



HAL
open science

The SgLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection

Charles-Elie Rabier, Céline Delmas

► To cite this version:

Charles-Elie Rabier, Céline Delmas. The SgLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection. 2019. hal-02123295v1

HAL Id: hal-02123295

<https://hal.science/hal-02123295v1>

Preprint submitted on 7 May 2019 (v1), last revised 10 Nov 2021 (v5)

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

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The SgLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection

C.E. Rabier^{a,b}, C. Delmas^c

^a*ISEM, Université de Montpellier, CNRS, France*

^b*LIRMM, Université de Montpellier, CNRS, France*

^c*INRA, UR875 Mathématiques et Informatique Appliquées Toulouse, F-313326 Castanet-Tolosan, France*

Abstract. We introduce a new variable selection method, called SgLasso, that handles extreme data, and 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. Besides, we prove that the signal to noise ratio is largely increased by considering the extremes. Our method relies on the construction of a specific statistical test, a transformation of the data and by the knowledge of the correlation between regressors. 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 point to the validity of our approach.

Keywords: Gaussian process, Extremes, Genomic Selection, High Dimension, Linear Model, Variable Selection, Prediction, Sparsity.

1. Preliminaries

1.1. Motivation

Today, more and more genomic data are available thanks to advances in molecular biology and to technology. This makes statistical science very exciting for geneticists, statisticians and mathematicians always eager to propose new methods (e.g. [Momen et al. \(2018\)](#)).

Genomics and mathematics, two fields not expanding at the same speed, are sometimes complementary. Old-fashioned tools, studied deeply by mathematicians, may be of importance for the genomic community. In this context, we propose to introduce here a new variable selection method, relying on an old concept, called selective genotyping, and that meets big data needs. Although genotyping costs have largely dropped recently, selective genotyping or extreme sampling, is still a relevant concept in the modern genomic era. It was first introduced by [Lebowitz et al. \(1987\)](#) who noticed that most of the information about Quantitative Trait Loci, so-called QTL (genes influencing a quantitative trait which is able to be measured) is present in the extreme phenotypes (i.e. extreme traits). Later, [Lander and Bostein \(1989\)](#) formalized this approach and called it selective genotyping. Today, application fields of selective genotyping lie in Genome Wide Association Study (GWAS), and in Genomic Selection (GS). We can also find applications in biotechnology [Zou et al. \(2016\)](#).

The aim of GWAS is to find associations between locations (i.e. loci) of the genome and a trait of interest. We denote some recent association studies using selective genotyping in plants (e.g. sugarcane [Gutierrez et al. \(2018\)](#)), soybean [Phansak et al. \(2016\)](#); [Yan et al. \(2017\)](#); [Vuong et al. \(2016\)](#)), chickpea [Upadhyaya et al. \(2016\)](#), tomatoes [Ohlson et al. \(2018\)](#)), in animals (e.g. dairy cattle [Kurz et al. \(2019\)](#)),

drosophila [Bastide et al. \(2013\)](#), sow [Cordoba et al. \(2015\)](#), mouse [Fernandes et al. \(2016\)](#)), and in humans (e.g. on KashinBeck disease [Zhang et al. \(2014\)](#) and on intelligence [Zabaneh et al. \(2018\)](#)). Selective genotyping is particularly rewarding for finding QTLs: by considering the extremes, the signal is significantly increased.

Genomic Selection (GS), motivated by the seminal paper of [Hayes et al. \(2001\)](#), can be considered as the most promising application field of selected genotyping in years to come. GS is today a very hot topic in genomics (e.g. strawberry [Genzan et al. \(2017\)](#), banana [Nyine et al. \(2018\)](#)) and it consists in predicting breeding values of selection candidates using a large number of genetic markers (see [Rabier et al. \(2019\)](#)). The goal is to predict the future phenotype of young candidates as soon as their DNA has been collected. GS considers many generations and in order to maintain the reliability of the prediction model over time, the model has to be updated. Then, the model is learned on extreme individuals, selected at the previous generation because of their favorable genomic predictions (see our Section 7 for more details).

1.2. On high dimensional data

There are many issues related to high dimensional data. As mentioned in [Fan and Lv \(2008\)](#), one of the challenges lies in the fact 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 [Wu et al. \(2007\)](#) 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 our new variable selection method, called SgLasso.

The number of selected variables by SgLasso, is bounded by the number of predictors, instead of being bounded by the number of observations as in the classical Lasso ([Tibshirani \(1996\)](#)). 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. Moreover, in high dimensional problem, it is well known that the sparse coefficient should be large enough (see for instance [Bühlmann and Van de Geer \(2011\)](#)) in order to recover the true model (β -min condition). We prove that the signal to noise ratio is largely increased thanks to the selective genotyping approach. Last but not least, SgLasso, that handles extreme data, enjoys all known statistical properties of Lasso since the problem has been replaced in a L1 penalized regression framework. Typically, it is not the case for Lasso in presence of extreme data.

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. [Cierco \(1998\)](#)). Note that we could have focused on an evolutionary process such as the Wright Fisher model. In what follows, although we consider the backcross framework for the sake of readability, results are also suitable for phased data with biallelic genetic markers (alleles +1 or -1).

2. Introduction

As in [Broman and Speed \(2002\)](#), 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 \(1919\)](#) which states that:

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

where for any $b \in \mathbb{R}$, δ_b denotes the point mass at b and $N(\cdot)$ is a standard Poisson process on $[0, T]$. In a more practical point of view, the [Haldane \(1919\)](#) model assumes no crossover interference and the Poisson process represents the number of crossovers on $[0, T]$ which happen during meiosis.

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 [Hayes \(2007\)](#)). Let q_s and t_s^* denote respectively the QTL effect and the location of the s th QTL. Besides, we consider $0 < t_1^* < \dots < t_m^* < T$. We assume an “analysis of variance model” for the quantitative trait:

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

where ε is a Gaussian white noise.

In fact the “genome information” is 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 is

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

Moreover we consider the selective genotyping approach [Lebowitz et al. \(1987\)](#). At a given power, a large increase of the number of individuals leads to a decrease of the number of individuals genotyped. Note that in practice, the cutoffs for genotyping are based on quantiles. However, in most of the theoretical studies about selective genotyping (e.g. [Darvasi and Soller \(1992\)](#); [Muranty and Goffinet \(1997\)](#)), authors consider fixed thresholds. This approximation is reasonable when we deal with a large number of observations. In this context, let S_- and S_+ be two real thresholds, 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 $\bar{X}(t)$ the random variable such as

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

then, in our problem, one observation is now

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

Note that when $Y \notin [S_-, S_+]$, we have $\bar{X}(t_1) = X(t_1), \dots, \bar{X}(t_K) = X(t_K)$ and when $Y \in [S_-, S_+]$, we have $\bar{X}(t_1) = 0, \dots, \bar{X}(t_K) = 0$, which means that the genome information is missing at the marker locations. So, we observe n observations $(Y_j, \bar{X}_j(t_1), \dots, \bar{X}_j(t_K))$ i.i.d.

The main 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 . It will be helpful for building a prediction model in GS. For that, we consider the score statistic and the LRT statistic of the hypothesis “ $q_1 = 0$ ” at each location t , assuming $m = 1$. It leads

to a “score process”, $\bar{S}_n(\cdot)$, and to a “LRT process”, $\bar{\Lambda}_n(\cdot)$. This technique that consists in scanning the genome, is called the “Interval Mapping” (Lander and Bostein (1989)) and is very popular in genetics (see Wu et al. (2007); Siegmund and Yakir (2007)). In the complete data situation where all the individuals are genotyped (i.e. $S_- = S_+$) and assuming $m = 1$, the distributions of the score and LRT processes have been studied extensively by Rebaï et al. (1995, 1994); Cierco (1998); Azaïs and Cierco-Ayrolles (2002); Chen and Chen (2005); Azaïs and Wschebor (2009); Chang et al. (2009); Azaïs et al. (2012). In contrast, under selective genotyping, the result is relatively new for $m = 1$ since the asymptotic distribution was given recently in Rabier (2015, 2014c). These two studies were complementary to the work of Rabbee et al. (2004) and Manichaikul et al. (2007), relying on simulated data. In the past, the authors mainly focused on the theory of selective genotyping at only one fixed location t of the genome (e.g. Lander and Bostein (1989); Darvasi and Soller (1992); Muranty and Goffinet (1997); Rabier (2014a)).

Selective genotyping is challenging since some correlation is present between the errors ε and the genome of extreme individuals. In order to show the influence of this correlation, let us consider $m \geq 1$. At the marker location t_k , the score statistic, $\bar{S}_n(t_k)$, can be decomposed in the following way (cf. formula (9) in Section 2 of Supplement A):

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

where \mathcal{A} is a quantity linked to the choice of S_- and S_+ (see formula (9) in Section 3). By imposing $q_s = a_s/\sqrt{n}$, we can apply under this local alternative, the Law of Large Numbers and the Central Limit Theorem for the first and the second term, respectively (see for instance Van der Vaart (1998)). Then, according to a technical proof (cf. Section 4 of Supplement A), we have the relationship

$$\sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n \mathcal{A}}} \xrightarrow{\mathcal{L}} \mathcal{N}[\Omega, 1]$$

where Ω is a function of $a_1, \dots, a_m, t_1^*, \dots, t_m^*, t_k, S_-$ and S_+ . As a consequence, the correlation between ε and $\bar{X}(t_k)$ plays a role in the asymptotic theory. In contrast, under the complete data situation ($S_- = S_+$), the random variable $\bar{X}(t_k)$, equal to $X(t_k)$, is independent of ε : since ε is centered, Ω is the constant null function.

2.1. Roadmap

Section 3 introduces our main result, Theorem 1, that gives the asymptotic distribution of the score process and the LRT process under the alternative hypothesis that there exists m QTLs located at t_1^*, \dots, t_m^* with effects q_1, \dots, q_m . Under this general alternative, the LRT process is still asymptotically the square of a “non linear interpolated process”, as in Theorem 4.1 of Rabier (2015) where the focus was only on the case $m = 1$ under selective genotyping. Besides, as in Rabier (2015), the difference between the complete data situation and the selective genotyping approach is traducted by a difference between the mean functions of the asymptotic processes: they are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$. However, contrary to Rabier (2015) and Azaïs et al. (2012), the mean function depends here on the number of QTLs, their positions and their effects.

These theoretical results allow us to propose a new method, called SgLasso, to estimate the number of QTLs, their positions and their effects using the Lasso (Tibshirani (1996)). SgLasso differs from the classical Lasso since it models explicitly the extremes. As its famous ancestor Lasso, SgLasso has multiple cousins, each one imposing its own penalty on parameters: we can cite for instance SgElasticNet (a mixture of L1 and L2 penalties) and SgGroupLasso (penalty by group).

Note that Theorem 1 gives also the asymptotic distribution of the statistic $\sup \bar{\Lambda}_n(\cdot)$ when $m \geq 1$, since this test can be viewed as a global test or max test (see for instance [Arias-Castro et al. \(2011\)](#)). In this context, $\sup \bar{\Lambda}_n(\cdot)$ 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 SgLasso, in order to look for “some signal” on the chromosome.

Lemma 1 gives the Asymptotic Relative Efficiency (ARE) with respect to the complete data situation. Recall that the ARE determines the sample size required to obtain the same local asymptotic power as the one of the test under the complete data situation where all the genotypes are known. According to Lemma 1, we have exactly the same ARE as the one obtained in [Rabier \(2015\)](#) for $m = 1$. Theorem 2 shows that the signal is largely increased by genotyping extreme individuals, provided that the phenotyping is free. Corollary 1 deals with interactions between QTLs (so-called epistasis phenomenon). Indeed, it is well known that interactions can be responsible for a non-negligible part of the genetic variability of a quantitative trait (see for instance [Wu et al. \(2007\)](#)). According to Corollary 1 that gives the asymptotic properties of the LRT process, interaction effects are unidentifiable since they are not present in the mean function of the process. Last, Corollary 2 tackles the reverse configuration of selective genotyping, where only non extreme individuals are genotyped (i.e. the individuals for which $Y \in [S_-, S_+]$).

In Section 4, we describe our new gene mapping method, SgLasso, and one of its cousin SgElasticNet (SgEN). In particular, we explain how to combine the theoretical results of Section 3 and the chosen penalized likelihood method. Section 5 investigates theoretical properties of SgLasso, such as the rate of convergence for prediction and the consistency of the variable selection. Next, Section 6 illustrates performances of our new method and proposes a comparison with existing methods in a GWAS context. As expected, the signal to noise ratio is largely increased by considering extreme individuals. SgLasso and its cousins outperformed existing methods (Lasso, [Tibshirani \(1996\)](#), Group Lasso, [Yuan and Lin \(2006\)](#), Elastic Net, [Zhou and Hastie \(2005\)](#), RaLasso, [Fan et al. \(2017\)](#) and BayesianLasso, [Park and Casella \(2008\)](#)), specially when a unidirectional selective genotyping was performed (i.e. when only the best individuals were genotyped). Recall that SgLasso models explicitly the fact that \bar{X} and ε are not independent. Section 6.4 is devoted to a rice data analysis. Our study ends with Section 7 dedicated to GS: the SgLasso presented the best performances for genomic prediction.

