma = 1$ ), the L1 ratio is the smallest. It confirms our theoretical results presented in Section 6 and illustrated in Figure 1.

To conclude, selective genotyping is largely more rewarding for localizing genes.

**8.3. Comparison with existing methods.** In this section, we propose to compare our new method with existing methods. We will concentrate on the LASSO, the Group LASSO ([51]), the “Interval Mapping” (IM) with either a parametric threshold or a non parametric threshold, and Lee and al. [26]’s method. In what follows, IM with a parametric threshold will be called IM(parametric), whereas the non parametric version will be called IM(shuffle). Note that LASSO methods focus only on marker locations, whereas IM, Lee and al. [26] and our method present the advantage to consider also locations between markers.

Recall that the Group LASSO differs from his cousin LASSO, since it

| all $ q_s $ | $\gamma$ | (Sparse, $n = 100$ ) |           | (Sparse, $n = 200$ ) |           | (Dense, $n = 100$ ) |           | (Dense, $n = 200$ ) |           |
|-------------|----------|----------------------|-----------|----------------------|-----------|---------------------|-----------|---------------------|-----------|
|             |          | L1 ratio             | $\hat{m}$ | L1 ratio             | $\hat{m}$ | L1 ratio            | $\hat{m}$ | L1 ratio            | $\hat{m}$ |
| 0.2         | 0.1      | 96.83%               | 14.75     | 99.61%               | 15.54     | 99.81%              | 17.2      | 99.88%              | 16.7      |
|             | 0.2      | 90.32%               | 18.17     | 97.99%               | 15.3      | 99.78%              | 17.35     | 99.64%              | 16.96     |
|             | 0.3      | 88.03%               | 17.45     | 95.84%               | 17.22     | 98.83%              | 17.25     | 99.72%              | 16.95     |
|             | 1        | 70.91%               | 18.47     | 82.57%               | 16.94     | 91.08%              | 16.69     | 98.36%              | 17.39     |
| 0.1         | 0.1      | 82.26%               | 14.74     | 91.29%               | 16.74     | 95.73%              | 17.15     | 98.39%              | 16.87     |
|             | 0.2      | 73.43%               | 15.64     | 85.43%               | 16.74     | 94.18%              | 17.61     | 96.26%              | 16.93     |
|             | 0.3      | 70.95%               | 16.59     | 83.48%               | 16.66     | 88.64%              | 16.70     | 96.50%              | 17.12     |
|             | 1        | 55.41%               | 18.57     | 62.35%               | 17.62     | 72.59%              | 16.23     | 88.37%              | 17.01     |
| 0.05        | 0.1      | 61.00%               | 15.06     | 68.66%               | 15.17     | 79.15%              | 16.08     | 87.25%              | 16.82     |
|             | 0.2      | 52.73%               | 15.07     | 63.70%               | 15.86     | 72.97%              | 16.47     | 80.58%              | 16.62     |
|             | 0.3      | 52.27%               | 15.38     | 68.24%               | 16.5      | 66.13%              | 17.39     | 79.91%              | 16.45     |
|             | 1        | 45.34%               | 15.64     | 46.49%               | 18.07     | 52.23%              | 16.8      | 67.40%              | 16.83     |

TABLE 1

Performances of the new method as a function of the percentage  $\gamma$  of genotyped individuals and as a function of the QTL effects (Mean over 100 samples,  $\gamma_+/\gamma = 1/2$ , on average  $n$  individuals genotyped,  $T = 10$ ,  $m = 16$ , QTLs randomly located only on  $[0M, 4M]$ ). Sparse map:  $K = 201$ ,  $t_k = 0.05(k - 1)$ ,  $L = 401$ ,  $t'_l = 0.025(k - 1)$ . Dense map:  $K = L = 10,001$ ,  $t_k = t'_l = 0.001(k - 1)$ . The L1 ratio corresponds to the quantity  $\sum_{i=1}^{161} |\hat{\Delta}_i| / \sum_{i=1}^{401} |\hat{\Delta}_i|$  for the sparse map, and to the quantity  $\sum_{i=1}^{4001} |\hat{\Delta}_i| / \sum_{i=1}^{10001} |\hat{\Delta}_i|$  for the dense map.  $\hat{m}$  denotes the estimated QTL number.

allows to handle a group structure. In the context of genomic prediction, the Bayesian LASSO was used in Boligon et al. [8] under selective genotyping. However, instead of studying the Bayesian LASSO, we decided to focus on the LASSO since it is one of the most famous method in the statistical literature and it makes our study more general. Indeed, our new method is based on the LASSO, and we found interesting to compare its performances with the ones of the classical LASSO (i.e. without transforming the data). Note that since the genome information was unknown for the non extreme individuals, we performed the LASSO only on the extreme individuals. The Group LASSO was based on groups of 10 consecutive markers.

On the other hand, IM(shuffle) was recently used in two recent studies dealing with selective genotyping experiments ([31, 49]). In order to perform the same permutation test as in these studies, we considered 1000 random permutations and we shuffled the phenotypes within the extremes as advised in Manichaikul et al. [28]. Recall that in our simulation framework, the number of QTLs  $m$  was set to 16. Then, in order to compare IM(shuffle) in a satisfactory manner with other studied methods, we adopted the following rules. First, as in [31, 49], an inferred QTL location was a location where the observed value of the test statistic was higher than the permutation threshold. However, when more than 16 observed values (along the genome) were higher than the permutation threshold, we kept only the 16 largest

values. Last, the same rules were used for IM(parametric) and the parametric threshold was chosen according to Delong [16]. Indeed, according to Theorem 6.1, the threshold remains the same under selective genotyping as under the complete data situation. Furthermore, it has been proved by Cierco [14] that under a dense map and the complete data situation, the limiting process of  $\Lambda_n(\cdot)$  is an Ornstein-Uhlenbeck process. The method proposed by Lee and al. [26] is available at [www.stat.sinica.edu.tw/chkao](http://www.stat.sinica.edu.tw/chkao). This method was originally developed for a F2 population and we adapted it to a backcross population: instead of dealing with a mixture model with three components, we had to handle a mixture model with two components. Furthermore, this method is a sequential procedure whereas our method is simultaneous. That is to say, the Lee and al. [26]’s method first finds one QTL, and then looks for a second one conditionally on the first one, and so on until the 16th QTL ( $m = 16$ ). In view of the results on simulated data (cf. tables below), it is clear that this approach is not satisfactory. Then, we will not comment it below.

Table 2 focuses on the same dense map as previously. In order to propose a sharp comparison of the methods, we placed the QTLs on the interval  $[0M, 1M]$ , still considering a genome of size 10M. We considered different ways of performing the selective genotyping, by letting the ratio  $\gamma_+/\gamma$  vary. All the QTL effects were chosen such as  $|q_s| = 0.1$ . According to the table, the performances of the different methods were fair when the ratio  $\gamma_+/\gamma$  took the values  $1/2$ ,  $3/4$  or  $7/8$ . However, when the selective genotyping was performed unilaterally ( $\gamma_+/\gamma = 1$ ), the LASSO, Group LASSO, and IM(shuffle) deteriorate heavily, which was not the case of our method and IM(parametric). For instance, when  $\gamma = 0.1$ , the power associated to the LASSO, Group LASSO, and IM(shuffle) was found to be equal to 20.78%, 16.73% and 17%, respectively. LASSO and Group LASSO suffer from the fact that the tails of errors are not light, and that the conditional distribution is asymmetric around 0 (see for instance [23]).

In view of the fair results of IM(parametric) (power of 94% on the same example), we can infer that the permutation threshold must be too large, and must be responsible for the bad performances of IM(shuffle). Table 3 of Supplement B deals with the case  $|q_s|$  equal to 0.2: although the signal has been increased, LASSO and Group LASSO still present a very small empirical power. It confirms our original thoughts: the LASSO and Group LASSO do not model explicitly the extremes, in contrast to our new method.

Last, Table 3 is dedicated to the sparse map. The selective genotyping was performed unilaterally. Since the distance between markers became larger, we computed the Azaïs et al. [2]’s threshold, suitable for any genetic map.



As expected, we can observe the superiority of our method over the LASSO, Group LASSO and IM(shuffle). Besides, it is clear that our method is more appropriate than IM(parametric), due to the low number of QTLs found by this method. A major drawback of IM methods is in the fact that they rely on a threshold: in order to find a given QTL, the value of the associated test statistic has to be above the threshold. On the example studied in Table 3, it limits drastically the number of found QTLs. For instance, when  $\gamma = 0.1$  and  $n = 100$ , IM(parametric) found  $\hat{m} = 1.64$  QTLs whereas our method found  $\hat{m} = 16.64$  QTLs. Besides, we will show on real data (cf. next section) that, when the number of QTLs is unknown, the number of found QTLs (above the threshold) can become very large. Recall that at this time  $\hat{m}$  is bounded by 16, because of the adopted rule.

*8.4. Real data analysis.* To conclude this article, we analyzed data from the recent joint papers Spindel et al. [44] and Begum et al. [7] dealing respectively with genomic prediction and association mapping in rice.

We considered the dataset of 13,101 SNPs, randomly chosen by the authors from their 73,147 collected SNPs (cf. p20 of Spindel et al. [44]), and we decided to focus on the flowering date during the dry season 2012. In this context, we propose to compare the performances of the different methods.

Assuming that the 13,101 markers are spread out along the rice genome of length 13.101M (cf. Section “GS using marker subsets” of [44]), we can infer that a marker is located every 0.1cM. Then, we used Delong [16]’s threshold to compute IM(parametric). The percentage  $\gamma$  of genotyped individuals was set to either 1, 0.3 or 0.2, and the selective genotyping was performed symmetrically ( $\gamma_+/\gamma_- = 1/2$ ). Since Begum et al. [7] considered the complete data situation ( $\gamma = 1$ ), we removed data to mimick selective genotyping experiments. In particular, for  $\gamma = 1$  we kept the original data from Spindel et al. [44] ( $n = 312$  by averaging the replicates), whereas for  $\gamma = 0.2$  (resp.  $\gamma = 0.3$ ), we kept the genome information of only 62 (resp. 93) extreme individuals. In what follows, in order to make the reading easier for non specialists, a gene will refer to a marker selected by a method.

The 10 genes found by Begum et al. [7] (cf. their S1 Table), and obtained after fitting a linear mixed model, are given at the top of Table 4. According to Table 4, our method selected 15 genes under the complete data situation ( $\gamma = 1$ ). All the genes found by [7] and present on chromosome 3, were either perfectly found by our method or were tagged by a marker located nearby (at less than a distance of 4 markers, i.e. 0.4cM). In contrast, LASSO selected 20 genes and none of them were close to the findings of [7].

Let us now move on to selective genotyping. For  $\gamma = 0.3$ , our method

selected 12 genes and as previously, all genes lying on chromosome 3 and suggested by [7], were either found or tagged by a close marker. The LASSO kept 46 regions that were not considered as regions of interest by [7]. Last, IM(shuffle) (resp. IM(parametric)) proposed 49 (resp. 45) genes. 5 (resp. 4) of these matched exactly the proposed genes, and 4 (resp. 5) were located nearby to a suggested gene location. However, 40 (resp. 36) selected regressors were false positives. Same kind of results were obtained for  $\gamma = 0.2$ . This large number of false positives is not surprising since IM requires the use of a threshold and is not an appropriate way of selecting multiple genes: it is rewarding for testing and localizing only one gene. In opposite, our method can be viewed as a L1 penalization method dedicated to the extremes. It should be of interest for geneticists.

| $\gamma$ | $\gamma_+/\gamma$ | Our method |           | LASSO    |           | Group LASSO |           | IM(parametric) |           | IM(shuffle) |           | Lee and al. [26] |
|----------|-------------------|------------|-----------|----------|-----------|-------------|-----------|----------------|-----------|-------------|-----------|------------------|
|          |                   | L1 ratio   | $\hat{m}$ | L1 ratio | $\hat{m}$ | L1 ratio    | $\hat{m}$ | L1 ratio       | $\hat{m}$ | L1 ratio    | $\hat{m}$ | L1 ratio         |
| 0.1      | 1/2               | 97.24%     | 17.22     | 94.21%   | 16.82     | 99.01%      | 19.4      | 99%            | 15.84     | 99.18%      | 16        | 50.32%           |
|          | 3/4               | 96.62%     | 17.45     | 92.22%   | 16.33     | 95.88%      | 19.1      | 98.10%         | 15.3      | 96.01%      | 15.6      | 49.13%           |
|          | 7/8               | 96.89%     | 17.58     | 82.32%   | 16.78     | 95.19%      | 22.9      | 97.75%         | 15.36     | 95%         | 15.2      | 45.10%           |
|          | 1                 | 93.97%     | 17.13     | 20.78%   | 16.66     | 16.73%      | 22.3      | 94%            | 14.89     | 17%         | 2.08      | 32.92%           |
| 0.2      | 1/2               | 94.19%     | 17.39     | 91.69%   | 16.95     | 97.46%      | 19.4      | 95.87%         | 15.46     | 95.82%      | 15.68     | 49.28%           |
|          | 3/4               | 91.52%     | 16.3      | 84.75%   | 16.54     | 95.88%      | 19.1      | 92.74%         | 15.13     | 93.75%      | 15.36     | 45.70%           |
|          | 7/8               | 92.38%     | 16.29     | 75.46%   | 16.55     | 94.67%      | 17.3      | 95%            | 15.25     | 93%         | 15.09     | 44.20%           |
|          | 1                 | 85.03%     | 17.09     | 21.14%   | 16.81     | 21.86%      | 26.2      | 77.88%         | 12.45     | 29%         | 3.59      | 20.12%           |
| 0.3      | 1/2               | 91.62%     | 17.55     | 83.45%   | 16.51     | 92.87%      | 18.6      | 82%            | 13.06     | 82%         | 13.91     | 46.12%           |
|          | 3/4               | 90.88%     | 17.59     | 76.18%   | 16.56     | 89.59%      | 21.6      | 81%            | 12.9      | 84%         | 13.36     | 42.10%           |
|          | 7/8               | 86.22%     | 16.82     | 65.03%   | 16.73     | 78.00%      | 17.3      | 77.00%         | 11.9      | 77%         | 12.23     | 30.20%           |
|          | 1                 | 78.00%     | 17.28     | 20.92%   | 16.57     | 20.82%      | 22.1      | 54.38%         | 8.8       | 23%         | 3.22      | 19.20%           |

TABLE 2

Performances of different methods, under the dense map, as a function of the percentage  $\gamma$  of genotyped individuals and as a function of the ratio  $\gamma_+/\gamma$ . (Mean over 100 samples, on average  $n = 100$  individuals genotyped,  $m = 16$ ,  $|q_1| = \dots = |q_{16}| = 0.1$ ,  $T = 10$ , QTLs randomly located only on  $[0M, 1M]$ ). Dense map:  $K = L = 10,001$ ,  $t_k = t'_k = 0.001(k - 1)$ . The L1 ratio, regarding our method, corresponds to the quantity  $\sum_{i=1}^{1001} |\hat{\Delta}_i| / \sum_{i=1}^{10001} |\hat{\Delta}_i|$ .  $\hat{m}$  denotes the estimated QTL number. IM(parametric) refers to the Interval Mapping method with a parametric threshold (14.07 by Delong [16]), whereas IM(shuffle) is the same approach based on a permutation threshold.

