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

Charles-Elie Rabier, Céline Delmas

Vol. 00, No. 00, February 2020, 1–30

RESEARCH ARTICLE

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

Charles-Elie Rabier [abc*], Céline Delmas [d]

[a]*ISEM, Université de Montpellier, CNRS, EPHE, IRD, France*; [b]*IMAG, Université de Montpellier, CNRS, France*; [c]*LIRMM, Université de Montpellier, CNRS, France*; [d]*INRA, UR875 MIAT, F-313326 Castanet-Tolosan, France*;

(*v February 2020*)

We introduce a new variable selection method, called SgLasso, that handles extreme data. 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. It is appropriate in genomics since once the genetic map has been built, the correlation is perfectly known. This new technique is inspired by stochastic processes arising from statistical genetics. We prove that the signal to noise ratio is largely increased by considering the extremes. Our approach and existing methods are compared on simulated and real data, and the results point to the validity of our approach.

Keywords: Gaussian process, Selective Genotyping, Genomic Selection, High-Dimensional Linear Model, Variable Selection, Sparsity

AMS Subject Classification: Primary 60G15; 62F03; 62F05

1. Introduction

1.1. *Context and goal of the study*

As in [1, 2], 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 $Y_j,\ j=1,...,n$, 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 (see for instance [3, 4]). 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

*Corresponding author. Email: ce.rabier@gmail.com

-1 if it is originated from B. The admitted model for the stochastic structure of $X(.)$ is called the Haldane model [5]. It states that:

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

where for any $b \in \mathrm{R}$, δ_b denotes the point mass at b and $N(.)$ is a standard Poisson process on $[0, T]$. In a more practical point of view, the Haldane model [5] assumes no crossover interference and the Poisson process represents the number of crossovers on $[0, T]$ which happen during meiosis. 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

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

we set in addition

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

We assume an "analysis of variance model" for the quantitative trait (see [3] for instance):

$$Y = \mu \; + \; \sum_{s=1}^{m} X(t_s^\star)\, q_s \; + \; \sigma\varepsilon \tag{1}$$

where μ is the global mean, ε is a Gaussian white noise independent of $X(.)$, σ^2 is the environmental variance, m is the number of QTLs, and q_s and $t_s^\star$ denote respectively the QTL effect and the location of the sth QTL. Indeed, it is well known that there is a finite number of loci underlying the variation in quantitative traits (e.g. in aquaculture and livestock, see [6]). Besides, we will consider $0 < t_1^\star < ... < t_m^\star < T$. We denote by $\vec{t^\star}$ the vector $(t_1^\star, \cdots, t_m^\star)$.

We will study the concept of QTL mapping: we will look for associations between allele variations at the QTLs and variation in the quantitative trait of interest.

Usually, in the classical problem of QTL mapping, the "genome information" is available only at fixed locations $t_1 = 0 < t_2 < ... < t_K = T$, called genetic markers. Note that in the following, the word "genotype" will refer to the genome information at all the marker locations. So, usually an observation is

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

and the challenge is that the number m of QTLs and their locations $t_1^\star$, ..., $t_m^\star$ are unknown.

In this paper, we consider the classical problem, but in order to reduce the costs of genotyping, a selective genotyping has been performed: we consider two real thresholds S_- and S_+, with $S_- \leqslant S_+$ and we genotype if and only if the phenotype Y is extreme, that is to say $Y \leqslant S_-$ or $Y \geqslant S_+$. Selective genotyping was first introduced by [7] who noticed that most of the information about QTL is present in the extreme phenotypes (i.e. extreme traits). Later, [8] formalized this approach and called it selective genotyping (cf. our studies [9, 10] for more details). Although genotyping costs have largely dropped recently, selective genotyping or extreme sampling, is still a relevant concept in the modern genomic era. Today, application fields of selective genotyping lie in Genome Wide Association Study

(GWAS) (e.g. plants [11], animals [12], humans [13]), and in Genomic Selection (GS). GWAS is a popular QTL mapping technique and GS is a very hot topic in genomics (e.g. plants [14, 15]), that consists in selecting individuals on the basis of genomic predictions. We can also find applications of selective genotyping in biotechnology [16].

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

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

then, in our problem, one observation is now

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

Note that with our notations:

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

We consider that we have n observations $\big(Y_j, \, \overline{X}_j(t_1), \, ..., \, \overline{X}_j(t_K)\big)$, $j = 1, \ldots, n$ which are i.i.d. . The main aim of our study is to propose a new variable selection method to estimate the number m (i.e. $m \geqslant 1$) of QTLs, their locations $t_1^\star$, ..., $t_m^\star$ and their effects q_1, ..., q_m. Since our new method is built on a very popular technique in genetics (see [3, 4]), called the "Interval Mapping" ([8]), let us briefly recall this concept. "Interval Mapping" consists in scanning the genome, and leads to the study of stochastic processes along the genome. When there is only one QTL (i.e. $m = 1$), the location $t_1^\star$ is considered as an unknown parameter t and the likelihood process will also depend on this parameter t. The absence of a QTL is given by the null hypothesis H_0:"$q_1 = 0$," and the likelihood ratio test (LRT) of H_0 against its alternative, has test statistic $\sup_t \overline{\Lambda}_n(t)$, where $\overline{\Lambda}_n(t)$ is the LRT statistic at location t. Note that $\arg\sup_t \Lambda_n(t)$ is a natural estimator of the QTL location $t_1^\star$. In statistics, the distributions of the "score process", $\overline{S}_n(\cdot)$, and of the "LRT process", $\overline{\Lambda}_n(\cdot)$, have been studied extensively by [2, 17–23] under the complete data situation (i.e. $S_- = S_+$), and more recently by [10, 24] under selective genotyping. However, although the use of the test statistic $\sup \overline{\Lambda}_n(.)$ is appropriate for testing and localizing one QTL on $[0, T]$, it is not so rewarding when more than one QTL (i.e. $m > 1$) lie on $[0, T]$.

So, the main aim of our study is to propose a new variable selection method to estimate the number m (i.e. $m \geqslant 1$) of QTLs, their locations and their effects, with the help of the "score process" and the "LRT process". Our new method, suitable under selective genotyping and under the complete data situation, will be helpful for building a prediction model in GS. The programs used in this study are available at "http://charles-elie.rabier.pagesperso-orange.fr/doc/articles.html".

1.2. *Roadmap*

In Section 2, we present our theoretical results. Theorem 2.2 gives the asymptotic distribution of the score process and the LRT process under the alternative hypothesis that there exist m QTLs located at $t_1^\star$, ..., $t_m^\star$ with effects q_1, ..., q_m. Lemma 2.3 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. Theorem 2.4 shows that the signal is largely increased by genotyping extreme individuals, provided that the phenotyping is free. Corollary 2.5 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 [3]). According to Corollary 2.5, interaction effects are unidentifiable since they are not present in the mean function of the process. Last, Corollary 2.6 tackles the reverse configuration of selective genotyping, where only non extreme individuals are genotyped (i.e. the individuals for which $Y \in [S_-, S_+]$).

The theoretical results of Section 2 allow us to propose a new method, called SgLasso (for Selective genotyping Lasso), to estimate the number of QTLs, their positions and their effects using the Lasso ([25]). This method is described in Section 3. SgLasso differs from the classical Lasso since it models explicitly the extremes (see the appendix for some intuition on asymptotic theory). SgLasso 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. 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). Section 4 investigates theoretical properties of SgLasso, such as the rate of convergence for prediction and the consistency of the variable selection.

Next, Section 5 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, [25], Group Lasso, [26], Elastic Net, [27], RaLasso, [28] and BayesianLasso, [29]), specially when a unidirectional selective genotyping was performed (i.e. only the individuals for which $Y > S_+$ are genotyped, i.e. the so-called best individuals). Section 5.5 is devoted to a rice data analysis. Our study ends with Section 6 where we show that SgLasso presents the best performances for genomic prediction.

2. Theoretical results

For $t \in [t_1, t_K] \backslash \mathrm{T}_K$ where $\mathrm{T}_K = \{t_1, ..., t_K\}$, let us define t^ℓ and t^r as :

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

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

Let us consider the case $m = 1$ (i.e. one QTL located at $t_1^\star$), 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 $\big(Y,\ \overline{X}(t^\ell),\ \overline{X}(t^r)\big)$. According to [10], the likelihood of $\big(Y,\ \overline{X}(t^\ell),\ \overline{X}(t^r)\big)$ with respect to the measure $\lambda \otimes N \otimes N$, λ being the Lebesgue measure, N the counting measure on N, is $\forall t \in [t_1, t_K] \backslash \mathrm{T}_K$:

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

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

$\mathrm{P}\left\{X(t)=1 \mid X(t^\ell), X(t^r)\right\}$ and

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

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, in formula (2), $\overline{g}(t)$ is the following quantity:

$$\overline{g}(t) = \mathrm{P}\left\{X(t^\ell), X(t^r)\right\}\ 1_{Y\notin[S_-,S_+]}\ +\ 1_{Y\in[S_-,S_+]} \tag{3}$$

with

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

As a result, the likelihood is a function of Y, $\overline{X}(t^\ell)$, $\overline{X}(t^r)$, 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 $\overline{L}_{t_1^\star}(\theta^1)$.
The score statistic of the hypothesis "$q_1 = 0$" at t, for n independent observations, is defined as

$$\overline{S}_n(t) = \frac{\frac{\partial \bar{l}_t^n}{\partial q_1}\ |_{\theta_0^1}}{\sqrt{Var\left(\frac{\partial \bar{l}_t^n}{\partial q_1}\ |_{\theta_0^1}\right)}}\ , \tag{4}$$

where *Var* is the variance, $\bar{l}_t^n$ 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

