On statistical inference for selective genotyping

Charles-Elie Rabier

sankhya manuscript No.
(will be inserted by the editor)

On statistical inference for selective genotyping

Charles-Elie Rabier

Received: date / Accepted: date

Abstract In Quantitative Trait Locus detection, selective genotyping is a way to reduce costs due to genotyping : only individuals with extreme phenotypes are genotyped. We focus here on statistical inference for selective genotyping. We study, in a very large framework, the performances of different tests suitable for selective genotyping. We proof that we have to genotype symetrically, that is to say the same percentage of large and small phenotypes whatever the proportions of the two genotypes in the population. Besides, we proof that the non extreme phenotypes (ie. the phenotypes for which genotypes are missing) don't bring any information for statistical inference. Same results are obtained in the case of a selective genotyping with two phenotypes correlated.

Keywords Hypothesis testing · Asymptotic properties of tests · Asymptotic Relative Efficiency · Selective genotyping · Quantitative Trait Locus detection.

PACS 62F03 · 62F05 · 62F12 · 62P10

1 Introduction

1.1 Introducing our study

We address the problem of detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured). In this study, we will focus only on a single locus on the genome, called genetic marker (see Lynch and Walsh (1997), Weller (2001), Wu et al. (2007)). X refers to a genetic marker with two possible genotypes : $+1$ with probability p and -1 with probability $1-p$. Y denotes the Quantitative Trait (ie. phenotype). Y and X are linked by the following relationship : $Y = \mu + qX + \varepsilon$ where ε is a Gaussian noise with mean 0 and variance σ^2. We will say that there is a QTL if the

Charles-Elie Rabier
Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., Toulouse, France
INRA UR631, Station d'Amélioration Génétique des Animaux, Auzeville, France
Tel.: +(608)-265-8765, Fax: +(608)-262-0032
E-mail: rabier@stat.wisc.edu
Present address: University of Wisconsin-Madison, Statistic Department, Medical Science Center, 1300 University Avenue, Madison, WI 53706-1532, USA

QTL effect q is different from zero. Indeed, in such a situation, the individuals for which $X = +1$ will tend to have a larger Y than the individuals for which $X = -1$.

The problem is that genotyping (ie. having the marker information X) is very expensive. In such a context, Lebowitz and al. (1987) had a very good idea based on the observation that most of the information about the QTL is present in the extreme phenotypes (ie. the smallest and the largest Y). So, they propose to genotype only the individuals who present an extreme phenotype. This way, at a given power, a large increase of the number of individuals leads to a decrease of the number of individuals genotyped. Later, Lander and Botstein (1989), formalized this approach and called it "selective genotyping".
More recently, Muranty and Goffinet (1997) focused on the estimation of the QTL effect for selective genotyping. However, although there have been many papers on selective genotyping, the theory of statistical inference for selective genotyping is still missing. In a very famous article, Darvasi and Soller (1992) proposed to perform a comparison of means between the extreme individuals (ie with extreme phenotypes) for which $X = +1$ at the marker and those for which $X = -1$. It is such a nice idea since it is very intuitive. However, some errors are present in this paper. In this context, the aim of this article is to study statistical inference for selective genotyping in a mathematical point of view. Our study justifies some practice of geneticists and gives new ways of analysing data. Selective genotyping has been motivated by agronomy but there are many areas where the data analysis is crucial but under economic pressures (aeronautics for instance). That's why, we study selective genotyping here in a large framework (in genetics, we mainly consider $p = 1/2$ which corresponds to the backcross). Besides, we present a study as a function of the unknown parameters μ, q, σ. Obviously, the most interesting situation is when all these parameters are unknown, like in real life. However, in some articles on selective genotyping (for instance Darvasi and Soller (1992)), people consider that without loss of generality, the global mean μ and the variance σ^2 are known. In fact, is there a loss of generality ?

We will also focus on selective genotyping in presence of two correlated phenotypes Y and Z, and when it is difficult to measure Z for some biological reasons (see Medugorac and Soller (2001)). In such a context, the costs due to genotyping and due to phenotyping can be reduced : a selective genotyping is performed on Y, and Z is measured only on the genotyped individuals (ie. with extreme phenotypes Y). Obviously, in such a situation, the interest is on finding a QTL which has an effect on Z. We will answer same kinds of questions as for a selective genotyping with only one phenotype. Finally, we will establish the link between selective genotyping with one and two phenotypes.

1.2 Roadmap and main results

Our study begins with only one phenotype Y (Sections 2 and 3). In Section 2, we consider the classical situation where no genotypes are missing. We call it "oracle situation" since we know all the genotypes. We propose a simple test ("oracle test") which is optimal and which will be considered as the test of reference. In Section 3, starts our study of selective genotyping. We study different strategies for the data analysis. The different tests (corresponding to the different strategies) are compared in terms of Asymptotic Relative Efficiciency (ARE), which determines for each test, the sample size required to obtain same local asymptotic power as the oracle test. Theorem 1, which gives the different ARE for the different tests, is the main result of the first part dealing with one phenotype. It says that we have the same ARE if we keep or not the non extreme phenotypes Y (ie the phenotypes for which the genotype is missing) in the data analysis. We have to keep in mind that these non extreme

phenotypes are available when we collect data in selective genotyping. Lemma 1 is a direct consequence of Theorem 1. An easy and optimal test is presented. It is based on the comparison of means of the extreme phenotypes. An other very important result of Theorem 1 is that, if we want to genotype only a percentage γ of the population, we have to genotype symetrically, that is to say the $\gamma/2\%$ individuals with the largest phenotypes and the $\gamma/2\%$ individuals with the smallest phenotypes. This result holds whatever the proportion p (ie the probability that $X = +1$). When $p = 1/2$, this result was expected : it confirms by the theory what geneticists do in practice. However, when $p \neq 1/2$, this result is original : we didn't know how to analyze such data.

Sections 4 and 5 are related to the second part : we deal now with two correlated phenotypes Y and Z. Same kind of analysis is given as in the first part which deals with one phenotype. Theorem 2 and Lemma 2 are the main results. Theorem 2 says that we still have to genotype symetrically and that the non extreme phenoypes Y still don't bring any information for statistical inference. Theorem 2 also establishes the relationship between the ARE of a selective genotyping with two phenotypes and a selective genotyping with one phenotype. On the other hand, Lemma 2 presents optimal tests.

Section 6 is an illustration of the theoretical results of this paper : we check the asymptotic validity of our tests. Note that this paper deals with Le Cam (1986)'s work on contiguity. We refer to the book of Van der Vaart (1998) for elements of asymptotic statistics used in proofs. We join "Online Ressource 1" which contains some proofs not needed at first reading of this paper.

2 Oracle situation : all the genotypes are known (ie no selective genotyping)

To begin, we consider the situation with no missing genotypes : the oracle situation. The study of such a situation will be interesting in order to quantify the lost of information due to missing genotypes. We present here a simple test (oracle test), which is optimal and which will be considered as our reference test for our future study on selective genotyping.

2.1 Model

X denotes the random variable (r.v.) which corresponds to the genotype at the QTL. We consider 2 genotypes at the QTL :

$$X = \begin{cases} -1 & \text{with probability } 1-p \\ 1 & \text{with probability } p. \end{cases}$$

We suppose $p \neq \{0, 1\}$. Y is the r.v. refering to the phenotype :

$$Y = \mu + qX + \varepsilon$$

where ε is a Gaussian r.v. centered with variance σ^2. q is the QTL effect. We consider a sample of n observations (X_j, Y_j) independent and equally distributed.