| $n$ | $\gamma$ | (Our method) |           | (LASSO)  |           | (Group LASSO) |           | IM(parametric) |           | IM(shuffle) |           | Lee and al. [26] |
|-----|----------|--------------|-----------|----------|-----------|---------------|-----------|----------------|-----------|-------------|-----------|------------------|
|     |          | L1 ratio     | $\hat{m}$ | L1 ratio | $\hat{m}$ | L1 ratio      | $\hat{m}$ | L1 ratio       | $\hat{m}$ | L1 ratio    | $\hat{m}$ | L1 ratio         |
| 100 | 0.1      | 41.07%       | 16.75     | 12.77%   | 16.82     | 8.89%         | 13.76     | 26.93%         | 1.64      | 5%          | 0.08      | 18.30%           |
|     | 0.2      | 33.53%       | 16.67     | 14.37%   | 16.74     | 16.85%        | 15.47     | 13.81%         | 0.91      | 6%          | 0.14      | 16.81%           |
|     | 0.3      | 25.07%       | 16.85     | 12.10%   | 16.80     | 14.60%        | 15.4      | 7.76%          | 0.56      | 3%          | 0.14      | 15.43%           |
| 200 | 0.1      | 55.84%       | 16.73     | 18.55%   | 16.66     | 21.54%        | 14.99     | 77.58%         | 7.61      | 23%         | 0.43      | 21.22%           |
|     | 0.2      | 45.49%       | 16.70     | 16.01%   | 16.71     | 11.46%        | 14.4      | 44.16%         | 2.67      | 9.75%       | 0.23      | 19.76%           |
|     | 0.3      | 40.79%       | 16.75     | 17.81%   | 16.79     | 21.88%        | 14.4      | 28.69%         | 2.38      | 12.05%      | 0.36      | 18.60%           |

TABLE 3

Performances of different methods, under the sparse map, when  $\gamma_+/\gamma = 1$  and as a function of the average  $n$  of genotyped individuals. (Mean over 100 samples,  $m = 16$ ,  $|q_1| = \dots = |q_{16}| = 0.1$ ,  $T = 10$ , QTLs randomly located only on  $[0M, 1M]$ ). Sparse map:  $K = 201$ ,  $t_k = 0.05(k - 1)$ ,  $L = 401$ ,  $t'_i = 0.025(k - 1)$ . The L1 ratio, regarding our method, corresponds to the quantity  $\sum_{i=1}^{41} |\hat{\Delta}_i| / \sum_{i=1}^{401} |\hat{\Delta}_i|$ . Tests are only performed on markers for LASSO and Group LASSO. IM(parametric) refers to the Interval Mapping method with a parametric threshold (12.25 by Azaïs et al. [2]), whereas IM(shuffle) is the same approach based on a permutation threshold.

**Acknowledgements.** We thank Professor Jean-Marc Azaïs from university Paul Sabatier Toulouse (FR) for fruitful discussions.

## SUPPLEMENTARY MATERIAL

**Supplement A: Supplementary proofs for the paper “On gene mapping with the mixture model and the extremes”**

(<http://www.e-publications.org/ims/support/download/suppmaterial>). We give the mathematical proofs of Theorems 3.1, 5.1, 6.1 and 6.2.

**Supplement B: Supplementary illustrations for the paper “On gene mapping with the mixture model and the extremes”**

(<http://www.e-publications.org/ims/support/download/suppmaterial>). We illustrate our theoretical results regarding the max test, relying on the test statistic,  $\sup \Lambda_n(\cdot)$ .

**Supplement C: Supplementary models for the paper “On gene mapping with the mixture model and the extremes”**

(<http://www.e-publications.org/ims/support/download/suppmaterial>). We give results and proofs regarding the interference model. We also state Lemma 6.3 based on the reverse configuration of selective genotyping.

## References.

- [1] AZAÏS, J.M. and CIERCO-AYROLLES, C. (2002). An asymptotic test for quantitative gene detection. *Ann. Inst. Henri Poincaré (B)*, **38(6)** 1087-1092.
- [2] AZAÏS, J.M., DELMAS, C., and RABIER, C.E. (2012). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*, **48(4)** 787-801.
- [3] AZAÏS, J.M., GASSIAT, E., and MERCADIER, C. (2006). Asymptotic distribution and local power of the likelihood ratio test for mixtures. *Bernoulli*, **12(5)** 775-799.
- [4] AZAÏS, J.M., GASSIAT, E., and MERCADIER, C. (2009). The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM*, **13** 301-327.
- [5] AZAÏS, J.M. and WSCHEBOR, M. (2009). *Level sets and extrema of random processes and fields*. Wiley, New-York.
- [6] ARIAS-CASTRO, E., CANDES, E.J., and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, **39(5)** 2533-2556.
- [7] BEGUM, H., SPINDEL, J.E., LALUSIN, A., BORROMEO, T., GREGORIO, G., HERNANDEZ, J., ..., and MCCOUCH, S.R. (2015). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PloS one*, **10(3)** e0119873.
- [8] BOLIGON, A.A., LONG, N., ALBUQUERQUE, L.G.D., WEIGEL, K.A., GIANOLA, D., and ROSA, G.J.M. (2012). Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. *Journal of animal science*, **90(13)** 4716-4722.
- [9] BROMAN, K. and SPEED T. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64(4)** 641-656.

| $\gamma$ | Method           | Selected genes  |
|----------|------------------|---|
| 1        | Begum et al. [7] | S3-1125848, S3-1165376, S3-1221494, S3-1269941, S3-1394477, S6-2900101, S6-2961503, S6-3057752, S8-4137990, S8-4138023  |
| 1        | Our method       | S3-565311, S3-608947, S3-673934, S3-744281, S3-754676<br>S3-813574, S3-905839, S3-1003372, S3-1030333, <i>S3-1070111</i><br><b>S3-1125848</b> , <i>S3-1179404</i> , <b>S3-1221494</b> , <i>S3-1320779</i> , S3-1711393  |
| 0.3      | Our method       | S3-565311, S3-673934, S3-905839, S3-963933, S3-1003372,<br>S3-1030333, <i>S3-1070111</i> , <b>S3-1125848</b> , <i>S3-1179404</i> , <b>S3-1221494</b><br><b>S3-1269941</b> , <i>S3-1320779</i>   |
| 0.2      | Our method       | S3-565311, S3-673934, S3-905839, S3-1030333, <b>S3-1125848</b><br><i>S3-1179404</i> , <b>S3-1221494</b> , <i>S3-1320779</i> , S3-1711393  |
| 1        | LASSO            | S1 $\times$ 9, S3-1604923, S3-7291971, S3-16735327, S4 $\times$ 2<br>S6 $\times$ 2, S7 $\times$ 1, S8 $\times$ 1, S9 $\times$ 1, S11 $\times$ 1   |
| 0.3      | LASSO            | S1 $\times$ 6, S2 $\times$ 3, S3-8247017, S3-22163479, S3-31852244<br>S4 $\times$ 9, S5 $\times$ 6, S6 $\times$ 11, S7 $\times$ 3, S8 $\times$ 3, S9 $\times$ 1, S10 $\times$ 1   |
| 0.2      | LASSO            | S1 $\times$ 9, S2 $\times$ 7, S3-2091924, S3-4160463, S3-4980730<br>S3-5046485, S3-6577917, S3-6608757, S3-13478393, S3-14678425<br>S3-20036925, S3-24525448, S3-25713904, S3-32162755, S3-32181235<br>S4 $\times$ 7, S5 $\times$ 1, S7 $\times$ 3, S8 $\times$ 1, S11 $\times$ 1 |
| 0.3      | IM(shuffle)      | S3 $\times$ 37, <i>S3-1094192</i> , <i>S3-1123429</i> , <b>S3-1125848</b> , <b>S3-1165376</b><br><i>S3-1179404</i> , <b>S3-1221494</b> , <b>S3-1269941</b> , <i>S3-1320779</i> , <i>S3-1342244</i><br><b>S3-1394477</b> , S8 $\times$ 2   |
| 0.2      | IM(shuffle)      | S3 $\times$ 33, <i>S3-1094192</i> , <i>S3-1123429</i> , <b>S3-1125848</b> , <b>S3-1165376</b><br><i>S3-1179404</i> <b>S3-1221494</b> , <b>S3-1269941</b> , <i>S3-1320779</i> , <i>S3-1342244</i><br><b>S3-1394477</b> , S12 $\times$ 1  |
| 0.3      | IM(parametric)   | S2 $\times$ 4, S3 $\times$ 29, <i>S3-1094192</i> , <i>S3-1123429</i> , <b>S3-1125848</b><br><b>S3-1165376</b> , <i>S3-1179404</i> , <b>S3-1221494</b> , <b>S3-1269941</b> , <i>S3-1320779</i> ,<br><i>S3-1342244</i> , <b>S3-1394477</b> , S8 $\times$ 2                          |
| 0.2      | IM(parametric)   | S1 $\times$ 1, S3 $\times$ 31, <i>S3-1094192</i> , <i>S3-1123429</i> , <b>S3-1125848</b><br><b>S3-1165376</b> , <i>S3-1179404</i> , <b>S3-1221494</b> , <b>S3-1269941</b> , <i>S3-1320779</i><br><i>S3-1342244</i> , <b>S3-1394477</b> , S12 $\times$ 1                           |

TABLE 4

Comparison, on rice data (Begum et al. [7], Spindel et al. [44]), of the selected genes as a function of the methods and as function of the percentage  $\gamma$  of genotyped individuals.

The considered trait is the flowering date during the dry season 2012. The selective genotyping is performed symmetrically ( $\gamma_+/\gamma = 1/2$ ) and  $K = 13,101$  markers lie on the rice genome ( $T = 13,101$ ). IM(parametric) refers to the Interval Mapping method with a parametric threshold (14.66 by Delong [16]), whereas IM(shuffle) is the same approach based on a permutation threshold (16.54 for  $\gamma = 0.3$ , and 16.37 for  $\gamma = 0.2$ ). Markers in bold match exactly one of the genes selected by Begum et al. [7]. A marker in italic refers to a marker which is located at a maximum distance of 0.4cM from a gene inferred by Begum et al. [7]. SA-B refers to a marker on chromosome A with id B. SA $\times$ N refers to N markers on chromosome A, and these markers are located further than 0.4cM from a gene found by Begum et al. [7].

- [10] BUHLMANN, P. and VAN DE GEER, S. (2011). Statistics for high-dimensional data: methods, theory and applications, Springer Science.
- [11] CHANG, M.N., WU, R., WU, S.S., and CASELLA, G. (2009). Score statistics for mapping quantitative trait loci. *Stat. Appl. Genet. Mol. Biol.*, **8**(1) 16.
- [12] CHEN, Z., and CHEN, H. (2005). On some statistical aspects of the interval mapping for QTL detection. *Statistica Sinica*, **15** 909-925.
- [13] CHURCHILL, G.A. and DOERGE, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138** 963-971.
- [14] CIERCO, C. (1998). Asymptotic distribution of the maximum likelihood ratio test for gene detection. *Statistics*, **31** 261-285.
- [15] DARVASI D. and SOLLER M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor. Appl. Genet.*, **85** 353-359.
- [16] DE LONG, D.M. (1981). Crossing probabilities for a square root boundary by a Bessel process. *Communications in Statistics-Theory and Methods*, **10**(21) 2197-2213.
- [17] FAN, J., and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space *Journal of the Royal Statistical Society: Series B*, **70**(5) 849-911.
- [18] GASSIAT, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré (B)*, **6** 897-906.
- [19] GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Stat.*, **1** 141-149.
- [20] HALDANE, J.B.S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*, **8** 299-309.
- [21] HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2001). *The elements of statistical learning theory*. Springer, New York.
- [22] HAYES, B (2007). QTL Mapping, MAS, and Genomic Selection. *Short course organized by Iowa State University*.
- [23] FAN, J., LI, Q., and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(1) 247-265.
- [24] LANDER, E.S. and BOTSTEIN, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138** 235-240.
- [25] LEBOWITZ, R.J., SOLLER, M., and BECKMANN, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theor. Appl. Genet.*, **73** 556-562.
- [26] LEE, H.I., HO, H.A., and KAO, C.H. (2014). A new simple method for improving QTL mapping under selective genotyping. *Genetics*, **198**(4) 1685-1698.
- [27] LOBO, I., and SHAW, K. (2008). Thomas Hunt Morgan, genetic recombination, and gene mapping. *Nat. Educ.*, 1:1.
- [28] MANICHAIKUL, A., PALMER, A., SEN, S., and BROMAN, K. (2007). Significance thresholds for Quantitative Trait Locus mapping under selective genotyping. *Genetics*, **177** 1963-1966.
- [29] MARTINI, E., DIAZ, L.D., HUNTER, N., and KEENEY, S. (2006). Crossover homeostasis in yeast meiosis. *Cell*, **126** 285-295.
- [30] MURANTY, H. and GOFFINET, B. (1997). Selective genotyping for location and estimation of the effect of the effect of a quantitative trait locus. *Biometrics*, **53** 629-643.
- [31] PHANSACK, P., SOONSUWON, W., HYTEN, D.L., ..., and SPECHT, J.E. (2016). Multi-population selective genotyping to identify soybean (*Glycine max* (L.) Merr.) seed protein and oil QTLs. *G3: Genes, Genomes, Genetics*, **6** 1635.

- [32] RABBEE, N., SPECA, D., ARMSTRONG, N., and SPEED, T. (2004). Power calculations for selective genotyping in QTL mapping in backcross mice. *Genet. Res. Camb.*, **84** 103-108.
- [33] RABIER, C.E. (2014). On statistical inference for selective genotyping. *J. Stat. Plan. Infer.*, **147** 24-52.
- [34] RABIER, C.E. (2014). On Quantitative Trait Locus mapping with an interference phenomenon. *TEST*, **23(2)** 311-329.
- [35] RABIER, C.E. (2014). An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes. *Annales de la faculté des sciences de Toulouse*, **6(23)** 755-778.
- [36] RABIER, C.E. (2014). On empirical processes for Quantitative Trait Locus mapping under the presence of a selective genotyping and an interference phenomenon. *J. Stat. Plan. Infer.*, **153** 42-55.
- [37] RABIER, C.E. (2015). On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. *Statistics*, **49(1)** 19-34.
- [38] RABIER, C.E. and DELMAS, C. (2017). Supplement A: Supplement to “On gene mapping with the mixture model and the extremes”. **DOI ...**
- [40] RABIER, C.E. and DELMAS, C. (2017). Supplement B: Supplement to “On gene mapping with the mixture model and the extremes”. **DOI ...**
- [40] RABIER, C.E. and DELMAS, C. (2017). Supplement C: Supplement to “On gene mapping with the mixture model and the extremes”. **DOI ...**
- [41] REBAÏ, A., GOFFINET, B., and MANGIN, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138** 235-240.
- [42] REBAÏ, A., GOFFINET, B., and MANGIN, B. (1995). Comparing power of different methods for QTL detection. *Biometrics*, **51** 87-99.
- [43] SIEGMUND, D. and YAKIR, B. (2007). *The statistics of gene mapping*. Springer, New York.
- [44] SPINDEL, J., BEGUM, H., AKDEMIR, D., VIRK, P., COLLARD, B., REDOÑA, E., ..., and McCOUCH, S.R. (2015). Genomic Selection and Association Mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics*. **11(2)**, e1004982.
- [45] STURTEVANT, A.H. (1915). The behavior of the chromosomes as studied through linkage. *Z. Indukt. Abstammungs. Vererbungs.*, **13** 234-287.
- [46] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58(1)** 267-288.
- [47] TORTEREAU, F., SERVIN, B., ..., and GROENEN, M.A.M. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC genomics*, **13**, 586.
- [48] VAN DER VAART, A.W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.
- [49] VUONG, T. D., WALKER, D. R., NGUYEN, B. T., NGUYEN, T. T., DINH, H. X., HYTEN, D. L., ..., and NGUYEN, H. T. (2016). Molecular Characterization of Resistance to Soybean Rust (*Phakopsora pachyrhizi* Syd. Syd.) in Soybean Cultivar DT 2000 (PI 635999). *PloS one*, **11(12)** e0164493.
- [50] WU, R., MA, C.X., and CASELLA, G. (2007). *Statistical Genetics of Quantitative Traits*. Springer, New York.
- [51] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68(1)**, 49-67.