3. Some theoretical results

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

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

For $t \in [t_1, t_K] \setminus \mathbb{T}_K$ where $\mathbb{T}_K = \{t_1, \dots, t_K\}$, we define t^ℓ and t^r as :

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

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

Let us consider the case $m = 1$ (i.e. one QTL located at t_1^*), and let $\theta^1 = (q_1, \mu, \sigma)$ be the parameter of the model at t fixed. Since all the information is contained in the flanking markers of the putative QTL location t , the focus is only on the triplet $(Y, \bar{X}(t^\ell), \bar{X}(t^r))$. According to [Rabier \(2015\)](#), the likelihood

of $(Y, \bar{X}(t^\ell), \bar{X}(t^r))$ with respect to the measure $\lambda \otimes N \otimes N$, λ being the Lebesgue measure, N the counting measure on \mathbb{N} , is $\forall t \in [t_1, t_K] \setminus \mathbb{T}_K$:

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

where $f_{(\mu, \sigma)}$ is the Gaussian density with parameters (μ, σ) and

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

with

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

We can notice that we have

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

Moreover we have:

$$\bar{g}(t) = \mathbb{P} \{ X(t^\ell), X(t^r) \} 1_{Y \notin [S_-, S_+]} + 1_{Y \in [S_-, S_+]} \quad (3)$$

with

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

As a result, 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 (2) will be very convenient for the generalization to several QTLs. Note that the true probability distribution is $\bar{L}_{t_1^*}(\theta^1)$. The score statistic of the hypothesis “ $q_1 = 0$ ” at t , for n independent observations, is defined as

$$\bar{S}_n(t) = \frac{\frac{\partial \bar{l}_t^n}{\partial q_1} \Big|_{\theta_0^1}}{\sqrt{\mathbb{V} \left(\frac{\partial \bar{l}_t^n}{\partial q_1} \Big|_{\theta_0^1} \right)}}, \quad (4)$$

where \bar{l}_t^n denotes the log likelihood at t , associated to n observations, and $\theta_0^1 = (0, \mu, \sigma)$ refers to the parameter θ_1 under \mathcal{H}_0 .

In the same way, the LRT statistic at t , for n independent observations, is defined as

$$\bar{\Lambda}_n(t) = 2 \left\{ \bar{l}_t^n(\hat{\theta}^1) - \bar{l}_t^n(\hat{\theta}^1|_{H_0}) \right\}, \quad (5)$$

where $\hat{\theta}^1$ is the maximum likelihood estimator (MLE) of the parameters (q_1, μ, σ) , and $\hat{\theta}^1|_{\mathcal{H}_0}$ the MLE under H_0 . As previously said, the processes $\bar{S}_n(\cdot)$ and $\bar{\Lambda}_n(\cdot)$ respectively defined by (4) and (5) for $t \in [0, T]$ are respectively called the score process and the LRT process.

3.1. Main results

Before giving our first main result, let us define the following quantities:

$$\gamma := \mathbb{P}_{\mathcal{H}_0}(Y \notin [S_-, S_+]) \quad (6)$$

$$\gamma_+ := \mathbb{P}_{\mathcal{H}_0}(Y > S_+) \quad (7)$$

$$\gamma_- := \mathbb{P}_{\mathcal{H}_0}(Y < S_-) \quad (8)$$

$$\mathcal{A} := \sigma^2 \{ \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) \} \quad (9)$$

where $\varphi(x)$ and z_α denote respectively the density of a standard normal distribution taken at the point x , and the quantile of order $1 - \alpha$ of a standard normal distribution.

Remark: When there is no selective genotyping (complete data situation): $\gamma = 1$, $\gamma_+ + \gamma_- = 1$ and $\mathcal{A} = \sigma^2$.

Notations 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 first main result is given in the following theorem. We obtain the asymptotic distribution of the score process $\bar{S}_n(\cdot)$ and the LRT process $\bar{\Lambda}_n(\cdot)$ under the null hypothesis that there is no QTL on $[0, T]$ and under the general hypothesis that there exist m QTLs on $[0, T]$. The originality is that the test processes are constructed under the hypothesis that there is a QTL at t and we look for their asymptotic distributions under the general hypothesis that there exist m QTLs on $[0, T]$. This leads to asymptotic processes with mean function depending on the locations and effects of the m QTLs. Using a variable selection method we will propose in the next section a new QTL detection procedure.

Theorem 1. *Suppose that the parameters $(q_1, \dots, q_m, \mu, \sigma^2)$ vary in a compact and that σ^2 is bounded away from zero, and also that m is finite. Let \mathcal{H}_0 be the null hypothesis of no QTL on $[0, T]$, and let define the following local alternatives $\mathcal{H}_{a\vec{t}^*}$: “there are m QTLs located respectively at t_1^*, \dots, t_m^* 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, as n tends to infinity,*

$$\bar{S}_n(\cdot) \Rightarrow V(\cdot) \quad , \quad \bar{\Lambda}_n(\cdot) \xrightarrow{F.d.} V^2(\cdot) \quad , \quad \sup \bar{\Lambda}_n(\cdot) \xrightarrow{\mathcal{L}} \sup V^2(\cdot) \quad (10)$$

under \mathcal{H}_0 and $\mathcal{H}_{a\vec{t}^*}$ where $V(\cdot)$ is the Gaussian process with unit variance such as

$$V(t) = \frac{\alpha(t) V(t^\ell) + \beta(t) V(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} \quad ,$$

$$\text{Cov}\{V(t_k), V(t_{k'})\} = \rho(t_k, t_{k'}) = e^{-2|t_k - t_{k'}|} \quad \forall (k, k') \in \mathbb{T}_K \times \mathbb{T}_K$$

with $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$, $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$. The mean function of $V(\cdot)$ is such that:

- under \mathcal{H}_0 , $\bar{m}(t) = 0$
- under $\mathcal{H}_{a\vec{t}^*}$,

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

where

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

and \mathcal{A} is defined in (9). When there is no selective genotyping (complete data situation), $\mathcal{A} = \sigma^2$.

The case $m > 1$ differs from the case $m = 1$ since the true probability distribution is the one of $(Y, \bar{X}(t_1^{\star\ell}), \bar{X}(t_1^{\star r}), \dots, \bar{X}(t_m^{\star\ell}), \bar{X}(t_m^{\star r}))$. Indeed, all the information is contained in the flanking markers of all QTLs locations. This probability distribution and the proof of Theorem 1 are given respectively in Section 1 and Section 2 of Supplement A.

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 1. *Let κ denote the Asymptotic Relative Efficiency, then we have*

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

where $\varphi(x)$ and z_α denote respectively the density of a standard normal distribution taken at the point x , and the quantile of order $1 - \alpha$ of a standard normal distribution.

This lemma is a generalization of Theorem 4.2 of Rabier (2015) where the focus was only on the case $m = 1$. To prove Lemma 1, just use the same proof as the one of Theorem 4.2 of Rabier (2015).

According to i) of Lemma 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{A}/σ , it is obvious that the ARE does not depend on those parameters. On the other hand, according to ii) of Lemma 1, if we want to genotype only a percentage γ of the population, we should genotype the $\gamma/2\%$ individuals with the largest phenotypes and $\gamma/2\%$ individuals with the smallest phenotypes.

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, following the same lines as the proof of Theorem 1, we obtain:

Theorem 2. *Suppose that the parameters $(q_1, \dots, q_m, \mu, \sigma^2)$ vary in a compact and that σ^2 is bounded away from zero, and also that m is finite. Assume that $n^* = n/\gamma$. Let \mathcal{H}_0 be the null hypothesis of no QTL on $[0, T]$, and let define the following local alternatives $\mathcal{H}_{a\vec{t}^*}$: “there are m QTLs located respectively at t_1^*, \dots, t_m^* 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, as n^* tends to infinity,*

$$\bar{S}_{n^*}(\cdot) \Rightarrow V_\star(\cdot) \quad , \quad \bar{\Lambda}_{n^*}(\cdot) \xrightarrow{F.d.} V_\star^2(\cdot) \quad , \quad \sup \bar{\Lambda}_{n^*}(\cdot) \xrightarrow{\mathcal{L}} \sup V_\star^2(\cdot) \quad (11)$$

under \mathcal{H}_0 and $\mathcal{H}_{a\vec{t}^*}$ where $V_\star(\cdot)$ is the Gaussian process with unit variance such as

$$V_\star(t) = \frac{\alpha(t) V_\star(t^\ell) + \beta(t) V_\star(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} \quad ,$$

$$Cov\{V_\star(t_k), V_\star(t_{k'})\} = \rho(t_k, t_{k'}) = e^{-2|t_k - t_{k'}|} \quad \forall (k, k') \in \mathbb{T}_K \times \mathbb{T}_K$$

with $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$, $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$. The mean function of $V_\star(\cdot)$ is such that:

- under \mathcal{H}_0 , $\bar{m}^*(t) = 0$
- under $\mathcal{H}_{a\bar{t}^*}$,

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

where

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

and \mathcal{A} is defined in (9). When there is no selective genotyping (complete data situation), $\mathcal{A} = \sigma^2$ and $\gamma = 1$.

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

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

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

3.2. Some corollaries

3.2.1. Model with interactions

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 Wu et al. (2007)). 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) for the quantitative trait becomes

$$Y = \mu + \sum_{s=1}^m X(t_s^*) q_s + \sum_{s=1}^{m-1} \sum_{\bar{s}=s+1}^m X(t_s^*) X(t_{\bar{s}}^*) q_{s,\bar{s}} + \sigma \varepsilon \quad (12)$$

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

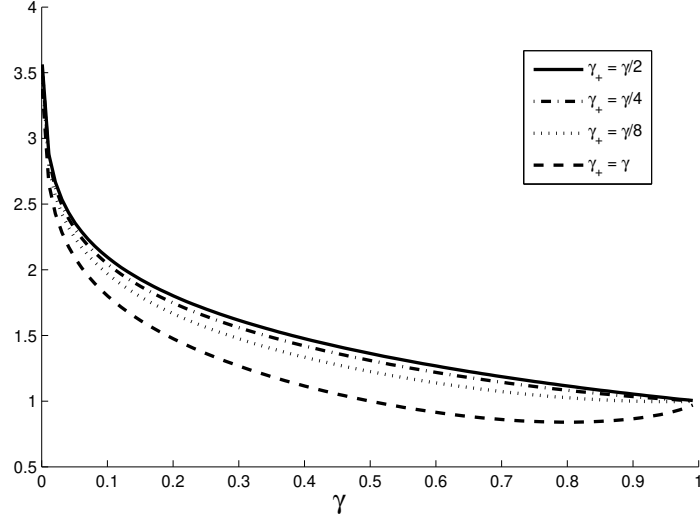


Figure 1: Function $\sqrt{z_{\gamma+} \varphi(z_{\gamma+})/\gamma - z_{1-\gamma-} \varphi(z_{1-\gamma-})/\gamma + 1}$ as a function of the percentage γ of individuals genotyped and as a function of the ratio γ_+/γ .

Corollary 1. *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 σ^2 is bounded away from zero, and also that m is finite. Let define the local alternative*

- $\mathcal{H}_{a\vec{t}^*, b\vec{t}^*}$: “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 $\mathcal{H}_{a\vec{t}^*, b\vec{t}^*}$, as n or n^* tends to infinity, results (10) and (11) of Theorem 1 and Theorem 2 hold.

The proof is given in Section 3 of Supplement A. 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.

3.2.2. The reverse configuration

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 result.

Corollary 2. *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 1, Theorem 2 and Corollary 1 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 of the proof of Theorem 1, Theorem 2, Corollary 1, and also from Rabier (2014b) where this configuration is studied under the local alternative of one QTL at t^* on $[0, T]$.

4. 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 Theorems 1 and 2 and a penalized likelihood method.

Notations 2. $\mathcal{G}_{\gamma,\sigma}$ denotes respectively $\frac{\sqrt{A}}{\sigma}$ or $\frac{\sqrt{A}}{\sqrt{\gamma}\sigma}$ when the total number of phenotypic observations is n or $n^* = n/\gamma$.