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

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

2.1. *Main results*

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

$$\gamma := \mathrm{P}_{\mathcal{H}_0}\left(Y \notin [S_-,\ S_+]\right) \tag{6}$$

$$\gamma_+ := \mathrm{P}_{\mathcal{H}_0}\left(Y > S_+\right) \tag{7}$$

$$\gamma_- := \mathrm{P}_{\mathcal{H}_0}\left(Y < S_-\right) \tag{8}$$

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

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

Notation 2.1: *$\Rightarrow$ is the weak convergence, $\overset{F.d.}{\to}$ 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 $\overline{S}_n(\cdot)$ and the LRT process $\overline{\Lambda}_n(\cdot)$ 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 2.2: *Suppose that the parameters $(q_1, ..., q_m, \mu, \sigma^2)$ 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 $\mathcal{H}_{a\vec{t}^\star}$: "there are m QTLs located respectively at $t_1^\star, \cdots, t_m^\star$ with effect $q_1 = a_1/\sqrt{n}, \cdots, q_m = a_m/\sqrt{n}$ where $a_1 \neq 0, \cdots, a_m \neq 0$". Then, as $n \to +\infty$,*

$$\overline{S}_n(.) \Rightarrow V(.)\ ,\ \ \overline{\Lambda}_n(.) \overset{F.d.}{\to} V^2(.)\ \ ,\ \ \sup \overline{\Lambda}_n(.) \xrightarrow{\mathcal{L}} \sup V^2(.) \tag{10}$$

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

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

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

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

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

where

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

and $\mathcal{A}$ *is defined in* (9).

The proof of Theorem 2.2 is given in Section 2 of Supplement A. It is based on [10, 30–32]. The case $m > 1$ differs from the case $m = 1$ since the true probability distribution is the one of $\big(Y, \overline{X}(t_1^{\star\ell}), \overline{X}(t_1^{\star r}), \ldots, \overline{X}(t_m^{\star\ell}), \overline{X}(t_m^{\star r})\big)$. Indeed, all the information is contained in the flanking markers of all QTLs locations. This probability distribution is given in Section 1 of Supplement A.

According to Theorem 2.2, under this general alternative, the LRT process is still asymptotically the square of a "non linear interpolated process", as in Theorem 4.1 of [10] where the focus was only on the case $m = 1$ under selective genotyping. Besides, as in [10], the difference between the complete data situation and the selective genotyping approach is translated by a difference between the mean functions of the asymptotic processes: they are proportional of a factor linked to the selective genotyping. However, contrary to [10] and [23], the mean function depends here on the number of QTLs, their positions and their effects.

Note that Theorem 2.2 gives also the asymptotic distribution of the statistic $\sup \overline{\Lambda}_n(.)$ when $m \geqslant 1$, since this test can be viewed as a global test or max test (see for instance [33]). In this context, $\sup \overline{\Lambda}_n(.)$ matches the test statistic corresponding to the statistical test with the smallest p-value 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.

In the following lemma, we study the Asymptotic Relative Efficiency (ARE). Recall that the 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 2.3: *Let* κ *denote the Asymptotic Relative Efficiency, then we have*

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

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 [10] where the focus was only on the case $m = 1$. To prove Lemma 2.3, just use the same proof as the one of Theorem 4.2 of [10].

According to i) of Lemma 2.3, the ARE with respect to the complete data situation, does not depend on the number of QTLs m, the constants a_1, ..., a_m linked to the QTL effects, and the QTLs locations $t_1^\star$, ..., $t_m^\star$. Indeed, since the mean functions (complete data situation and selective genotyping) are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$, it is obvious that the ARE does not depend on those parameters. As a consequence, we have exactly the same ARE as the one obtained in [10] for $m = 1$. On the other hand, according to ii) of Lemma 2.3, 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 $n^\star$ individuals for a selective genotyping experiment, and let us assume that we have the relationship $n = n^\star\gamma$. In other words, we focus on the case where, for economical reasons, we are allowed to genotype only n individuals. By considering $n = n^\star\gamma$, we are allowed to genotype n extreme individuals, provided that the overall population size has been increased to $n^\star$. In this context, following the same lines as the proof of Theorem 2.2, we obtain:

Theorem 2.4 : *Suppose that the parameters $(q_1, ..., q_m, \mu, \sigma^2)$ vary in a compact and that σ^2 is bounded away from zero, and also that m is finite. Assume that $n^\star = n/\gamma$. Let $\mathcal{H}_0$ be the null hypothesis of no QTL on $[0, T]$, and let define the following local alternatives $\mathcal{H}_{a\vec{t}^\star}$: "there are m QTLs located respectively at $t_1^\star, \cdots, t_m^\star$ with effect $q_1 = a_1/\sqrt{n}, \cdots, q_m = a_m/\sqrt{n}$ where $a_1 \neq 0, \cdots, a_m \neq 0$." Then, as $n \to +\infty$,*

$$\overline{S}_{n^\star}(.) \Rightarrow V_\star(.) \ , \ \overline{\Lambda}_{n^\star}(.) \stackrel{F.d.}{\to} V_\star^2(.) \quad , \quad \sup \overline{\Lambda}_{n^\star}(.) \stackrel{\mathcal{L}}{\longrightarrow} \sup V_\star^2(.) \tag{11}$$

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

$$V_\star(t) = \frac{\alpha(t)\, V_\star(t^\ell) \ + \ \beta(t)\, V_\star(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) \ + \ 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} \ ,$$
$$Cov\{V_\star(t_k), V_\star(t_{k'})\} = \rho(t_k, t_{k'}) = e^{-2|t_k - t_{k'}|} \quad \forall (t_k, t_{k'}) \in \mathrm{T}_K \times \mathrm{T}_K$$

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

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

where

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

and $\mathcal{A}$ is defined in (9).

As a result, the ratio between the signal corresponding to selective genotyping and the one matching the complete data situation is equal to $\sqrt{\dfrac{\mathcal{A}}{\gamma\sigma^2}}$. 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

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

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

Figure 1 around here

2.2. *Some corollaries*

2.2.1. 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 [3]). 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

$$Y = \mu \; + \; \sum_{s=1}^{m} X(t_s^\star)\, q_s \; + \; \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^{m} X(t_s^\star) X(t_{\tilde{s}}^\star)\, q_{s,\tilde{s}} \; + \; \sigma\varepsilon \tag{12}$$

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

Corollary 2.5: *Suppose that the parameters* $(q_1, ..., q_m, q_{1,2}, ..., q_{m-1,m}, \mu, \sigma^2)$ *vary in a compact and that* σ^2 *is bounded away from zero, and also that* m *is finite. Let define the local alternative*

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

then, with the previous notations, under $\mathcal{H}_{a\vec{t}^\star, b\vec{t}^\star}$ *, as* n *or* $n^\star$ *tends to infinity, results* (10) *and* (11) *of Theorem 2.2 and Theorem 2.4 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.

2.2.2. 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 2.6: *Under the reverse configuration, that is to say if* $\overline{X}(t_k) = X(t_k)\, 1_{Y \in [S_-\, ,\, S_+]}$*, then we have the same results as in Theorem 2.2, Theorem 2.4*

and Corollary 2.5 provided that we replace the quantity $\mathcal{A}$ by the quantity $\mathcal{B}$ defined in the following way

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

The proof is largely inspired of the proof of Theorem 2.2, Theorem 2.4, Corollary 2.5, and also from [34] where this configuration is studied under the local alternative of one QTL at $t^\star$ on $[0, T]$.

3. 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 2.2 and 2.4 and a penalized likelihood method.

Notation 3.1: *$\mathcal{G}_{\gamma,\sigma}$ denotes respectively $\dfrac{\sqrt{\mathcal{A}}}{\sigma}$ or $\dfrac{\sqrt{\mathcal{A}}}{\sqrt{\gamma}\sigma}$ when the total number of phenotypic observations is n or $n^\star = n/\gamma$.*

In the sequel $\tilde{n}$ denotes the total number of phenotypic observations. It may be n or $n^\star$. According to Theorems 2.2 and 2.4, as soon as we discretize the score process at markers positions, we have the following relationship when $\tilde{n}$ is large:

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

where $\vec{S}_{\tilde{n}} = \left(\overline{S}_{\tilde{n}}(t_1) \; , \; \overline{S}_{\tilde{n}}(t_2) \; , \; ... \; , \; \overline{S}_{\tilde{n}}(t_K)\right)^\top$, $\vec{\varepsilon} \sim N(0, \Sigma)$ with $\Sigma_{kk'} = \rho(t_k, t_{k'})$ and $\vec{m}_{\vec{t}^\star} = \left(\overline{m}_{\vec{t}^\star}(t_1) \; , \; \overline{m}_{\vec{t}^\star}(t_2) \; , \; ..., \overline{m}_{\vec{t}^\star}(t_K)\right)^\top$.

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 $\overline{S}_{\tilde{n}}(.)$ is an "interpolated process".