2.2 Oracle statistical test (μ, q, σ)

We consider a statistical model with 3 unknown parameters (μ, q, σ). In order to test the presence of a QTL, we consider the two following hypotheses :

$$H_0 : \ q = 0 \ \ vs \ \ H_1 : \ q \neq 0.$$

We will consider in particular, a local alternative $H_a : q = \frac{a}{\sqrt{n}}$ where a is a constant different from zero.

In this context, an easy test to perform is based on the test statistic

$$T = \sqrt{p\,(1-p)} \left\{ \frac{\sum_{j=1}^{n} \ \frac{1}{p}(Y_j - \overline{Y})\ 1_{X_j=1} \ - \ \frac{1}{1-p}(Y_j - \overline{Y})\ 1_{X_j=-1}}{\hat{\sigma}\ \sqrt{n}} \right\}$$

where $\hat{\sigma} = \frac{1}{\sqrt{n}} \left\{ \sum_{j=1}^{n} (Y_j - \overline{Y})^2 \right\}^{1/2}$ and $\overline{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j$.
The asymptotic laws are :

$$T \overset{H_0}{\to} N\,(0,\ 1) \quad \text{and} \quad T \overset{H_a}{\to} N\left(\frac{2a\ \sqrt{p\,(1-p)}}{\sigma},\ 1 \right).$$

This test, which is almost a comparison of means between the two genotypes at the QTL, is the most powerful test we can perform : it has the same asymptotic properties as the Wald test. A proof is given in Section 7. Note that in this paper, we will use the terminology "comparison of means" even if our tests are only almost "comparison of means".

3 Selective genotyping

3.1 Motivation

In this section, we mainly want to give answers to the following questions for selective genotyping :

- What is the loss of information due to missing genotypes in a general framework ?
- Do the non extreme phenotypes (ie for which the genotype is missing) bring any extra information for statistical inference ?
- Is it possible to propose an easy and optimal test for selective genotying ?
- If we want to genotype only a percentage γ of the individuals, how should we genotype ? Should we genotype only the $\gamma\%$ individuals with the largest phenotypes? Or the $\gamma\%$ with the smallest phenotypes? Or some individuals with the largest phenotypes and some with the smallest phenotypes ?
- Do we have the same results when the number of unknown parameters varies ?

3.2 Model and strategies

We consider two real thresholds (constant) S_- and S_+ such as $S_- \leqslant S_+$. We consider that the genotype X is known if and only if the phenotype Y is extreme, ie. if and only if $Y \leqslant S_-$ or $Y \geqslant S_+$.

In order to make the reading easier, we define a new r.v. $\overline{X}$ such as :

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

In other words, $\overline{X} = 0$ refers to the case where the genotype is missing.

As in the oracle situation, we want to test the presence of a QTL ($q = 0$ vs $q \neq 0$) and we deal with a local alternative $H_a : q = \frac{a}{\sqrt{n}}$. We consider here 3 different strategies suitable for the data analysis in selective genotyping :

- 1. we keep all the phenotypes (even the phenotypes which are non extremes, ie the phenotypes for which the genotype is missing) and we perform a Wald test
- 2. we keep only the extreme phenotypes (ie. the phenotypes for which the genotype is available) and we perform a comparison of means between the two genotypes at the QTL
- 3. we keep only the extreme phenotypes (ie. the phenotypes for which the genotype is available) and we perform a Wald test

Each test corresponding to each strategy will be compared to the oracle test in terms of ARE, which determines for each test, the sample size required to obtain same local asymptotic power as the oracle test. The study of such strategies will help us to give answers to our questions of Section 3.1. Note that strategy 2 (inspired by Darvasi and Soller (1992)) is the easiest to compute.

3.3 Results

To begin, we present our main theorem :

Theorem 1 *Let κ_1, κ_2 and κ_3 be the efficiencies corresponding respectively to strategies one, two and three. Let γ, γ_+ and γ_- be respectively the following quantities $\mathbb{P}_{H_0}(Y \notin [S_-, \ S_+])$, $\mathbb{P}_{H_0}(Y > S_+)$ and $\mathbb{P}_{H_0}(Y < S_-)$. Then, if we consider a statistical model with 3 unknown parameters (μ, q, σ), $\forall p \in]0,1[$:*

i) $\kappa_1 = \kappa_2 = \kappa_3 = \gamma \ + \ z_{\gamma_+}\varphi(z_{\gamma_+}) \ - \ z_{1-\gamma_-} \ \varphi(z_{1-\gamma_-})$

ii) *κ_1, κ_2 and κ_3 reach their maximum, M, when $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$, with*

$$M = \gamma \ + \ 2 \ z_{\gamma/2} \ \varphi(z_{\gamma/2})$$

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

The proof is given in Section 8.

Before interpreting this theorem, we have to give some precisions on the quantities γ, γ_+ γ_-. According to the law of large numbers, under the null hypothesis H_0 and under the local alternative H_a, $\frac{1}{n}\sum 1_{\overline{X}_j \neq 0} \to \gamma$. So, γ corresponds asymptotically to the percentage of individuals genotyped. In the same way, γ_+ (resp. γ_-) corresponds asymptotically to the percentage of individuals genotyped with the largest (resp. the smallest) phenotypes.

Let's explain now Theorem 1. According to i), the three strategies have exactly the same ARE. We can deduce of it two consequences. First, since $\kappa_1 = \kappa_3$, the non extreme phenotypes don't bring any extra information for statistical inference. Secondly, since $\kappa_2 =$

κ_3, there is no loss of power between a comparison of means and the Wald test based on the extreme phenotypes. In other words, we should perform the comparison of means : it is an easy and optimal test. However, we will see in Lemma 1, that a little adjustment has to be done in order to make this test easy. On the other hand, i) presents the ARE in a general framework. We can see that the ARE is independent of p (ie the probability that $X = +1$) and a (ie. the constant linked to the QTL effect). It only depends on γ, γ_+ and γ_-.
ii) of Theorem 1 says that the ARE is maximum for $\gamma_+ = \gamma_- = \gamma/2$. That is to say, 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. It is true for any p. When $p = 1/2$, this result was expected : it confirms by the theory what geneticists do in practice. However, when $p \neq 1/2$, this result is original : we didn't know how to analyse such data.

We introduce now Lemma 1, which presents explicitly, contrary to Theorem 1, the different tests corresponding to the different strategies.

Lemma 1 *If we consider a statistical model with* 3 *unknown parameters* (μ, q, σ)*, the Wald test statistic* W_1*, the test statistic of comparison of means* T_2*, and the Wald test statistic* W_3*, which correspond respectively to strategies one, two and three :*

$$W_1 := \frac{2\sqrt{n}}{\hat{\sigma}^2}\ \sqrt{\hat{\mathcal{A}}_1\ p(1-p)}\ \ \hat{q}_1$$

$$T_2 := \sqrt{p(1-p)}\ \left\{ \frac{\sum_{j=1}^n \frac{1}{p}(Y_j - \hat{\mu}_3)1_{\overline{X}_j=1}\ -\ \frac{1}{1-p}(Y_j - \hat{\mu}_3)1_{\overline{X}_j=-1}}{\sqrt{n\ \hat{\mathcal{A}}_3}} \right\}$$

$$W_3 := \frac{2\sqrt{n}}{\hat{\sigma}_3^2}\ \sqrt{\hat{\mathcal{A}}_3\ p(1-p)}\ \ \hat{q}_3$$

have the same asymptotic laws under H_0 *and under* H_a*, that is to say :*

$$N(0,\ 1) \quad \textit{and} \quad N\left(\frac{2a\ \sqrt{\mathcal{A}\ p(1-p)}}{\sigma^2},\ 1 \right)$$

where $\hat{q}_1$ *and* $\hat{q}_3$ *denote the MLE respective of* q *for strategies one and three,* $\hat{\mu}_3$ *and* $\hat{\sigma}_3^2$ *the MLE repective of* μ *and* σ^2 *for strategy three,*
$\mathcal{A} = \sigma^2\ \{\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\}$, $\hat{\mathcal{A}}_1 = \frac{1}{n}\sum_{j=1}^n (Y_j - \overline{Y})^2 1_{\overline{X}_j \neq 0}$
$\hat{\mathcal{A}}_3 = \frac{1}{n}\sum_{j=1}^n (Y_j - \hat{\mu}_3)^2 1_{\overline{X}_j \neq 0}$, $\hat{\sigma}^2$ *is given in Section 2.2.*

For the proof, we refer to the proof of Theorem 1 in Section 8. Note that the estimators $\hat{\sigma}^2$ and $\hat{\sigma}_3^2$ are also consistent under H_a by contiguity. Same remark for $\hat{\mathcal{A}}_1$ and $\hat{\mathcal{A}}_3$, which are estimators of $\mathcal{A}$.