In the sequel \tilde{n} denotes the total number of phenotypic observations. It may be n or n^* . According to Theorems 1 and 2, as soon as we discretize the score process at markers positions, we have the following relationship when \tilde{n} is large:

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

where $\vec{S}_{\tilde{n}} = (\bar{S}_{\tilde{n}}(t_1), \bar{S}_{\tilde{n}}(t_2), \dots, \bar{S}_{\tilde{n}}(t_K))'$, $\vec{m}_{\tilde{t}^*} = (\bar{m}_{\tilde{t}^*}(t_1), \bar{m}_{\tilde{t}^*}(t_2), \dots, \bar{m}_{\tilde{t}^*}(t_K))'$ 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 $\vec{S}_{\tilde{n}}$ keeping only points of the process taken at marker positions. Recall that $\bar{S}_{\tilde{n}}(\cdot)$ is an ‘‘interpolated process’’. Let us consider the Cholesky decomposition $\Sigma = AA'$. We have

$$A^{-1}\vec{S}_{\tilde{n}} = A^{-1}B \left(\frac{a_1\mathcal{G}_{\gamma,\sigma}}{\sigma}, \dots, \frac{a_m\mathcal{G}_{\gamma,\sigma}}{\sigma} \right)' + A^{-1}\vec{\varepsilon} + o_P(1)$$

where B is a matrix of size $K \times m$ such as $B_{ks} = e^{-2|t_k - t_s^*|}$, $k = 1, \dots, K$ and $s = 1, \dots, m$.

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: $0 \leq t'_1 < t'_2 < \dots < t'_L \leq T$. Note that although we focus only on the discretized process at markers locations, we look for QTL not only on markers. We note Δ_l the putative effect at location t_l . The model can be rewritten in the following way:

$$A^{-1}\vec{S}_{\tilde{n}} = A^{-1}C(\Delta_1, \dots, \Delta_L)' + A^{-1}\vec{\varepsilon} + o_P(1) \quad (13)$$

where C is a matrix of size $K \times L$ such as $C_{kl} = e^{-2|t_k - t'_l|}$, $k = 1, \dots, K$ and $l = 1, \dots, L$.

Last, in order to find the non zero Δ_l , a natural approach is to use a penalized regression and estimate Δ by:

$$\hat{\Delta}_{\text{Sg}}(\lambda, \alpha) = \arg \min_{\Delta} \left(\left\| A^{-1}\vec{S}_{\tilde{n}} - A^{-1}C\Delta \right\|_2^2 + \lambda \text{pen}(\alpha) \right) \quad (14)$$

where:

$$\text{pen}(\alpha) = \frac{1-\alpha}{2} \|\Delta\|_2^2 + \alpha \|\Delta\|_1 \quad (15)$$

and $\|\cdot\|_2$ is the L2 norm, $\|\cdot\|_1$ is the L1 norm, $\Delta = (\Delta_1, \dots, \Delta_L)'$ and λ and α denote tuning parameters. We define:

$$\hat{\Delta}_{\text{SgLasso}}(\lambda) = \hat{\Delta}_{\text{Sg}}(\lambda, 1) \text{ and } \hat{\Delta}_{\text{SgEN}}(\lambda, \alpha) = \hat{\Delta}_{\text{Sg}}(\lambda, \alpha). \quad (16)$$

Another estimator, based on the group Lasso penalty, will be studied. We leave the study of the Ridge estimator, $\hat{\Delta}_{\text{Sg}}(\lambda, 0)$, for future research, since this estimator is only helpful for prediction.

Our estimators will be compared in section 6.3 with the classical estimators such as the Lasso (Tibshirani (1996))

and its cousins (e.g. [Zhou and Hastie \(2005\)](#); [Yuan and Lin \(2006\)](#)). These classical estimators consider exclusively marker locations. In order to describe a few of them under selective genotyping, let us define β_0 the global mean and β_k the putative effect of marker k . We set $\beta = (\beta_0, \beta_1, \dots, \beta_K)'$. In addition, let M_{ext} denote the matrix, where each row contains the multivariate random variable $(1, X(t_1), \dots, X(t_K)) \mid Y \notin [S_-, S_+]$ associated to an extreme individual. In the same way, Y_{ext} refers to the column vector containing the phenotypes of the extreme individuals. Indeed, since the genome information is unknown for the non extreme individuals, the classical estimators are built only on extreme individuals. According to these notations, the classical Lasso estimator $\hat{\beta}_{\text{Lasso}}(\lambda)$, and the classical Elastic Net estimator $\hat{\beta}_{\text{EN}}(\lambda, \alpha)$ are the following under selective genotyping:

$$\hat{\beta}_{\text{Lasso}}(\lambda) = \arg \min_{\beta} \left(\|Y_{\text{ext}} - M_{\text{ext}}\beta\|_2^2 + \lambda \|\beta\|_1 \right) \quad (17)$$

$$\hat{\beta}_{\text{EN}}(\lambda, \alpha) = \arg \min_{\beta} \left(\|Y_{\text{ext}} - M_{\text{ext}}\beta\|_2^2 + \lambda \left\{ \frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right\} \right). \quad (18)$$

Note that the Elastic Net penalty is described here in its version implemented in the R package GLMNet that will be used on simulated data.

5. Asymptotic theory for SgLasso under complete Linkage Disequilibrium

Before studying the theory of SgLasso, we have to give precisions regarding prediction and variable selection of SgLasso. As its cousin Lasso, SgLasso is able to select variables and these findings are considered as QTLs. Recall that SgLasso presents the advantage over its cousin to handle extreme data. On the other hand, in terms of prediction, we have to highlight the fact that SgLasso (in its version declined in formula (16)) will only predict values of a decorrelated score process. In what follows, we propose to investigate the rate of convergence for this prediction and we will also give conditions for consistent variable selection. We refer to Section 7 for the prediction of the phenotypes Y .

Let us assume that we are under complete Linkage Disequilibrium, i.e. the m QTLs are located on some markers. Furthermore, let us consider exclusively marker locations, i.e. $L = K$ and $t'_l = t_k$. We have the relationships $C = \Sigma$, $A^{-1}C = A'$ and $\Delta = (\Delta_1, \dots, \Delta_K)'$. When Δ_k is null, the corresponding marker is not a QTL, whereas a non-null Δ_k refers to a QTL.

According to formulas (14) and (16), our L1 penalized regression is:

$$\hat{\Delta}_{\text{SgLasso}}(\lambda) = \arg \min_{\Delta} \left(\|A^{-1}\vec{S}_n - A'\Delta\|_2^2 + \lambda \|\Delta\|_1 \right). \quad (19)$$

Let us normalize all covariables on the same scale. It will replace our problem in the classical setting where the theory for Lasso is well known (cf. [Bühlmann and Van de Geer \(2011\)](#) page 108). Since $\hat{\sigma}_k^2 := \frac{1}{K}(A'A)_{kk} = \frac{\Sigma_{kk}}{K} = \frac{\rho(t_k, t_k)}{K} = \frac{1}{K}$, let us set $A'_{\text{scal}} := \sqrt{K}A'$. Then, let us define

$$\hat{\Delta}_{\text{SgLasso}_{\text{scal}}}(\lambda) := \arg \min_{\Delta} \left(\frac{\|A^{-1}\vec{S}_n - A'_{\text{scal}}\Delta/\sqrt{K}\|_2^2}{K} + \lambda \left\| \frac{\Delta}{\sqrt{K}} \right\|_1 \right).$$

As soon as we set $\Delta' := \Delta/\sqrt{K}$, this problem can be rewritten in the following way:

$$\hat{\Delta}_{\text{SgLasso}_{\text{scal}}}(\lambda) := \arg \min_{\Delta'} \left(\frac{\|A^{-1}\vec{S}_n - A'_{\text{scal}}\Delta'\|_2^2}{K} + \lambda \|\Delta'\|_1 \right). \quad (20)$$

We can apply Corollary 6.1 of [Bühlmann and Van de Geer \(2011\)](#) with $\hat{\sigma} = 1$ (cf. our linear model in formula (13)), that establishes the slow rate of convergence

$$\left\| \frac{A'_{\text{scal}}(\hat{\Delta}_{\text{SgLasso}_{\text{scal}}} - \Delta')}{K} \right\|_2^2 = O_P \left(\sqrt{\frac{\log(K)}{K}} \sum_{s=1}^m \frac{|a_s| \mathcal{G}_{\gamma, \sigma}}{\sigma \sqrt{K}} \right) \quad (21)$$

where $O_P(1)$ denotes a sequence that is bounded in probability when $K \rightarrow +\infty$.

On the other hand, assuming that the “compatibility condition” holds, Corollary 6.2 of [Bühlmann and Van de Geer \(2011\)](#) applies and we obtain the fast rate of convergence:

$$\left\| \frac{A'_{\text{scal}}(\hat{\Delta}_{\text{SgLasso}_{\text{scal}}} - \Delta')}{K} \right\|_2^2 = O_P \left(\frac{\log(K) m}{K \Phi_0^2} \right) \quad (22)$$

where Φ_0^2 is a compatibility constant. Recall that the number of QTLs m is the factor linked to the sparsity.

Last, in order to make things clearer for future users, we propose to state the classical Lasso conditions in the “SgLasso” context.

The β -min condition:

$$\min_{1 \leq s \leq m} \frac{|a_s| \mathcal{G}_{\gamma, \sigma}}{\sigma \sqrt{K}} \gg \Phi^{-2} \sqrt{\frac{m \log(K)}{K}}$$

where Φ^2 is a restricted eigen value of the design matrix A'_{scal} .

Recall that $\mathbb{T}_K = \{t_1, \dots, t_K\}$ and that Σ is the $K \times K$ matrix, where $\Sigma_{kk'} = \rho(t_k, t_{k'})$. Note that $A_{\text{scal}} A'_{\text{scal}} / K = AA' = \Sigma$.

The bounded pairwise correlation:

$$\frac{\sqrt{m} \max_{k \in \mathbb{T}_K \setminus \mathbb{T}_m^*} \sqrt{\sum_{s \in \mathbb{T}_m^* | t_s^* \neq t_k} \rho^2(t_k, t_s^*)}}{d_{\min}^2(\Sigma^{(\cdot, \cdot)})} \leq C < 1 \quad (23)$$

where $\mathbb{T}_m^* = \{t_1^*, \dots, t_m^*\}$, C is a constant, $\Sigma^{(\cdot, \cdot)}$ is the submatrix of Σ restricted to QTL loci, and $d_{\min}^2(\Sigma^{(\cdot, \cdot)})$ refers to the smallest eigenvalue of $\Sigma^{(\cdot, \cdot)}$.

The irrepresentable condition:

$$\left\| \Sigma^{(\cdot, \cdot)} (\Sigma^{(\cdot, \cdot)})^{-1} \text{Sign}(a_1, \dots, a_m) \right\|_{\infty} \leq C < 1$$

where $\|x\|_{\infty} = \max_j |x_j|$, $\text{Sign}(a_1, \dots, a_m) = (\text{Sign}(a_1), \dots, \text{Sign}(a_m))'$, and $\Sigma^{(\cdot, \cdot)}$ is a matrix of size $(K - m) \times m$. $\Sigma^{(\cdot, \cdot)}$ is the submatrix of Σ where rows refers to markers not matching QTL locations, and where columns refers to QTL loci.

Note that according to [Bühlmann and Van de Geer \(2011\)](#), the bounded pairwise correlation implies the irrepresentable condition, which implies the compatibility condition. This compatibility condition ensures the fast rate of convergence for prediction (cf. formula (22)). On the other hand, the β -min condition and the irrepresentable condition, ensure consistent variable selection for SgLasso under selective genotyping.

6. Illustrations regarding max test and GWAS

In what follows, the variance σ^2 is set to 1 in all simulated data.

6.1. 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 γ is equal to 0.3 or 1). This validates our theoretical results presented in Theorem 1.

6.2. 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 4) 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 $[0,4]$ whereas no QTLs are present on the rest of the genome (i.e. $[6,10]$). 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 (13)). 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. We considered n^* individuals under selective genotyping and n individuals under the complete data situation. In other words, our simulation set up follows Theorem 2.

For instance, when n was equal to 100 ($\gamma = 1$), n^* took the values 1000, 500 and 333 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 Δ_l , present in formula (13), 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 γ 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 $[0,4]$ to L1 norm of estimated effects on $[0,10]$. 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 $[0,4]$. 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.