Remark 1 : In genomics, once the genetic map is built (see [3] for instance), the correlation between (the genome information at) markers is perfectly known. As a consequence, since the correlation between $X(t_k)$ and $X(t_{k'})$ is $\rho(t_k, t_{k'})$, the matrix Σ is known.

Let us consider the Cholesky decomposition $\Sigma = AA^\top$. We have

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

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

Since the number m of QTLs and their positions $t_1^\star$,...,$t_m^\star$ are unknown, we propose to focus on a new discretization of $[0, T]$ corresponding to all the putative QTL locations: $0 \leqslant t'_1 < t'_2 < ... < t'_L \leqslant T$. 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:

$$A^{-1}\vec{S}_{\tilde{n}} = A^{-1}C \left(\Delta_1 \; , \; ... \; , \; \Delta_L\right)^\top \; + \; A^{-1}\vec{\varepsilon} \; + \; o_P(1) \tag{13}$$

where C is a matrix of size $K \times L$ such as $C_{kl} = e^{-2|t_k - t'_l|}$, $k = 1, \cdots, K$ and $l = 1, \cdots, L$.
Last, in order to find the non zero Δ_l, a natural approach is to use a penalized regression and estimate Δ by:

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

where:

$$\mathrm{pen}(\alpha) = \frac{1-\alpha}{2} \left\| \Delta \right\|_2^2 + \alpha \left\| \Delta \right\|_1 \tag{15}$$

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

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

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

Our estimators will be compared in section 5.4 with the classical estimators such as the Lasso ([25]) and its cousins (e.g. [26, 27]). 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 $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^\top$. In addition, let M_{ext} denote the matrix, where each row contains the multivariate random variable $(1, X(t_1), \ldots, X(t_K)) \mid Y \notin [S_-, S_+]$ 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 $\hat{\beta}_{\mathrm{Lasso}}(\lambda)$, and the classical Elastic Net estimator $\hat{\beta}_{\mathrm{EN}}(\lambda, \alpha)$ are the following under selective genotyping:

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

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

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.

4. 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 6 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 $C = \Sigma$, $A^{-1}C = A^\top$ and $\Delta = (\Delta_1, ..., \Delta_K)^\top$. 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:

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

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. [35] page 108). Since $\hat{\sigma}_k^2 := \frac{1}{K}(AA^\top)_{kk} = \frac{\Sigma_{kk}}{K} = \frac{\rho(t_k,t_k)}{K} = \frac{1}{K}$, let us set $A_{\text{scal}} := \sqrt{K}A^\top$. Then, let us define

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

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

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

We can apply Corollary 6.1 of [35] with $\hat{\sigma} = 1$ (cf. our linear model in formula (13)), that establishes the slow rate of convergence

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

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

On the other hand, assuming that the "compatibility condition" holds, Corollary 6.2 of [35] applies and we obtain the fast rate of convergence:

$$\frac{\left\| A_{\text{scal}}(\hat{\widetilde{\Delta}}_{\text{SgLasso}_{\text{scal}}} - \widetilde{\Delta}) \right\|_2^2}{K} = O_P\left(\frac{\log(K)\, m}{K\, \Phi_0^2} \right) \tag{22}$$

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:

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

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

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

The bounded pairwise correlation:

$$\frac{\sqrt{m}\ \max_{k\in \mathrm{T}_K\setminus \mathrm{T}^\star_m} \sqrt{\sum_{s\in \mathrm{T}^\star_m | t^\star_s\neq t_k} \rho^2(t_k, t^\star_s)}}{d^2_{\min}(\Sigma^{(\star,\star)})} \leqslant C < 1 \tag{23}$$

where $\mathrm{T}^\star_m = \{t^\star_1, ..., t^\star_m\}$, C is a constant, $\Sigma^{(\star,\star)}$ is the submatrix of Σ restricted to QTL loci, and $d^2_{\min}(\Sigma^{(\star,\star)})$ refers to the smallest eigenvalue of $\Sigma^{(\star,\star)}$.

The irrepresentable condition:

$$\left\|\Sigma^{(.,\star)}(\Sigma^{(\star,\star)})^{-1}\mathrm{Sign}(a_1,\ldots,a_m)\right\|_\infty \leqslant C < 1$$

where $\|x\|_\infty = \max_j |x_j|$, $\mathrm{Sign}(a_1,\ldots,a_m) = (\mathrm{Sign}(a_1),\ldots,\mathrm{Sign}(a_m))^\top$, and $\Sigma^{(.,\star)}$ is a matrix of size $(K-m)\times m$. $\Sigma^{(.,\star)}$ 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 [35], 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.

5. Illustrations regarding max test and GWAS

5.1. *Simulation framework*

Data were simulated thanks to the Matlab software. The genome of one individual was simulated according to Haldane [52]. In particular, the random variable $X(0)$ representing the genome information at location 0, was drawn from a Bernoulli distribution (taking values +1 or -1 in our case) with parameter 1/2, and the recombinations events were obtained by sampling along the genome, independent random variables from the exponential distribution with parameter 1. This way, the process N(.) representing the number of recombination events, corresponds to the standard Poisson process on [0,T]. At each location matching a recombination event, the sign of the genome information was switched. The phenotype Y was generated according to formula (1). The variance σ^2 was set to 1 in all simulated data. Next, the score test was obtained at each marker located at t_k by computing the following test statistic:

$$T_n(t_k) = \frac{\sum_{j=1}^n (Y_j - \overline{Y})1_{\overline{X}_j(t_k)=1} \ - \ (Y_j - \overline{Y})1_{\overline{X}_j(t_k)=-1}}{\sqrt{\sum_{j=1}^n (Y_j - \overline{Y})^2 1_{\overline{X}_j(t_k)\neq 0}}}$$

where $\overline{Y}$ is the global mean. Note that this test statistic was introduced in Rabier [9] in a more general framework, and was later used in Rabier [10].

In order to compute the maximum of the process (cf. Section 5.2 below), simulated data were analyzed using Lemma 1 of Azaïs et al. [23], that is to say performing LRT on markers (by computing the square of score statistics on markers) and performing only one test in each marker interval if the ratio of the score statistics on markers fulfills the given condition (cf. Rabier [10]).

In order to compute the SgLasso and its cousins, score statistics on markers were considered and decorrelated thanks to the "chol" function of Matlab that implements the Cholesky decomposition. Last, data were analyzed using the R software, because of the large number of available packages dedicated to penalized regressions (cf. Section 5.4).

5.2. *About the max test*

We propose to illustrate here our theoretical results regarding the max test. Recall that it relies on the test statistic, $\sup \overline{\Lambda}_n(.)$. 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. [23] and based on Genz [36]. Recall that the threshold remains the same under selective genotyping as under the complete data situation (cf. Theorem 2.2).

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. $\gamma_+ = \gamma/2$). 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 2.2.

Last, the power of the test is reported as a function of the QTL effect signs. We can see that when the two QTLs at 3cM and 28cM have the same signs, the power is almost equal to 1 whereas it largely decreases ($\approx 15\%$ for $\gamma = 1$) when the signs are opposite. In this case, the max test is clearly not the most appropriate test to perform. We refer to the recent study of [33] where the authors compared performances of the max test and the ANOVA in another context.

5.3. *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 3) 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 $t'_l = 0.025(l - 1)$, $l = 1$,

..., 401. Recall that this grid is different from the one corresponding to marker locations: $t_k = 0.05(k-1)$, $k = 1, \ldots, 201$. 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 $n^\star$ individuals under selective genotyping and n individuals under the complete data situation. In other words, our simulation set up follows Theorem 2.4.

For instance, when n was equal to 100 ($\gamma = 1$), $n^\star$ 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 $n^\star = n/\gamma$), the quantities Δ_l, present in formula (13), are increased by a factor $\sqrt{\mathcal{A}}/\sqrt{\gamma}$ at each gene location. Then, under the configuration studied, the quantities $|a|\sqrt{\mathcal{A}}/\sqrt{\gamma}$ 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.

Figure 2 around here

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 2 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 an 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 1 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 smallest) individuals are genotyped ($\gamma_+/\gamma = 1$), the L1 ratio is the smallest. It confirms our theoretical results presented in Section 2 and illustrated in Figure 1.

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

Table 2 around here

5.4. *Comparison with existing methods*

In this section, we propose to compare our new method with existing methods. We will concentrate on the Lasso ([25]), the Group Lasso ([26]), the Elastic Net ([27]), the Bayesian Lasso ([29]), and the RaLasso ([28]).

Recall that the Group Lasso differs from his cousin Lasso, since it allows to handle a group structure (see [37]). In the context of genomic prediction, the Bayesian Lasso was used in [38] 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 ([28]). The RaLasso can be viewed as a method that models the dependency between regressors and errors: the loss function can be either quadratic or linear, depending on the regressor values.

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 : $\text{loss}(t) = \frac{t^2}{2M} 1_{|t|\leqslant M} + (|t| - M/2)\, 1_{|t|\geqslant M}$, 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 (2.2) of [28]. Last, we have to mention that the RaLasso incorporates the Huber loss 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, we chose the model for which the number of parameters was the closest to 16, thanks to the GLMNet package. For the Bayesian Lasso, we considered a grid search for the learning rate η and chose the η that maximized the L1 ratio (optimal setting). Last, in order to compute the RaLasso, we ran a grid search to find the best pair $(M,\, \lambda)$. For each value of M, we chose the λ matching the model with a number of parameters close to 16. Next, the best pair $(M,\, \lambda)$ was the one maximizing the L1 ratio.