As said previously, we want to propose an easy and optimal test. In order to compute the MLE $\hat{q}_1$ and $\hat{q}_3$, we need to use respectively an EM algorithm and a Newton method (cf. Rabier (2010)). As a consequence, the tests corresponding to strategies one and three are difficult to perform. According to Lemma 1, the test based on T_2, ie the comparison of mean between the two genotypes at the QTL, is not so easy to perform. Indeed, we have to compute the estimator $\hat{\mu}_3$ which is not straightforward. However, instead of using $\hat{\mu}_3$, we can use the empirical mean $\overline{Y}$, because this estimator is $\sqrt{n}$ consistent. In the same way, we can also replace $\hat{\mathcal{A}}_3$ by $\hat{\mathcal{A}}_1$. This way, the test is very easy to compute :

$$T_2 = \sqrt{p(1-p)n}\ \left\{ \frac{\sum_{j=1}^n \frac{1}{p}(Y_j - \overline{Y})1_{\overline{X}_j=1}\ -\ \frac{1}{1-p}(Y_j - \overline{Y})1_{\overline{X}_j=-1}}{\sqrt{\sum_{j=1}^n (Y_j - \overline{Y})^2 1_{\overline{X}_j \neq 0}}} \right\}.$$

The asymptotic laws are unchanged. Note that we use now the non extreme phenotypes in this expression of T_2 (contrary to the definition of strategy 2). Besides, we can see that this test statistic is a generalization of our oracle test statistic introduced in Section 2.2. To conclude, when we analyse data, we should use this test and genotype symetrically.

Until now, we have focused on the most interesting configuration : all the parameters (ie. μ, q, σ) were unknown. Let's focus now on statistical models with respectively one unknown parameter (q) and two unknown parameters (μ, q). The idea is to see if we obtain the same results as previously. We will consider the same strategies as previously. For strategy 2, when just q is unknown, we have to keep in mind that $\mathcal{A}$ is known. Indeed, according to the proof of Theorem 1 (see Section 8.2.2), we have $\mathcal{A} = \mathbb{E}_{H_0}\left\{(Y-\mu)^2 1_{Y \notin [S_-,\ S_+]}\right\}$. As a consequence, we will consider the test statistic T_2 of Lemma 1 except that we replace $\hat{\mu}_3$ by μ and $\hat{\mathcal{A}}_3$ by $\mathcal{A}$. Note that when we consider (μ, q) unknown, we will use same test statistic T_2 as in Lemma 1. Besides, in order to calculate the different ARE for the different strategies, we will obviously consider the appropriate oracle test (ie. the oracle test with only q unknown, and the one with (μ, q) unknown).

Corollary 1 *If we consider a statistical model with one unknown parameter (q), then (with the previous notations) :*

i) $\kappa_1 = \gamma \ + \ z_{\gamma_+}\varphi(z_{\gamma_+}) \ - \ z_{1-\gamma_-} \ \varphi(z_{1-\gamma_-}) \ + \ (2p-1)^2 \ \left\{1-\gamma - z_{\gamma_+}\varphi(z_{\gamma_+}) + z_{1-\gamma_-} \ \varphi(z_{1-\gamma_-})\right\}$

ii) $\kappa_2 = 4 \ p \ (1-p) \ \left\{\gamma \ - \ z_{1-\gamma_-} \ \varphi(z_{1-\gamma_-}) \ + \ z_{\gamma_+} \ \varphi(z_{\gamma_+})\right\}$

iii) $\kappa_3 = \gamma \ + \ z_{\gamma_+}\varphi(z_{\gamma_+}) \ - \ z_{1-\gamma_-} \ \varphi(z_{1-\gamma_-}) \ + \ \dfrac{(2p-1)^2}{1 \ -\gamma} \left\{\varphi(z_{1-\gamma_-}) \ - \ \varphi(z_{\gamma_+})\right\}^2 \quad \forall \gamma \neq 1$

iv) $\kappa_1 = \kappa_2 = \kappa_3 \Leftrightarrow p = \dfrac{1}{2}$

v) $\forall p \in]0,1[\quad \kappa_1$, κ_2 *and* κ_3 *reach their maximum for* $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$.

Corollary 2 *If we consider a statistical model with two unknown parameters (μ, q), then the results are the same as in Theorem 1.*

The proof of Corollary 1 is given in Section 2 of "Online Ressource 1". The proof of Corollary 2 is obvious according to the proof of Theorem 1.

According to Corollary 2, when only the variance σ^2 is known, we have same results as previously. So, there is no loss of generality to consider the variance known. However, according to Corollary 1, there is a loss of generality to consider the mean μ known. Indeed, when we consider only q unknown, the three strategies have same ARE if and only if $p = 1/2$ (ie. backcross in genetics). In other words, when $p \neq 1/2$, the non extreme phenotypes Y bring some extra information for statistical inference. So, in this case, we have to use strategy 1. Note that we still have to genotype symetrically for all strategies.

3.4 Remark on the work of Darvasi and Soller (1992)

In our study, in order to model selective genotyping, two real thresholds (constant) S_- and S_+ have been considered. An individual is genotyped if and only if $Y \notin [S_-, S_+]$ (ie. $\overline{X} \neq 0$). As said previously, under H_0 and H_a, $\frac{1}{n}\sum 1_{\overline{X}_j \neq 0} \to \gamma$ where $\gamma = \mathbb{P}_{H_0}(Y \notin [S_-, S_+])$. This way, our modelization agrees with the usual definition of selective genotyping : selective genotyping consists in genotyping only the $\gamma\%$ individuals with extreme phenotypes.

In Darvasi and Soller (1992), the authors focus on a comparison of means, between the extreme individuals, only when $p = 1/2$. They consider μ and σ known without loss of generality (which is true according to our study since $p = 1/2$). Besides, the main difference with our approach, is that they consider thresholds which vary with the QTL effect. Indeed, they consider $\gamma = \mathbb{P}(Y \notin [S_-, S_+])$. The problem is that since the QTL effect is such as $q = a/\sqrt{n}$, S_- and S_+ depend on n. As a consequence, the authors make an error when they use classical central limit theorem : they should use Lindeberg-Feller central limit theorem. Furthermore, they use approximations about thresholds (see their formulae (1) and (2)), and results about sample sizes (see their formula (24)), which are not suitable for models with local alternatives.