CHARLES-ELIE RABIER  
ISE-M, UMR 5554  
CNRS, IRD  
UNIVERSITÉ DE MONTPELLIER, FRANCE  
E-MAIL: [ce.rabier@gmail.com](mailto:ce.rabier@gmail.com)

CELINE DELMAS  
INRA, UR 875 MIAT  
BP 52627, CASTANET-TOLOSAN  
31326 CEDEX, FRANCE  
E-MAIL: [celine.delmas@toulouse.inra.fr](mailto:celine.delmas@toulouse.inra.fr)

1.

## SUPPLEMENT A FOR: “ON GENE MAPPING WITH THE MIXTURE MODEL AND THE EXTREMES”

BY CHARLES-ELIE RABIER<sup>\*,†,‡</sup>, CÉLINE DELMAS<sup>‡</sup>

*ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France* <sup>\*</sup>

*LIRMM, UMR 5506, CNRS, Université de Montpellier, France* <sup>†</sup>

*MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France* <sup>‡</sup>

**1. Proof of Theorem 3.1.** The proof is divided into four parts:

- Preliminaries (i.e. computation of the Fisher Information Matrix)
- Weak convergence of the score process under  $H_0$
- Study of the score process under the local alternative  $H_{a\vec{\tau}^*}$
- Study of the supremum of the LRT process.

Note that under  $H_0$ , the proof has already been given in Azaïs et al. [1]. However, the weak convergence of the score process has not been proved in details. Indeed, the authors only mentioned the continuous mapping theorem, after having proved the convergence of finite-dimensional. As a consequence, we propose to give here a more rigorous proof by showing the tightness 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 Azaïs and Wschebor [4]).

In what follows, we will consider values  $t, t_1^*, \dots, t_m^*$  of the parameters that are distinct of the markers positions (i.e.  $t_1$  and  $t_2$ ), and the result will be extended by continuity at the markers positions.

**1.1. Preliminaries.** The proof starts with the computation of the Fisher Information Matrix. As a result, calculations are exactly the same as in Azaïs et al. [1], see Section “Study of the score process under the null hypothesis” of the proof of their Theorem 2.1. We propose to recall here the key elements of the proof.

First, the authors compute the score function at a point  $\theta_0^1 = (0, \mu, \sigma)$  that belongs to  $H_0$ :

$$(1.1) \quad \frac{\partial l_t}{\partial q_1} \big|_{\theta_0^1} = \frac{Y - \mu}{\sigma^2} x(t)$$

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

Then, they introduce their key Lemma (Lemma 2.3 of Azaïs et al. [1]), which states that

$$(1.2) \quad x(t) = \alpha(t)X(t_1) + \beta(t)X(t_2)$$

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

As a result, the Fisher information at  $\theta_0^1$ , denoted  $I_{\theta_0^1}$ , verifies

$$(1.3) \quad I_{\theta_0^1} = \text{Diag} \left\{ \frac{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1, t_2)}{\sigma^2}, \frac{1}{\sigma^2}, \frac{2}{\sigma^2} \right\}.$$

1.2. *Weak convergence of the score process under  $H_0$ .*

*Convergence of finite-dimensional.* We have  $\forall k = 1, 2$ :

$$S_n(t_k) = \frac{\frac{\partial l_{t_k}^n}{\partial q_1} |_{\theta_0^1}}{\sqrt{\mathbb{V}_{H_0} \left( \frac{\partial l_{t_k}^n}{\partial q_1} |_{\theta_0^1} \right)}} = \sum_{j=1}^n \frac{\varepsilon_j X_j(t_k)}{\sqrt{n}}.$$

Since  $\frac{\partial l_{t_k}^n}{\partial q_1} |_{\theta_0^1}$  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 (1.2))

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

the continuous mapping theorem implies that

$$S_n(t) \xrightarrow{\mathcal{L}} Z(t).$$

It proves the convergence of finite-dimensional.

*Tightness.* 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  $S_n(\cdot)$  is a continuous function on  $[t_1, t_2]$ . Recall the modulus of continuity of a continuous function  $h(t)$  on  $[t_1, t_2]$ :

$$\varpi_h(\delta) = \sup_{|t'-t|<\delta} |h(t') - h(t)| \quad \text{where } t_1 < \delta \leq t_2.$$

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

1. the sequence  $S_n(t_1)$  is tight.
2. For each positive  $\varepsilon$  and  $\eta$ , there exists a  $\delta$ , with  $t_1 < \delta < t_2$ , and an integer  $n_0$  such that  $\mathbb{P}(\varpi_{S_n}(\delta) \geq \eta) \leq \varepsilon \quad \forall n \geq n_0$ .

According to Prohorov, the sequence  $S_n(t_1)$  is tight. Then, Condition 1 is verified. Let us define the functions  $\alpha'(t)$  and  $\beta'(t)$  in the following way:

$$\begin{aligned}\alpha'(t) &= \alpha(t) / \sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1, t_2)}, \\ \beta'(t) &= \beta(t) / \sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1, t_2)}.\end{aligned}$$

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

$$\begin{aligned}\varpi_{S_n}(\delta) &= \sup_{|t'-t|<\delta} |S_n(t') - S_n(t)| \\ &= \sup_{|t'-t|<\delta} |(\alpha'(t') - \alpha'(t)) S_n(t_1) + (\beta'(t') - \beta'(t)) S_n(t_2)| \\ (1.4) \quad &\leq \max(|S_n(t_1)|, |S_n(t_2)|) (\varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta)).\end{aligned}$$

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

$$(1.5) \quad \forall \varepsilon > 0 \quad \exists M > 0 \quad \forall n \geq 1 \quad \mathbb{P}(\max(|S_n(t_1)|, |S_n(t_2)|) \geq M) \leq \varepsilon.$$

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

$$(1.6) \quad \forall v > 0 \quad \exists \delta \text{ such as } t_1 < \delta < t_2, \quad \varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta) < v.$$

Let  $\eta$  be a positive quantity. Using formulae (1.5) and (1.6) and imposing  $v = \eta/M$ , we have

$$\mathbb{P}(\max(|S_n(t_1)|, |S_n(t_2)|) (\varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta)) \geq \eta) \leq \varepsilon.$$

As a consequence, according to formula (1.4), we have

$$\forall n \geq 1 \quad \mathbb{P}(\varpi_{S_n}(\delta) \geq \eta) \leq \varepsilon.$$

It proves Condition 2 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.

1.3. *Study of the score process under the local alternative  $H_{a\vec{t}^*}$ .* There are  $m$  QTLs located on  $[0, T]$  and the model for the quantitative trait is the following:

$$(1.7) \quad Y = \mu + \sum_{s=1}^m X(t_s^*) q_s + \sigma \varepsilon$$

where  $\varepsilon$  is a Gaussian white noise.

Since the score test statistic at  $t$  can be obtained using the following non linear interpolation

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

, the mean function will be also a non linear interpolation

$$m_{\vec{t}^*}(t) = \frac{\alpha(t) m_{\vec{t}^*}(t_1) + \beta(t) m_{\vec{t}^*}(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1, t_2)}} .$$

Let us compute the quantities  $m_{\vec{t}^*}(t_1)$  and  $m_{\vec{t}^*}(t_2)$ .

Without loss of generality, let's consider location  $t_k$  which refers to the location of marker  $k$ . According to formulae (1.1) and (1.7), we have

$$\begin{aligned} S_n(t_k) &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j X_j(t_k) + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^m X_j(t_s^*) a_s \right\} X_j(t_k) \\ (1.8) \quad &= S_n^0(t_k) + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^m X_j(t_s^*) a_s \right\} X_j(t_k) \end{aligned}$$

where  $S_n^0(t_k)$  is the score obtained under  $H_0$  at location  $t_k$ .

By the law of large number :

$$\frac{1}{n} \sum_{j=1}^n \left\{ \sum_{s=1}^m X_j(t_s^*) a_s \right\} X_j(t_k) \rightarrow \mathbb{E} \left[ \left\{ \sum_{s=1}^m X(t_s^*) a_s \right\} X(t_k) \right] .$$

We have

$$\mathbb{E} \left[ \left\{ \sum_{s=1}^m X(t_s^*) a_s \right\} X(t_k) \right] = \sum_{s=1}^m a_s e^{-2|t_s^* - t_k|} = \sum_{s=1}^m a_s \rho(t_s^*, t_k) .$$

Then,

$$m_{\vec{t}^*}(t_k) = \frac{1}{\sigma} \sum_{s=1}^m a_s \rho(t_s^*, t_k) .$$

As a consequence, if we consider  $t_k = t_1$  or  $t_k = t_2$ , we have

$$m_{\vec{t}^*}(t_1) = \frac{1}{\sigma} \sum_{s=1}^m a_s \rho(t_s^*, t_1) , \quad m_{\vec{t}^*}(t_2) = \frac{1}{\sigma} \sum_{s=1}^m a_s \rho(t_s^*, t_2) .$$

1.4. *Study of the supremum of the LRT process.* At fixed  $t$ , the model is regular and it is well known that we have the following relationship under  $H_0$  (i.e. no QTL on the whole interval studied)

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

and where  $o_P(1)$  is short for a sequence of random vectors that converges to zeros in probability. Let us consider now  $t$  as an extra parameter. It easy to check that at  $H_0$  the Fisher Information relative to  $t$  is zero so that the model is not regular. As a consequence, Azaïs et al. [1] studied this irregular model and proved that

$$(1.9) \quad \sup \Lambda_n(t) = \sup S_n^2(t) + o_P(1) .$$

Note that the proof is based on results of Azaïs et al. [3], Azaïs et al. [2] and Gassiat [5] on empirical process theory. This result has been obtained under  $H_0$  and under the local alternative of only one QTL (i.e.  $m = 1$ ), located at  $t_1^*$  on  $[0, T]$ . This way, our goal is now to show that the remainder converges also to zero under  $H_{a\vec{t}^*}$ .

Recall that the parameters  $\theta^m$  and  $\theta_0^m$  are defined in the following way :  $\theta^m = (q_1, \dots, q_m, \mu, \sigma)$  and  $\theta_0^m = (0, \dots, 0, \mu, \sigma)$ . Recall also that the full likelihood of the triplet  $(Y, X(t_1), X(t_2))$ , with respect to the measure  $\lambda \otimes N \otimes N$ , is given in formula (3.5) of the main text. According to Bayes rules,

$$w_{\vec{t}^*}(u_1, \dots, u_m) = \frac{\mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m, X(t_1), X(t_2)\}}{\mathbb{P}\{X(t_1), X(t_2)\}} .$$

Besides, we have the relationships

$$\begin{aligned} & \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m, X(t_1), X(t_2)\} \\ &= \mathbb{P}\{X(t_1)\} \mathbb{P}\{X(t_1^*) = u_1 \mid X(t_1)\} \\ &\times \mathbb{P}\{X(t_2^*) = u_2 \mid X(t_1^*) = u_1\} \cdots \mathbb{P}\{X(t_m^*) = u_m \mid X(t_{m-1}^*) = u_{m-1}\} \\ &\times \mathbb{P}\{X(t_2) \mid X(t_m^*) = u_m\} \\ &= \frac{1}{2} \left\{ r(t_1, t_1^*) 1_{X(t_1)u_1=-1} + \bar{r}(t_1, t_1^*) 1_{X(t_1)u_1=1} \right\} \\ &\times \left\{ r(t_1^*, t_2^*) 1_{u_1u_2=-1} + \bar{r}(t_1^*, t_2^*) 1_{u_1u_2=1} \right\} \\ &\times \dots \times \left\{ r(t_{m-1}^*, t_m^*) 1_{u_{m-1}u_m=-1} + \bar{r}(t_{m-1}^*, t_m^*) 1_{u_{m-1}u_m=1} \right\} \\ &\times \left\{ r(t_m^*, t_2) 1_{u_mX(t_2)=-1} + \bar{r}(t_m^*, t_2) 1_{u_mX(t_2)=1} \right\} \end{aligned}$$

and in the same way,

$$\mathbb{P}\{X(t_1), X(t_2)\} = \frac{1}{2} \left\{ r(t_1, t_2) 1_{X(t_1)X(t_2)=-1} + \bar{r}(t_1, t_2) 1_{X(t_1)X(t_2)=1} \right\} .$$

As a result,

$$\begin{aligned}
& w_{\vec{t}^\star}(u_1, \dots, u_m) \\
&= \left\{ r(t_1, t_1^\star) \mathbf{1}_{X(t_1)u_1=-1} + \bar{r}(t_1, t_1^\star) \mathbf{1}_{X(t_1)u_1=1} \right\} \\
&\times \left\{ r(t_1^\star, t_2^\star) \mathbf{1}_{u_1u_2=-1} + \bar{r}(t_1^\star, t_2^\star) \mathbf{1}_{u_1u_2=1} \right\} \\
&\times \dots \times \left\{ r(t_{m-1}^\star, t_m^\star) \mathbf{1}_{u_{m-1}u_m=-1} + \bar{r}(t_{m-1}^\star, t_m^\star) \mathbf{1}_{u_{m-1}u_m=1} \right\} \\
&\times \left\{ r(t_m^\star, t_2) \mathbf{1}_{u_mX(t_2)=-1} + \bar{r}(t_m^\star, t_2) \mathbf{1}_{u_mX(t_2)=1} \right\} \\
&/ \left\{ r(t_1, t_2) \mathbf{1}_{X(t_1)X(t_2)=-1} + \bar{r}(t_1, t_2) \mathbf{1}_{X(t_1)X(t_2)=1} \right\}.
\end{aligned}$$

The likelihood  $L_{\vec{t}^\star}^{m,n}(\theta^m)$  for  $n$  observations is obtained by the product of  $n$  terms as in formula (3.5) of the main text. Let  $Q_n$  and  $P_n$  be two sequences of probability measures defined on the same space  $(\Omega_n, \mathcal{A}_n)$ .  $Q_n$  (respectively  $P_n$ ) is the probability distribution with density  $L_{\vec{t}^\star}^{m,n}(\theta^m)$  (respectively  $L_{\vec{t}^\star}^{m,n}(\theta_0^m)$ ).

In what follows,  $\log \frac{dQ_n}{dP_n}$  will denote the log likelihood ratio. By definition, we have the relationship,

$$(1.10) \quad \log \frac{dQ_n}{dP_n} = \log \left\{ \frac{L_{\vec{t}^\star}^{m,n}(\theta^m)}{L_{\vec{t}^\star}^{m,n}(\theta_0^m)} \right\}.$$

Since the model is differentiable in quadratic mean at  $\theta^m$  and according to the central limit theorem :

$$\log \left( \frac{dQ_n}{dP_n} \right) \xrightarrow{H_0} N\left(-\frac{1}{2}\vartheta^2, \vartheta^2\right) \text{ with } \vartheta^2 \in \mathbb{R}^{+\star}.$$

As a result, according to iii) of Le Cam's first lemma, we have  $Q_n \triangleleft P_n$ , that is to say the sequence  $Q_n$  is contiguous with respect to the sequence  $P_n$ . Then, formula (1.9) is also true under the alternative  $H_{a\vec{t}^\star}$ .