Table 3 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 [28]). 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 4 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 5 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. This can be explained by the fact that we chose groups of 10 consecutive markers, and we should have probably considered groups of 5 markers in order to increase the accuracy. However, for the same framework of 10 markers, in view of Tables 4 and 5, SgGroupLasso outperformed the GroupLasso when the selective genotyping was unidirectional.

Tables 3, 4, 5, 6 and 7 around here

5.5. *Real data analysis*

To illustrate performances of our new method on real data, we analyzed data from the joint papers [39] and [40] 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 [39]), 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 [39]), 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 $\lambda = 0.1, 10.1, \ldots, 1000.1$ and $M = 0.1, 0.2, \ldots, 1$. 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 [40] 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 [39] ($n = 312$ 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 [40] (cf. their S1 Table), and obtained after fitting a linear mixed model, are given at the top of Table 8. Note that the most significant SNPs for the flowering date are located on chromosome 3 (see [40]). Indeed, the p-values associated to 5 SNPs on chromosome 3 and reported by [40], 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 8. FP refers to the number of falsely selected variables whereas FN is the number of genes that are not selected.

According to Table 8, SgLasso and SgEN selected respectively 26 and 33 genes under the complete data situation ($\gamma = 1$). All the genes found by [40] 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 [40] 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 [40] 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 [40]. 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.

Table 8 around here

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

Genomic Selection (GS) ([41]) 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 (e.g. [42, 43]) of young candidates as soon as their DNA has been collected. GS was first applied to animal breeding (see [44] for a review), and it is nowadays extensively investigated in plants. We can mention recent genomic prediction studies on apple ([45]), eucalyptus ([46]), japanese pears ([47]), strawberry ([14]), banana ([15]) and coffea ([48]). 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. [49–53]) 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 [54]). [55] 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, [54] 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 [54]). 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 5.4 (cf. Tables 3 and 4).

6.1. *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 $R(.)$ denote a standard poisson process on $[0, T]$ representing the number of recombinations. $W(.)$, 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: $U = \mu + \sum_{s=1}^{m} W(t_s^\star)q_s + \sigma\varepsilon$, where ε is a Gaussian white noise.

In what follows, the notation "new" will refer to the progeny of an extreme individual ; U_{new}, $W_{\text{new}}(.)$, ε_{new}, $R_{\text{new}}(.)$ 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 [56]). The *phenotypic accuracy*, ρ_{ph}, also called predictive ability, is defined as the correlation between the predictor $\hat{U}_{\text{new}}$ and the trait U_{new}, i.e. $\text{Cor}\left(\hat{U}_{\text{new}},\, U_{\text{new}}\right)$ (see for instance [57]). 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:

$$\widehat{U}_{\text{new}} = (1, W_{\text{new}}(t_1), \ldots, W_{\text{new}}(t_K))\, \hat{\beta}_{\text{Lasso}}\,,$$
$$\widehat{U}_{\text{new}} = (W_{\text{new}}(t_1), \ldots, W_{\text{new}}(t_K))\, \hat{\Delta}_{\text{SgLasso}}(\lambda) \frac{\sigma\sqrt{\gamma}}{\sqrt{n\mathcal{A}}}.$$

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 6 and 7 report the average Pearson correlation computed over 100 data sets containing 100 progenies.

According to Table 6, 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 3 and 4). Recall that [54] 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 7 compares SgLasso and its cousins : SgLasso, SgEN and SgGroupLasso, presented an accuracy of same order.

Acknowledgements

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

Supplementary materials

Supplement A: We give the mathematical proofs of Theorem 2.2 and Corollary 2.5.
Supplement B: It contains supplementary illustrations. Figure 1 describes the simulation framework, regarding genomic selection (cf. Section 6.1 of the main text). Table 1 illustrates the performances of SgLasso as a function of the ratios γ_+/γ (cf. Section 5.3 of the main text).

References

[1] K. Broman, T. Speed, *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) (2002), pp. 641–656.
[2] J.M. Azaïs and M. Wschebor, *Level sets and extrema of random processes and fields*, Wiley, New-York (2009).
[3] R. Wu, C.X. Ma, G. Casella, *Statistical Genetics of Quantitative Traits*, Springer, New York (2007).

[4] D. Siegmund, B. Yakir, *The statistics of gene mapping*, Springer, New York (2007).
[5] J.B.S. Haldane, *The combination of linkage values and the calculation of distance between the loci of linked factors*, Journal of Genetics, 8 (1919), pp. 299–309.
[6] B. Hayes, *QTL Mapping, MAS, and Genomic Selection*, Short course organized by Iowa State University, (2007).
[7] R.J. Lebowitz, M. Soller, J.S. Beckmann, *Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines*, Theor. Appl. Genet., 73 (1987), pp. 556–562.
[8] E.S. Lander, D. Botstein, *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps*, Genetics, 138 (1989), pp. 235–240.
[9] C.E. Rabier, *On statistical inference for selective genotyping*, J. Stat. Plan. Infer., 147 (2014), pp. 24–52.
[10] C.E. Rabier, *On stochastic processes for Quantitative Trait Locus mapping under selective genotyping*, Statistics, 49(1) (2015), pp. 19–34.
[11] A. Gutierrez, J. Hoy, C. Kimbeng, N. Baisakh, *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 (2018), 877.
[12] J.P. Kurz, Z. Yang, R.B. Weiss, D.J. Wilson, K.A. Rood, G.E. Liu, Z. Wang, *A genome-wide association study for mastitis resistance in phenotypically well-characterized Holstein dairy cattle using a selective genotyping approach*, Immunogenetics, 71(1) (2019), pp. 35–47.
[13] D. Zabaneh et al., *A genome-wide association study for extremely high intelligence*, Molecular psychiatry, 23(5) (2018), 1226.
[14] S.A. Gezan, L.F. Osorio, S. Verma, V.M. Whitaker, *An experimental validation of genomic selection in octoploid strawberry*, Horticulture research, 4 (2017), 16070.
[15] M. Nyine et al., *Genomic prediction in a multiploid crop: genotype by environment interaction and allele dosage effects on predictive ability in banana*, The Plant Genome, 11(2) (2018), 170090.
[16] C. Zou, P. Wang, Y. Xu, *Bulked sample analysis in genetics, genomics and crop improvement,* Plant biotechnology journal, 14(10) (2016), pp. 301–320.
[17] A. Rebaï, B. Goffinet, B. Mangin, *Comparing power of different methods for QTL detection*, Biometrics, 51 (1995), pp. 87–99.
[18] A. Rebaï, B. Goffinet, B. Mangin, *Approximate thresholds of interval mapping tests for QTL detection*, Genetics, 138 (1994), pp. 235–240.
[19] C. Cierco, *Asymptotic distribution of the maximum likelihood ratio test for gene detection*, Statistics 31 (1998), pp. 261–285.
[20] J.M. Azaïs and C. Cierco-Ayrolles, *An asymptotic test for quantitative gene detection*, Ann. Inst. Henri Poincaré (B) 38(6) (2002), pp. 1087–1092.
[21] Z. Chen, H. Chen, *On some statistical aspects of the interval mapping for QTL detection*, Statistica Sinica 15 (2005), pp. 909–925.
[22] M.N. Chang, R. Wu, S.S. Wu, G. Casella (2009), *Score statistics for mapping quantitative trait loci*, Stat. Appl. Genet. Mol. Biol. 8(1) (2009), pp. 1–35.
[23] J.M. Azaïs, C. Delmas, C.E Rabier, *Likelihood ratio test process for Quantitative Trait Locus detection*, Statistics 48(4) (2012) pp. 787–801.
[24] C.E. Rabier, *On empirical processes for Quantitative Trait Locus mapping under the presence of a selective genotyping and an interference phenomenon*, J. Stat. Plan. Infer., 153 (2014), pp. 42–55.
[25] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society B, 58(1) (1996), pp. 267–288.
[26] M. Yuan, Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society Series B, 68(1) (2006), pp. 49–67.
[27] H. Zou, T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2) (2005), pp. 301–320.
[28] J. Fan, Q. Li, Y. Wang, *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) (2017), pp. 247–265.
[29] T. Park, G. Casella, *The bayesian lasso*, Journal of the American Statistical Association, 103(482) (2008), pp. 681–686.
[30] J.M. Azaïs, E. Gassiat, C. Mercadier, *Asymptotic distribution and local power of the likelihood ratio test for mixtures*, Bernoulli 12(5) (2006), pp. 775–799.
[31] J.M. Azaïs, E. Gassiat, C. Mercadier, *The likelihood ratio test for general mixture models with possibly structural parameter*, ESAIM 13 (2009), pp. 301–327.
[32] A.W. Van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (1998).
[33] E. Arias-Castro, E.J. Candes, Y. Plan, *Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism*, The Annals of Statistics 39(5) (2011), pp. 2533–2556.
[34] C.E. Rabier, *An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes*, Annales de la faculté des sciences de Toulouse, 6(23) (2014), pp. 755–778.
[35] P. Bühlmann, S. Van de Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science (2011).
[36] A. Genz, *Numerical computation of multivariate normal probabilities*, J. Comp. Graph. Stat., 1 (1992), pp. 141–149.
[37] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning theory*, Springer, New York (2001).
[38] A.A. Boligon, N. Long, L.G.D. Albuquerque, K.A. Weigel, D. Gianola, G.J.M. Rosa, *Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection*, Journal of animal science 90(13) (2012), pp. 4716–4722.
[39] J. Spindel et al., *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) (2015), e1004982.