Last, Table 2 of Supplement B focuses on different ways of performing the selective genotyping: different ratios γ_+/γ are investigated under both maps. As expected, when only the largest (or the

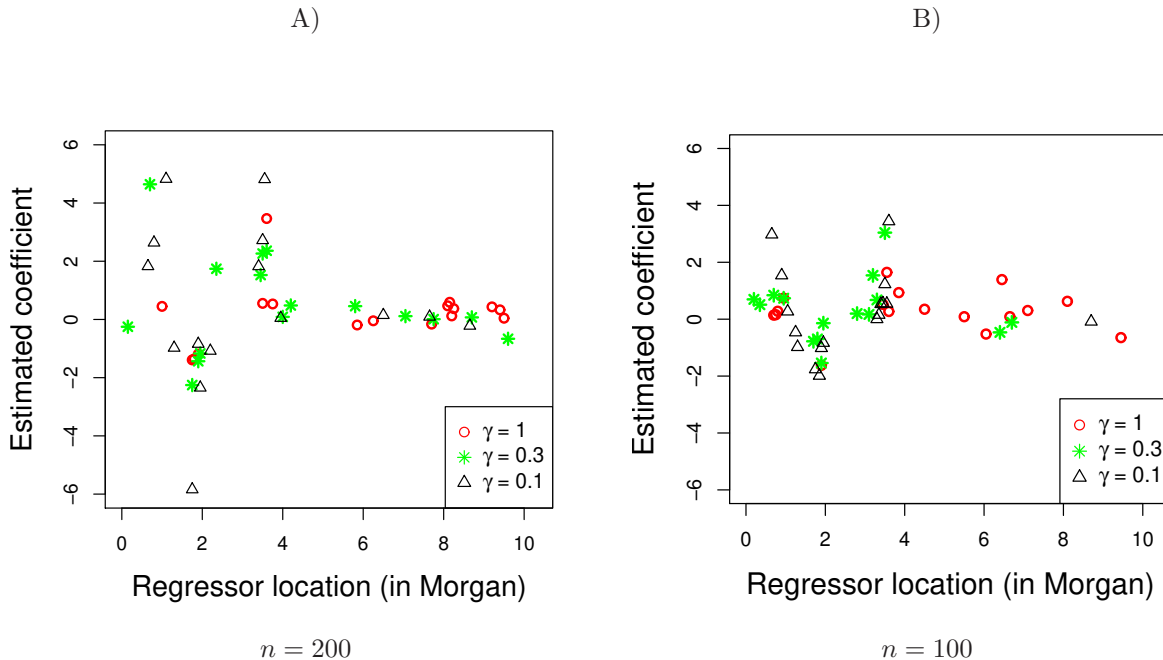


Figure 2: Estimated coefficients according to our new method as a function of the percentage γ of genotyped individuals (1 sample, $m = 16$, $T = 10$, $|q_1| = \dots = |q_{16}| = 0.2$, QTLs randomly located only on $[0,4]$, $\sigma = 1$, $K = 201$, $t_k = 0.05(k - 1)$, $L = 401$, $t'_l = 0.025(k - 1)$,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

smallest) individuals are genotyped ($\gamma_+/\gamma = 1$), the L1 ratio is the smallest. It confirms our theoretical results presented in Section 3 and illustrated in Figure 1.

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

6.3. Comparison with existing methods

In this section, we propose to compare our new method with existing methods. We will concentrate on the Lasso (Tibshirani (1996)), the Group Lasso (Yuan and Lin (2006)), the Elastic Net (Zhou and Hastie (2005)), the Bayesian Lasso (Park and Casella (2008)), and the RaLasso (Fan et al. (2017)).

Recall that the Group Lasso differs from his cousin Lasso, since it allows to handle a group structure (see Hastie et al. (2001)). In the context of genomic prediction, the Bayesian Lasso was used in Boligon et al. (2012) under selective genotyping. Contrary to the Lasso, the Bayesian LASSO guarantees an unimodal full posterior, since it relies on a conditional Laplace prior. The last method studied here, is the so-called RaLasso (Fan et al. (2017)). It presents the advantage of handling conditional errors ϵ that are asymmetric and that follow heavy-tailed distributions.

In what follows, the Group Lasso is based on groups of 10 consecutive markers. For Elastic Net, the value of the parameter α was set to 0.5 (cf. formula (18)). The Elastic Net, Group Lasso, Bayesian Lasso and RaLasso were computed with the help of the R packages, GLMNet, gglasso, SafeBayes and hqreg, respectively.

Recall the Huber loss considered in the package hqreg : $\text{loss}(t) = \frac{t^2}{2M} 1_{|t| \leq M} + (|t| - M/2) 1_{|t| > M}$, where M is a tuning parameter. Huber loss is quadratic for absolute values less than M and linear for those greater than M . As soon as we multiply by $2M$ and that we replace M by α^{-1} , we obtain formula

all $ q_s $	γ	(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 SgLasso as a function of the percentage γ 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 $[0,4]$, $\sigma = 1$). 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.

(2.2) of [Fan et al. \(2017\)](#). Last, we have to mention that the RaLasso incorporates the Huberloss and a L1 penalty.

Recall that in our simulation framework, the number of QTLs m was set to 16. For Lasso, Elastic Net and Group Lasso, the tuning parameter was selected in two steps. We first chose a few λ values leading to models with approximately 16 parameters. Then, we performed cross validation in order to select the best model. Same remark for the Bayesian Lasso except that the learning rate η replaces λ . In order to compute the RaLasso, we ran a grid search to find the best pair (M, λ) .

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 $[0,1]$, still considering a genome of size 10M. We considered different ways of performing the selective genotyping, by letting the ratio γ_+/γ 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 γ_+/γ took the values $1/2$, $3/4$ or $7/8$. However, when a unidirectional selective genotyping was performed ($\gamma_+/\gamma = 1$), the Lasso, Group Lasso, Elastic Net and RaLasso deteriorated heavily, which was not the case of our SgLasso method. For instance, when γ was set to 0.1, the power associated to the Lasso, Group Lasso, Elastic Net, Bayesian Lasso was found to be equal to 20.78%, 16.73% and 21.00%, respectively. The Lasso and its cousins suffer from the fact that the tails of errors are not light, and that the conditional distribution is asymmetric around 0 (see for instance [Fan et al. \(2017\)](#)). The RaLasso, that models heavy tails and asymmetry, gave better results (47.01%) than these methods but was still far from performances of SgLasso (93.97%). Last, the Bayesian Lasso performed badly in all the configurations studied. Table 3 deals with the case $|q_s|$ equal to 0.2: although the signal had been increased, we observed the same behaviour of the different methods. Table 4 compares performances of SgLasso and its cousins. SgLasso and SgEN presented similar results, whereas the SgGroupLasso seemed to select too many genes under this simulation setting.

6.4. Real data analysis

To illustrate performances of our new method on real data, we analyzed data from the joint papers [Spindel et al. \(2015\)](#) and [Begum et al. \(2015\)](#) 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. \(2015\)](#)), 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 [Spindel et al. \(2015\)](#)), we can infer that a marker is located every 0.1cM. Then, we performed 5 fold cross validation for all methods. As previously, a grid search was used for RaLasso in order to find the best pair (M, λ) . In particular, we considered the values $\lambda = 0.1, 10.1, \dots, 1000.1$ and $M = 0.1, 0.2, \dots, 1$. The percentage γ of genotyped individuals was set to either 1 or 0.3 and the selective genotyping was performed symmetrically ($\gamma_+/\gamma = 1/2$). Since [Begum et al. \(2015\)](#) 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. \(2015\)](#) ($n = 312$ by averaging the replicates), whereas for $\gamma = 0.3$, we kept the genome information of only 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. \(2015\)](#) (cf. their S1 Table), and obtained after fitting a linear mixed model, are given at the top of Table 7. Note that the most significant SNPs for the flowering date are located on chromosome 3 (see [Begum et al. \(2015\)](#)). Indeed, the pvalues associated to 5 SNPs on chromosome 3 and reported by [Begum et al. \(2015\)](#), are the following: 5.02×10^{-27} for the so-called gene S3-1269941, 1.47×10^{-24} for S3-1165376, 1.82×10^{-23} for S3-1125848, 2.80×10^{-22} for S3-1394477, and 1.49×10^{-21} for S3-1221494. The number of false positives (FP) and the number of false negatives (FN) are also reported in Table 7. FP refers to the number of falsely selected variables whereas FN is the number of genes that are not selected.

According to Table 7, SgLasso and SgEN selected respectively 26 and 33 genes under the complete data situation ($\gamma = 1$). All the genes found by [Begum et al. \(2015\)](#) and present on chromosome 3, were either perfectly found by SgLasso and SgEN or were tagged by a marker located nearby (at less than a distance of 4 markers, i.e. 0.4cM). In contrast, SgGroupLasso’s performances were not as fair since SgGroupLasso was unable to select the gene S3-1394477, even when a tolerance level of 0.4cM was used. Classical methods such as Lasso, EN and Group Lasso, found respectively 3, 4 and 3 (or 4 with the tolerance level) genes matching the findings of [Begum et al. \(2015\)](#) on chromosome 3. In that sense, when γ was set to 1, SgLasso and SgEN performed better than traditional methods. We can also highlight the fact that RaLasso was unsatisfactory, exhibiting thousands of False Positives.

Let us now move on to selective genotyping. The selective genotyping was performed symmetrically ($\gamma_+/\gamma = 1/2$). For $\gamma = 0.3$, SgLasso, SgGroupLasso and SgEN selected 4, 5 and 5 genes, respectively, corresponding to those suggested by [Begum et al. \(2015\)](#) on chromosome 3. Lasso, Group Lasso and EN were able to recover 2, 3 and 5 genes, respectively. In other words, we observed the superiority of SgLasso (resp. SgGroupLasso) over Lasso (resp. GroupLasso). SgEN and EN presented both fair results, with a slight advantage to EN that exhibited only 2 FP. Moreover, as previously, RaLasso gave poor results on this dataset.

To conclude, in order to show the strength of our methods, we tackled the case $\gamma_+/\gamma = 1$. However, due to a lack of signal and a small sample size, all methods were unable to recover the findings of [Begum et al. \(2015\)](#). Recall that the unidirectional selective genotyping is the worst configuration. Contrary to our simulation studies, we were unable to increase the sample size to compensate this small amount of signal. We leave it for future research.

7. A promising application field of SgLasso in the future : Genomic Selection

Genomic Selection (GS) (Hayes et al. (2001)) can be considered as the most promising application field of SgLasso in years to come. Recall that it consists in predicting breeding values of selection candidates using a large number of genetic markers: the goal is to predict the future phenotype of young candidates as soon as their DNA has been collected. GS was first applied to animal breeding (see Hayes et al. (2009) for a review), and it is nowadays extensively investigated in plants. We can mention recent genomic prediction studies on apple (Muranty et al. (2015)), eucalyptus (Tan et al. (2017)), japanese pears (Minamikawa et al. (2018)), strawberry (Genzan et al. (2017)), banana (Nyine et al. (2018)) and coffea (Ferrao et al. (2008)). GS allows to consider a large number of generations without having to observe the future adult phenotype. For instance, in citrus, 25 years are required to obtain fruits of interest. In bananas, the waiting time can reach 8 months, in order to figure out the production capacity.

Many studies (e.g. Wolc et al. (2011); Pszczola and Calus (2016); Rabier et al. (2016); Auinger et al. (2016); Neyhart et al. (2017)) have shown that it is essential to update the learning model during GS cycles in order to maintain the reliability of the prediction model over time. When updating the calibration model, the model is learned on extreme individuals, selected at the previous generation because of their favorable genomic predictions. In that sense, this area of research in GS is highly linked to selective genotyping. GS differs slightly from selective genotyping because individuals are selected on the basis of genomic prediction, instead of being selecting according to their phenotypes. However, in practice, there is only a very small difference in considering predicted or true phenotypes (cf. experiments 1 and 2 of Brandariz and Bernardo (2018)). Zhao et al. (2012) highlighted the “drastic reduction” in terms of predictive ability when only the best individuals (i.e. with the largest phenotypes) were used in the learning model in GS. Interestingly, Brandariz and Bernardo (2018) has shown recently that it is crucial to include a few worst individuals in the training set, to keep GS efficient. As soon as only the best individuals were included in the training set, the model was not reliable anymore (see Table 1 of Brandariz and Bernardo (2018)). However, keeping the poorest lines in a breeding program has a non negligible cost. In this context, we will show below on simulated data that SgLasso and its cousins do not suffer from this drawback: they give satisfactory results even when only best individuals are considered. In other words, there is a strong agreement with results from our association study in Section 6.3 (cf. Tables 2 and 3).

7.1. Mathematical model and comparison with existing methods

As mentioned in introduction, A and B are homozygous lines. In order to generate candidates, let us cross the extreme backcross individuals to their parent A , that is to say performing the cross $(A \times (A \times B))_{\text{ext}} \times A$ where $(A \times (A \times B))_{\text{ext}}$ refers to the backcrossed individuals that are extremes (cf. Figure 1 in Supplement B). From a theoretical point of view, let $X_{\text{ext}}(t)$ denote the random variable $X(t) \mid Y \notin [S_-, S_+]$, i.e. the genome at t of an extreme individual, and let $R(\cdot)$ denote a standard poisson process on $[0, T]$ representing the number of recombinations. $W(\cdot)$, the random process such as $W(t) = X_{\text{ext}}(t)1_{R(t) \text{ even}} - 1_{R(t) \text{ odd}}$, will refer to the genome of the progeny of an extreme individual (taken at random among all extreme individuals). The quantitative trait of this progeny, noted U , is based on the ANOVA model: $U = \mu + \sum_{s=1}^m W(t_s^*)q_s + \sigma\varepsilon$, where ε is a Gaussian white noise.