Note that in their paper, Darvasi and Soller (1992) suppose symetry, that is to say $\mathbb{P}(Y > S_+) = \mathbb{P}(Y < S_-) = \gamma/2$. Anyway, if we consider the same configuration as Darvasi and Soller (1992) (ie $p = 1/2$ and symetry), our study gives the same ARE as presented in formula (27) of Darvasi and Soller (1992). However, we have to keep in mind that our comparison of means based on the test statistic T_2 is totally new and was not present in

Darvasi and Soller (1992). Indeed, we consider $p \in]0,1[$, not only symetry, and μ and σ unknown.

4 Introducing a second phenotype

We don't observe only one phenotype Y anymore, but two correlated phenotypes, Y and Z. The aim is to detect a QTL which has an effect on Z. As previously, we begin by considering the situation with no missing genotypes. We present here our optimal oracle test, which will be considered as our reference test for our future study on selective genotyping.

4.1 Model

X is still the r.v. corresponding to the genotype at the QTL. We consider the following model :

$$\begin{pmatrix} \mathrm{Y} \\ \mathrm{Z} \end{pmatrix} = \begin{pmatrix} \mu_Y + q_Y X \\ \mu_Z + q_Z X \end{pmatrix} + \varepsilon$$

where

$$\varepsilon \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & r\,\sigma^2 \\ r\,\sigma^2 & \sigma^2 \end{pmatrix} \right).$$

We suppose $r \in]-1,1[$. Besides, we consider that r and σ^2 are known. μ_{YX} and μ_{ZX} will be the following quantities : $\mu_{YX} = \mu_Y + q_Y X$ and $\mu_{ZX} = \mu_Z + q_Z X$. We consider a sample of n observations (X_j, Y_j, Z_j) independent and equally distributed. Note that q_Z and q_Y are respectively the QTL effects on phenotypes Z and Y.

4.2 Oracle statistical test (μ_Z, q_Z)

In order to test the presence of a QTL with effect on the phenotype Z, we consider the two following hypotheses :

$$H_{0Z} : \; q_Z = 0 \;\; vs \;\; H_{1Z} : \; q_Z \neq 0.$$

We will consider in particular, a local alternative $H_{bZ} : q_Z = \frac{b}{\sqrt{n}}$ where b is a constant different from zero.

According to what has been done with only one phenotype (cf Sections 2.2 and 7), an easy and optimal test to perform is based on the following statistic

$$T = \frac{\sum_{j=1}^{n} \frac{1}{p}(Z_j - \overline{Z}) 1_{X_j=1} \; - \; \frac{1}{1-p}(Z_j - \overline{Z}) 1_{X_j=-1}}{\sigma \sqrt{\frac{n}{p(1-p)}}}.$$

The asymptotic laws are :

$$T \overset{H_{0Z}}{\rightarrow} N(0,\, 1) \qquad T \overset{H_{bZ}}{\rightarrow} N\left(\frac{2\, b\, \sqrt{p(1-p)}}{\sigma},\, 1 \right)$$

where $\overline{Z} = \frac{1}{n} \sum_{j=1}^{n} Z_j$.

5 Selective genotyping with two correlated phenotypes

When it is difficult to measure Z for some biological reasons (see Medugorac and Soller (2001)), the costs due to genotyping and due to phenotyping can be reduced : a selective genotyping is performed on Y, and Z is measured only on the genotyped individuals (ie. with extreme phenotypes Y). In such a situation, the interest is on finding a QTL which has an effect on Z. Obviously, Y and Z has to be correlated otherwise this selective genotyping has no sense.

5.1 Motivation

We will try to answer same kinds of questions as for a selective genotyping with only one phenotype :

- What is the loss of information due to missing genotypes in a general framework ?
- Do the non extreme phenotypes Y (ie for which the genotype is missing) bring any extra information for statistical inference on the QTL effect q_Z ?
- If we want to genotype only a percentage γ of the individuals, how should we genotype ?
- Do we have the same results when the number of unknown parameters varies ?

5.2 Model and strategies

We consider the same model as previously (see Section 3.2). As in the oracle situation, we want to test the presence of a QTL which affects Z ($q_Z = 0$ vs $q_Z \neq 0$) and we deal with a local alternative $H_{bZ} : q_Z = \frac{b}{\sqrt{n}}$.
Since Z and Y are correlated, we will have to deal with hypotheses on q_Y. So, the new notations will be, H_{0Y} for $q_Y = 0$, and H_{aY} for $q_Y = \frac{a}{\sqrt{n}}$.
We consider here 2 strategies suitable for the data analysis :

- 1. we keep all the phenotypes Y (even the phenotypes which are non extremes, ie the phenotypes for which the genotype is missing) and we perform a Wald test.
- 2. we keep only the extreme phenotypes Y (ie. the phenotypes for which the genotype is available) and we perform a Wald test.

Each test corresponding to each strategy will be compared to the oracle test in terms of ARE, which determines for each strategy, the sample size required to obtain same local asymptotic power as the oracle test. The study of such strategies will help us to give answers to our questions of Section 5.1. Note that we don't consider the comparison of means on Z : it is obvious that this test won't be optimal. As a consequence, here, strategy 2 is analoguous to strategy 3 of the first part.

5.3 Results

To begin, we present our main theorem, Theorem 2, which is the analogous of Corollary 2 for two phenotypes (the covariance matrix is known here). However, since Corollary 2 and Theorem 1 give same results, Theorem 2 can be also viewed as the analogous of Theorem 1.

Theorem 2 *Let $\tilde{\kappa}_1$ and $\tilde{\kappa}_2$ be the efficiencies corresponding to strategies one and two. Let γ, γ_+ and γ_- be respectively the following quantities $\mathbb{P}_{H_{0Y}}\left(Y \notin [S_-,\ S_+]\right)$, $\mathbb{P}_{H_{0Y}}\left(Y > S_+\right)$ and $\mathbb{P}_{H_{0Y}}\left(Y < S_-\right)$. Then, if we consider a statistical model with 4 unknown parameters (μ_Z, q_Z, μ_Y, q_Y), we have under H_{0Y} and under H_{aY} , $\forall p \in]0,1[$:*

i) $\tilde{\kappa}_1 = \tilde{\kappa}_2 = \left\{ \dfrac{1-r^2}{\gamma} + \dfrac{r^2}{\kappa_1} \right\}^{-1}$

ii) *$\tilde{\kappa}_1$ and $\tilde{\kappa}_2$ reach their maximum, $\tilde{M}$, for $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$, with*

$$\tilde{M} = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{M} \right\}^{-1}$$

where κ_1 and M are the quantities of Theorem 1.

The proof is given in Section 9. As expected, the ARE increase with r and γ. As previously, the non extreme phenotypes Y (ie. for which the genotype is missing) don't bring any extra information for statistical inference on q_Z. Besides, we still have to genotype symetrically for a selective genotyping with two phenotypes. Note that Theorem 2 establishes the relationship between the ARE of selective genotyping with one and two phenotypes. Lemma 2 presents the different tests corresponding to the different strategies.

Lemma 2 *If we consider a statistical model with 4 unknown parameters (μ_Z, q_Z, μ_Y, q_Y) and that we are under H_{0Y} or H_{aY}, then the Wald test statistic $\tilde{W}_1$ and the Wald test statistic $\tilde{W}_2$, which correspond respectively to strategy one and two :*

$$\tilde{W}_1 := \sqrt{n}\ \hat{q}_Z^1\ \left\{ \frac{\sigma^2\ (1-r^2)}{4\ p\ (1-p)\ \gamma} \ +\ \frac{\sigma^4\ r^2}{4\ p\ (1-p)\ \hat{\mathcal{A}}_1} \right\}^{-1/2}$$

$$\tilde{W}_2 := \sqrt{n}\ \hat{q}_Z^2\ \left\{ \frac{\sigma^2\ (1-r^2)}{4\ p\ (1-p)\ \gamma} \ +\ \frac{\sigma^4\ r^2}{4\ p\ (1-p)\ \hat{\mathcal{A}}_3} \right\}^{-1/2}$$

have the same asymptotic laws under H_{0Z} and H_{bZ}, that is to say :

$$N(0,\ 1) \quad \textit{and} \quad N\left(b \left\{ \frac{\sigma^2\ (1-r^2)}{4\ p\ (1-p)\ \gamma} \ +\ \frac{\sigma^4\ r^2}{4\ p\ (1-p)\ \mathcal{A}} \right\}^{-1/2},\ 1 \right)$$

with $\hat{q}_Z^i$ MLE of q_Z for strategy i. $\mathcal{A}$, $\hat{\mathcal{A}}_1$ and $\hat{\mathcal{A}}_3$ are given in Lemma 1.