[40] H. Begum et al., *Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa)*, PloS one 10(3) (2015), e0119873.

[41] T.H. Meuwissen, B. Hayes, M.E. Goddard, *Prediction of total genetic value using genome-wide dense marker maps.* Genetics, 157(4) (2001), pp. 1819–1829.

[42] M. Momen, A.A. Mehrgardi, A. Sheikhi, A. Kranis, A. Tusell, L. Morota, G. Rosa , D. Gianola, *Predictive ability of genome-assisted statistical models under various forms of gene action*, Scientific reports, 8 (2018).

[43] C.E. Rabier, B. Mangin, S. Grusea, *On the accuracy in high dimensional linear models and its application to genomic selection*, Scandinavian Journal of Statistics, **46**(1) (2019), pp. 289–313.

[44] B. Hayes, P. Bowman, A. Chamberlain, M. Goddard, *Invited review: Genomic selection in dairy cattle: Progress and challenges*, Journal of dairy science, 92(2) (2009), pp. 433–443.

[45] H. Muranty et al., *Accuracy and responses of genomic selection on key traits in apple breeding*, Horticulture research, 2 (2015), 15060.

[46] B. Tan, D. Grattapaglia, G.S. Martins, K.Z. Ferreira, B. Sundberg, P.K. Ingvarsson, *Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids*, BMC plant biology, 17(1) (2017), 110.

[47] M.F. Minamikawa et al., *Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (Pyrus pyrifolia Nakai)*, Scientific reports, 8(1) (2018), 11994.

[48] L.F.V. Ferrao, R.G. Ferrao, M.A.G. Ferrao, A. Fonseca, P. Carbonetto, M. Stephens, A.A.F. Garcia, *Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models*, Heredity, (2018).

[49] A. Wolc, *Persistence of accuracy of genomic estimated breeding values over generations in layer chickens*, Genetics Selection Evolution, 43(1) (2011), 23.

[50] M. Pszczola, M.P.L. Calus, *Updating the reference population to achieve constant genomic prediction reliability across generations*, animal, 10(6) (2016), pp. 1018–1024.

[51] C.E. Rabier, P. Barre, T. Asp, G. Charmet, B. Mangin, *On the Accuracy of Genomic Selection*, PloS One, 11(6) (2016), e0156086. doi:10.1371/ journal.pone.0156086.

[52] H.J. Auinger et al., *Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.)*, Theor. Appl. Genet. 129(11) (2016), pp. 2043–2053.

[53] J.L. Neyhart, T. Tiede, A.J. Lorenz, K.P. Smith, *Evaluating methods of updating training data in long-term genomewide selection*, G3: Genes, Genomes, Genetics, 7(5) (2017), pp. 1499–1510.

[54] S.P. Brandariz, R. Bernardo, *Maintaining the Accuracy of Genomewide Predictions when Selection Has Occurred in the Training Population*, Crop Science 58(3) (2018), pp. 1226–1231.

[55] Y. Zhao, M. Gowda, F.H. Longin, T. Würschum, N. Ranc, J.C. Reif, *Impact of selective genotyping in the training population on accuracy and bias of genomic selection,* Theoretical and Applied Genetics, 125(4) (2012), pp. 707–713.

[56] M. Lynch, B. Walsh, *Genetics and analysis of quantitative traits.* Sinauer Sunderland, MA (1998).

[57] P.M. Visscher, J. Yang, M.E. Goddard (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) (2010), pp. 517–524.

Appendix

Intuition on asymptotic theory

Selective genotyping is challenging since some correlation is present between the errors ε and the genome of extreme individuals. Indeed, by definition, $\overline{X}(.)$ depends on Y, that contains the noise ε. In order to show the influence of this correlation, let us consider $m \geqslant 1$. At the marker location t_k, the score statistic, $\overline{S}_n(t_k)$, can be decomposed in the following way (cf. formula (2.8) in Section 2 of Supplement A):

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

where $\mathcal{A}$ is a quantity linked to the choice of S_- and S_+ (see formula (9) in Section 2). 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 [32]). Then, according to a technical proof , we have

the relationship

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

where Ω is a function of a_1, ..., a_m, $t_1^\star$, ..., $t_m^\star$, t_k, S_- and S_+. The proof is given in Section 2.2.1 of Supplement A (cf. lines below formula (2.9) until formula (2.10), and see also Sections 2.3 and 4). As a consequence, the correlation between ε and $\overline{X}(t_k)$ plays a role in the asymptotic theory. In contrast, under the complete data situation ($S_- = S_+$), the random variable $\overline{X}(t_k)$, equal to $X(t_k)$, is independent of ε by definition: since ε is centered, Ω is the constant null function.

Details about SgGroupLasso

We give here some details about SgGroupLasso which is based on the Group Lasso penalty. Recall that the column vector Δ is equal to $(\Delta_1, ..., \Delta_L)^\top$, where Δ_l is the putative effect at location t_l. Then, the SgGroupLasso estimator is the following:

$$\hat{\Delta}_{\text{SgGroupLasso}}(\lambda) = \arg\min_{\Delta} \left(\left\| A^{-1}\vec{S}_{\tilde{n}} - A^{-1}C\Delta \right\|_2^2 + \lambda \sum_{i=1}^{\text{nbGroup}} \sqrt{L_i} \left\| \vec{\Delta}_i \right\|_2 \right)$$

where L_i is the number of locations considered in the i-th group, and $\vec{\Delta}_i$ is the column vector containing the components of Δ referring to the i-th group.
Note that in our illustrations, a group is a set of consecutive locations. Under the dense map, since only marker locations are considered, a group is a set of consecutive markers.

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, for different values of the ratio γ_+/γ.

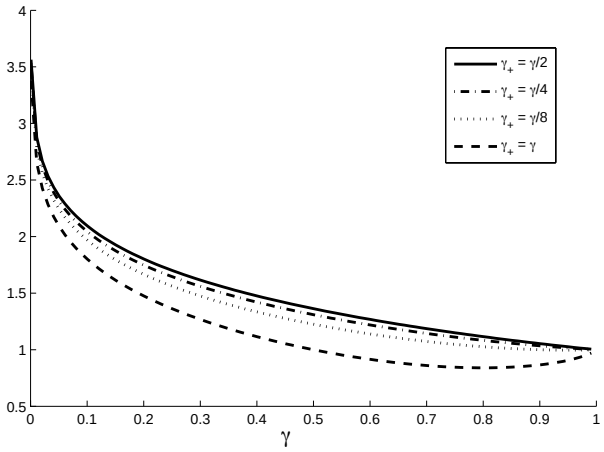
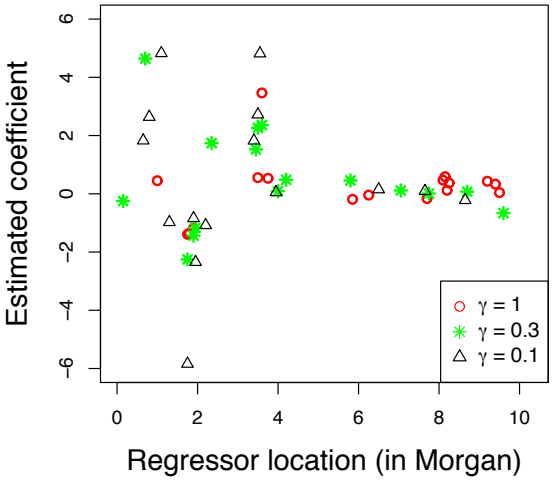
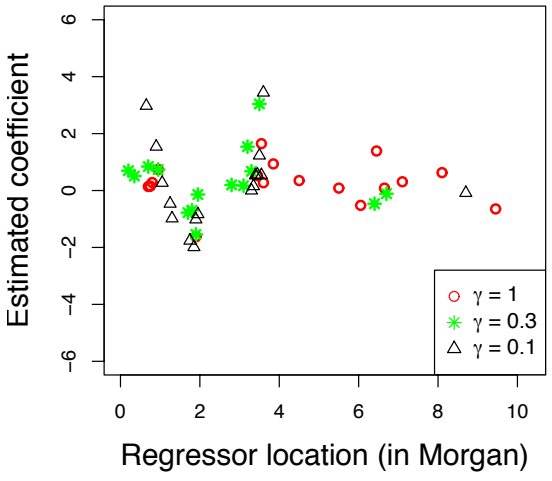


Figure 2. Estimated coefficients according to our new method as a function of the percentage γ of genotyped individuals (1 sample, $m = 16$, $T = 10$, $|q_1| = \ldots = |q_{16}| = 0.2$, QTLs randomly located only on [0,4], $\sigma = 1$, $K = 201$, $t_k = 0.05(k-1)$, $L = 401$, $t'_l = 0.025(k-1)$, $\gamma_+/\gamma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

A) B)



$n = 200$



$n = 100$

Table 1. Theoretical power and empirical power associated to the test statistic $\sup \Lambda_n(.)$, and as a function of the number m of QTLs and the percentage γ of genotyped individuals ($T = 1$, $K = 21$, $t_k = 0.05(k-1)$, $(m = 1,\ t_1^\star = 0.03)$, $(m = 2,\ t_1^\star = 0.03,\ t_2^\star = 0.80)$, $(m = 3,\ t_1^\star = 0.03,\ t_2^\star = 0.28,\ t_3^\star = 0.72)$, 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, $\gamma_+/\gamma = 1/2$).