In what follows, the notation “new” will refer to the progeny of an extreme individual ; $U_{\text{new}}, W_{\text{new}}(\cdot), \varepsilon_{\text{new}}, R_{\text{new}}(\cdot)$ are random variables or processes associated to this new individual. In GS, the quality of the prediction is evaluated according to some accuracy criteria, i.e. the correlation between predicted and true values. This criterion is a key element in genetics: it plays a role in the rate of genetic gain (see for instance Lynch and Walsh (1998)). The *phenotypic accuracy*, ρ_{ph} , also called predictive ability, is defined as the correlation between the predictor \hat{U}_{new} and the trait U_{new} , i.e. $\text{Cor}(\hat{U}_{\text{new}}, U_{\text{new}})$ (see for instance Visscher et al. (2010)). We propose to compare here the accuracy associated to the classical predictor

and the one relying on our method. These two estimators have respectively the following expressions: $\widehat{U}_{\text{new}} = (1, W_{\text{new}}(t_1), \dots, W_{\text{new}}(t_K)) \widehat{\beta}_{\text{LASSO}}$ and $\widehat{U}_{\text{new}} = (W_{\text{new}}(t_1), \dots, W_{\text{new}}(t_K)) \widehat{\Delta}_{\text{SgLasso}}(\lambda) \frac{\sigma \sqrt{\gamma}}{\sqrt{n_A}}$. We will also investigate accuracies of the cousins of the different predictors. To clarify, each simulated data set rely on 100 progenies and each progeny is a descendent of an extreme individual taken at random among all extremes. The model is learned on all extreme individuals and evaluated on the progenies. Pearson correlation was computed between predicted values and true values. In this context, Tables 5 and 6 report the average Pearson correlation computed over 100 data sets containing 100 progenies.

According to Table 5, when the model was learned on the best individuals ($\gamma_+/\gamma = 1$), we clearly observed the superiority of the SgLasso over other methods, regarding the predictive ability. As soon as a few worst individuals were included in the learning model ($\gamma_+/\gamma = 7/8$), all the different methods gave similar results. As mentioned before, these results were expected in view of our previous association study (Tables 2 and 3). Recall that Brandariz and Bernardo (2018) already observed, using classical methods, that it was crucial to include a few worst individuals in the model. In contrast, our method presents good prediction abilities even when only best individuals are considered. Last, Table 6 compares SgLasso and its cousins : SgLasso, SgEN and SgGroupLasso, presented an accuracy of same order.

Supplementary files

Supplement A : Supplementary proofs

We give the mathematical proofs of Theorem 1 and Corollary 1.

Supplement B : Supplementary illustrations

We illustrate our theoretical results regarding the max test, relying on the test statistic, $\sup \Lambda_n(\cdot)$.

Acknowledgements

We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources.

References

- Azaïs, J.M. and Cierco-Ayrolles, C., 2002. An asymptotic test for quantitative gene detection. *Ann. Inst. Henri Poincaré (B)*, **38(6)** 1087-1092.
- Azaïs, J.M., Delmas, C., and Rabier, C.E. (2012). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*, **48(4)** 787-801.
- 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.
- 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.
- Azaïs, J.M. and Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. Wiley, New-York.
- 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.

- Auinger, H. J., Schönleben, M., Lehermeier, C., Schmidt, M., Korzun, V., Geiger, H. H., ... Schön, C.C. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). *Theor. Appl. Genet.*, **129**(11), 20432053.
- Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A., Schlotterer, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS genetics*, **9**(6), e1003534.
- 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.
- 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.
- Brandariz, S. P., Bernardo, R. (2018). Maintaining the Accuracy of Genomewide Predictions when Selection Has Occurred in the Training Population. *Crop Science*, **58**, (3), 1226-1231.
- 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.
- Bühlmann, P. and Van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*, Springer Science.
- 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.
- Chen, Z., and Chen, H. (2005). On some statistical aspects of the interval mapping for QTL detection. *Statistica Sinica*, **15** 909-925.
- Cierco, C. (1998). Asymptotic distribution of the maximum likelihood ratio test for gene detection. *Statistics*, **31** 261-285.
- Córdoba, S., Balcells, I., Castelló, A., Ovilo, C., Noguera, J. L., Timoneda, O., Sanchez, A. (2015). Endometrial gene expression profile of pregnant sows with extreme phenotypes for reproductive efficiency. *Scientific reports*, **5**, 14416.
- 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.
- 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.
- Ferrao, L. F. V., Ferrao, R. G., Ferrao, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., Garcia, A. A. F. (2018). Accurate genomic prediction of *Coffea canephora* in multiple environments using whole-genome statistical models. *Heredity*.
- Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré (B)*, **6** 897-906.
- Gezan, S. A., Osorio, L. F., Verma, S., Whitaker, V. M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Horticulture research*. **4**, 16070.

- Gutierrez, A., Hoy, J., Kimbeng, C., Baisakh, N. (2018). Identification of genomic regions controlling leaf scald resistance in sugarcane using a bi-parental mapping population and selective genotyping by sequencing. *Frontiers in plant science*, **9**, 877.
- 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.
- Hayes, B., Bowman, P., Chamberlain, A. & Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*. **92**, (2), 433-443.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning theory*. Springer, New York.
- Hayes, B (2007). QTL Mapping, MAS, and Genomic Selection. *Short course organized by Iowa State University*.
- 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.
- Fernandes, G. R., Massironi, S. M., Pereira, L. V. (2016). Identification of Loci Modulating the Cardiovascular and Skeletal Phenotypes of Marfan Syndrome in Mice. *Scientific reports*, **6**, 22426.
- Kurz, J. P., Yang, Z., Weiss, R. B., Wilson, D. J., Rood, K. A., Liu, G. E., Wang, Z. (2019). A genome-wide association study for mastitis resistance in phenotypically well-characterized Holstein dairy cattle using a selective genotyping approach. *Immunogenetics*, **71**, (1), 35-47.
- Lander, E.S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138** 235-240.
- 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.
- Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA.
- Manichaikul, A., Palmer, A., Sen, S., and Broman, K. (2007). Significance thresholds for Quantitative Trait Locus mapping under selective genotyping. *Genetics*, **177** 1963-1966.
- Meuwissen, T.H., Hayes, B. & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. **157**, (4), 1819-1829.
- Minamikawa, M. F., Takada, N., Terakami, S., Saito, T., Onogi, A., Kajiya-Kanegae, H., ... Iwata, H. (2018). Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (*Pyrus pyrifolia* Nakai). *Scientific reports*. **8(1)**, 11994.
- Momen, M., Mehrgardi, A. A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., ... Gianola, D. (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Scientific reports*. **8**.
- 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.

- Muranty, H., Troglio, M., Sadok, I. B., ... Kumar, S. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture research*, **2**, 15060.
- Neyhart, J. L., Tiede, T., Lorenz, A. J., Smith, K. P. (2017). Evaluating methods of updating training data in long-term genomewide selection. *G3: Genes, Genomes, Genetics*, **7**, (5), 1499-1510.
- Nyine, M., Uwimana, B., Blavet, N., ... Dolezel, J. (2018). Genomic prediction in a multiploid crop: genotype by environment interaction and allele dosage effects on predictive ability in banana. *The Plant Genome*. **11(2)**, 170090.
- Ohlson, E. W., Ashrafi, H., Foolad, M. R. (2018). Identification and Mapping of Late Blight Resistance Quantitative Trait Loci in Tomato Accession PI 163245. *The plant genome*.
- Park, T., Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**, (482), 681-686.
- Phansak, 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.
- Pszczola, M., Calus, M. P. L. (2016). Updating the reference population to achieve constant genomic prediction reliability across generations. *animal*, **10**, (6), 1018-1024.
- 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.
- Rabier, C.E., Barre, P., Asp, T., Charmet, G. & Mangin, B. (2016). On the Accuracy of Genomic Selection. *PLoS One*. **11**, (6), e0156086. doi:10.1371/ journal.pone.0156086.
- Rabier, C.E., Mangin, B. & Grusea, S. (2019). On the accuracy in high dimensional linear models and its application to genomic selection. *Scandinavian Journal of Statistics*. **46**, (1), 289-313.
- Rabier, C.E. (2014a). On statistical inference for selective genotyping. *J. Stat. Plan. Infer.*, **147** 24-52.
- Rabier, C.E. (2014b). 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.
- Rabier, C.E. (2014c). 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.
- Rabier, C.E. (2015). On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. *Statistics*, **49(1)** 19-34.
- Rebaï, A., Goffinet, B., and Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138** 235-240.
- Rebaï, A., Goffinet, B., and Mangin, B. (1995). Comparing power of different methods for QTL detection. *Biometrics*, **51** 87-99.
- Siegmund, D. and Yakir, B. (2007). *The statistics of gene mapping*. Springer, New York.
- 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.

- Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC plant biology*. **17**, (1),110.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**(1) 267-288.
- Upadhyaya, H. D., Bajaj, D., Narnoliya, L., Das, S., Kumar, V., Gowda, C. L. L., ... Parida, S. K. (2016). Genome-wide scans for delineation of candidate genes regulating seed-protein content in chickpea. *Frontiers in Plant Science*, **7**, 302.
- Van der Vaart, A.W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.
- Visser, P.M., Yang, J. & Goddard, M.E. (2010). A commentary on “common SNPs explain a large proportion of the heritability for human height” by Yang et al.(2010). *Twin Research and Human Genetics*. **13**, (06), 517-524.
- 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.
- Wolc, A., Arango, J., Settar, P., Fulton, J. E., O’Sullivan, N. P., Preisinger, R., ... Dekkers, J. C. (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genetics Selection Evolution*, **43**, (1), 23.
- Wu, R., Ma, C.X., and Casella, G. (2007). *Statistical Genetics of Quantitative Traits*. Springer, New York.
- Yan, L., Hofmann, N., Li, S., Ferreira, M. E., Song, B., Jiang, G., ... Song, Q. (2017). Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC genomics*, **18**, (1), 529.
- 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.
- Zabaneh, D., Krapohl, E., Gaspar, H. A., Curtis, C., Lee, S. H., Patel, H., ... Lubinski, D. (2018). A genome-wide association study for extremely high intelligence. *Molecular psychiatry*, **23**, (5), 1226.
- Zhang, F., Guo, X., Zhang, Y., Wen, Y., Wang, W., Wang, S., ... Tan, L. (2014). Genome-wide copy number variation study and gene expression analysis identify ABI3BP as a susceptibility gene for KashinBeck disease. *Human genetics*, **133**, (6), 793-799.
- Zhao, Y., Gowda, M., Longin, F. H., Würschum, T., Ranc, N., Reif, J. C. (2012). Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theoretical and Applied Genetics*, **125**, (4), 707-713.
- Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. **67**, (2), 301-320.
- Zou, C., Wang, P., Xu, Y. (2016). Bulk sample analysis in genetics, genomics and crop improvement. *Plant biotechnology journal*. **14**, (10), 301-320.

Charles-Elie Rabier (ce.rabier@gmail.com)

ISEM, Université de Montpellier, CNRS, France.

Céline Delmas (celine.delmas@inra.fr)

INRA, UR875 Mathématiques et Informatique Appliquées Toulouse, F-313326 Castanet-Tolosan, France.

γ	γ_+/γ	SgLasso		Lasso		Group Lasso		EN		RaLasso		Bayesian Lasso
		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	99.06%	17.94	99.91%	15.89	11.66%
	3/4	96.62%	17.45	92.22%	16.33	95.88%	19.1	97.64%	17.57	98.25%	16.74	11.53%
	7/8	96.89%	17.58	82.32%	16.78	95.19%	22.9	96.09%	16.03	91.05%	16.23	11.33%
	1	93.97%	17.13	20.78%	16.66	16.73%	22.3	21.00%	16.94	47.01%	15.83	10.70%
0.2	1/2	94.19%	17.39	91.69%	16.95	97.46%	19.4	97.44%	16.21	98.09%	16.35	11.39%
	3/4	91.52%	16.3	84.75%	16.54	95.88%	19.1	96.02%	17.21	95.08%	15.44	11.20%
	7/8	92.38%	16.29	75.46%	16.55	94.67%	17.3	95.23%	16.90	89.33%	15.33	11.07%
	1	85.03%	17.09	21.14%	16.81	21.86%	26.2	27.37%	17.91	44.93%	15.48	10.64%
0.3	1/2	91.62%	17.55	83.45%	16.51	92.87%	18.6	93.67%	17.5	95.36%	16.67	11.19%
	3/4	90.88%	17.59	76.18%	16.56	89.59%	21.6	91.10%	17.67	91.13%	15.84	11.08%
	7/8	86.22%	16.82	65.03%	16.73	78.00%	17.3	82.84%	17.40	80.32%	15.11	10.91%
	1	78.00%	17.28	20.92%	16.57	20.82%	22.1	24.92%	17.62	48.25%	16.10	10.66%

Table 2: Performances of different methods, under the dense map, as a function of the percentage γ of genotyped individuals and as a function of the ratio γ_+/γ . (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 $[0,1]$, $\sigma = 1$). Dense map: $K = L = 10,001$, $t_k = t'_l = 0.001(k - 1)$. 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.