**2. Proof of Theorem 5.1.** To begin with, let us recall the epistatic model, given in formula (5.1) of the manuscript:

$$(2.1) \quad Y = \mu + \sum_{s=1}^m X(t_s^\star) q_s + \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m X(t_s^\star) X(t_{\tilde{s}}^\star) q_{s,\tilde{s}} + \sigma \varepsilon$$

where  $\varepsilon$  is a Gaussian white noise, and  $q_{s,\tilde{s}}$  is the interaction effect between loci  $t_s^\star$  and  $t_{\tilde{s}}^\star$ .

Since the process  $S_n(\cdot)$  is an interpolated process, we can focus, without loss of generality, only on location  $t_k$  (i.e. the location of marker  $k$ ). According to formulae (1.1) and (2.1), we have

$$\begin{aligned} S_n(t_k) &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j X_j(t_k) + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^m X_j(t_s^*) a_s \right\} X_j(t_k) \\ &\quad + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m X_j(t_s^*) X_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X_j(t_k) \\ &= S_n^0(t_k) + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^m X_j(t_s^*) a_s \right\} X_j(t_k) \\ &\quad + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m X_j(t_s^*) X_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X_j(t_k) \end{aligned}$$

where  $S_n^0(t_k)$  is the score obtained under  $H_0$  at location  $t_k$ .

As in the previous proofs, by the law of large number

$$\frac{1}{n} \sum_{j=1}^n \left\{ \sum_{s=1}^m X_j(t_s^*) a_s \right\} X_j(t_k) \rightarrow \sum_{s=1}^m a_s \rho(t_s^*, t_k) .$$

In the same way,

$$\begin{aligned} &\frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m X_j(t_s^*) X_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X_j(t_k) \\ &\rightarrow \mathbb{E} \left[ \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m X(t_s^*) X(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X(t_k) \right] . \end{aligned}$$

We have the relationship

$$\begin{aligned} \mathbb{E} \{ X(t_s^*) X(t_{\tilde{s}}^*) X(t_k) \} &= \mathbb{E} [ X(t_s^*) X(t_{\tilde{s}}^*) \{ 21_{X(t_k)=1} - 1 \} ] \\ &= 2 \mathbb{E} [ X(t_s^*) X(t_{\tilde{s}}^*) 1_{X(t_k)=1} ] - \rho(t_s^*, t_{\tilde{s}}^*) . \end{aligned}$$

Then, if  $t_k < \min(t_s^*, t_{\tilde{s}}^*)$  or  $t_k > \max(t_s^*, t_{\tilde{s}}^*)$ , we have

$$\begin{aligned} \mathbb{E} [ X(t_s^*) X(t_{\tilde{s}}^*) 1_{X(t_k)=1} ] &= \mathbb{E} [ X(t_s^*) X(t_{\tilde{s}}^*) \mid X(t_k) = 1 ] / 2 \\ &= \mathbb{E} [ X(t_s^*) X(t_{\tilde{s}}^*) ] / 2 = \rho(t_s^*, t_{\tilde{s}}^*) / 2 . \end{aligned}$$



Besides, if  $\min(t_s^*, t_{\tilde{s}}^*) < t_k < \max(t_s^*, t_{\tilde{s}}^*)$ ,

$$\begin{aligned} \mathbb{E} [X(t_s^*)X(t_{\tilde{s}}^*)1_{X(t_k)=1}] &= \mathbb{E} [X(t_s^*)X(t_{\tilde{s}}^*) \mid X(t_k) = 1] / 2 \\ &= \mathbb{E} [X(t_s^*) \mid X(t_k) = 1] \mathbb{E} [X(t_{\tilde{s}}^*) \mid X(t_k) = 1] / 2 \\ &= \rho(t_s^*, t_k) \rho(t_k, t_{\tilde{s}}^*) / 2 = \rho(t_s^*, t_{\tilde{s}}^*) / 2 . \end{aligned}$$

As a result, we always have

$$\mathbb{E} \{X(t_s^*)X(t_{\tilde{s}}^*)X(t_k)\} = 0 .$$

To conclude,

$$\begin{aligned} (2.2) \quad & \frac{1}{n} \sum_{j=1}^n \left\{ \sum_{s=1}^m X_j(t_s^*) a_s \right\} X_j(t_k) + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m X_j(t_s^*) X_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X_j(t_k) \\ & \rightarrow \sum_{s=1}^m a_s \rho(t_s^*, t_k) . \end{aligned}$$

We can notice that the interaction effects have disappeared and that we have exactly the same mean function as in Theorem 4.1:

$$m_{\vec{t}^*}(t_k) = \sum_{s=1}^m a_s \rho(t_k, t_s^*) / \sigma .$$

Since the model (based on formula 5.2 of the main text) is differentiable in quadratic mean at  $(0, \dots, 0, \mu, \sigma^2)$  (i.e. no additive and no epistatic effect), this result is also suitable for the LRT process.

### 3. Proof of Theorem 6.1.

3.0.1. *Proof of formula (6.3) of the main text.* Recall that  $K$  genetic markers are located at  $0 = t_1 < t_2 < \dots < t_K = T$ . Besides,  $m$  QTLs lie on  $[0, T]$  at locations  $t_1^*, t_2^*, \dots, t_m^*$ , that are distinct of marker locations. By definition  $t_1^* < t_2^* < \dots < t_m^*$ . Let us compute the probability distribution of  $(Y, \bar{X}(t_1^{\star\ell}), \bar{X}(t_1^{\star r}), \dots, \bar{X}(t_m^{\star\ell}), \bar{X}(t_m^{\star r}))$ .

We have

$$\begin{aligned} & \mathbb{P}(Y \in [y, y + dy] , Y \notin [S_-, S_+] , \bar{X}(t_1^{\star\ell}), \bar{X}(t_1^{\star r}), \dots, \bar{X}(t_m^{\star\ell}), \bar{X}(t_m^{\star r})) \\ &= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{P}(Y \in [y, y + dy] \mid \bar{X}(t_1^*) = u_1, \bar{X}(t_2^*) = u_2, \dots, \bar{X}(t_m^*) = u_m) \\ & \times \mathbb{P}(\bar{X}(t_1^*) = u_1, \bar{X}(t_2^*) = u_2, \dots, \bar{X}(t_m^*) = u_m, \bar{X}(t_1^{\star\ell}), \bar{X}(t_1^{\star r}), \dots, \bar{X}(t_m^{\star\ell}), \bar{X}(t_m^{\star r})) . \end{aligned}$$

Besides,

$$\begin{aligned}
& \mathbb{P}(Y \in [y, y + dy] \mid \bar{X}(t_1^*) = u_1, \bar{X}(t_2^*) = u_2, \dots, \bar{X}(t_m^*) = u_m) \\
&= \frac{\mathbb{P}(Y \in [y, y + dy], Y \notin [S_-, S_+] \mid X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m)}{\mathbb{P}(Y \notin [S_-, S_+] \mid X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m)} \\
&= \frac{f_{(\mu+u_1q_1+u_2q_2+\dots+u_mq_m, \sigma)}(y) \mathbf{1}_{Y \notin [S_-, S_+]}}{\mathbb{P}(Y \notin [S_-, S_+] \mid X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m)}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
& \mathbb{P}(\bar{X}(t_1^*) = u_1, \bar{X}(t_2^*) = u_2, \dots, \bar{X}(t_m^*) = u_m, \bar{X}(t_1^{\star\ell}), \bar{X}(t_1^{\star r}), \dots, \bar{X}(t_m^{\star\ell}), \bar{X}(t_m^{\star r})) \\
&= \mathbb{P}(Y \notin [S_-, S_+], X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m, X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r})) \\
&= \mathbb{P}(Y \notin [S_-, S_+] \mid X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m) \\
& \mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m, X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r})).
\end{aligned}$$

As a result,

$$\begin{aligned}
& \mathbb{P}(Y \in [y, y + dy], Y \notin [S_-, S_+], \bar{X}(t_1^{\star\ell}), \bar{X}(t_1^{\star r}), \dots, \bar{X}(t_m^{\star\ell}), \bar{X}(t_m^{\star r})) \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} f_{(\mu+u_1q_1+u_2q_2+\dots+u_mq_m, \sigma)}(y) \mathbf{1}_{Y \notin [S_-, S_+]} \\
& \times \mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m, X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r})).
\end{aligned}$$

In the same way, when the genome information is missing at marker locations (i.e. the phenotype is not extreme), we find

$$\begin{aligned}
& \mathbb{P}(Y \in [y, y + dy], \bar{X}(t_1^{\star\ell}) = 0, \bar{X}(t_1^{\star r}) = 0, \dots, \bar{X}(t_m^{\star\ell}) = 0, \bar{X}(t_m^{\star r}) = 0) \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{P}(Y \in [y, y + dy], Y \in [S_-, S_+], X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m) \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} f_{(\mu+u_1q_1+\dots+u_mq_m, \sigma)}(y) \mathbf{1}_{Y \in [S_-, S_+]} \mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m).
\end{aligned}$$

Then, the probability distribution of  $(Y, \bar{X}(t_1^{\star\ell}), \bar{X}(t_1^{\star r}), \dots, \bar{X}(t_m^{\star\ell}), \bar{X}(t_m^{\star r}))$ , with respect to the measure  $\lambda \otimes N \otimes \dots \otimes N$ , is

$$\begin{aligned}
\bar{L}_{\vec{t}^*}^m(\theta^m) &= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} [w_{\vec{t}^*}(u_1, \dots, u_m) f_{(\mu+u_1q_1+\dots+u_mq_m, \sigma)}(Y) \mathbf{1}_{Y \notin [S_-, S_+]} \\
&+ v_{\vec{t}^*}(u_1, \dots, u_m) f_{(\mu+u_1q_1+\dots+u_mq_m, \sigma)}(Y) \mathbf{1}_{Y \in [S_-, S_+]}] \bar{g}^m(t_1^*, \dots, t_m^*)
\end{aligned}$$

with

$$w_{\vec{t}^*}(u_1, \dots, u_m) = \mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r})),$$

$$v_{\vec{t}^*}(u_1, \dots, u_m) = \mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m)$$

and

$$\bar{g}^m(t_1^*, \dots, t_m^*) = \mathbb{P}(X(t_1^*), X(t_1^{*r}), \dots, X(t_m^*), X(t_m^{*r})) \mathbf{1}_{Y \notin [S_-, S_+]} + \mathbf{1}_{Y \in [S_-, S_+]} .$$

3.0.2. *Study of the score process under  $H_0$ .* The score process  $\bar{S}_n(\cdot)$  has already been studied in Rabier [8]. Let us recall the key elements of the proof. By definition, the score statistic at  $t$  is the following

$$\bar{S}_n(t) = \frac{\frac{\partial \bar{l}_t^n}{\partial q_1} |_{\theta_0^1}}{\sqrt{\mathbb{V}\left(\frac{\partial \bar{l}_t^n}{\partial q_1} |_{\theta_0^1}\right)}} \text{ where } \theta_0^1 = (0, \mu, \sigma) .$$

The score function verifies

$$\begin{aligned} \frac{\partial \bar{l}_t^n}{\partial q_1} |_{\theta_0^1} &= \sum_{j=1}^n \frac{Y_j - \mu}{\sigma^2} \{2p_j(t) - 1\} \mathbf{1}_{Y_j \notin [S_-, S_+]} \\ &= \frac{\alpha(t)}{\sigma} \sum_{j=1}^n \varepsilon_j \bar{X}_j(t^\ell) + \frac{\beta(t)}{\sigma} \sum_{j=1}^n \varepsilon_j \bar{X}_j(t^r) . \end{aligned}$$

As a result, the limiting process is a non linear interpolated process.

On the other hand, at location  $t_k$ :

$$\bar{S}_n(t_k) = \frac{\frac{\partial \bar{l}_{t_k}^n}{\partial q} |_{\theta_0}}{\sqrt{\mathbb{V}\left(\frac{\partial \bar{l}_{t_k}^n}{\partial q} |_{\theta_0}\right)}} = \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} .$$

According to the Central Limit Theorem,

$$\bar{S}_n(t_k) \xrightarrow{\mathcal{L}} N(0, 1) .$$

Besides, we have the relationship

$$\text{Cov}_{H_0} \{\bar{S}_n(t_k), \bar{S}_n(t_{k'})\} = \rho(t_k, t_{k'}) .$$

3.0.3. *Study of the score process under the local alternative  $H_{a\vec{t}^*}$ .* There are  $m$  QTLs located on  $[0, T]$  and that the model for the quantitative trait is the following:

$$(3.1) \quad Y = \mu + \sum_{s=1}^m X(t_s^*) q_s + \sigma \varepsilon$$

where  $\varepsilon$  is a Gaussian white noise.

Since the score test statistic at  $t$  can be obtained using the following non linear interpolation

$$\bar{S}_n(t) = \frac{\alpha(t) \bar{S}_n(t^\ell) + \beta(t) \bar{S}_n(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} ,$$

, the mean function will be also a non linear interpolation

$$\bar{m}_{t^\star}(t) = \frac{\alpha(t) \bar{m}_{t^\star}(t^\ell) + \beta(t) \bar{m}_{t^\star}(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} .$$

Let us compute the quantities  $\bar{m}_{t^\star}(t^\ell)$  and  $\bar{m}_{t^\star}(t^r)$ .