γ	n \ m	1 (+)	2 (++)	2 (+-)	3 (+-+)
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 2. 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.

all $\lvert q_s\rvert$	γ	(Sparse, $n = 100$) L1 ratio	$\hat{m}$	(Sparse, $n = 200$) L1 ratio	$\hat{m}$	(Dense, $n = 100$) L1 ratio	$\hat{m}$	(Dense, $n = 200$) 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 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, $m = 16$, $|q_1| = \ldots = |q_{16}| = 0.1$, $T = 10$, 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.

		SgLasso		Lasso		Group Lasso		EN		RaLasso		Bayesian Lasso
γ	γ_+/γ	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 4. 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, $m = 16$, $|q_1| = \ldots = |q_{16}| = 0.2$, $T = 10$, 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.

		SgLasso		Lasso		Group Lasso		EN		RaLasso		Bayesian Lasso
γ	γ_+/γ	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 5. 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 , $t_k = t'_l = 0.001(k-1)$. 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 $\|q_s\| = 0.2$					
		SgLasso		SgGroupLasso		SgEN		SgLasso		SgGroupLasso		SgEN	
γ	γ^+/γ	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 6. 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 $\lvert q_s\rvert$	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 7. 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 indivduals (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 $\lvert q_s \rvert$	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 8. Comparison, on rice data ([39, 40]), 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 [40]. A marker in italic refers to a marker which is located at a maximum distance of 0.4cM from a gene inferred by [40]. 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 [40]. 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.

γ	Method	FP	FN	Selected genes
1	[40]			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)	*S3-1094192*, **S3-1125848**, **S3-1165376**, **S3-1269941** **S3-1394477**, S3×21
1	SgEN	28 (24)	5 (5)	*S3-1030333*, *S3-1094192*, *S3-1123429*, **S3-1125848**, **S3-1165376** *S3-1179404*, **S3-1221494** ,**S3-1269941** **S3-1394477**, S3×24
1	SgGroupLasso	37 (23)	7 (6)	S3×31, *S3-1030333*, *S3-1070111*, *S3-1094192*, *S3-1123429* **S3-1125848**, **S3-1165376**, *S3-1179404*, **S3-1221494**, *S3-1225693*
0.3	SgLasso	28 (23)	6 (5)	*S3-1070111*, *S3-1094192*, **S3-1165376**, **S3-1221494**, *S3-1225693*, **S3-1269941** *S3-1298550*, *S3-1354306*, **S3-1394477**, S3×23
0.3	SgEN	26 (23)	5 (5)	*S3-1030333*, *S3-1094192*, *S3-1123429* , **S3-1125848**, **S3-1165376**, **S3-1221494** **S3-1269941**, **S3-1394477**, S3×23
0.3	SgGroupLasso	65 (51)	5 (5)	*S3-1030333*, *S3-1070111*, *S3-1094192*, *S3-1123429*, **S3-1125848** **S3-1165376**, *S3-1179404*, **S3-1221494**, *S3-1225693*, **S3-1269941** *S3-1298550*, *S3-1320779*, *S3-1342244*, *S3-1354306*, **S3-1394477**, *S3-1403300* *S3-1439520*, *S3-1462159*, *S3-1495153*, 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, *S3-1030333*, *S3-1070111*, *S3-1094192* *S3-1123429*, **S3-1125848**, **S3-1165376**, *S3-1179404*, **S3-1221494** *S3-1225693*, 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)	*S3-1094192*, *S3-1123429*, **S3-1125848**, **S3-1165376** **S3-1221494**, **S3-1269941**, **S3-1394477**
0.3	Group Lasso	7 (2)	7 (6)	S3×2, *S3-1070111*, *S3-1094192*, *S3-1123429*, **S3-1125848** **S3-1165376**, *S3-1179404*, **S3-1221494**, *S3-1225693*
1	RaLasso	2600 (2568)	5 (0)	S1×704, S2×220, *S3-1123429*, **S3-1125848**, **S3-1165376**, *S3-1179404* **S3-1221494**, *S3-1225693*, **S3-1269941**, *S3-1298550* *S3-1320779*, *S3-1342244*, *S3-1354306*, **S3-1394477** *S3-1403300*, *S3-1439520*, *S3-1462159*, *S3-1495153*, S3×203, S4×192, S5×174 *S6-2848386*, *S6-2866608*, *S6-2899016*, *S6-2913729*, *S6-2941202* *S6-2913729*, *S6-2941202*, *S6-2958750*, *S6-2980225*, *S6-3001176*, *S6-3041790* *S6-3041790*, *S6-3056545*, *S6-3076966*, *S6-3112878*, S6×160, S7×168 *S8-4063097*, *S8-4082527*, *S8-4101244*, *S8-4147562*, *S8-4150777*, *S8-4188989*, S8×162 S9×133, S10×140, S11×165, S12×147
0.3	RaLasso	782 (775)	10 (4)	S1×219, S2×74, S3×64 *S3-1354306*, *S3-1403300*, S4×59, S5×49 S6×52, *S6-2913729*, *S6-2958750*, *S6-2980225*, *S6-3056545* S7×41, S8×52, *S8-4101244*, S9×36, S10×39, S11×48, S12×42

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

Charles-Elie Rabier

ISEM, Université de Montpellier, CNRS, EPHE, IRD, France
IMAG, Université de Montpellier, CNRS, France
LIRMM, Université de Montpellier, CNRS, France
e-mail: ce.rabier@gmail.com

Céline Delmas

INRA, UR875 MIAT, F-313326 Castanet-Tolosan, France
e-mail: celine.delmas@inra.fr

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

Recall that K genetic markers are located at $0 = t_1 < t_2 < \ldots < t_K = T$. Besides, m QTLs lie on $[0, T]$ at locations $t_1^\star$, $t_2^\star$, ..., $t_m^\star$, that are distinct of marker locations. By definition $t_1^\star < t_2^\star < ... < t_m^\star$.

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 $\big(Y, \overline{X}(t_1^{\star\ell}), \overline{X}(t_1^{\star r}), \ldots, \overline{X}(t_m^{\star\ell}), \overline{X}(t_m^{\star r})\big)$.

We have

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

Besides,

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

On the other hand,

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

As a result,

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

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

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

Let $\theta^m=(q_1,...,q_m,\mu,\sigma)$ denote the new parameter. Then, the probability distribution of $\big(Y,\overline{X}(t_1^{\star\ell}),\overline{X}(t_1^{\star r}),\dots,\overline{X}(t_m^{\star\ell}),\overline{X}(t_m^{\star r})\big)$, with respect to the measure $\lambda\otimes N\otimes\dots\otimes N$, is

$$\begin{aligned}
\overline{L}^m_{\vec{t}^\star}(\theta^m)=\sum_{(u_1,\dots,u_m)\in\{-1,1\}^m}&\big[w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\dots+u_mq_m,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]}\\
&+\ v_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\dots+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]}\big]\ \overline{g}^m(t_1^\star,\dots,t_m^\star)
\end{aligned}\tag{1.1}$$

with

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

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

and

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

Note that as soon as we set $m=1$ in formula (1.1), we obtain $\overline{L}_{t_1^\star}(\theta^1)$ given in formula (2) of the main manuscript.

2. Proof of Theorem 2.2

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 $\mathcal{H}_{a\vec{t}^\star}$
- 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, $t_1^\star$, ..., $t_m^\star$ 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

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

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

$$\{2p(t)-1\}\,1_{Y\notin[S_-,S_+]} = \alpha(t)\overline{X}(t^\ell) + \beta(t)\overline{X}(t^r) \tag{2.1}$$

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 :

$$I_{\theta_0} = Diag\left[\mathcal{A}\left\{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)\right\}/\sigma^4\ ,\ \frac{1}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right] \quad . \tag{2.2}$$

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:

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

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

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

Then, since we have the relationship (cf. formula (2.1))

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

the continous mapping theorem implies that

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

It proves the convergence of finite-dimensional.

Note also that we have the relationship

$$\mathrm{Cov}_{H_0}\left\{\overline{S}_n(t^\ell), \overline{S}_n(t^r)\right\} = \rho(t^\ell, t^r)\ .$$

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 $\overline{S}_n(.)$ is a continuous function on $[t^\ell, t^r]$. Recall the modulus of continuity of a continuous function $h(t)$ on $[t^\ell, t^r]$:

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

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 $\overline{S}_n(t^\ell)$ 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 $\mathbb{P}\left(\varpi_{\overline{S}_n}(\delta) \geq \eta\right) \leq \varepsilon \quad \forall n \geq n_0$.