For the proof and also how to compute the MLE $\hat{q}_Z^i$, we refer to the proof of Theorem 2 in Section 9. We introduce now Corollary 3 which is the analogous of Corollary 1. Only q_Z and q_Y are now unknown.

Corollary 3 *If we consider a statistical model with two unknown parameters (q_Z, q_Y), then under H_{0Y} and under H_{aY}:*

i) $\tilde{\kappa_1} = \left\{ \dfrac{1-r^2}{\gamma} + \dfrac{r^2}{\kappa_1} \right\}^{-1}$

ii) $\tilde{\kappa}_2 = \left\{ \dfrac{1-r^2}{\gamma} + \dfrac{r^2}{\kappa_3} \right\}^{-1}$

iii) $\tilde{\kappa}_1 = \tilde{\kappa}_2 \Leftrightarrow p = \dfrac{1}{2}$

iv) $\forall p \in]0,1[$ *$\tilde{\kappa}_1$ and $\tilde{\kappa}_2$ reach their maximum for $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$*

where κ_1 and κ_3 are the quantities of Corollary 1.

The proof is given in Section 3 of "Online Ressource 1". According to this Corollary, the two strategies have same ARE if and only if $p = 1/2$. When $p \neq 1/2$, the non extreme phenotypes Y bring some extra information for statistical inference on q_Z. As a consequence, there is a loss of generality to consider the parameters μ_Y and μ_Z known. However, we still have to genotype symetrically. Note that Corollary 3 establishes a link with the ARE of Corollary 1.

To conclude, in the following Corollary 4, we consider all the parameters known except q_Z.

Corollary 4 *If we consider a statistical model with one unknown parameter (q_Z), then $\forall p \in]0,1[$:*

$$\tilde{\kappa}_1 = \tilde{\kappa}_2 = \frac{\mathbb{P}\left(Y \notin [S_-, S_+]\right)}{1 - r^2}.$$

The proof is given in Section 4 of "Online Ressource 1". Here, q_Y is a known constant : contrary to Theorem 2 and Corollary 3, q_Y does not depend on n. The quantity $\mathbb{P}\left(Y \notin [S_-, S_+]\right)$ depends on q_Y, and is asymptotically the percentage of individuals genotyped. According to Corollary 4, we don't have to genotype symetrically anymore when q_Y is known : we can genotype only the individuals with the largest (resp. smallest) phenotypes. Another interesting result is that, when $\mathbb{P}\left(Y \notin [S_-, S_+]\right) > 1 - r^2$, selective genotyping becomes more powerful than the oracle test. This surprising result is due to the fact that q_Y is known.

6 Illustration

In this Section, we propose to illustrate our theoretical results. To begin, Figure 1 represents the efficiencies with respect to the oracle test, for a selective genotyping with one phenotype (left-side) and for two phenotypes (right-side). These efficiencies correspond to the two main theorems of this article : Theorem 1 for a selective genotyping with one phenotype, and Theorem 2 for a selective genotyping with two phenotypes. In other words, it corresponds to the situation where all the parameters are unknown. Note that the efficiencies do not depend on the QTL effects (see Theorem 1 and 3) and p. We study here the efficiencies as a function of the percentage of individuals genotyped γ and also as a function of the ratio γ_+/γ (ie the percentage of individuals genotyped with large phenotypes among the individuals genotyped). For instance, $\gamma_+/\gamma = 1/2$ refers that we genotype symetrically whereas $\gamma_+/\gamma = 1/4$ means that we genotype three times more individuals with small phenotypes than with large phenotypes. According to the graphs, we can see that we have to genotype symetrically. The worst configuration is to genotype only the large phenotypes (see $\gamma_+/\gamma = 1$) or to genotype only the small phenotypes (same curve as the one for $\gamma_+/\gamma = 1$). Obviously, we can remark that when $\gamma = 1$, all the efficiencies are equal to one, since all the individuals are genotyped.

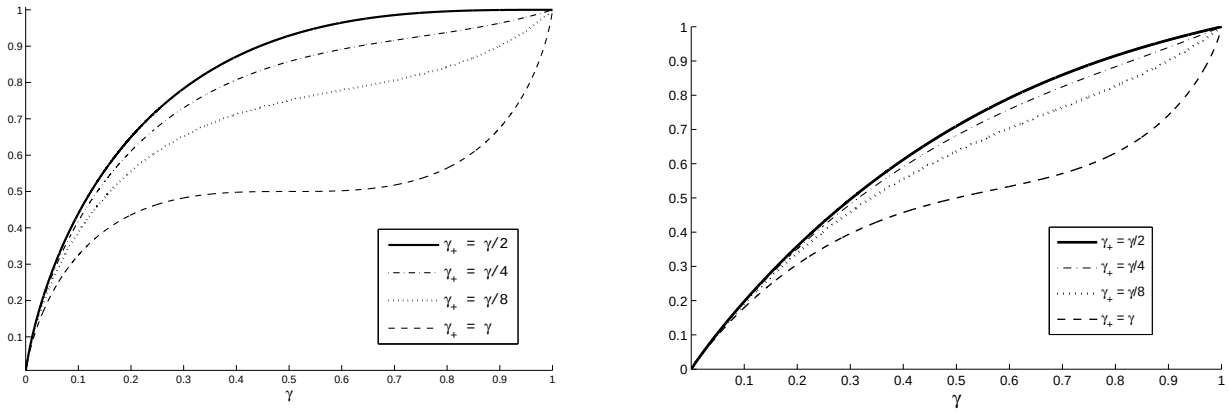
In Tables 1, 2 and 3, we will study the performances of our tests on simulated data in order to see if our tests which are based on asymptotic results, are suitable in real life. We will consider one-sided tests at the 5% level. In Table 1, we consider a selective genotyping with one phenotype. We focus on the most interesting situation : all the parameters are unknown. We consider the test based on the statistic T_2. It is very easy to perform since it is a comparison of means, between the two genotypes at the QTL. Note that we consider the easier expression of T_2 (see the remark below Lemma 1). We genotype symetrically

($\gamma_+/\gamma = 1/2$) and we consider $p = 1/2$ which corresponds in genetics to the backcross. Besides, $a = 2$ and $n = 100$. We remind that $q = a/\sqrt{n}$, so we have $q = 0.2$ in our case. β_2 refers to the theoretical power whereas β_{MC} to the Monte-Carlo power based on 10000 samples. CI refers to a 95% confidence interval for the true value of the power :

$$CI \quad = \quad \left[\beta_{MC} - 1.96 \sqrt{\frac{\beta_{MC}(1-\beta_{MC})}{10000}} \quad ; \quad \beta_{MC} + 1.96 \sqrt{\frac{\beta_{MC}(1-\beta_{MC})}{10000}} \right].$$

According to Table 1, we can see that β_2 is always in the confidence interval, whatever the value of γ. As a consequence, our test is suitable for $n = 100$.