Without loss of generality, let's consider location  $t_k$  which refers to the location of marker  $k$ .

$$\begin{aligned} \bar{S}_n(t_k) &= \sum_{j=1}^n \frac{(Y_j - \mu) \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} \\ (3.2) \quad &= \sum_{j=1}^n \sum_{s=1}^m \frac{q_s \bar{X}_j(t_s^\star) \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} + \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} . \end{aligned}$$

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

$$\begin{aligned} \mathbb{E} \{ \bar{X}(t_s^\star) \bar{X}(t_k) \} &= \\ \mathbb{E} [ 1_{Y \notin [S_-, S_+]} \{ 1_{X(t_s^\star)=1} 1_{X(t_k)=1} + 1_{X(t_s^\star)=-1} 1_{X(t_k)=-1} \} ] \\ &- \mathbb{E} [ 1_{Y \in [S_-, S_+]} \{ 1_{X(t_s^\star)=-1} 1_{X(t_k)=1} + 1_{X(t_s^\star)=1} 1_{X(t_k)=-1} \} ] . \end{aligned}$$

According to calculations present in Section 5,

$$\begin{aligned} &\mathbb{E} [ 1_{Y \notin [S_-, S_+]} \{ 1_{X(t_s^\star)=1} 1_{X(t_k)=1} + 1_{X(t_s^\star)=-1} 1_{X(t_k)=-1} \} ] \\ &= \bar{r}(t_k, t_s^\star) \left\{ 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) + \Phi \left( \frac{S_- - \mu}{\sigma} \right) \right\} + o(1) , \end{aligned}$$

where  $\Phi$  is the cumulative distribution of a standard normal distribution. In the same way,

$$\begin{aligned} &\mathbb{E} [ 1_{Y \in [S_-, S_+]} \{ 1_{X(t_s^\star)=-1} 1_{X(t_k)=1} + 1_{X(t_s^\star)=1} 1_{X(t_k)=-1} \} ] \\ &= r(t_k, t_s^\star) \left\{ 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) + \Phi \left( \frac{S_- - \mu}{\sigma} \right) \right\} + o(1) . \end{aligned}$$

Since we have the relationships

$$1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) = \gamma \text{ and } \bar{r}(t_k, t_s^*) - r(t_k, t_s^*) = \rho(t_k, t_s^*),$$

then we have

$$\mathbb{E}\{\bar{X}(t_s^*) \bar{X}(t_k)\} = \rho(t_k, t_s^*) \gamma + o(1).$$

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

$$(3.3) \quad \sum_{j=1}^n \sum_{s=1}^m \frac{q_s \bar{X}_j(t_s^*) \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} \rightarrow \sum_{s=1}^m \frac{a_s \rho(t_k, t_s^*) \gamma}{\sqrt{\mathcal{A}}}.$$

Let us now focus on the second term of formula (3.2). According to a technical proof present in Section 5, we have

$$\mathbb{E}\{\sigma \varepsilon \bar{X}(t_k)\} = \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \sum_{s=1}^m \rho(t_s^*, t_k) q_s + o(\max_{1 \leq s \leq m} |q_s|).$$

Besides, according to iii) of Lemma 5 of Rabier [6],

$$\begin{aligned} \mathbb{E}\left[\{\sigma \varepsilon \bar{X}(t_k)\}^2\right] &= \mathbb{E}(\sigma^2 \varepsilon^2 1_{Y \notin [S_-, S_+]}) \\ &= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{E}\{\sigma^2 \varepsilon^2 1_{Y \notin [S_-, S_+]} \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\ &\quad \times \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\ &\rightarrow \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathcal{A} \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \rightarrow \mathcal{A}. \end{aligned}$$

As a result,

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

Then, according to the Central Limit Theorem,

$$(3.4) \quad \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} \xrightarrow{\mathcal{L}} N\left[\frac{\sum_{s=1}^m \rho(t_s^*, t_k) a_s}{\sqrt{\mathcal{A}}} \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\}, 1\right].$$

Finally, according to formulae (3.3) and (3.4),

$$\bar{S}_n(t_k) \xrightarrow{\mathcal{L}} N\left[\sum_{s=1}^m \rho(t_k, t_s^*) a_s \sqrt{\mathcal{A}} / \sigma^2, 1\right].$$

3.0.4. *Study of the supremum of the LRT process.* At fixed  $t$ , the model is regular and it is well known that we have the following relationship under  $H_0$  (i.e. no QTL on the whole interval studied)

$$\bar{\Lambda}_n(t) = \bar{S}_n^2(t) + o_P(1)$$

and where  $o_P(1)$  is short for a sequence of random vectors that converges to zeros in probability. The problem is that, when  $t$  is not fixed, the Fisher Information relative to  $t$  at  $H_0$  is zero so that the model is not regular. As a result, let us consider now  $t$  as an extra parameter. Let  $t_1^*$  and  $\theta^{1*}$ , be the true parameters that will be assumed to belong to  $H_0$ . Note that  $t_1^*$  makes no sense for  $\theta$  belonging to  $H_0$ .

Without loss of generality, let us consider that  $\mu$  and  $\sigma$  are known ( $\mu = 0$  and  $\sigma = 1$ ) and as previously, let us consider values of  $t$  distinct of the markers positions. Besides, let us consider only two genetic markers located at  $t_1 = 0$  and  $t_2 = T$ . Note that in order to make the reading easier, we will use the notation  $f_q(\cdot)$  instead of  $f_{(q,1)}(\cdot)$  to denote a Gaussian density with mean  $q$  and unit variance.

For computing scores at each  $t$  separately, the likelihood of  $(Y, X(t_1), X(t_2))$  can be considered in the following way  $\forall t \in ]t_1, t_2[$  :

(3.5)

$$\begin{aligned} \bar{L}(\psi_{q_1}, t(q_1)) = & \left[ p\{t(q_1)\} f_{\psi_{q_1}}(Y) 1_{Y \notin [S_-, S_+]} + [1 - p\{t(q_1)\}] f_{-\psi_{q_1}}(Y) 1_{Y \notin [S_-, S_+]} \right. \\ & \left. + \frac{1}{2} f_{\psi_{q_1}}(Y) 1_{Y \in [S_-, S_+]} + \frac{1}{2} f_{-\psi_{q_1}}(Y) 1_{Y \in [S_-, S_+]} \right] \bar{g}(t) \end{aligned}$$

where  $t(q_1)$  is a continuous function on  $[0, 1]$  into  $]t_1, t_2[$ ,  $p(t)$  is the classical weight (cf. Sections 3 and 4 of the manuscript) and  $\bar{g}(t)$  is given in formula (6.2) of the manuscript. Then, at each value of  $q_1$  corresponds a value of  $t$ , denoted  $t(q_1)$ .

Let us compute the score function corresponding to  $\tilde{L}$  at  $q_1 = 0$ . Recall

that the null hypothesis is reached if and only if the QTL effect is null.

$$\begin{aligned}
\frac{\partial \log \bar{L}}{\partial q_1} \Big|_{q_1=0} &= \psi \left\{ 2Q_{t(0)}^{1,1} - 1 \right\} Y 1_{\bar{X}(t_1)=1} 1_{\bar{X}(t_2)=1} \\
&+ \psi \left\{ 2Q_{t(0)}^{1,-1} - 1 \right\} Y 1_{\bar{X}(t_1)=1} 1_{\bar{X}(t_2)=-1} \\
&+ \psi \left\{ 2Q_{t(0)}^{-1,1} - 1 \right\} Y 1_{\bar{X}(t_1)=-1} 1_{\bar{X}(t_2)=1} \\
&+ \psi \left\{ 2Q_{t(0)}^{-1,-1} - 1 \right\} Y 1_{\bar{X}(t_1)=-1} 1_{\bar{X}(t_2)=-1} \\
&= \psi [\alpha \{t(0)\} + \beta \{t(0)\}] Y 1_{\bar{X}(t_1)=1} 1_{\bar{X}(t_2)=1} \\
&+ \psi [\alpha \{t(0)\} - \beta \{t(0)\}] Y 1_{\bar{X}(t_1)=1} 1_{\bar{X}(t_2)=-1} \\
&+ \psi [\beta \{t(0)\} - \alpha \{t(0)\}] Y 1_{\bar{X}(t_1)=-1} 1_{\bar{X}(t_2)=1} \\
&- \psi [\beta \{t(0)\} + \alpha \{t(0)\}] Y 1_{\bar{X}(t_1)=-1} 1_{\bar{X}(t_2)=-1} \\
&= \psi \alpha \{t(0)\} Y \bar{X}(t_1) + \psi \beta \{t(0)\} Y \bar{X}(t_2)
\end{aligned}$$

where  $\alpha(\cdot)$  and  $\beta(\cdot)$  are the classical quantities introduced in Theorems 3.1 and 4.1. We have

$$\mathbb{E}_{H_0} \left\{ \left( \frac{\partial \log \bar{L}}{\partial q_1} \Big|_{q_1=0} \right)^2 \right\} = \frac{\mathcal{A}}{\sigma^4} \psi^2 [\alpha^2 \{t(0)\} + \beta^2 \{t(0)\} + 2\alpha \{t(0)\} \beta \{t(0)\} \rho(t_1, t_2)] .$$

This quantity is always different from 0. As previously, for any function  $t(q_1)$ , each sub-model  $\tilde{L}(q_1, t(q_1))_{q_1 \in \mathbb{R}}$  is differentiable in quadratic mean and Assumption 2 of Azaïs et al. [3] is verified. Besides, the set of log likelihood is Glivenko-Cantelli (cf. example 19.7 with  $r = 1$  of Van der Vaart [9]), so Assumption 1 of Azaïs et al. [3] holds.

As in Azaïs et al. [3], let us define the set of scores renormalized  $\mathcal{D}$ . In our case :

$$\mathcal{D} = \left\{ \text{sign}(\psi) \frac{\alpha \{t(0)\} Y \bar{X}(t_1) + \beta \{t(0)\} Y \bar{X}(t_2)}{\sqrt{\alpha^2 \{t(0)\} + \beta^2 \{t(0)\} + 2\alpha \{t(0)\} \beta \{t(0)\} \rho(t_1, t_2)}} ; t(0) \in ]t_1, t_2[ ; \psi \in \mathbb{R} \right\}$$

which can be rewritten

$$\mathcal{D} = \left\{ \psi' \frac{\alpha(t) Y \bar{X}(t_1) + \beta(t) Y \bar{X}(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t) \rho(t_1, t_2)}} ; t \in ]t_1, t_2[ ; \psi' \in \{-1, 1\} \right\} .$$

Since we have already shown the tightness of the score process  $S_n(\cdot)$ ,  $\tilde{\mathcal{D}}$  is Donsker (cf. [9]). In particular it proves that Theorem 1 of Azaïs et al. [2] applies in the sense that

$$(3.6) \quad \sup_{(t, \theta)} \bar{l}_t^n(\theta) - \bar{l}_{t_1^*}^n(\theta^{1*}) = \sup_{d \in \mathcal{D}} \left[ \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right\}^2 1_{\sum_{j=1}^n d(X_j) \geq 0} \right] + o_P(1)$$

where the observation  $X_j$  stands for  $(Y_j, X_j(t_1), X_j(t_2))$ . Note that since  $\psi' \in \{-1, 1\}$ , the indicator function can be removed in formula (3.6).

Since the model (based on formula 6.3 of the manuscript) is differentiable in quadratic mean at  $\theta_0^m = (0, \dots, 0, \mu, \sigma)$ , we can apply Le Cam first lemma and formula (3.6) is also suitable under the contiguous alternative  $H_{a\tilde{t}^*}$ .

**4. Proof of Theorem 6.2.** Since the process  $\bar{S}_n(\cdot)$  is an interpolated process, we can focus, without loss of generality, only on location  $t_k$  (i.e. the location of marker  $k$ ). According to formulae (2.1) and (3.2), we have

$$(4.1) \quad \begin{aligned} \bar{S}_n(t_k) = & \sum_{j=1}^n \sum_{s=1}^m \frac{a_s \bar{X}_j(t_s^*) \bar{X}_j(t_k)}{n \sqrt{\mathcal{A}}} + \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n \mathcal{A}}} \\ & + \frac{1}{n \sqrt{\mathcal{A}}} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m \bar{X}_j(t_s^*) \bar{X}_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} \bar{X}_j(t_k) . \end{aligned}$$

According to calculations present in Section 5, when  $1 \leq s \leq m-1$  and  $s+1 \leq \tilde{s} \leq m$ ,

$$\mathbb{E} \{ \bar{X}(t_s^*) \bar{X}(t_{\tilde{s}}^*) \bar{X}(t_k) \} = o(1) .$$

Then, according to the law of large numbers,

$$\bar{S}_n(t_k) = \sum_{j=1}^n \sum_{s=1}^m \frac{a_s \bar{X}_j(t_s^*) \bar{X}_j(t_k)}{n \sqrt{\mathcal{A}}} + \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n \mathcal{A}}} + o_P(1) .$$

As a result, using formulae (3.3) and (3.4),

$$\bar{S}_n(t_k) \xrightarrow{\mathcal{L}} N \left[ \sum_{s=1}^m \rho(t_k, t_s^*) a_s \sqrt{\mathcal{A}} / \sigma^2, 1 \right] .$$



### 5. Study of quantities present in the proofs.

5.1. *Preliminaries.* To begin with, let us recall Lemma 5 of Rabier [6]. It will be very useful for our theoretical calculations since it is related to truncated normal distributions.

LEMMA 5 (Rabier [6]). *Let  $W \sim N(\mu, \sigma^2)$ , then*

$$\begin{aligned}
i) \quad & \mathbb{E}(W^2 1_{W \notin [S_-, S_+]}) = (\mu^2 + \sigma^2) \mathbb{P}(W \notin [S_-, S_+]) + \sigma (S_+ + \mu) \varphi\left(\frac{S_+ - \mu}{\sigma}\right) \\
& - \sigma (S_- + \mu) \varphi\left(\frac{S_- - \mu}{\sigma}\right) \\
ii) \quad & \mathbb{E}(W 1_{W \notin [S_-, S_+]}) = \mu \mathbb{P}(W \notin [S_-, S_+]) + \sigma \varphi\left(\frac{S_+ - \mu}{\sigma}\right) - \sigma \varphi\left(\frac{S_- - \mu}{\sigma}\right) \\
iii) \quad & \mathbb{E}\{(W - \mu)^2 1_{W \notin [S_-, S_+]}\} = \sigma^2 \mathbb{P}(W \notin [S_-, S_+]) + \sigma (S_+ - \mu) \varphi\left(\frac{S_+ - \mu}{\sigma}\right) \\
& - \sigma (S_- - \mu) \varphi\left(\frac{S_- - \mu}{\sigma}\right) \\
iv) \quad & \mathbb{E}\{(W - \mu) 1_{W \notin [S_-, S_+]}\} = \sigma \varphi\left(\frac{S_+ - \mu}{\sigma}\right) - \sigma \varphi\left(\frac{S_- - \mu}{\sigma}\right) \\
v) \quad & \mathbb{E}\{(W - \mu)^2 1_{W \in [S_-, S_+]}\} = \sigma^2 - \sigma^2 \mathbb{P}(W \notin [S_-, S_+]) - \sigma (S_+ - \mu) \varphi\left(\frac{S_+ - \mu}{\sigma}\right) \\
& + \sigma (S_- - \mu) \varphi\left(\frac{S_- - \mu}{\sigma}\right).
\end{aligned}$$

Recall that  $\varphi(\cdot)$  and  $\Phi(\cdot)$  denote respectively the density and the cumulative distribution of a standard normal distribution.