γ	γ_+/γ	SgLasso		Lasso		Group Lasso		EN		RaLasso		Bayesian Lasso
		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	99.70%	18.60	99.73%	16.84	100%	19.1	100%	18.73	100%	17.28	14.30%
	3/4	99.83%	17.28	99.69%	16.89	100%	20	100%	17.73	100%	16.03	13.88%
	7/8	99.55%	16.62	99.24%	16.69	100%	20.9	100%	17.63	100%	17.11	13.37%
	1	99.69%	16.64	31.43%	16.83	18.30%	22.61	33.33%	16.34	60.55%	16.60	10.75%
0.2	1/2	99.23%	17.56	98.99%	16.81	100%	18.4	100%	17.77	99.99%	17.96	13.41%
	3/4	99.60%	17.41	98.47%	16.82	100%	19.2	100%	18.41	100%	16.51	13.38%
	7/8	99.27%	17.48	98.35%	16.90	100%	18.9	100%	17.13	99.73%	16.00	12.59%
	1	99.36%	17.79	24.53%	17.15	11.97%	29.1	25.71%	17.26	54.22%	17.32	10.69%
0.3	1/2	99.20%	17.96	97.50%	16.90	100%	19.6	99.99%	16.88	100%	17.39	12.89%
	3/4	99.60%	17.31	97.5%	16.81	100%	18.9	100%	16.96	99.59%	17.56	12.69%
	7/8	99.66%	17.86	96.50%	16.99	99.82%	22.8	99.90%	18.05	99.95%	17.07	12.22%
	1	98.69%	17.50	42.93%	17	38.45%	19.1	48.13%	17.36	72.39%	15.58	10.78%

Table 3: Performances of different methods, under the dense map, as a function of the percentage γ of genotyped individuals and as a function of the ratio γ_+/γ . (Mean over 100 samples, on average $n = 100$ individuals genotyped, $m = 16$, $|q_1| = \dots = |q_{16}| = 0.2$, $T = 10$, QTLs randomly located only on $[0,1]$, $\sigma = 1$). Dense map: $K = L = 10,001$, $t_k = t'_l = 0.001(k - 1)$. 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.

		all $ q_s = 0.1$						all $ q_s = 0.2$					
γ	γ^+/γ	SgLasso		SgGroupLasso		SgEN		SgLasso		SgGroupLasso		SgEN	
		L1 ratio	\hat{m}	L1 ratio	\hat{m}	L1 ratio	\hat{m}	L1 ratio	\hat{m}	L1 ratio	\hat{m}	L1 ratio	\hat{m}
0.1	1/2	97.24%	17.22	99.25%	25	98.19%	17.59	99.70%	18.60	99.90%	27.9	99.88%	18.37
	3/4	96.62%	17.15	99.41%	22.5	97.17%	18.12	99.83%	17.28	99.80%	28.1	100%	16.94
	7/8	96.89%	17.58	99.15%	24.4	98.37%	18.22	99.55%	16.62	100%	27.6	99.98%	16.93
	1	93.97%	17.13	97.29%	24.4	95.31%	17.46	99.69%	16.64	100%	27	99.88%	17.37
0.2	1/2	94.19%	17.39	98.33%	24.9	96.03%	16.90	99.23%	17.56	100%	28.5	99.69%	17.81
	3/4	91.52%	16.3	95.38%	24.3	92.59%	17.41	99.60%	17.41	99.94%	29	99.72%	19.27
	7/8	92.38%	16.29	96.83%	24.6	93.19%	17.13	99.27%	17.48	100%	26.5	99.67%	18.61
	1	85.03%	17.09	90.53%	22.8	84.93%	17.67	99.36%	17.79	100%	27.2	99.69%	18.33
0.3	1/2	91.62%	17.55	92.35%	24.6	86.53%	17.87	99.20%	17.96	99.60%	28.1	99.24%	18.55
	3/4	90.88%	17.59	94.84%	30.9	91.84%	15.43	98.60%	17.31	100%	30.5	99.88%	19.02
	7/8	86.22%	16.82	89.96%	29.3	86.68%	17.30	98.69%	17.50	99.89%	31.9	99.92%	18.29
	1	78.00%	17.28	82.61%	28.6	77.23%	17.89	98.69%	17.50	99.86%	26.5	99.18%	18.44

Table 4: Performances of our method, under the dense map, as a function of the penalization used. (mean over 100 samples, on average $n = 100$ individuals genotyped, $m = 16$, $T = 10$, QTLs randomly located only on $[0,1]$, $\sigma = 1$). 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.

γ	all $ q_s $	γ^+/γ	SgLasso	Lasso	Group Lasso	EN	RaLasso	Bayesian Lasso
0.1	0.1	1	30.97%	6.49%	3.17%	4.38%	10.43%	7.12%
		7/8	31.25%	30.55%	29.87%	29.74%	28.78%	25.50%
	0.2	1	56.85%	27.96%	7.57%	21.17%	33.09%	31.30%
		7/8	57.89%	56.96%	54.95%	55.26%	54.66%	57.24%
	0.3	1	70.64%	46.54%	5.35%	19.89%	39.38%	49.30%
		7/8	72.34%	70.16%	68.07%	68.17%	67.63%	72.59%
0.2	0.1	1	27.88%	7.12%	4.05%	5.41%	11.08%	8.97%
		7/8	28.26%	27.98%	27.86%	28.09%	26.28%	22.11%
	0.2	1	54.37%	31.70%	13.85%	24.73%	36.39%	29.68%
		7/8	54.72%	55.30%	53.08%	53.44%	53.20%	55.71%
	0.3	1	67.74%	57.21%	16.33%	39.61%	49.63%	50.41%
		7/8	68.49%	68.64%	66.00%	65.93%	66.18%	72.09%
0.3	0.1	1	26.79%	9.02%	6.89%	7.48%	11.96%	9.13%
		7/8	28.13%	27.85%	26.59%	28.25%	26.05%	21.09%
	0.2	1	52.83%	38.15%	21.23%	33.17%	42.96%	31.38%
		7/8	54.07%	54.04%	51.96%	51.46%	51.39%	51.24%
	0.3	1	66.73%	57.51%	26.08%	46.30%	55.06%	50.47%
		7/8	67.13%	67.43%	64.91%	65.08%	63.99%	69.57%

Table 5: Predictive abilities of the different methods, under the dense map, as a function of the percentage γ of genotyped individuals and as a function of the ratio γ_+/γ . The model is learned on the genotyped individuals, and evaluated on 100 progenies of the training individuals. (mean over 100 samples, on average $n = 100$ individuals genotyped, $m = 16$, $T = 10$, QTLs randomly located only on $[0,1]$, $\sigma = 1$). Dense map: $K = L = 10,001$, $t_k = t'_l = 0.001(k - 1)$.

γ	all $ q_s $	γ^+/γ	SgLasso	SgGroupLasso	SgEN
0.1	0.1	1	30.97%	30.31%	30.89%
		7/8	31.25%	30.60%	31.12%
	0.2	1	56.85%	54.13%	55.44%
		7/8	57.89%	55.38%	55.81%
	0.3	1	70.64%	66.91%	67.56%
		7/8	72.34%	68.47%	69.06%
0.2	0.1	1	27.88%	27.84%	27.86%
		7/8	28.26%	27.80%	28.03%
	0.2	1	54.37%	52.76%	53.62%
		7/8	54.72%	52.77%	53.79%
	0.3	1	67.74%	65.07%	65.91%
		7/8	68.49%	65.86%	66.57%
0.3	0.1	1	26.79%	27.05%	26.85%
		7/8	28.13%	27.82%	28.14%
	0.2	1	52.83%	52.22%	52.54%
		7/8	54.07%	52.47%	53.64%
	0.3	1	66.73%	64.59%	65.90%
		7/8	67.43%	65.11%	66.25%

Table 6: Predictive ability of our method, under the dense map, as a function of the penalization used, and as a function of the percentage γ of genotyped individuals. The model is learned on the genotyped individuals, and evaluated on 100 progenies of the training individuals (mean over 100 samples, on average $n = 100$ individuals genotyped, $m = 16$, $T = 10$, QTLs randomly located only on $[0,1]$, $\sigma = 1$). Dense map: $K = L = 10,001$, $t_k = t'_l = 0.001(k - 1)$.

γ	Method	FP	FN	Selected genes
1	Begum et al. (2015)			S3-1125848, S3-1165376, S3-1221494, S3-1269941, S3-1394477, S6-2900101, S6-2961503, S6-3057752, S8-4137990, S8-4138023
1	SgLasso	22 (21)	6 (5)	<i>S3-1094192</i> , S3-1125848 , S3-1165376 , S3-1269941 S3-1394477 , S3×21
1	SgEN	28 (24)	5 (5)	<i>S3-1030333</i> , <i>S3-1094192</i> , <i>S3-1123429</i> , S3-1125848 , S3-1165376 <i>S3-1179404</i> , S3-1221494 , S3-1269941 , S3-1394477 , S3×24
1	SgGroupLasso	37 (23)	7 (6)	S3×31, <i>S3-1030333</i> , <i>S3-1070111</i> , <i>S3-1094192</i> , <i>S3-1123429</i> S3-1125848 , S3-1165376 , <i>S3-1179404</i> , S3-1221494 , <i>S3-1225693</i>
0.3	SgLasso	28 (23)	6 (5)	<i>S3-1070111</i> , <i>S3-1094192</i> , S3-1165376 , S3-1221494 , <i>S3-1225693</i> , S3-1269941 <i>S3-1298550</i> , <i>S3-1354306</i> , S3-1394477 , S3×23
0.3	SgEN	26 (23)	5 (5)	<i>S3-1030333</i> , <i>S3-1094192</i> , <i>S3-1123429</i> , S3-1125848 , S3-1165376 , S3-1221494 S3-1269941 , S3-1394477 , S3×23
0.3	SgGroupLasso	65 (51)	5 (5)	<i>S3-1030333</i> , <i>S3-1070111</i> , <i>S3-1094192</i> , <i>S3-1123429</i> , S3-1125848 S3-1165376 , <i>S3-1179404</i> , S3-1221494 , <i>S3-1225693</i> , S3-1269941 <i>S3-1298550</i> , <i>S3-1320779</i> , <i>S3-1342244</i> , <i>S3-1354306</i> , S3-1394477 , <i>S3-1403300</i> <i>S3-1439520</i> , <i>S3-1462159</i> , <i>S3-1495153</i> , S3×41, S8×10
1	Lasso	17 (17)	7 (6)	S1×2, S2×3, S3-1165376 , S3-1221494 , S3-1269941 S3×3, S7×2, S8×2, S9×2 S10×1, S11×1, S12×1
1	EN	34 (34)	6 (6)	S1×5, S2×4, S3-1125848 , S3-1165376 , S3-1221494 , S3-1269941 S3×7, S7×4, S8×3, S9×3, S10×2, S11×3 S12×3
1	Group Lasso	134 (128)	7 (6)	S1×30, S2×20, <i>S3-1030333</i> , <i>S3-1070111</i> , <i>S3-1094192</i> <i>S3-1123429</i> , S3-1125848 , S3-1165376 , <i>S3-1179404</i> , S3-1221494 <i>S3-1225693</i> , S3×11, S3×7, S7×10, S8×20, S9×10, S11×20
0.3	Lasso	0 (0)	8 (6)	S3-1221494 , S3-1269941
0.3	EN	2 (0)	5 (5)	<i>S3-1094192</i> , <i>S3-1123429</i> , S3-1125848 , S3-1165376 S3-1221494 , S3-1269941 , S3-1394477
0.3	Group Lasso	7 (2)	7 (6)	S3×2, <i>S3-1070111</i> , <i>S3-1094192</i> , <i>S3-1123429</i> , S3-1125848 S3-1165376 , <i>S3-1179404</i> , S3-1221494 , <i>S3-1225693</i>
1	RaLasso	2600 (2568)	5 (0)	S1×704, S2×220, <i>S3-1123429</i> , S3-1125848 , S3-1165376 , <i>S3-1179404</i> S3-1221494 , <i>S3-1225693</i> , S3-1269941 , <i>S3-1298550</i> <i>S3-1320779</i> , <i>S3-1342244</i> , <i>S3-1354306</i> , S3-1394477 <i>S3-1403300</i> , <i>S3-1439520</i> , <i>S3-1462159</i> , <i>S3-1495153</i> , S3×203, S4×192, S5×174 <i>S6-2848386</i> , <i>S6-2866608</i> , <i>S6-2899016</i> , <i>S6-2913729</i> , <i>S6-2941202</i> <i>S6-2913729</i> , <i>S6-2941202</i> , <i>S6-2958750</i> , <i>S6-2980225</i> , <i>S6-3001176</i> , <i>S6-3041790</i> <i>S6-3041790</i> , <i>S6-3056545</i> , <i>S6-3076966</i> , <i>S6-3112878</i> , S6×160, S7×168 <i>S8-4063097</i> , <i>S8-4082527</i> , <i>S8-4101244</i> , <i>S8-4147562</i> , <i>S8-4150777</i> , <i>S8-4188989</i> , S8×162 S9×133, S10×140, S11×165, S12×147
0.3	RaLasso	782 (775)	10 (4)	S1×219, S2×74, S3×64 <i>S3-1354306</i> , <i>S3-1403300</i> , S4×59, S5×49 S6×52, <i>S6-2913729</i> , <i>S6-2958750</i> , <i>S6-2980225</i> , <i>S6-3056545</i> S7×41, S8×52, <i>S8-4101244</i> , S9×36, S10×39, S11×48, S12×42