In Tables 2 and 3, we consider a selective genotyping with two phenotypes. (μ_Z, q_Z, μ_Y, q_Y) are the unknown parameters. In this context, we focus on the test based on the test statistic $\tilde{W}_1$ of Lemma 2. Indeed, in order to obtain the the MLE $\hat{q}_Z$, we need to compute the MLE $\hat{q}_Y$, which can be obtained by EM (resp. Newton method) for strategy 1 (resp. strategy 2) (see Section 9 for details). As a consequence, the test based on strategy 1 is the easiest to compute. As previously, we consider $p = 1/2$ and $\gamma_+/\gamma = 1/2$. To begin, in Table 2, we study the situation where the QTL has no effect on the phenotype Z (ie. $q_Z = 0$). We compute the percentage of false positive (FP) and the confidence interval (CI) for the true value of FP (in the same way as previously). According to the table, we can see that for $n = 50$, 5% is always in the confidence interval, whatever the value of q_Y and r. In Table 3, we focus on the alternative. We consider $b = 4$, so $q_Z = 0.5657$. We can see that the theoretical power $\tilde{\beta}_1$ is always in the confidence interval, despite the fact that q_Z is not so close to 0. As a consequence, our test gives good performances for $n = 50$. That's why, it must be interesting for geneticists.



Selective genotyping with one phenotype

Selective genotyping with two phenotypes ($r = 0.8$)

Fig. 1 Efficiency as a function of γ and as a function of the ratio γ_+/γ

γ	β_{MC}	β_2	CI in %
0.1	36.74%	37.45%	[35.80 ; 37.68]
0.2	48.01%	48.61%	[47.03 ; 48.99]
0.3	54.28%	54.77%	[53.30 ; 55.26]
0.4	58.00%	58.58%	[57.03 ; 58.97]
0.5	60.10%	60.93%	[59.14 ; 61.06]
0.6	62.19%	62.33%	[61.24 ; 63.14]
0.7	62.26%	63.13%	[61.31 ; 63.21]
0.8	62.67%	63.52%	[61.72 ; 63.62]
0.9	63.30%	63.68%	[62.36 ; 64.24]
1	63.02%	63.68%	[62.07 ; 63.97]

Table 1 Study of strategy 2 for a selective genotyping with one phenotype. Theoretical power (β_2) and Monte-Carlo power (β_{MC}) as a function of γ (10000 samples, $n = 100$, $a = 2$, $q = \frac{2}{\sqrt{100}} = 0.2$, $\mu = 0$, $\sigma = 1$, $p = 1/2$, $\gamma_+/\gamma = 1/2$)

a	$q_Y = \frac{a}{\sqrt{n}}$	r	FP	CI in %
0	0	0.4	5.81 %	[5.35 ; 6.27]
0	0	0.7	5.76 %	[5.30 ; 6.22]
0	0	0.9	5.62 %	[5.17 ; 6.07]
2	0.2828	0.4	4.87 %	[4.45 ; 5.29]
2	0.2828	0.7	5.13 %	[4.70 ; 5.56]
2	0.2828	0.9	4.71 %	[4.29 ; 5.13]

Table 2 Study of strategy 1 for a selective genotyping with two phenotypes. Percentage of false positives (FP) as a function of a and r ($b = 0$, $q_Z = 0$, $\mu_Y = 0$, $\mu_Z = 0$, $\sigma = 1$, $p = 1/2$, $\gamma = 0.30$, $n = 50$, 10000 samples)

a	$q_Y = \frac{a}{\sqrt{n}}$	r	β_{MC}	$\tilde{\beta}_1$	CI in %
0	0	0.4	73.96 %	74.47 %	[73.10 ; 74.82]
0	0	0.7	82.23 %	82.31 %	[81.48 ; 82.98]
0	0	0.9	92.24 %	92.61 %	[91.72 ; 92.76]
2	0.2828	0.4	74.72 %	74.47 %	[73.87 ; 75.57]
2	0.2828	0.7	83.18 %	83.47 %	[82.45 ; 83.91]
2	0.2828	0.9	92.21 %	92.61 %	[91.68 ; 92.74]

Table 3 Study of strategy 1 for selective genotyping with two phenotypes. Theoretical power ($\tilde{\beta}_1$) and Monte-Carlo power (β_{MC}) ($b = 4$, $q_Z = 0.5657$, $\mu_Y = 0$, $\mu_Z = 0$, $\sigma = 1$, $p = 1/2$, $\gamma = 0.30$, $n = 50$, 10000 samples)

7 Proof for the oracle statistical test (μ, q, σ)

A natural estimator of the QTL effect q is the following comparison of means :

$$\frac{1}{2}\left\{\frac{\sum_{j=1}^{n} Y_j \; 1_{X_j=1}}{\sum_{j=1}^{n} 1_{X_j=1}} \; - \; \frac{\sum_{j=1}^{n} Y_j \; 1_{X_j=-1}}{\sum_{j=1}^{n} 1_{X_j=-1}}\right\}$$

However, this estimator is not convenient because of the random denominators. So, we want to build an easier estimator. Let $\eta = qX + \varepsilon$, we can remark that under the local alternative H_a :

$\mathbb{E}_{H_a}\left\{\frac{1}{2n}\ \left(\sum_{j=1}^{n}\ \frac{\eta_j}{p}\ 1_{X_j=1}\ -\ \frac{\eta_j}{1-p}\ 1_{X_j=-1}\right)\right\} = q.$

Besides under H_0, $\mathbb{E}_{H_0}\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right) = 0$ and

$\mathbb{E}_{H_0}\left\{\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right)^2\right\} = \mathbb{E}_{H_0}\left(\frac{\eta^2}{p^2}\ 1_{X=1}\ +\ \frac{\eta^2}{(1-p)^2}\ 1_{X=-1}\right) = \frac{\sigma^2}{p(1-p)}.$

It comes, $\mathbb{V}_{H_0}\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right) = \frac{\sigma^2}{p(1-p)}$.

Besides, under the local alternative H_a :

$$\mathbb{E}_{H_a}\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right) = 2q\ , \tag{1}$$

$\mathbb{E}_{H_a}\left\{\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right)^2\right\} = \frac{1}{p}(\sigma^2+q^2)\ +\ \frac{1}{1-p}(\sigma^2+q^2)\ \rightarrow\ \frac{\sigma^2}{p(1-p)}\ ,$

$\mathbb{V}_{H_a}\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right) = \frac{1}{p}(\sigma^2+q^2)\ +\ \frac{1}{1-p}(\sigma^2+q^2)\ -\ 4q^2.$

We remark that $\mathbb{V}_{H_a}\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right) \rightarrow \mathbb{V}_{H_0}\left(\frac{\eta}{p}\ 1_{X=1}\ -\ \frac{\eta}{1-p}\ 1_{X=-1}\right)$.

As a consequence, let $\widetilde{T}$ be the following test statistic :

$$\widetilde{T} = \frac{\sum_{j=1}^{n}\ \frac{\eta_j}{p}\ 1_{X_j=1}\ -\ \frac{\eta_j}{1-p}\ 1_{X_j=-1}}{\sigma\ \sqrt{\frac{n}{p\ (1-p)}}}.$$

The asymptotic laws are : $\widetilde{T} \overset{H_0}{\rightarrow} N\left(0,\ 1\right)$ and $\widetilde{T} \overset{H_a}{\rightarrow} N\left(\frac{2a\ \sqrt{p\ (1-p)}}{\sigma},\ 1\right)$.