Since we consider  $q_1, \dots, q_m$  small, using a Taylor expansion at first order, we obtain for instance :

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

Since

$$\begin{aligned}
& \mathbb{P}\{Y \notin [S_-, S_+] \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\
& = \Phi\left(\frac{S_- - \mu - \sum_{s=1}^m u_s q_s}{\sigma}\right) + 1 - \Phi\left(\frac{S_+ - \mu - \sum_{s=1}^m u_s q_s}{\sigma}\right),
\end{aligned}$$

using the Taylor expansions and after some work on integrals, we obtain

$$\begin{aligned}
& \mathbb{P}\{Y \notin [S_-, S_+] \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\
& = \Phi\left(\frac{S_- - \mu}{\sigma}\right) - \frac{\sum_{s=1}^m u_s q_s}{\sigma} \varphi\left(\frac{S_- - \mu}{\sigma}\right) + 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) \\
& + \frac{\sum_{s=1}^m u_s q_s}{\sigma} \varphi\left(\frac{S_+ - \mu}{\sigma}\right) + o\left(\sum_{s=1}^m u_s q_s\right).
\end{aligned}$$

5.2. *Formulas for*  $\mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=1} 1_{X(t_k)=1} + 1_{X(t_s^*)=-1} 1_{X(t_k)=-1}\}]$   
*and*  $\mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=-1} 1_{X(t_k)=1} + 1_{X(t_s^*)=1} 1_{X(t_k)=-1}\}]$ .  
 First, let us recall that by definition we have  $t_1^* < t_2^* < \dots < t_m^*$ . Besides, let us consider a genetic marker located at  $t_k$ . We have

$$\begin{aligned}
 & \mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=1} 1_{X(t_k)=1}\}] \\
 &= \sum_{(u_1, \dots, u_{s-1}, u_{s+1}, \dots, u_m) \in \{-1, 1\}^{m-1}} \\
 & \mathbb{E} [1_{Y \notin [S_-, S_+]} 1_{X(t_1^*)=u_1} \dots 1_{X(t_{s-1}^*)=u_{s-1}} 1_{X(t_s^*)=1} 1_{X(t_{s+1}^*)=u_{s+1}} \dots 1_{X(t_m^*)=u_m} 1_{X(t_k)=1}] \\
 &= \sum_{(u_1, \dots, u_{s-1}, u_{s+1}, \dots, u_m) \in \{-1, 1\}^{m-1}} \\
 & \mathbb{P} \{Y \notin [S_-, S_+] \mid X(t_1^*) = u_1, \dots, X(t_{s-1}^*) = u_{s-1}, X(t_s^*) = 1, X(t_{s+1}^*) = u_{s+1}, \dots, X(t_m^*) = u_m\} \\
 & \mathbb{P} \{X(t_1^*) = u_1, \dots, X(t_{s-1}^*) = u_{s-1}, X(t_s^*) = 1, X(t_{s+1}^*) = u_{s+1}, \dots, X(t_m^*) = u_m, X(t_k) = 1\} \\
 &= \sum_{(u_1, \dots, u_{s-1}, u_{s+1}, \dots, u_m) \in \{-1, 1\}^{m-1}} \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) + o(1) \right\} \\
 & \mathbb{P} \{X(t_1^*) = u_1, \dots, X(t_{s-1}^*) = u_{s-1}, X(t_s^*) = 1, X(t_{s+1}^*) = u_{s+1}, \dots, X(t_m^*) = u_m, X(t_k) = 1\} \\
 &= \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) + o(1) \right\} \mathbb{P} \{X(t_s^*) = 1, X(t_k) = 1\} \\
 &= \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) \right\} \bar{r}(t_s^*, t_k)/2 + o(1).
 \end{aligned}$$

Using the same kind of proof, we have

$$\begin{aligned}
 & \mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=-1} 1_{X(t_k)=-1}\}] = \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) \right\} \bar{r}(t_s^*, t_k)/2 + o(1), \\
 & \mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=-1} 1_{X(t_k)=1}\}] = \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) \right\} r(t_s^*, t_k)/2 + o(1), \\
 & \mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=1} 1_{X(t_k)=-1}\}] = \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) \right\} r(t_s^*, t_k)/2 + o(1).
 \end{aligned}$$

As a result, we have the relationships

$$\begin{aligned}
 & \mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=1} 1_{X(t_k)=1} + 1_{X(t_s^*)=-1} 1_{X(t_k)=-1}\}] \\
 &= \bar{r}(t_k, t_s^*) \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) \right\} + o(1), \\
 & \mathbb{E} [1_{Y \notin [S_-, S_+]} \{1_{X(t_s^*)=-1} 1_{X(t_k)=1} + 1_{X(t_s^*)=1} 1_{X(t_k)=-1}\}] \\
 &= r(t_k, t_s^*) \left\{ 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right) \right\} + o(1).
 \end{aligned}$$

5.3. *Formula for  $\mathbb{E} \{ \sigma \varepsilon \overline{X}(t_k) \}$ .* We have

$$\begin{aligned}
& \mathbb{E} \{ \sigma \varepsilon \overline{X}(t_k) \} \\
&= \mathbb{E} \{ \sigma \varepsilon 1_{X(t_k)=1} 1_{Y \notin [S_-, S_+]} \} - \mathbb{E} \{ \sigma \varepsilon 1_{X(t_k)=-1} 1_{Y \notin [S_-, S_+]} \} \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{E} \{ \sigma \varepsilon 1_{X(t_k)=1} 1_{X(t_1^*)=u_1} \dots 1_{X(t_m^*)=u_m} 1_{Y \notin [S_-, S_+]} \} \\
&\quad - \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{E} \{ \sigma \varepsilon 1_{X(t_k)=-1} 1_{X(t_1^*)=u_1} \dots 1_{X(t_m^*)=u_m} 1_{Y \notin [S_-, S_+]} \} \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{E} \{ \sigma \varepsilon 1_{Y \notin [S_-, S_+]} \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&\quad [2\mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1 \dots X(t_m^*) = u_m \} - 1] \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \left\{ \sigma \varphi(z_{\gamma^+}) + z_{\gamma^+} \varphi(z_{\gamma^+}) \sum_{s=1}^m u_s q_s - \sigma \varphi(z_{1-\gamma^-}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-}) \sum_{s=1}^m u_s q_s \right\} \\
&\quad [2\mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} - 1] \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&\quad + o(\max_{1 \leq s \leq m} |q_s|) .
\end{aligned} \tag{5.1}$$

Note that in order to obtain the last expression, we used iv) of Lemma 5 of Rabier [6] (cf. Section 5.1). Recall that  $z_\alpha$  denotes the quantile of order  $1 - \alpha$  of a standard normal distribution. Let us focus on the quantity

$$\begin{aligned}
& \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \{ \sigma \varphi(z_{\gamma^+}) - \sigma \varphi(z_{1-\gamma^-}) \} [2\mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} - 1] \\
& \times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&= \{ \sigma \varphi(z_{\gamma^+}) - \sigma \varphi(z_{1-\gamma^-}) \} \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} 2 \mathbb{P} \{ X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&\quad - \{ \sigma \varphi(z_{\gamma^+}) - \sigma \varphi(z_{1-\gamma^-}) \} \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&= \{ \sigma \varphi(z_{\gamma^+}) - \sigma \varphi(z_{1-\gamma^-}) \} 2 \mathbb{P} \{ X(t_k) = 1 \} - \{ \sigma \varphi(z_{\gamma^+}) - \sigma \varphi(z_{1-\gamma^-}) \} = 0 .
\end{aligned} \tag{5.2}$$

Let us focus on the quantity

$$\begin{aligned}
& \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \left\{ z_{\gamma^+} \varphi(z_{\gamma^+}) \sum_{s=1}^m u_s q_s - z_{1-\gamma^-} \varphi(z_{1-\gamma^-}) \sum_{s=1}^m u_s q_s \right\} \\
& [2\mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} - 1] \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} .
\end{aligned}$$

Let  $\xi$  denote a given QTL. We have

$$\begin{aligned}
& \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} u_\xi q_\xi \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \\
& [2\mathbb{P}\{X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} - 1] \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\
& = \sum_{(u_1, \dots, u_{\xi-1}, u_{\xi+1}, \dots, u_m) \in \{-1, 1\}^{m-1}} q_\xi \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \\
& \times [2\mathbb{P}\{X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = 1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\} - 1] \\
& \times \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = 1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\} \\
& - q_\xi \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \\
& \times [2\mathbb{P}\{X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = -1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\} - 1] \\
& \times \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = -1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\} \\
& = q_\xi \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \sum_{(u_1, \dots, u_{\xi-1}, u_{\xi+1}, \dots, u_m) \in \{-1, 1\}^{m-1}} \\
& [2\mathbb{P}\{X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = 1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\} \\
& - \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = 1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\} \\
& - 2\mathbb{P}\{X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = -1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\} \\
& - \mathbb{P}\{X(t_1^*) = u_1, \dots, X(t_{\xi-1}^*) = u_{\xi-1}, X(t_\xi^*) = -1, X(t_{\xi+1}^*) = u_{\xi+1}, \dots, X(t_m^*) = u_m\}] \\
& = q_\xi \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \\
& \times [-\mathbb{P}\{X(t_\xi^*) = 1\} + \mathbb{P}\{X(t_\xi^*) = -1\} + 2\mathbb{P}\{X(t_k) = 1, X(t_\xi^*) = 1\} - 2\mathbb{P}\{X(t_k) = 1, X(t_\xi^*) = -1\}] \\
& = q_\xi \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \{\bar{r}(t_k, t_\xi^*) - r(t_k, t_\xi^*)\} \\
& = q_\xi \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \rho(t_k, t_\xi^*).
\end{aligned}
\tag{5.3}$$

As a result, according to formulae (5.1), (5.2) and (5.3), we have

$$\mathbb{E}\{\sigma_\varepsilon \bar{X}(t_k)\} = \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\} \sum_{s=1}^m \rho(t_s^*, t_k) q_s + o\left(\max_{1 \leq s \leq m} |q_s|\right).$$

5.4. *Formula for the quantity  $\mathbb{E} \{ \overline{X}(t_s^*) \overline{X}(t_s^*) \overline{X}(t_k) \}$ .* We have

$$\begin{aligned}
& \mathbb{E} \{ \overline{X}(t_s^*) \overline{X}(t_s^*) \overline{X}(t_k) \} \\
&= \mathbb{E} \{ 1_{X(t_s^*)X(t_s^*)X(t_k)=1} 1_{Y \notin [S_-, S_+]} \} - \mathbb{E} \{ 1_{X(t_s^*)X(t_s^*)X(t_k)=-1} 1_{Y \notin [S_-, S_+]} \} \\
&= \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = -u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) + o(1) \right\} \\
&\times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \mathbb{P} \{ X(t_k) = -1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&+ \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) + o(1) \right\} \\
&\times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&- \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = -u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) + o(1) \right\} \\
&\times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&- \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) + o(1) \right\} \\
&\times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \mathbb{P} \{ X(t_k) = -1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&= -2 \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = -u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) \right\} \mathbb{P} \{ X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&+ 2 \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) \right\} \mathbb{P} \{ X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&+ \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = -u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) \right\} \times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&- \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = u_s}} \left\{ \Phi \left( \frac{S_- - \mu}{\sigma} \right) + 1 - \Phi \left( \frac{S_+ - \mu}{\sigma} \right) \right\} \times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} + o(1) .
\end{aligned}
\tag{5.4}$$

Besides,

$$\begin{aligned}
& \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = -u_s}} \mathbb{P} \{ X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} = \mathbb{P} \{ X(t_k) = 1, X(t_s^*)X(t_s^*) = -1 \} \\
&= \mathbb{P} \{ X(t_s^*)X(t_s^*) = -1 \mid X(t_k) = 1 \} / 2
\end{aligned}$$

and

$$\begin{aligned} & \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = u_s}} \mathbb{P} \{X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\ &= \mathbb{P} \{X(t_k) = 1, X(t_s^*)X(t_{\bar{s}}^*) = 1\} = \mathbb{P} \{X(t_s^*)X(t_{\bar{s}}^*) = 1 \mid X(t_k) = 1\} / 2 . \end{aligned}$$

As a result,

$$\begin{aligned} & 2 \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = u_s}} \mathbb{P} \{X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\ & - 2 \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = -u_s}} \mathbb{P} \{X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\ &= 2\mathbb{P} \{X(t_s^*)X(t_{\bar{s}}^*) = 1 \mid X(t_k) = 1\} - 1 = 2\mathbb{P} \{X(t_s^*)X(t_{\bar{s}}^*) = 1\} - 1 = \rho(t_s^*, t_{\bar{s}}^*) . \end{aligned}$$

In the same way,

$$\begin{aligned} & \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = -u_s}} \mathbb{P} \{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\ & - \sum_{\substack{(u_1, \dots, u_m) \in \{-1, 1\}^m \\ u_{\bar{s}} = u_s}} \mathbb{P} \{X(t_1^*) = u_1, \dots, X(t_m^*) = u_m\} \\ &= \mathbb{P} \{X(t_s^*)X(t_{\bar{s}}^*) = -1\} - \mathbb{P} \{X(t_s^*)X(t_{\bar{s}}^*) = 1\} = -\rho(t_s^*, t_{\bar{s}}^*) . \end{aligned}$$

Then, according to formula (5.4), we have

$$\mathbb{E} \{ \overline{X}(t_s^*) \overline{X}(t_{\bar{s}}^*) \overline{X}(t_k) \} = o(1) .$$

It concludes the proof.

## References.

- [1] AZAÏS, J.M., DELMAS, C., and RABIER, C.E. (2012). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*, **48(4)** 787-801.
- [2] AZAÏS, J.M., GASSIAT, E., and MERCADIER, C. (2006). Asymptotic distribution and local power of the likelihood ratio test for mixtures. *Bernoulli*, **12(5)** 775-799.
- [3] AZAÏS, J.M., GASSIAT, E., and MERCADIER, C. (2009). The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM*, **13** 301-327.
- [4] AZAÏS, J.M., and WSCHBOR, M. (2009). *Level sets and extrema of random processes and fields*. Wiley, New-York.
- [5] GASSIAT, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré (B)*, **6** 897-906.
- [6] RABIER, C.E. (2014). On statistical inference for selective genotyping. *J. Stat. Plan. Infer.*, **147** 24-52.
- [7] RABIER, C.E. (2014). An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes. *Annales de la faculté des sciences de Toulouse*, **6(23)** 755-778.

- [8] RABIER, C.E. (2015). On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. *Statistics*, **49(1)** 19-34.
- [9] VAN DER VAART, A.W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.

CHARLES-ELIE RABIER  
ISE-M, UMR 5554  
CNRS, IRD  
UNIVERSITÉ DE MONTPELLIER, FRANCE  
E-MAIL: [ce.rabier@gmail.com](mailto:ce.rabier@gmail.com)

CELINE DELMAS  
INRA, UR 875 MIAT  
BP 52627, CASTANET-TOLOSAN  
31326 CEDEX, FRANCE  
E-MAIL: [celine.delmas@toulouse.inra.fr](mailto:celine.delmas@toulouse.inra.fr)

1.

## SUPPLEMENT B FOR: “ON GENE MAPPING WITH THE MIXTURE MODEL AND THE EXTREMES”

BY CHARLES-ELIE RABIER<sup>\*,†,‡</sup>, CÉLINE DELMAS<sup>‡</sup>

*ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France* <sup>\*</sup>

*LIRMM, UMR 5506, CNRS, Université de Montpellier, France* <sup>†</sup>

*MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France* <sup>‡</sup>

We propose to illustrate here our theoretical results regarding the max test. Recall that it relies on the test statistic,  $\sup \Lambda_n(\cdot)$ . The focus is on a sparse map: a chromosome of length 1M ( $T = 1$ ), with 21 markers ( $K = 21$ ) equally spaced every 5cM. In this context, Table 1 compares the theoretical power and the empirical power, under different configurations: either 1 QTL ( $m = 1$ ) at 3cM, either 2 QTLs ( $m = 2$ ) at 3cM and 28cM, or 3 QTLs ( $m = 3$ ) at 3cM, 28cM and 72cM. For all cases, the absolute value of the constant linked to the QTL effect was equal to 2.8284 (i.e.  $|a_s| = 2.8284$ ), allowing to deal with a small QTL effect of 0.2 when  $n = 200$ . The theoretical power was obtained by generating 10,000 paths of the asymptotic process, whereas 1,000 samples of size  $n$  equal to 1,000, 200 or 100 were considered for the empirical power. The threshold (i.e. critical value) at the 5% level was set to 7.84 using the Monte-Carlo Quasi Monte-Carlo method, proposed by Azaïs et al. [2] and based on Genz [3]. In order to compute the maximum of the process, simulated data were analyzed using Lemma 1 of Azaïs et al. [2], that is to say performing LRT on markers and performing only one test in each marker interval if the ratio of the score statistics on markers fulfills the given condition.