Table 7: Comparison, on rice data ([Begum et al. \(2015\)](#); [Spindel et al. \(2015\)](#)), of the selected genes as a function of the methods and as function of the percentage γ 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 ($T = 13,101$). Markers in bold match exactly one of the genes selected by [Begum et al. \(2015\)](#). 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. \(2015\)](#). SA-B refers to a marker on chromosome A with id B. SA×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. \(2015\)](#). FP and FN refer to the number of false positives and the number of false negatives, respectively. In brackets, are also given FP and FN, assuming a tolerance level of 0.4cM.

Supplementary A: “The SgLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection” ☆

C.E. Rabier^{a,b,c,*}, C. Delmas^c

^a*ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France.*

^b*LIRMM, UMR 5506, CNRS, Université de Montpellier, France.*

^c*INRA, UR875 Mathématiques et Informatique Appliquées Toulouse, F-31326 Castanet-Tolosan, France.*

1. True probability distribution when m QTLs lie on $[0, T]$ (with $m > 1$)

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 $t_1^* < t_2^* < \dots < t_m^*$.

All the information is contained in the flanking markers of the QTLs locations, because of the Poisson process. As a consequence, let us compute the probability distribution of $(Y, \bar{X}(t_1^{\ell}), \bar{X}(t_1^{*r}), \dots, \bar{X}(t_m^{\ell}), \bar{X}(t_m^{*r}))$.

We have

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

Besides,

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

*Corresponding author. Tel.: +33 4 67 14 36 97

Email addresses: ce.rabier@gmail.com (C.E. Rabier), celine.delmas@inra.fr (C. Delmas)

On the other hand,

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

As a result,

$$\begin{aligned}
& \mathbb{P}(Y \in [y, y + dy], Y \notin [S_-, S_+], \bar{X}(t_1^{*\ell}), \bar{X}(t_1^{*r}), \dots, \bar{X}(t_m^{*\ell}), \bar{X}(t_m^{*r})) \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} f_{(\mu + u_1 q_1 + u_2 q_2 + \dots + u_m q_m, \sigma)}(y) \mathbf{1}_{y \notin [S_-, S_+]} \\
& \times \mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m, X(t_1^{*\ell}), X(t_1^{*r}), \dots, X(t_m^{*\ell}), X(t_m^{*r})) .
\end{aligned}$$

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

$$\begin{aligned}
& \mathbb{P}(Y \in [y, y + dy], \bar{X}(t_1^{*\ell}) = 0, \bar{X}(t_1^{*r}) = 0, \dots, \bar{X}(t_m^{*\ell}) = 0, \bar{X}(t_m^{*r}) = 0) \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{P}(Y \in [y, y + dy], Y \in [S_-, S_+], X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m) \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} f_{(\mu + u_1 q_1 + \dots + u_m q_m, \sigma)}(y) \mathbf{1}_{y \in [S_-, S_+]} \mathbb{P}(X(t_1^*) = u_1, X(t_2^*) = u_2, \dots, X(t_m^*) = u_m) .
\end{aligned}$$

Let $\theta^m = (q_1, \dots, q_m, \mu, \sigma)$ denote the new parameter. 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

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

with

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

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

and

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

Note that as soon as we set $m = 1$ in formula (1), we obtain $\bar{L}_{t_1^*}(\theta^1)$ given

in formula (2) of the main manuscript.

2. Proof of Theorem 1

The proof is divided into five parts (the first four parts rely on the case $K = 2$ markers):

- Preliminaries (i.e. computation of the Fisher Information Matrix)
- Weak convergence of the score process under \mathcal{H}_0
- Study of the score process under the local alternative $\mathcal{H}_{a\vec{t}^*}$
- Study of the supremum of the LRT process
- Generalization to $K > 2$

Note that under \mathcal{H}_0 , the proof has already been given in Rabier (2015). However, the weak convergence of the score process has not been proved in details. Indeed, the author 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 (2009)).

Let us consider the case $K = 2$, that is to say two markers are located at $t_1 = 0$ and $t_2 = T$. 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. As a consequence, in what follows, $t^\ell = t_1$ and $t^r = t_2$. The notations t^ℓ and t^r will be convenient for the generalization to the case $K > 2$.

2.1. Preliminaries

The proof starts with the computation of the Fisher Information Matrix. As a result, calculations are exactly the same as in Rabier (2015), see Section "Study of the score process under the null hypothesis" of the proof of Theorem 2.5. We propose to recall here the key elements of the proof.

First, the author computes the score function at a point $\theta_0^1 = (0, \mu, \sigma)$ that belongs to \mathcal{H}_0 . We have the relationship

$$\begin{aligned} \frac{\partial \bar{l}_t}{\partial q_1} \Big|_{\theta_0^1} &= \frac{Y - \mu}{\sigma^2} \{2p(t) - 1\} 1_{Y \notin [S_-, S_+]} \\ &= \frac{\alpha(t)}{\sigma} \varepsilon \bar{X}(t^\ell) + \frac{\beta(t)}{\sigma} \varepsilon \bar{X}(t^r) \end{aligned}$$

because of the key Lemma (Lemma 2.6 of Rabier (2015)), which states that

$$\{2p(t) - 1\} 1_{Y \notin [S_-, S_+]} = \alpha(t) \bar{X}(t^\ell) + \beta(t) \bar{X}(t^r) \quad (2)$$

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

To conclude, after some easy calculations, he finds that the Fisher information is diagonal :

$$I_{\theta_0} = \text{Diag} \left[\mathcal{A} \{ \alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r) \} / \sigma^4, \frac{1}{\sigma^2}, \frac{2}{\sigma^2} \right]. \quad (3)$$

2.2. Weak convergence of the score process under \mathcal{H}_0

Convergence of finite-dimensional

At a marker location t_k with $k \in \{1, 2\}$, we have:

$$\bar{S}_n(t_k) = \frac{\frac{\partial \bar{I}_{t_k}^n}{\partial q_1} | \theta_0^1}{\sqrt{\mathbb{V} \left(\frac{\partial \bar{I}_{t_k}^n}{\partial q_1} | \theta_0^1 \right)}} = \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n \mathcal{A}}}.$$

Since $\frac{\partial \bar{I}_{t_k}^n}{\partial q_1} | \theta_0^1$ is centered under H_0 , a direct application of the central limit theorem implies that

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

Then, since we have the relationship (cf. formula (2))

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

the continuous mapping theorem implies that

$$\bar{S}_n(t) \xrightarrow{\mathcal{L}} V(t).$$

It proves the convergence of finite-dimensional.

Note also that we have the relationship

$$\text{Cov}_{H_0} \{ \bar{S}_n(t^\ell), \bar{S}_n(t^r) \} = \rho(t^\ell, t^r).$$

Tightness

Since we have already proved the convergence of finite-dimensional, let us focus on the tightness of the score process. Since $\alpha(t)$, $\beta(t)$ and $\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)$ are continuous functions, each path of the process $\bar{S}_n(\cdot)$ is a continuous function on $[t^\ell, t^r]$. Recall the modulus of continuity of a continuous function $h(t)$ on $[t^\ell, t^r]$:

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

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 $\bar{S}_n(t^\ell)$ is tight.
2. For each positive ε and η , there exists a δ , with $t^\ell < \delta < t^r$, and an integer n_0 such that $\mathbb{P}(\varpi_{\bar{S}_n}(\delta) \geq \eta) \leq \varepsilon \quad \forall n \geq n_0$.

According to Prohorov's theorem, the sequence $\bar{S}_n(t^\ell)$ is tight. Then, Condition 1 is verified. Let us define the functions $\alpha'(t)$ and $\beta'(t)$ in the following way:

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

First, we can notice that $\forall \delta$ such as $t^\ell < \delta \leq t^r$,

$$\begin{aligned}\varpi_{\bar{S}_n}(\delta) &= \sup_{|t'-t|<\delta} |\bar{S}_n(t') - \bar{S}_n(t)| \\ &= \sup_{|t'-t|<\delta} |(\alpha'(t') - \alpha'(t))\bar{S}_n(t^\ell) + (\beta'(t') - \beta'(t))\bar{S}_n(t^r)| \\ &\leq \max(|\bar{S}_n(t^\ell)|, |\bar{S}_n(t^r)|) (\varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta)).\end{aligned}\tag{4}$$

Furthermore, the sequence $\max(|\bar{S}_n(t^\ell)|, |\bar{S}_n(t^r)|)$ is uniformly tight. This way,

$$\forall \varepsilon > 0 \quad \exists M > 0 \quad \forall n \geq 1 \quad \mathbb{P}(\max(|\bar{S}_n(t^\ell)|, |\bar{S}_n(t^r)|) \geq M) \leq \varepsilon.\tag{5}$$

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

$$\forall v > 0 \quad \exists \delta \text{ such as } t^\ell < \delta < t^r, \quad \varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta) < v.\tag{6}$$

Let η be a positive quantity. Using formulae (5) and (6) and imposing $v = \eta/M$, we have

$$\mathbb{P}(\max(|\bar{S}_n(t^\ell)|, |\bar{S}_n(t^r)|) (\varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta)) \geq \eta) \leq \varepsilon.$$

As a consequence, according to formula (4), we have

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

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 on $[t^\ell, t^r]$, i.e. on $[t_1, t_2]$.

2.2.1. Study of the score process under the local alternative $\mathcal{H}_{a\bar{t}^*}$

There are m QTLs located on $[0, T]$ and the model for the quantitative trait is the following:

$$Y = \mu + \sum_{s=1}^m X(t_s^*) q_s + \sigma \varepsilon \quad (7)$$

where ε 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

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

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

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

$$\bar{S}_n(t_k) = \sum_{j=1}^n \frac{(Y_j - \mu) \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} \quad (8)$$

$$= \sum_{j=1}^n \sum_{s=1}^m \frac{q_s \bar{X}_j(t_s^*) \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}} + \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n} \mathcal{A}}. \quad (9)$$

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

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

According to calculations present in Section 4,

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

where Φ is the cumulative distribution of a standard normal distribution. In

the same way,

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

Since we have the relationships

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

then we have

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

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

$$\sum_{j=1}^n \sum_{s=1}^m \frac{q_s \bar{X}_j(t_s^*) \bar{X}_j(t_k)}{\sqrt{n \mathcal{A}}} \rightarrow \sum_{s=1}^m \frac{a_s \rho(t_k, t_s^*) \gamma}{\sqrt{\mathcal{A}}}. \quad (10)$$

Let us now focus on the second term of formula (9). According to a technical proof present in Section 4, we have

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

Besides, according to iii) of Lemma 5 of Rabier (2014a),

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

As a result,

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

Then, according to the Central Limit Theorem,

$$\sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n \mathcal{A}}} \xrightarrow{\mathcal{L}} \mathcal{N} \left[\frac{\sum_{s=1}^m \rho(t_s^*, t_k) a_s}{\sqrt{\mathcal{A}}} \{z_{\gamma^+} \varphi(z_{\gamma^+}) - z_{1-\gamma^-} \varphi(z_{1-\gamma^-})\}, 1 \right]. \quad (11)$$

Finally, according to formulae (10) and (11),

$$\bar{S}_n(t_k) \xrightarrow{\mathcal{L}} \mathcal{N} \left[\sum_{s=1}^m \rho(t_k, t_s^*) a_s \sqrt{\mathcal{A}} / \sigma^2, 1 \right]. \quad (12)$$

2.2.2. 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 \mathcal{H}_0 (i.e. no QTL on the whole interval studied)

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

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.