However, we don't observe the r.v. η but the phenotypes Y. Let $\overline{Y}$ and $\overline{\eta}$ be the empirical means : $\overline{Y} = \frac{1}{n}\sum Y_j$ and $\overline{\eta} = \frac{1}{n}\sum \eta_j$. Then, $\overline{Y} = \mu + \overline{\eta}$ and $Y - \overline{Y} = \eta - \overline{\eta}$. Let T be the following test statistic :

$$T = \frac{\sum_{j=1}^{n}\ \frac{1}{p}(Y_j-\overline{Y})\ 1_{X_j=1}\ -\ \frac{1}{1-p}(Y_j-\overline{Y})\ 1_{X_j=-1}}{\sigma\ \sqrt{\frac{n}{p\ (1-p)}}}. \tag{2}$$

We have

$$T = \widetilde{T}\ +\ \overline{\eta}\ \frac{\sum_{j=1}^{n}\ \frac{1}{1-p}\ 1_{X_j=-1}\ -\ \frac{1}{p}\ 1_{X_j=1}}{\sigma\ \sqrt{\frac{n}{p\ (1-p)}}}.$$

Notation 1 $o_P(1)$ will be a sequence of random vectors which tend to 0 in probability and $O_P(1)$ will be a sequence bounded in probability.

According to Prohorov, $\overline{\eta} = O_P(\frac{1}{\sqrt{n}})$ and $\sum_{j=1}^{n}\ \frac{1}{1-p}1_{X_j=-1} - \frac{1}{p}\ 1_{X_j=1} = O_P(\sqrt{n})$. It comes,

$$\overline{\eta}\ \frac{\sum_{j=1}^{n}\ \frac{1}{1-p}\ 1_{X_j=-1}\ -\ \frac{1}{p}\ 1_{X_j=1}}{\sigma\ \sqrt{\frac{n}{p\ (1-p)}}}\ \rightarrow\ 0.$$

As a consequence (we remind that we are under H_0 or under H_a):

$$T = \widetilde{T}\ +\ o_P(1).$$

It comes T has the same asymptotic laws as $\widetilde{T}$. We need now to estimate the variance σ^2 which is unknown in the model studied. We will consider the empirical variance $\hat{\sigma}^2 = \frac{1}{n}\left\{\sum_{j=1}^{n}(Y_j - \overline{Y})^2\right\}$ with $\overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$. $\hat{\sigma}^2$ is a consistant estimator under H_0 and H_a by contiguity. We just have to adapt the previous test statistic T. T is now such as :

$$T = \frac{\sum_{j=1}^{n} \frac{1}{p}(Y_j - \overline{Y})\ 1_{X_j=1} - \frac{1}{1-p}(Y_j - \overline{Y})\ 1_{X_j=-1}}{\hat{\sigma}\ \sqrt{\frac{n}{p\ (1-p)}}}.$$

The asymptotic laws are unchanged : $T \overset{H_0}{\rightarrow} N(0,\ 1)$ and $T \overset{H_a}{\rightarrow} N\left(\frac{2a\ \sqrt{p\ (1-p)}}{\sigma},\ 1\right)$. This test has the same asymptotic laws as the Wald test (proof given in Section 1 of "Online Ressource 1").

8 Proof of Theorem 1

Notation 2 I_θ will be the Fisher information matrix taken at the point θ . $I_{ij}(\theta)$ refers to the element ij of I_θ. $I_{ij}^{-1}(\theta)$ refers to the element ij of I_θ^{-1}, the inverse of I_θ.

8.1 Theoretical elements needed for the study

To begin, we introduce a theorem. It will be very convenient to calculate the power for the Wald tests.

Theorem 3 *Let $C_1, ..., C_n$ be an independent and equally distributed sample from a probability distribution P_θ. We suppose that Θ is an open subset of $\mathbb{R}^d$ and that the model $(P_\theta\ :\ \theta\ \in\ \Theta)$ is regular. Let $\hat{\theta}$ be the Maximum Likelihood Estimator (MLE) of θ and $\theta_0 \in \Theta$, then for every converging sequence $h_n \rightarrow h$, as $n \rightarrow +\infty$, we have :*

i) under P_{θ_0} , $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0,\ I^{-1}(\theta_0))$

ii) under $P_{\theta_0 + h_n/\sqrt{n}}$, $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(h,\ I^{-1}(\theta_0))$.

Proof Let P_n be the law corresponding to $P_{\theta_0}^{\otimes n}$, Q_n the law corresponding to $P_{\theta_0 + h_n/\sqrt{n}}^{\otimes n}$ and $\frac{dQ_n}{dP_n}$ the likelihood ratio.
Since the model is regular, we have i). Besides, we can use Theorem 7.2 of Van der Vaart (1998) which gives an explicit expression of the log likelihood under P_n. According to the central limit theorem, the law of large numbers and the properties of the Fisher Information matrix, we have (with h^t the transpose of h):

$$\log\left(\frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightarrow} N(-\frac{1}{2}\nu^2,\ \nu^2) \quad \text{with } \nu^2 = h^t I_{\theta_0} h.$$

Notation 3 $Q_n \lhd P_n$ will mean the sequence Q_n is contiguous with the respect to the sequence P_n.

By the iii) of Le Cam's first lemma, we have $Q_n \triangleleft P_n$. So, we can use Le Cam's third lemma. Since the model is regular, we can use Theorem 5.39 of Van der Vaart (1998) :

$$\sqrt{n}(\hat{\theta} - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{\ell}_{\theta_0}(C_j) \ + \ o_{P_{\theta_0}}(1)$$

where $\dot{\ell}_{\theta_0}(C_j)$ denotes the score function taken at θ_0, for an observation C_j.
According to Theorem 7.2 of Van der Vaart (1998) :

$$\log\left(\frac{dQ_n}{dP_n}\right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h^t \dot{\ell}_{\theta_0}(C_j) \ - \ \frac{1}{2} h^t I_{\theta_0} h \ + \ o_{P_{\theta_0}}(1).$$

Let $h_{(i)}$ be the ith component of h. At the ith line, we have :

$$\begin{aligned} \text{Cov}\left(\log\left(\frac{dQ_n}{dP_n}\right), \ \sqrt{n}(\hat{\theta} - \theta_0)\right) &= \sum_{k=1}^{d} h_{(k)} \left\{ I_{i1}^{-1}(\theta_0) I_{1k}(\theta_0) + ... + I_{id}^{-1}(\theta_0) I_{dk}(\theta_0) \right\} \ + \ o_{P_{\theta_0}}(1) \\ &= h_{(i)} \ + \ o_{P_{\theta_0}}(1). \end{aligned}$$

Then, according to Le Cam's third lemma :

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{Q_n}{\to} N(h, \ I^{-1}(\theta_0)).$$

This gives the result.

8.2 First strategy (Wald test using all the phenotypes)

8.2.1 Likelihood

To begin, we remind that the r.v. $\overline{X}$ is such as :

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

So, $\overline{X} = 0$ refers to the case where the genotype is missing. $(\overline{X} \ , Y)$ has a density with respect to the Lebesgue measure $\times$ the counting measure.

Notation 4 $\forall\, i \in \{-1, 1\}$ and $\forall\, k \in \{-1, 0, 1\}$, $\overline{\mathbb{P}}\{i \mid k\}$ and $\mathbb{P}\{k \mid i\}$ are the quantities such as :

$$\overline{\mathbb{P}}\{i \mid k\} = \mathbb{P}(X = i \mid \overline{X} = k) \quad \text{and} \quad \mathbb{P}\{k \mid i\} = \mathbb{P}(\overline{X} = k \mid X = i).$$

Notation 5 q_{-1}, q_1 and q_0 are the quantities such as :
$q_{-1} = \mathbb{P}(\overline{X} = -1)$, $q_1 = \mathbb{P}(\overline{X} = 1)$ and $q_0 = \mathbb{P}(\overline{X} = 0)$.