According to Table 1, we can notice a good agreement between the empirical power and the theoretical power for  $n = 200$ . However, the asymptotic seems to be really reached for  $n = 1,000$ . We also investigated the behavior of the test under a selective genotyping performed symmetrically (i.e.  $\gamma_+ = \gamma/2$ ). Recall that the threshold remains the same under selective genotyping (cf. Theorem 6.1 of the main text). We can observe that when  $\gamma = 0.3$ , the empirical power still matches the theoretical power for  $n = 1,000$ . This validates our theoretical results presented in Theorems 4.1 and 6.1 of the main text.



| $\gamma$ | $m$       |  | 1 (+)  | 2 (++) | 2 (+-) | 3 (++-) |
|----------|-----------|--|--------|--------|--------|---------|
|          | $n$       |  |        |        |        |         |
| 1        | $+\infty$ |  | 60.20% | 99.35% | 15.27% | 49.74%  |
|          | 1,000     |  | 59.7%  | 98.90% | 15.70% | 49.00%  |
|          | 200       |  | 60.00% | 98.80% | 15.50% | 47.30%  |
|          | 100       |  | 53.90% | 98.50% | 13.70% | 45.80%  |
| 0.3      | $+\infty$ |  | 48.21% | 97.47% | 12.71% | 39.36%  |
|          | 1,000     |  | 47.90% | 97.10% | 12.20% | 39.50%  |
|          | 200       |  | 47.70% | 96.80% | 10.50% | 37.50%  |
|          | 100       |  | 46.10% | 96.50% | 9.40%  | 32.80%  |

TABLE 1

*Theoretical power and empirical power associated to the test statistic  $\sup \Lambda_n(\cdot)$ , and as a function of the number  $m$  of QTLs and the percentage  $\gamma$  of genotyped individuals ( $T = 1$ ,  $K = 21$ ,  $t_k = 0.05(k - 1)$ , ( $m = 1$ ,  $t_1^* = 0.03$ ), ( $m = 2$ ,  $t_1^* = 0.03$ ,  $t_2^* = 0.80$ ), ( $m = 3$ ,  $t_1^* = 0.03$ ,  $t_2^* = 0.28$ ,  $t_3^* = 0.72$ ), all  $|a_s| = 2.828$ , + for positive effect, - for negative effect, 10,000 paths for the theoretical power, 1,000 samples of size  $n$  for the empirical power,  $\gamma_+/\gamma = 1/2$ ).*

Last, the power of the test is reported as a function of the QTL effect signs. We can see that when the two QTLs at 3cM and 28cM have the same signs, the power is almost equal to 1 whereas it largely decreases ( $\approx 15\%$  for  $\gamma = 1$ ) when the signs are opposite. In this case, the max test is clearly not the most appropriate test to perform. We refer to the recent study of [1] where the authors compared performances of the max test and the ANOVA in another context.

| $\gamma$ | $\gamma_+/\gamma$ | (Sparse, $n = 100$ ) |           | (Sparse, $n = 200$ ) |           | (Dense, $n = 100$ ) |           | (Dense, $n = 200$ ) |           |
|----------|-------------------|----------------------|-----------|----------------------|-----------|---------------------|-----------|---------------------|-----------|
|          |                   | L1 ratio             | $\hat{m}$ | L1 ratio             | $\hat{m}$ | L1 ratio            | $\hat{m}$ | L1 ratio            | $\hat{m}$ |
| 0.1      | 1/2               | 82.86%               | 14.74     | 91.29%               | 16.74     | 95.73%              | 17.15     | 98.39%              | 16.87     |
|          | 3/4               | 79.17%               | 15.35     | 90.91%               | 16.87     | 94.59%              | 16.52     | 98.26%              | 16.39     |
|          | 7/8               | 74.61%               | 15.89     | 89.85%               | 16.85     | 93.63%              | 17.11     | 98.69%              | 16.77     |
|          | 1                 | 68.87%               | 16.26     | 86.71%               | 16.69     | 92.77%              | 16.99     | 98.08%              | 16.63     |
| 0.2      | 1/2               | 73.43%               | 15.64     | 85.43%               | 16.74     | 94.18%              | 17.61     | 96.26%              | 16.93     |
|          | 3/4               | 71.27%               | 16.36     | 85.19%               | 16.80     | 94.01%              | 17.65     | 95.79%              | 16.53     |
|          | 7/8               | 68.19%               | 17.15     | 83.69%               | 16.77     | 93.43%              | 18.16     | 93.80%              | 17.25     |
|          | 1                 | 63.80%               | 16.95     | 81.04%               | 16.72     | 90.09%              | 17.15     | 92.18%              | 16.91     |
| 0.3      | 1/2               | 70.95%               | 16.59     | 83.48%               | 16.66     | 88.64%              | 16.70     | 96.50%              | 17.12     |
|          | 3/4               | 68.84%               | 15.39     | 81.77%               | 16.67     | 85.72%              | 17.71     | 95.24%              | 16.09     |
|          | 7/8               | 65.36%               | 15.75     | 79.48%               | 16.83     | 84.67%              | 16.93     | 94.17%              | 16.98     |
|          | 1                 | 61.76%               | 16.63     | 74.09%               | 16.74     | 79.96%              | 16.85     | 91.63%              | 16.56     |

TABLE 2

Performances of the new method as a function of the ratio  $\gamma_+/\gamma$  (Mean over 100 samples, on average  $n = 100$  individuals genotyped,  $m = 16$ ,  $|q_1| = \dots = |q_{16}| = 0.1$ ,  $T = 10$ , QTLs randomly located only on  $[0M, 4M]$ ). Sparse map:  $K = 201$ ,  $t_k = 0.05(k - 1)$ ,  $L = 401$ ,  $t'_l = 0.025(k - 1)$ . Dense map:  $K = L = 10,001$ ,  $t_k = t'_l = 0.001(k - 1)$ . The L1 ratio corresponds to the quantity  $\sum_{i=1}^{161} |\hat{\Delta}_i| / \sum_{i=1}^{401} |\hat{\Delta}_i|$  for the sparse map, and to the quantity  $\sum_{i=1}^{4001} |\hat{\Delta}_i| / \sum_{i=1}^{10001} |\hat{\Delta}_i|$  for the dense map.  $\hat{m}$  denotes the estimated QTL number.

| $\gamma$ | Our method |           | LASSO    |           | Group LASSO |           | IM(parametric) |           | IM(shuffle) |           | Lee and al. [4] |
|----------|------------|-----------|----------|-----------|-------------|-----------|----------------|-----------|-------------|-----------|-----------------|
|          | L1 ratio   | $\hat{m}$ | L1 ratio | $\hat{m}$ | L1 ratio    | $\hat{m}$ | L1 ratio       | $\hat{m}$ | L1 ratio    | $\hat{m}$ | L1 ratio        |
| 0.1      | 99.69%     | 16.64     | 31.43%   | 16.83     | 18.30%      | 22.61     | 100%           | 16        | 38%         | 4.5       | 22.41%          |
| 0.2      | 99.36%     | 17.79     | 24.53%   | 17.15     | 11.97%      | 29.1      | 99%            | 16        | 34%         | 4.36      | 20.52%          |
| 0.3      | 98.69%     | 17.5      | 42.93%   | 17.00     | 38.46%      | 19.1      | 99%            | 16        | 58%         | 8.28      | 25.22%          |

TABLE 3

Same framework as Table 2 of the manuscript, except that  $\gamma_+/\gamma = 1$  and  $|q_1| = \dots = |q_{16}| = 0.2$ .

**References.**

- [1] ARIAS-CASTRO, E., CANDÉS, E.J., and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, **39(5)** 2533-2556.
- [2] AZAÏS, J.M., DELMAS, C., and RABIER, C.E. (2012). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*, **48(4)** 787-801.
- [3] GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Stat.*, **1** 141-149.
- [4] LEE, H.I., HO, H.A., and KAO, C.H. (2014). A new simple method for improving QTL mapping under selective genotyping. *Genetics*, **198(4)** 1685-1698.

CHARLES-ELIE RABIER  
ISE-M, UMR 5554  
CNRS, IRD  
UNIVERSITÉ DE MONTPELLIER, FRANCE  
E-MAIL: [ce.rabier@gmail.com](mailto:ce.rabier@gmail.com)

CELINE DELMAS  
INRA, UR 875 MIAT  
BP 52627, CASTANET-TOLOSAN  
31326 CEDEX, FRANCE  
E-MAIL: [celine.delmas@toulouse.inra.fr](mailto:celine.delmas@toulouse.inra.fr)

1.

## SUPPLEMENT C FOR: “ON GENE MAPPING WITH THE MIXTURE MODEL AND THE EXTREMES”

BY CHARLES-ELIE RABIER<sup>\*,†,‡</sup>, CÉLINE DELMAS<sup>‡</sup>

*ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France* <sup>\*</sup>

*LIRMM, UMR 5506, CNRS, Université de Montpellier, France* <sup>†</sup>

*INRA, UR 875 MIAT, Castanet-Tolosan, France* <sup>‡</sup>

**1. Introduction.** We propose to focus here on the model introduced by Rebaï et al. [5] (see in particular their Section 2) in which double recombination between the QTL and its flanking markers is not allowed. Note that in Rebaï et al. [4], the authors extended their previous model to several markers, keeping Haldane [1] modeling for the genetic information at marker locations. As a result, the probability distribution of  $(X(t_1), \dots, X(t_K))$  is unchanged. In order to extend the interference model to  $m$  QTLs, we will impose that the QTLs do not belong to the same marker intervals. Obviously, double recombination between each QTL and the corresponding flanking markers is not allowed. Then, under the interference model, the “analysis of variance model” for the quantitative trait is the following:

$$(1.1) \quad Y = \mu + \sum_{s=1}^m U(t_s^*) q_s + \sigma \varepsilon$$

where  $\varepsilon$  is a Gaussian white noise, and  $U(t_s^*)$  is the analogue of  $X(t_s^*)$  under interference (more details in the next section).

When there are  $m$  QTLs, conditionally to  $X(t_1), \dots, X(t_K)$ , we will show that  $Y$  obeys to a mixture of  $2^m$  components

$$\sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \tilde{w}_{t^*}(u_1, \dots, u_m) f_{(\mu + u_1 q_1 + \dots + u_m q_m, \sigma)}(Y)$$

where  $\tilde{w}_{t^*}(u_1, \dots, u_m)$  is the probability  $\mathbb{P}\{U(t_1^*) = u_1, \dots, U(t_m^*) = u_m\}$  conditionally to the observations of the markers. Note that the weights  $\tilde{w}_{t^*}(u_1, \dots, u_m)$  are different from the weights  $w_{t^*}(u_1, \dots, u_m)$  obtained under Haldane (cf. formula 1.3 of the manuscript).

**2. Interference model and asymptotic results.** Let us consider the case  $m = 1$ . Then, under the model considered by Rebaï et al. [5], if the QTL is lying for instance between the first two markers (i.e.  $t_1^* \in ]t_1, t_2[$ ), we can not have the scenario  $X(t_1) = 1_1$   $U(t_1^*) = -1$  and  $X(t_2) = 1$ . Indeed, this would have supposed that there had been a recombination between the first marker and the QTL, and also a recombination between the second marker and the QTL. In particular, the model considers that if we have a

recombination between the QTL and one of its flanking marker, we could not have a recombination between the QTL and the other flanking marker. In other words, if  $X(t_1) = 1$  and  $U(t_1^*) = -1$ , then we have automatically  $X(t_2) = -1$ . In the same way, if  $X(t_2) = 1$  and  $U(t_1^*) = -1$ , then we have automatically  $X(t_1) = -1$ . As said before, Rebaï et al. [4] extended their previous model to several markers, keeping Haldane [1] modeling for the genetic information at marker locations. In other words, as previously,  $X(0)$  is a random sign and  $X(t_k) = X(0)(-1)^{N(t_k)}$  where  $N(\cdot)$  is a standard Poisson process on  $[0, T]$ . We will first study the classical model under interference, and later we will consider the epistatic model. Note that the results can easily be generalized to selective genotyping experiments.

To begin with, let us consider the case  $m = 1$  (i.e. one QTL on  $[0, T]$ ). According to Rabier [2], the likelihood of the triplet  $(Y, X(t^\ell), X(t^r))$  with respect to the measure  $\lambda \otimes N \otimes N$ ,  $\lambda$  being the Lebesgue measure,  $N$  the counting measure on  $\mathbb{N}$ , is  $\forall t \in ]t^\ell, t^r[$  :

$$(2.1) \quad \tilde{L}_t(\theta^1) = [\tilde{p}(t)f_{(\mu+q_1, \sigma)}(Y) + \{1 - \tilde{p}(t)\}f_{(\mu-q_1, \sigma)}(Y)] g(t)$$

where the function

$$\tilde{p}(t) = \mathbb{P} \left\{ U(t) = 1 \mid X(t^\ell), X(t^r) \right\} , \quad g(t) = \mathbb{P} \left\{ X(t^\ell), X(t^r) \right\} .$$

In particular,

$$(2.2) \quad \tilde{p}(t) = 1_{X(t^\ell)=1}1_{X(t^r)=1} + \frac{t^r - t}{t^r - t^\ell} 1_{X(t^\ell)=1}1_{X(t^r)=-1} + \frac{t - t^\ell}{t^r - t^\ell} 1_{X(t^\ell)=-1}1_{X(t^r)=1} .$$

As in the manuscript, the score statistic of the hypothesis “ $q_1 = 0$ ” at  $t$ , for  $n$  independent observations, will be defined as

$$\tilde{S}_n(t) = \frac{\frac{\partial \tilde{l}_t^n}{\partial q_1} |_{\theta_0^1}}{\sqrt{\mathbb{V} \left( \frac{\partial \tilde{l}_t^n}{\partial q_1} |_{\theta_0^1} \right)}} ,$$

where  $\tilde{l}_t^n(\theta^1)$  denotes the log likelihood at  $t$ , associated to  $n$  observations.

In the same way, the LRT at  $t$ , for  $n$  independent observations, will be defined as

$$\tilde{\Lambda}_n(t) = 2 \left\{ \tilde{l}_t^n(\hat{\theta}^1) - \tilde{l}_t^n(\hat{\theta}^1|_{H_0}) \right\} .$$

Let us now move on to the case  $m > 1$ .