Let us consider now t as an extra parameter. Rabier (2015) studied this irregular model and proved that

$$\sup \bar{\Lambda}_n(t) = \sup \bar{S}_n^2(t) + o_P(1). \quad (13)$$

Note that the proof is based on results of Azaïs et al. (2009), Azaïs et al. (2006) and Gassiat (2002) 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 θ^m and θ_0^m are defined in the following way : $\theta^m = (q_1, \dots, q_m, \mu, \sigma)$ and $\theta_0^m = (0, \dots, 0, \mu, \sigma)$.

The likelihood $\bar{L}_{\vec{t}^*}^{m,n}(\theta^m)$ for n observations is obtained by the product of n terms as in formula (1) of this supplementary material, with $K = 2$. 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 $\bar{L}_{\vec{t}^*}^{m,n}(\theta^m)$ (respectively $\bar{L}_{\vec{t}^*}^{m,n}(\theta_0^m)$).

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

$$\log \frac{dQ_n}{dP_n} = \log \left\{ \frac{\bar{L}_{\vec{t}^*}^{m,n}(\theta^m)}{\bar{L}_{\vec{t}^*}^{m,n}(\theta_0^m)} \right\}. \quad (14)$$

Since the model is differentiable in quadratic mean at θ^m and according to the

central limit theorem :

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

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 (13) is also true under the alternative $\mathcal{H}_{a\bar{t}^\star}$.

It concludes the proof of Theorem 1 for $K = 2$.

2.3. Generalization to $K > 2$

K genetic markers are now located at $0 = t_1 < t_2 < \dots < t_K = T$. We consider a location t that is distinct of the markers positions.

Under H_0 , for a position t , we can limit our attention to the interval (t^ℓ, t^r) , due to Haldane model with Poisson increments. Recall the notation $\mathbb{T}_K = \{t_1, \dots, t_K\}$. Besides, according to Rabier (2015), we have

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

Under the local alternative $\mathcal{H}_{a\bar{t}^\star}$, we just have to use the fact that the mean function $\bar{m}_{\bar{t}^\star}(t)$ is an interpolated function between $\bar{m}_{\bar{t}^\star}(t^\ell)$ and $\bar{m}_{\bar{t}^\star}(t^r)$. Then, in order to characterize the mean function, we only have to compute the distribution of $\bar{S}_n(t_k)$ at a marker located at t_k . We still have the relationship (as in formula (12))

$$\bar{S}_n(t_k) \xrightarrow{\mathcal{L}} \mathcal{N} \left[\sum_{s=1}^m \rho(t_k, t_s^\star) a_s \sqrt{A} / \sigma^2, 1 \right] \quad \forall k \in \mathbb{T}_K$$

since the formulae (9), (10) and (11) are still valid for $K > 2$. Indeed, those formulae rely on calculations present in Section 4 suitable for $K \geq 2$.

The tightness of the score process $S_n(\cdot)$ is obvious because of the interpolations. Besides, formula (13) above is still true for $K > 2$ according to Rabier (2015). In order to prove that the remainder converges also to zero under $\mathcal{H}_{a\bar{t}^\star}$, just use the same kind of proof as above (based on Le Cam's first lemma). Note that the likelihood $\bar{L}_{\bar{t}^\star}^{m,n}(\theta^m)$ for n observations is now obtained by the product of n terms as in formula (1) with $K > 2$. Same remark for $\bar{L}_{\bar{t}^\star}^{m,n}(\theta_0^m)$.

3. Proof of Corollary 1

To begin with, let us recall the epistatic model, given in formula (12) of the manuscript:

$$Y = \mu + \sum_{s=1}^m X(t_s^\star) q_s + \sum_{s=1}^{m-1} \sum_{\bar{s}=s+1}^m X(t_s^\star) X(t_{\bar{s}}^\star) q_{s,\bar{s}} + \sigma \varepsilon \quad (15)$$

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

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 (15) and (8), we have

$$\begin{aligned} \bar{S}_n(t_k) &= \sum_{j=1}^n \sum_{s=1}^m \frac{a_s \bar{X}_j(t_s^*) \bar{X}_j(t_k)}{n \sqrt{\mathcal{A}}} + \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(t_k)}{\sqrt{n \mathcal{A}}} \\ &+ \frac{1}{n \sqrt{\mathcal{A}}} \sum_{j=1}^n \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^m \bar{X}_j(t_s^*) \bar{X}_j(t_{\tilde{s}}^*) b_{s,\tilde{s}} \right\} \bar{X}_j(t_k) . \end{aligned} \quad (16)$$

According to calculations present in Section 4, 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 (10) and (11),

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

4. Study of quantities present in the proofs

In this section, all calculations are valid for a number of markers $K \geq 2$.

4.1. Preliminaries

To begin with, let us recall Lemma 5 of Rabier (2014a). It will be very useful for our theoretical calculations since it is related to truncated normal distributions.

Lemma 5 (Rabier (2014a)). *Let $W \sim \mathcal{N}(\mu, \sigma^2)$, then*

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

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

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

4.2. *Formulas for $\mathbb{E}\left[1_{Y \notin [S_-, S_+]}\left\{1_{X(t_s^*)=1}1_{X(t_k)=1} + 1_{X(t_s^*)=-1}1_{X(t_k)=-1}\right\}\right]$ and $\mathbb{E}\left[1_{Y \notin [S_-, S_+]}\left\{1_{X(t_s^*)=-1}1_{X(t_k)=1} + 1_{X(t_s^*)=1}1_{X(t_k)=-1}\right\}\right]$*

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

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

Using the same kind of proof, we have

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

As a result, we have the relationships

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

4.3. Formula for $\mathbb{E} \{ \sigma \varepsilon \bar{X}(t_k) \}$

We have

$$\begin{aligned}
& \mathbb{E} \{ \sigma \varepsilon \bar{X}(t_k) \} \\
&= \mathbb{E} \{ \sigma \varepsilon \mathbf{1}_{X(t_k)=1} \mathbf{1}_{Y \notin [S_-, S_+]} \} - \mathbb{E} \{ \sigma \varepsilon \mathbf{1}_{X(t_k)=-1} \mathbf{1}_{Y \notin [S_-, S_+]} \} \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{E} \{ \sigma \varepsilon \mathbf{1}_{X(t_k)=1} \mathbf{1}_{X(t_1^*)=u_1} \dots \mathbf{1}_{X(t_m^*)=u_m} \mathbf{1}_{Y \notin [S_-, S_+]} \} \\
&- \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{E} \{ \sigma \varepsilon \mathbf{1}_{X(t_k)=-1} \mathbf{1}_{X(t_1^*)=u_1} \dots \mathbf{1}_{X(t_m^*)=u_m} \mathbf{1}_{Y \notin [S_-, S_+]} \} \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{E} \{ \sigma \varepsilon \mathbf{1}_{Y \notin [S_-, S_+]} \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&[2\mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} - 1] \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&= \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \left\{ \sigma \varphi(z_{\gamma+}) + z_{\gamma+} \varphi(z_{\gamma+}) \sum_{s=1}^m u_s q_s - \sigma \varphi(z_{1-\gamma-}) - z_{1-\gamma-} \varphi(z_{1-\gamma-}) \sum_{s=1}^m u_s q_s \right\} \\
&[2\mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} - 1] \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&+ o(\max_{1 \leq s \leq m} |q_s|).
\end{aligned} \tag{17}$$

Note that in order to obtain the last expression, we used iv) of Lemma 5 of Rabier (2014a) (cf. Section 4.1). Recall that z_α denotes the quantile of order $1 - \alpha$ of a standard normal distribution. Let us focus on the quantity

$$\begin{aligned}
& \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \{ \sigma \varphi(z_{\gamma+}) - \sigma \varphi(z_{1-\gamma-}) \} [2\mathbb{P} \{ X(t_k) = 1 \mid X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} - 1] \\
&\times \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&= \{ \sigma \varphi(z_{\gamma+}) - \sigma \varphi(z_{1-\gamma-}) \} \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} 2 \mathbb{P} \{ X(t_k) = 1, X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&- \{ \sigma \varphi(z_{\gamma+}) - \sigma \varphi(z_{1-\gamma-}) \} \sum_{(u_1, \dots, u_m) \in \{-1, 1\}^m} \mathbb{P} \{ X(t_1^*) = u_1, \dots, X(t_m^*) = u_m \} \\
&= \{ \sigma \varphi(z_{\gamma+}) - \sigma \varphi(z_{1-\gamma-}) \} 2 \mathbb{P} \{ X(t_k) = 1 \} - \{ \sigma \varphi(z_{\gamma+}) - \sigma \varphi(z_{1-\gamma-}) \} = 0.
\end{aligned} \tag{18}$$

Let us focus on the quantity

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

Let ξ denote a given QTL. We have

$$\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^*) . \tag{19}$$

As a result, according to formulae (17), (18) and (19), we have

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

4.4. Formula for the quantity $\mathbb{E} \{ \bar{X}(t_s^*) \bar{X}(t_s^*) \bar{X}(t_k) \}$

We have

$$\begin{aligned}
& \mathbb{E} \{ \bar{X}(t_s^*) \bar{X}(t_s^*) \bar{X}(t_k) \} \\
&= \mathbb{E} \left\{ 1_{X(t_s^*)X(t_s^*)X(t_k)=1} 1_{Y \notin [S_-, S_+]} \right\} - \mathbb{E} \left\{ 1_{X(t_s^*)X(t_s^*)X(t_k)=-1} 1_{Y \notin [S_-, S_+]} \right\} \\
&= \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{20}$$

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 (20), we have

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

It concludes the proof.

References

- Azaïs, J.M., Delmas, C., and Rabier, C.E. (2012). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*, **48(4)** 787-801.
- 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.
- 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.
- Azaïs, J.M., and Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. Wiley, New-York.
- Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré (B)*, **6** 897-906.
- Rabier, C.E. (2014a). On statistical inference for selective genotyping. *J. Stat. Plan. Inference*, **147** 24-52.

- Rabier, C.E. (2014b). 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.
- Rabier, C.E. (2015). On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. *Statistics*, **49(1)** 19-34.
- Van Der Vaart, A.W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.

Supplementary B: “The SgLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection” ☆

C.E. Rabier^{a,b,c,*}, C. Delmas^c

^a*ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France.*

^b*LIRMM, UMR 5506, CNRS, Université de Montpellier, France.*

^c*INRA, UR875 Mathématiques et Informatique Appliquées Toulouse, F-31326 Castanet-Tolosan, France.*

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. (2012) and based on Genz (1992). In order to compute the maximum of the process, simulated data were analyzed using Lemma 1 of Azaïs et al. (2012), 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 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 Theorem 1 of the main text.

Last, the power of the test is reported as a function of the QTL effect signs.

*Corresponding author. Tel.:+33 4 67 14 36 97

Email addresses: ce.rabier@gmail.com (C.E. Rabier), celine.delmas@inra.fr (C. Delmas)

γ	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 γ 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, $\gamma_+/\gamma = 1/2$).

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 Arias-Castro et al. (2011) where the authors compared performances of the max test and the ANOVA in another context.

γ	γ_+/γ	(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 SgLasso as a function of the ratio γ_+/γ (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.

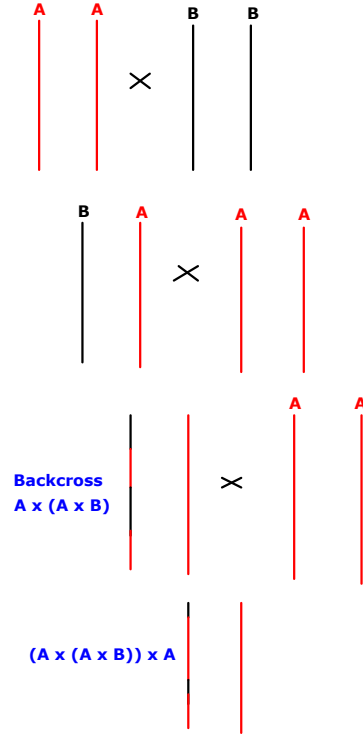


Figure 1: Backcross population $A \times (A \times B)$ and the progenies $(A \times (A \times B)) \times A$. Recall that A and B are purely homozygous lines. In the main manuscript, alleles from A (in red) are coded -1 and alleles from B (in black) are coded 1 .

References

- 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.
- Azaïs, J.M., Delmas, C., and Rabier, C.E. (2012). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*, **48**(4) 787-801.
- Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Stat.*, **1** 141-149.
- 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.