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

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

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

$$\begin{aligned}\varpi_{\overline{S}_n}(\delta) &= \sup_{|t'-t|<\delta} \left|\overline{S}_n(t') - \overline{S}_n(t)\right| \\ &= \sup_{|t'-t|<\delta} \left|\left(\alpha'(t') - \alpha'(t)\right)\overline{S}_n\left(t^\ell\right) \;+\; \left(\beta'(t') - \beta'(t)\right)\overline{S}_n\left(t^r\right)\right| \\ &\leq \max\left(\left|\overline{S}_n\left(t^\ell\right)\right|, \left|\overline{S}_n\left(t^r\right)\right|\right) \quad \left(\varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta)\right). \end{aligned} \tag{2.3}$$

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

$$\forall \varepsilon > 0 \;\; \exists M > 0 \;\; \forall n \geq 1 \;\; \mathbb{P}\left(\max\left(\left|\overline{S}_n(t^\ell)\right|, \left|\overline{S}_n(t^r)\right|\right) \geq M\right) \leq \varepsilon. \tag{2.4}$$

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,

$$\forall \upsilon > 0 \;\; \exists \delta \;\text{ such as }\; t^\ell < \delta < t^r, \;\; \varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta) < \upsilon. \tag{2.5}$$

Let η be a positive quantity. Using formulae (2.4) and (2.5) and imposing $\upsilon = \eta/M$, we have

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

As a consequence, according to formula (2.3), we have

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

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 $\mathcal{H}_{a\vec{t}^\star}$

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

$$Y = \mu \;+\; \sum_{s=1}^{m} X(t_s^\star)\, q_s \;+\; \sigma\varepsilon \tag{2.6}$$

where ε is a Gaussian white noise.

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

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

the mean function will be also a non linear interpolation

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

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

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

$$\overline{S}_n(t_k) = \sum_{j=1}^{n} \frac{(Y_j - \mu)\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}} \tag{2.7}$$

$$= \sum_{j=1}^{n}\sum_{s=1}^{m} \frac{q_s\ \overline{X}_j(t_s^\star)\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\ +\ \sum_{j=1}^{n} \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\ . \tag{2.8}$$

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

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

According to calculations present in Section 4,

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

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

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

Since we have the relationships

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

then we have

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

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

$$\sum_{j=1}^{n}\sum_{s=1}^{m}\frac{q_s\ \overline{X}_j(t_s^\star)\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}} \to \sum_{s=1}^{m}\frac{a_s\ \rho(t_k, t_s^\star)\ \gamma}{\sqrt{\mathcal{A}}}\ . \tag{2.9}$$

Let us now focus on the second term of formula (2.8). According to a technical proof present in Section 4, we have

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

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

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

As a result,

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

Then, according to the Central Limit Theorem,

$$\sum_{j=1}^{n}\frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\xrightarrow{\mathcal{L}}\mathcal{N}\left[\frac{\sum_{s=1}^{m}\rho(t_s^\star,t_k)\ a_s}{\sqrt{\mathcal{A}}}\ \left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})-z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\},\ 1\right]\ . \tag{2.10}$$

Finally, according to formulae (2.9) and (2.10),

$$\overline{S}_n(t_k)\xrightarrow{\mathcal{L}}\mathcal{N}\left[\sum_{s=1}^{m}\rho(t_k,t_s^\star)\ a_s\sqrt{\mathcal{A}}/\sigma^2,\ 1\right]\ . \tag{2.11}$$

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)

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

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 $\mathcal{H}_0$ 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

$$\sup \overline{\Lambda}_n(t) = \sup \overline{S}_n^2(t) + o_P(1) \ . \tag{2.12}$$

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 $t_1^\star$ on $[0, T]$. This way, our goal is now to show that the remainder converges also to zero under $\mathcal{H}_{a\vec{t}^\star}$.

Recall that the parameters θ^m and θ_0^m are defined in the following way : $\theta^m = (q_1, ..., q_m, \mu, \sigma)$ and $\theta_0^m = (0, ..., 0, \mu, \sigma)$.

The likelihood $\overline{L}_{\vec{t}^\star}^{m,n}(\theta^m)$ for n observations is obtained by the product of n terms as in formula (1.1) 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 $\overline{L}_{\vec{t}^\star}^{m,n}(\theta^m)$ (respectively $\overline{L}_{\vec{t}^\star}^{m,n}(\theta_0^m)$).

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

$$\log \frac{dQ_n}{dP_n} = \log \left\{ \frac{\overline{L}_{\vec{t}^\star}^{m,n}(\theta^m)}{\overline{L}_{\vec{t}^\star}^{m,n}(\theta_0^m)} \right\} \ . \tag{2.13}$$

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

$$\log \left(\frac{dQ_n}{dP_n} \right) \overset{\mathcal{H}_0}{\Rightarrow} \mathcal{N}(-\frac{1}{2}\vartheta^2, \ \vartheta^2) \ \text{ with } \vartheta^2 \in \mathbb{R}^{+\star} \ .$$

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

It concludes the proof of Theorem 2.2 for $K = 2$.

2.3. Generalization to $K > 2$

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

Under $\mathcal{H}_0$, 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 $\mathbb{T}_K = \{t_1, \ldots, t_K\}$. Besides, according to Rabier (2015), we have

$$\mathrm{Cov}_{H_0} \left\{ \overline{S}_n(t_k), \overline{S}_n(t_{k'}) \right\} = \rho(t_k, t_{k'}) \ .$$

Under the local aternative $\mathcal{H}_{a\vec{t}^\star}$, we just have to use the fact that the mean function $\overline{m}_{\vec{t}^\star}(t)$ is an interpolated function between $\overline{m}_{\vec{t}^\star}(t^\ell)$ and $\overline{m}_{\vec{t}^\star}(t^r)$. Then, in order to characterize the mean function, we only have to compute the distribution of $\overline{S}_n(t_k)$ at a marker located at t_k. We still have the relationship (as in formula (2.11))

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

since the formulae (2.8), (2.9) and (2.10) 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 $S_n(.)$ is obvious because of the interpolations. Besides, formula (2.12) above is still true for $K > 2$ according to Rabier (2015). In order to proove that the remainder converges also to zero under $\mathcal{H}_{a\vec{t}^\star}$, just use the same kind of proof as above (based on Le Cam's first lemma). Note that the likelihood $\overline{L}_{\vec{t}^\star}^{m,n}(\theta^m)$ for n observations is now obtained by the product of n terms as in formula (1.1) with $K > 2$. Same remark for $\overline{L}_{\vec{t}^\star}^{m,n}(\theta_0^m)$.

3. Proof of Corollary 2.5

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

$$Y = \mu \ + \sum_{s=1}^{m} X(t_s^\star)\, q_s \ + \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^{m} X(t_s^\star) X(t_{\tilde{s}}^\star)\, q_{s,\tilde{s}} \ + \ \sigma\varepsilon \tag{3.1}$$

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

Since the process $\overline{S}_n(.)$ is an interpolated process, we can focus, without loss of generality, only on location t_k (i.e. the location of marker k). According to formulae (3.1) and (2.7), we have

$$\begin{aligned}\overline{S}_n(t_k) = \sum_{j=1}^{n}\sum_{s=1}^{m} \frac{a_s\, \overline{X}_j(t_s^\star)\, \overline{X}_j(t_k)}{n\sqrt{\mathcal{A}}} \ + \sum_{j=1}^{n} \frac{\sigma\varepsilon_j\, \overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}} \\ + \frac{1}{n\sqrt{\mathcal{A}}} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^{m} \overline{X}_j(t_s^\star) \overline{X}_j(t_{\tilde{s}}^\star)\, b_{s,\tilde{s}} \right\} \overline{X}_j(t_k) \ .\end{aligned} \tag{3.2}$$

According to calculations present in Section 4, when $1 \leq s \leq m-1$ and $s+1 \leq \tilde{s} \leq m$,

$$\mathbb{E}\left\{\overline{X}(t_s^\star)\overline{X}(t_{\tilde{s}}^\star)\overline{X}(t_k)\right\} = o(1) \ .$$

Then, according to the Law of Large Numbers,

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

As a result, using formulae (2.9) and (2.10),

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

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*

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

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

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

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

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

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

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

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

Since

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

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

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

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

First, let us recall that by definition we have $t_1^\star < t_2^\star < ... < t_m^\star$. Besides, let us consider a genetic marker located at t_k. We have

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

Using the same kind of proof, we have

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

As a result, we have the relationships

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

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

We have

$$
\begin{aligned}
&\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\right\}\\
&= \mathbb{E}\left\{\sigma\varepsilon 1_{X(t_k)=1} 1_{Y\notin[S_-,S_+]}\right\} \ - \ \mathbb{E}\left\{\sigma\varepsilon 1_{X(t_k)=-1} 1_{Y\notin[S_-,S_+]}\right\}\\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \mathbb{E}\left\{\sigma\varepsilon 1_{X(t_k)=1} 1_{X(t_1^\star)=u_1}\ldots 1_{X(t_m^\star)=u_m} 1_{Y\notin[S_-,S_+]}\right\}\\
&- \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \mathbb{E}\left\{\sigma\varepsilon 1_{X(t_k)=-1} 1_{X(t_1^\star)=u_1}\ldots 1_{X(t_m^\star)=u_m} 1_{Y\notin[S_-,S_+]}\right\}\\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \mathbb{E}\left\{\sigma\varepsilon 1_{Y\notin[S_-,S_+]} \mid X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&\left[2\mathbb{P}\left\{X(t_k)=1 \mid X(t_1^\star)=u_1\ldots X(t_m^\star)=u_m\right\} \ - \ 1\right] \ \mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \left\{\sigma\varphi(z_{\gamma^+}) \ + \ z_{\gamma^+}\ \varphi(z_{\gamma^+}) \sum_{s=1}^m u_s q_s - \sigma\varphi(z_{1-\gamma^-}) \ - \ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-}) \sum_{s=1}^m u_s q_s\right\}\\
&\left[2\mathbb{P}\left\{X(t_k)=1 \mid X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\} \ - \ 1\right] \ \mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&+ o(\max_{1\le s\le m}|q_s|)\ .
\end{aligned}
\tag{4.1}
$$