Recall that we impose that the QTLs do not belong to the same marker intervals and that we consider Haldane modeling for the genome information

at genetic markers. As a result, since all the information is contained in the markers flanking the QTL locations, we have the relationship

$$\begin{aligned} & \mathbb{P} \{U(t_1^*), \dots, U(t_m^*) \mid X(t_1), \dots, X(t_K)\} \\ &= \mathbb{P} \left\{ U(t_1^*), \dots, U(t_m^*) \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r}) \right\} . \end{aligned}$$

The full likelihood of  $(Y, X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r}))$  is

$$(2.3) \quad \tilde{L}_{\vec{t}^*}^m(\theta^m) = \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \tilde{w}_{\vec{t}^*}(u_1, \dots, u_m) f_{(\mu + u_1 q_1 + \dots + u_m q_m, \sigma)}(Y) g^m(t_1^*, \dots, t_m^*)$$

where

$$\begin{aligned} \tilde{w}_{\vec{t}^*}(u_1, \dots, u_m) &= \mathbb{P} \left\{ U(t_1^*) = u_1, \dots, U(t_m^*) = u_m \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r}) \right\} , \\ g^m(t_1^*, \dots, t_m^*) &= \mathbb{P} \left\{ X(t_1^{\star\ell}), X(t_1^{\star r}), \dots, X(t_m^{\star\ell}), X(t_m^{\star r}) \right\} . \end{aligned}$$

**THEOREM 7.1.** *Suppose that the parameters  $(q_1, \dots, q_m, \mu, \sigma^2)$  vary in a compact and that  $\sigma^2$  is bounded away from zero, and also that  $m$  is finite. Then,*

$$\tilde{S}_n(\cdot) \Rightarrow W(\cdot) \quad , \quad \tilde{\Lambda}_n(\cdot) \xrightarrow{F.d.} W^2(\cdot) \quad , \quad \sup \tilde{\Lambda}_n(\cdot) \xrightarrow{\mathcal{L}} \sup W^2(\cdot)$$

as  $n$  tends to infinity, under  $H_0$  and  $H_{a\vec{t}^*}$  where  $W(\cdot)$  is the Gaussian process with unit variance such as

$$\begin{aligned} W(t) &= \frac{\tilde{\alpha}(t) W(t^\ell) + \tilde{\beta}(t) W(t^r)}{\sqrt{\tilde{\alpha}^2(t) + \tilde{\beta}^2(t) + 2\tilde{\alpha}(t)\tilde{\beta}(t)\rho(t^\ell, t^r)}} , \\ Cov \{W(t^\ell), W(t^r)\} &= \rho(t^\ell, t^r) = e^{-2|t^\ell - t^r|} \end{aligned}$$

where  $\tilde{\alpha}(t) = \frac{t^r - t}{t^r - t^\ell}$ ,  $\tilde{\beta}(t) = \frac{t - t^\ell}{t^r - t^\ell}$  and with mean function

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

$$\tilde{m}_{\vec{t}^*}(t) = \frac{\tilde{\alpha}(t) \tilde{m}_{\vec{t}^*}(t^\ell) + \tilde{\beta}(t) \tilde{m}_{\vec{t}^*}(t^r)}{\sqrt{\tilde{\alpha}^2(t) + \tilde{\beta}^2(t) + 2\tilde{\alpha}(t)\tilde{\beta}(t)\rho(t^\ell, t^r)}}$$

where

$$\begin{aligned}\tilde{m}_{\tilde{t}^*}(t^\ell) &= \sum_{s=1}^m a_s h(t^\ell, t_s^*) / \sigma \quad , \quad \tilde{m}_{\tilde{t}^*}(t^r) = \sum_{s=1}^m a_s h(t^r, t_s^*) / \sigma \quad , \\ \forall k \quad h(t_k, t_s^*) &= \rho(t_k, t_s^{\star\ell}) \left\{ \tilde{\alpha}(t_s^*) + \tilde{\beta}(t_s^*) \rho(t_s^{\star\ell}, t_s^{\star r}) \right\} 1_{t_s^* > t_k} \\ &\quad + \rho(t_k, t_s^{\star r}) \left\{ \tilde{\alpha}(t_s^*) \rho(t_s^{\star\ell}, t_s^{\star r}) + \tilde{\beta}(t_s^*) \right\} 1_{t_s^* < t_k} .\end{aligned}$$

The proof is given in the next section. Note that the functions  $\tilde{\alpha}(t)$  and  $\tilde{\beta}(t)$  are different from the ones of Theorems 3.1 and 4.1 of the manuscript. As in Rabier [2], the limiting process is the square of a linear interpolated process. As expected, the mean function depends now on the number of QTLs, their positions and their effects.

Let us now introduce Theorem 7.2, that deals with interference and epistasis.

**THEOREM 7.2.** *Suppose that the parameters  $(q_1, \dots, q_m, q_{1,2}, \dots, q_{m-1,m}, \mu, \sigma^2)$  vary in a compact and that  $\sigma^2$  is bounded away from zero, and also that  $m$  is finite. Then, with the previous notations, under  $H_{a\tilde{t}^*, b\tilde{t}^*}$ ,*

$$\tilde{S}_n(\cdot) \Rightarrow W(\cdot) \quad , \quad \tilde{\Lambda}_n(\cdot) \xrightarrow{F.d.} W^2(\cdot) \quad , \quad \sup \tilde{\Lambda}_n(\cdot) \xrightarrow{\mathcal{L}} \sup W^2(\cdot)$$

where  $W(\cdot)$  is the Gaussian process of Theorem 7.1 uncentered with mean function  $\tilde{m}_{\tilde{t}^*}(\cdot)$  defined in Theorem 7.1.

We can notice that the interaction effects are not present in the mean function.

### 3. Proofs.

#### 3.1. Proof of Theorem 7.1.

3.1.1. *Study of the score process under  $H_0$ .* The score process  $\tilde{S}_n(\cdot)$  has already been studied in details in Rabier [2], under  $H_0$ . We propose to recall here the main elements of the proof. By definition, the score statistic at  $t$  is the following

$$\tilde{S}_n(t) = \frac{\frac{\partial \tilde{l}_t^n}{\partial q_1} |_{\theta_0^1}}{\sqrt{\mathbb{V} \left( \frac{\partial \tilde{l}_t^n}{\partial q_1} |_{\theta_0^1} \right)}} \quad \text{where } \theta_0^1 = (0, \mu, \sigma) .$$



The score function verifies

$$\begin{aligned}
 \frac{\partial \tilde{l}_t^n}{\partial q_1} \Big|_{\theta_0^1} &= \sum_{j=1}^n \frac{Y_j - \mu}{\sigma^2} \{2\tilde{p}_j(t) - 1\} \\
 (3.1) \quad &= \frac{\tilde{\alpha}(t)}{\sigma} \sum_{j=1}^n \varepsilon_j X_j(t^\ell) + \frac{\tilde{\beta}(t)}{\sigma} \sum_{j=1}^n \varepsilon_j X_j(t^r) .
 \end{aligned}$$

Recall also that the score statistic at  $t_k$  which refers to the location of marker  $k$ , verifies:

$$\tilde{S}_n(t_k) = \sum_{j=1}^n \frac{(Y_j - \mu) X_j(t_k)}{\sqrt{n}} = \sum_{j=1}^n \frac{\sigma \varepsilon_j X_j(t_k)}{\sqrt{n}} .$$

Finally, according to formula (3.1) and the computation of the Fisher information matrix, the score statistic at  $t$  can be obtained using the following linear interpolation

$$\tilde{S}_n(t) = \frac{\tilde{\alpha}(t) \tilde{S}_n(t^\ell) + \tilde{\beta}(t) \tilde{S}_n(t^r)}{\sqrt{\tilde{\alpha}^2(t) + \tilde{\beta}^2(t) + 2\tilde{\alpha}(t)\tilde{\beta}(t)\rho(t^\ell, t^r)}} .$$

3.1.2. *Study of the score process under the local alternative  $H_{a\vec{t}^\star}$ .* Since the mean function is the following linear interpolation

$$\tilde{m}_{\vec{t}^\star}(t) = \frac{\tilde{\alpha}(t) \tilde{m}_{\vec{t}^\star}(t^\ell) + \tilde{\beta}(t) \tilde{m}_{\vec{t}^\star}(t^r)}{\sqrt{\tilde{\alpha}^2(t) + \tilde{\beta}^2(t) + 2\tilde{\alpha}(t)\tilde{\beta}(t)\rho(t^\ell, t^r)}} ,$$

we only need to compute the quantities  $\tilde{m}_{\vec{t}^\star}(t^\ell)$  and  $\tilde{m}_{\vec{t}^\star}(t^r)$ .

Without loss of generality, let us consider location  $t_k$ . According to formula (1.1),

$$\begin{aligned}
 \tilde{S}_n(t_k) &= \sum_{j=1}^n \frac{(Y_j - \mu) X_j(t_k)}{\sqrt{n}} \\
 &= \frac{1}{\sigma n} \sum_{j=1}^n \sum_{s=1}^m a_s U_j(t_s^\star) X_j(t_k) + \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j X_j(t_k) .
 \end{aligned}$$

As in the previous proofs, we can apply the law of large number to the first term. Then, we have

$$\frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^m U_j(t_s^\star) a_s \right\} X_j(t_k) \rightarrow \sum_{s=1}^m a_s h(t_k, t_s^\star) / \sigma$$

where

$$\begin{aligned} h(t_k, t_s^*) &= \rho(t_k, t_s^{\star\ell}) \left\{ \tilde{\alpha}(t_s^*) + \tilde{\beta}(t_s^*) \rho(t_s^{\star\ell}, t_s^{\star r}) \right\} 1_{t_s^* > t_k} \\ &\quad + \rho(t_k, t_s^{\star r}) \left\{ \tilde{\alpha}(t_s^*) \rho(t_s^{\star\ell}, t_s^{\star r}) + \tilde{\beta}(t_s^*) \right\} 1_{t_s^* < t_k} . \end{aligned}$$

Indeed, the quantity  $h(t_k, t_s^*)$  is equal to  $\mathbb{E}\{U_j(t_s^*)X_j(t_k)\}$ , which was computed in Theorem 2 of Rabier [2].

**3.2. Proof of Theorem 7.2.** Let us first introduce the model for the quantitative trait in presence of interference and epistasis:

$$(3.2) \quad Y = \mu + \sum_{s=1}^m U(t_s^*) q_s + \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m U(t_s^*) U(t_{\tilde{s}}^*) q_{s,\tilde{s}} + \sigma \varepsilon$$

where  $\varepsilon$  is a Gaussian white noise, and  $q_{s,\tilde{s}}$  is the interaction effect between loci  $t_s^*$  and  $t_{\tilde{s}}^*$ . Recall that we impose that the QTLs do not belong to the same marker intervals.

Since the process  $\tilde{S}_n(\cdot)$  is an interpolated process, we can focus, without loss of generality, only on location  $t_k$  (i.e. the location of marker  $k$ ). According to formulae (3.2), we have

$$\begin{aligned} \tilde{S}_n(t_k) &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j X_j(t_k) + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^m U_j(t_s^*) a_s \right\} X_j(t_k) \\ &\quad + \frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m U_j(t_s^*) U_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X_j(t_k) . \end{aligned}$$

According to the previous section, by the law of large number, we have

$$\frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^m U_j(t_s^*) a_s \right\} X_j(t_k) \rightarrow \sum_{s=1}^m a_s h(t_s^*, t_k) / \sigma .$$

In the same way,

$$\frac{1}{\sigma n} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m U_j(t_s^*) U_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X_j(t_k) \rightarrow \mathbb{E} \left[ \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m U(t_s^*) U(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} X(t_k) \right]$$

If  $t_k < \min(t_s^*, t_{\tilde{s}}^*)$  or  $t_k > \max(t_s^*, t_{\tilde{s}}^*)$ , since under the interference model  $t_s^*$  and  $t_{\tilde{s}}^*$  do not belong to the same marker interval, we have

$$\mathbb{E} [U(t_s^*) U(t_{\tilde{s}}^*) 1_{X(t_k)=1}] = \mathbb{E} [U(t_s^*) U(t_{\tilde{s}}^*) \mid X(t_k) = 1] / 2 = \mathbb{E} [U(t_s^*) U(t_{\tilde{s}}^*)] / 2 .$$

As a consequence,

$$\mathbb{E} \{U(t_s^*)U(t_s^*)X(t_k)\} = \mathbb{E} [U(t_s^*)U(t_s^*) \{21_{X(t_k)=1} - 1\}] = 0 .$$

Let us consider now the case  $\min(t_s^*, t_s^*) < t_k < \max(t_s^*, t_s^*)$ . By definition, conditionnally to  $X(t_k)$ ,  $U(t_s^*)$  and  $U(t_s^*)$  are independent. As a consequence, we have the relationship

$$\begin{aligned} & \mathbb{E} \{U(t_s^*)U(t_s^*)X(t_k)\} \\ &= \mathbb{E} \{U(t_s^*)U(t_s^*) \mid X(t_k) = 1\} \mathbb{P}\{X(t_k) = 1\} - \mathbb{E} \{U(t_s^*)U(t_s^*) \mid X(t_k) = -1\} \mathbb{P}\{X(t_k) = -1\} \\ &= \mathbb{E} \{U(t_s^*) \mid X(t_k) = 1\} \mathbb{E} \{U(t_s^*) \mid X(t_k) = 1\} / 2 \\ &\quad - \mathbb{E} \{U(t_s^*) \mid X(t_k) = -1\} \mathbb{E} \{U(t_s^*) \mid X(t_k) = -1\} / 2. \end{aligned}$$

Besides, we have

$$\begin{aligned} \mathbb{E} \{U(t_s^*)\} &= \frac{1}{2} \mathbb{E} \{U(t_s^*) \mid X(t_k) = 1\} + \frac{1}{2} \mathbb{E} \{U(t_s^*) \mid X(t_k) = -1\} \\ \mathbb{E} \{U(t_s^*)\} &= \frac{1}{2} \mathbb{E} \{U(t_s^*) \mid X(t_k) = 1\} + \frac{1}{2} \mathbb{E} \{U(t_s^*) \mid X(t_k) = -1\} . \end{aligned}$$

It is easy to check that  $U(t_s^*)$  and  $U(t_s^*)$  take value  $+1$  and  $-1$  with equal probability (cf. formula 2.2). Then,

$$\begin{aligned} \mathbb{E} \{U(t_s^*) \mid X(t_k) = 1\} &= -\mathbb{E} \{U(t_s^*) \mid X(t_k) = -1\} \\ \mathbb{E} \{U(t_s^*) \mid X(t_k) = 1\} &= -\mathbb{E} \{U(t_s^*) \mid X(t_k) = -1\} . \end{aligned}$$

As a result,

$$\mathbb{E} \{U(t_s^*)U(t_s^*)X(t_k)\} = 0 .$$

This gives the result.

**4. Reverse configuration of selective genotyping.** In this section, we will come back to Haldane modeling (i.e. without interference) and we propose to take the reverse configuration of selective genotyping. Indeed, sometimes, for some biological reasons, we are only able to genotype the non extreme individuals (i.e. the individuals for which  $Y \in [S_-, S_+]$ ). In this context, we present the following lemma.

**LEMMA 6.3.** *Under the reverse configuration, that is to say if  $\bar{X}(t_k) = X(t_k) 1_{Y \in [S_-, S_+]}$ , then we have the same results as in Theorem 6.1 and Theorem 6.2 provided that we replace the quantity  $\mathcal{A}$  by the quantity*

$\mathcal{B}$  defined in the following way

$$\mathcal{B} = \sigma^2 \{1 - \gamma - z_{\gamma_+} \varphi(z_{\gamma_+}) + z_{1-\gamma_-} \varphi(z_{1-\gamma_-})\} .$$

The proof is largely inspired from the proof of Theorem 6.1 and 6.2 of the manuscript, and also from Rabier [3] where this configuration is studied under the local alternative of one QTL at  $t^*$  on  $[0, T]$ .

### References.

- [1] HALDANE, J.B.S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*, **8**, 299-309.
- [2] RABIER, C.E. (2014). On Quantitative Trait Locus mapping with an interference phenomenon. *TEST*, **23(2)**, 311-329.
- [3] RABIER, C.E. (2014). An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes. *Annales de la faculté des sciences de Toulouse*, **6(23)** 755-778.
- [4] REBAÏ, A., GOFFINET, B., and MANGIN, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138**, 235-240.
- [5] REBAÏ, A., GOFFINET, B., and MANGIN, B. (1995). Comparing power of different methods for QTL detection. *Biometrics*, **51**, 87-99.

CHARLES-ELIE RABIER  
 ISE-M, UMR 5554  
 CNRS, IRD  
 UNIVERSITÉ DE MONTPELLIER, FRANCE  
 E-MAIL: [ce.rabier@gmail.com](mailto:ce.rabier@gmail.com)

CELINE DELMAS  
 INRA, UR 875 MIAT  
 BP 52627, CASTANET-TOLOSAN  
 31326 CEDEX, FRANCE  
 E-MAIL: [celine.delmas@toulouse.inra.fr](mailto:celine.delmas@toulouse.inra.fr)