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

$$
\begin{aligned}
&\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \left\{\sigma\varphi(z_{\gamma^+}) \ - \ \sigma\varphi(z_{1-\gamma^-})\right\} \left[2\mathbb{P}\left\{X(t_k)=1 \mid X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\} \ - \ 1\right]\\
&\times \mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&= \left\{\sigma\varphi(z_{\gamma^+}) \ - \ \sigma\varphi(z_{1-\gamma^-})\right\} \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} 2\ \mathbb{P}\left\{X(t_k)=1, X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&- \left\{\sigma\varphi(z_{\gamma^+}) \ - \ \sigma\varphi(z_{1-\gamma^-})\right\} \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&= \left\{\sigma\varphi(z_{\gamma^+}) \ - \ \sigma\varphi(z_{1-\gamma^-})\right. 2\ \left.\mathbb{P}\left\{X(t_k)=1\right\} \ - \left\{\sigma\varphi(z_{\gamma^+}) \ - \ \sigma\varphi(z_{1-\gamma^-})\right\} = 0\ .
\end{aligned}
\tag{4.2}
$$

Let us focus on the quantity

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

Let ξ denote a given QTL. We have

$$
\begin{aligned}
&\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} u_\xi\ q_\xi\ \left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})\ -\ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\\
&\left[2\mathbb{P}\left\{X(t_k)=1\mid X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\ -\ 1\right]\ \mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&=\sum_{(u_1,\ldots,u_{\xi-1},u_{\xi+1},\ldots u_m)\in\{-1,1\}^{m-1}} q_\xi\ \left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})\ -\ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\\
&\times\left[2\mathbb{P}\left\{X(t_k)=1\mid X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\ -\ 1\right]\\
&\times\mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\\
&-q_\xi\ \left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})\ -\ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\\
&\times\left[2\mathbb{P}\left\{X(t_k)=1\mid X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=-1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\ -\ 1\right]\\
&\times\mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=-1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\\
&=q_\xi\left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})\ -\ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\sum_{(u_1,\ldots,u_{\xi-1},u_{\xi+1},\ldots u_m)\in\{-1,1\}^{m-1}}\\
&\Big[\ 2\mathbb{P}\left\{X(t_k)=1,X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\\
&-\mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\\
&-2\mathbb{P}\left\{X(t_k)=1,X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=-1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\\
&-\mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_{\xi-1}^\star)=u_{\xi-1},X(t_\xi^\star)=-1,X(t_{\xi+1}^\star)=u_{\xi+1},\ldots,X(t_m^\star)=u_m\right\}\Big]\\
&=q_\xi\left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})\ -\ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\\
&\times\left[-\mathbb{P}\left\{X(t_\xi^\star)=1\right\}+\mathbb{P}\left\{X(t_\xi^\star)=-1\right\}+2\mathbb{P}\left\{X(t_k)=1,X(t_\xi^\star)=1\right\}-2\mathbb{P}\left\{X(t_k)=1,X(t_\xi^\star)=-1\right\}\right]\\
&=q_\xi\left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})\ -\ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\left\{\overline{r}(t_k,t_\xi^\star)-r(t_k,t_\xi^\star)\right\}\\
&=q_\xi\left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})\ -\ z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\ \rho(t_k,t_\xi^\star)\ .
\end{aligned}
\tag{4.3}
$$

As a result, according to formulae (4.1), (4.2) and (4.3), we have

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

4.4. Formula for the quantity $\mathbb{E}\left\{\overline{\boldsymbol{X}}(\boldsymbol{t_s^\star})\overline{\boldsymbol{X}}(\boldsymbol{t_{\tilde{s}}^\star})\overline{\boldsymbol{X}}(\boldsymbol{t_k})\right\}$

We have

$$
\begin{aligned}
&\mathbb{E}\left\{\overline{X}(t_s^\star)\overline{X}(t_{\tilde{s}}^\star)\overline{X}(t_k)\right\}\\
&= \mathbb{E}\left\{1_{X(t_s^\star)X(t_{\tilde{s}}^\star)X(t_k)=1}1_{Y\notin[S_-,S_+]}\right\} \;-\; \mathbb{E}\left\{1_{X(t_s^\star)X(t_{\tilde{s}}^\star)X(t_k)=-1}1_{Y\notin[S_-,S_+]}\right\}\\
&= \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{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}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\;\mathbb{P}\left\{X(t_k)=-1 \mid X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\\
&+ \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{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}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\;\mathbb{P}\left\{X(t_k)=1 \mid X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\\
&- \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{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}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\;\mathbb{P}\left\{X(t_k)=1 \mid X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\\
&- \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{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}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\;\mathbb{P}\left\{X(t_k)=-1 \mid X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\\
&= -2 \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{s}}=-u_s}} \left\{\Phi\left(\frac{S_- -\mu}{\sigma}\right) \;+\; 1 \;-\Phi\left(\frac{S_+ -\mu}{\sigma}\right)\right\}\mathbb{P}\left\{X(t_k)=1,\; X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\\
&+ 2 \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{s}}=u_s}} \left\{\Phi\left(\frac{S_- -\mu}{\sigma}\right) \;+\; 1 \;-\Phi\left(\frac{S_+ -\mu}{\sigma}\right)\right\}\mathbb{P}\left\{X(t_k)=1,\; X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\\
&+ \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{s}}=-u_s}} \left\{\Phi\left(\frac{S_- -\mu}{\sigma}\right) \;+\; 1 \;-\Phi\left(\frac{S_+ -\mu}{\sigma}\right)\right\}\times\mathbb{P}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}\\
&- \sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m\\ u_{\tilde{s}}=u_s}} \left\{\Phi\left(\frac{S_- -\mu}{\sigma}\right) \;+\; 1 \;-\Phi\left(\frac{S_+ -\mu}{\sigma}\right)\right\}\times\mathbb{P}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\} \;+o(1)\;.
\end{aligned}
\tag{4.4}
$$

Besides,

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

and

$$\sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m \\ u_{\tilde{s}}=u_s}} \mathbb{P}\left\{X(t_k)=1,\ X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}$$
$$=\mathbb{P}\left\{X(t_k)=1, X(t_s^\star)X(t_{\tilde{s}}^\star)=1\right\}=\mathbb{P}\left\{X(t_s^\star)X(t_{\tilde{s}}^\star)=1 \mid X(t_k)=1\right\}/2\ .$$

As a result,

$$2\sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m \\ u_{\tilde{s}}=u_s}} \mathbb{P}\left\{X(t_k)=1,\ X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}$$
$$-2\sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m \\ u_{\tilde{s}}=-u_s}} \mathbb{P}\left\{X(t_k)=1,\ X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}$$
$$=2\mathbb{P}\left\{X(t_s^\star)X(t_{\tilde{s}}^\star)=1 \mid X(t_k)=1\right\}-1=2\mathbb{P}\left\{X(t_s^\star)X(t_{\tilde{s}}^\star)=1\right\}-1=\rho(t_s^\star,t_{\tilde{s}}^\star)\ .$$

In the same way,

$$\sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m \\ u_{\tilde{s}}=-u_s}} \mathbb{P}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}$$
$$-\sum_{\substack{(u_1,\dots,u_m)\in\{-1,1\}^m \\ u_{\tilde{s}}=u_s}} \mathbb{P}\left\{X(t_1^\star)=u_1,\dots,X(t_m^\star)=u_m\right\}$$
$$=\mathbb{P}\left\{X(t_s^\star)X(t_{\tilde{s}}^\star)=-1\right\}-\mathbb{P}\left\{X(t_s^\star)X(t_{\tilde{s}}^\star)=1\right\}=-\rho(t_s^\star,t_{\tilde{s}}^\star)\ .$$

Then, according to formula (4.4), we have

$$\mathbb{E}\left\{\overline{X}(t_s^\star)\overline{X}(t_{\tilde{s}}^\star)\overline{X}(t_k)\right\}=o(1)\ .$$

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. 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. (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.

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

Charles-Elie Rabier

ISEM, Université de Montpellier, CNRS, EPHE, IRD, France
IMAG, Université de Montpellier, CNRS, France
LIRMM, Université de Montpellier, CNRS, France
e-mail: ce.rabier@gmail.com

Céline Delmas

INRA, UR875 MIAT, F-313326 Castanet-Tolosan, France
e-mail: celine.delmas@inra.fr

TABLE 1

Performances of the new method SgLasso as a function of the ratio γ_+/γ (Mean over 100 samples, on average $n = 100$ individuals genotyped, $m = 16$, $|q_1| = \ldots = |q_{16}| = 0.1$, $T = 10$, 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.

		(Sparse, $n = 100$)		(Sparse, $n = 200$)		(Dense, $n = 100$)		(Dense, $n = 200$)	
γ	γ_+/γ	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

FIG 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* $+1$.