It comes $\mathbb{P}\{i \mid i\} = \Phi\left(\frac{S_- - \mu - iq}{\sigma}\right) + 1 - \Phi\left(\frac{S_+ - \mu - iq}{\sigma}\right)$ where Φ is the cumulative distribution of a standard normal distribution, $q_{-1} = \mathbb{P}\{-1 \mid -1\}\ (1-p)\ ,\ q_1 = \mathbb{P}\{1 \mid 1\}\ p$ and
$q_0 = (1 - \mathbb{P}\{-1 \mid -1\})\ (1-p)\ +\ (1 - \mathbb{P}\{1 \mid 1\})\ p$.
As a consequence :

$$\overline{\mathbb{P}}\{-1 \mid k\} = \frac{\mathbb{P}\{k \mid -1\}\ (1-p)}{q_k} \quad , \quad \overline{\mathbb{P}}\{1 \mid k\} = \frac{\mathbb{P}\{k \mid 1\}\ p}{q_k}.$$

According to Bayes theorem, $\forall\ k \in \{-1, 1\}$, $\forall\ y \in\ \mathbb{R}$, we have

$$\mathbb{P}(Y \in [y\ ,\ y+dy] \mid \overline{X} = k) = \mathbb{P}(Y \in [y\ ,\ y+dy] \mid X = k\ \cap\ \overline{X} \neq 0) = \frac{\varphi(\frac{y-\mu-kq}{\sigma}) 1_{y \notin [S_-\ ,\ S_+]}}{\sigma\ \mathbb{P}\{k \mid k\}}\ dy\ ,$$

$$\mathbb{P}(Y \in [y\ ,\ y+dy]\ \cap\ \overline{X} = k) = \frac{\varphi(\frac{y-\mu-kq}{\sigma}) 1_{y \notin [S_-\ ,\ S_+]}}{\sigma\ \mathbb{P}\{k \mid k\}}\ q_k\ dy\ ,$$

where $\varphi(.)$ denotes the density of a standard normal distribution.
It comes :

$$\mathbb{P}(Y \in [y\ ,\ y+dy]\ \cap\ \overline{X} = -1) = \frac{1-p}{\sigma}\ \varphi\left(\frac{y-\mu+q}{\sigma}\right) 1_{y \notin [S_-\ ,\ S_+]}\ dy\ ,$$
$$\mathbb{P}(Y \in [y\ ,\ y+dy]\ \cap\ \overline{X} = 1) = \frac{p}{\sigma}\ \varphi\left(\frac{y-\mu-q}{\sigma}\right) 1_{y \notin [S_-\ ,\ S_+]}\ dy\ .$$

Besides,

$$\begin{aligned}\mathbb{P}(Y \in [y\ ,\ y+dy] \mid \overline{X} = 0) &= \sum_{i \in \{-1,1\}} \mathbb{P}(Y \in [y\ ,\ y+dy]\ \cap\ X = i\ \mid\ \overline{X} = 0) \\ &= \frac{p\ \varphi(\frac{y-\mu-q}{\sigma}) 1_{y \in [S_-\ ,\ S_+]}}{\sigma\ q_0}\ dy\ +\ \frac{(1-p)\ \varphi(\frac{y-\mu+q}{\sigma}) 1_{y \in [S_-\ ,\ S_+]}}{\sigma\ q_0}\ dy\ .\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}(Y \in [y\ ,\ y+dy]\ \cap\ \overline{X} = 0) = &\frac{p}{\sigma}\ \varphi\left(\frac{y-\mu-q}{\sigma}\right) 1_{y \in [S_-\ ,\ S_+]}\ dy \\ &+ \frac{1-p}{\sigma}\ \varphi\left(\frac{y-\mu+q}{\sigma}\right) 1_{y \in [S_-\ ,\ S_+]}\ dy\ .\end{aligned}$$

Finally, the likelihood L for an observation $\left(\overline{X}, Y\right)$ is such as :

$$\begin{aligned}L = &\frac{1-p}{\sigma}\ \varphi\left(\frac{y-\mu+q}{\sigma}\right) 1_{\overline{X}=-1}\ +\ \frac{p}{\sigma}\ \varphi\left(\frac{y-\mu-q}{\sigma}\right) 1_{\overline{X}=1} \\ &+ \left\{\frac{1-p}{\sigma}\ \varphi\left(\frac{y-\mu+q}{\sigma}\right)\ +\ \frac{p}{\sigma}\ \varphi\left(\frac{y-\mu-q}{\sigma}\right)\right\} 1_{\overline{X}=0}\ .\end{aligned}$$

8.2.2 Statistical test (μ, q)

We consider a statistical model with two unknown parameters (μ, q). We first introduce a useful lemma obtained mainly using integration by parts.

Lemma 3 *Let* $V \sim N(\mu,\ \sigma^2)$*, then :*

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

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

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

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

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

Notation 6 γ, γ_+ and γ_- are respectively the quantities $\mathbb{P}_{H_0}\left(Y \notin [S_-,\ S_+]\right), \mathbb{P}_{H_0}\left(Y > S_+\right)$ and $\mathbb{P}_{H_0}\left(Y < S_-\right)$. z_α denote the quantile of order $1-\alpha$ of a standard normal distribution. $\mathcal{A}$ is the quantity such as $\mathcal{A} = \sigma^2\ \left\{\gamma\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\right\}$.

According to this lemma, we have $\mathcal{A} = \mathbb{E}_{H_0}\left\{(Y-\mu)^2 1_{Y\notin[S_-,\ S_+]}\right\}$.
Let $\theta = (\mu, q)$ be the parameter of the model and $\theta_0 = (\mu, 0)$ be true value of the parameter under H_0. We first compute the score functions and the Fisher Information matrix. We have

$$\frac{\partial logL}{\partial q}\ |_{\theta_0} = -\left(\frac{y-\mu}{\sigma^2}\right) 1_{\overline{X}=-1}\ +\ \left(\frac{y-\mu}{\sigma^2}\right) 1_{\overline{X}=1}\ +\ \left(\frac{y-\mu}{\sigma^2}\right)(2p-1) 1_{\overline{X}=0}\ ,$$

$$\left(\frac{\partial logL}{\partial q}\ |_{\theta_0}\right)^2 = \frac{(y-\mu)^2}{\sigma^4} 1_{\overline{X}=-1}\ +\ \frac{(y-\mu)^2}{\sigma^4} 1_{\overline{X}=1}\ +\ \frac{(y-\mu)^2}{\sigma^4}\ (2p-1)^2\ 1_{\overline{X}=0}\ .$$

It comes $I_{22}(\theta_0) = \frac{\mathcal{A}}{\sigma^4}\ +\ \frac{(2p-1)^2}{\sigma^4}\ (\sigma^2 - \mathcal{A})$. Besides, $\frac{\partial logL}{\partial \mu}\ |_{\theta_0} = \ \frac{y-\mu}{\sigma^2}\ $. So, $I_{11}(\theta_0)\ =\ \frac{1}{\sigma^2}$. Furthermore,

$$\frac{\partial logL}{\partial q\ \partial \mu}\ |_{\theta_0} = \frac{1}{\sigma^2} 1_{\overline{X}=-1}\ -\ \frac{1}{\sigma^2} 1_{\overline{X}=1}\ -\ \frac{1}{\sigma^2}\ (2p-1) 1_{\overline{X}=0}\ .$$

Since we are under H_0, $P_{H_0}\left\{-1 \mid -1\right\} = P_{H_0}\left\{1 \mid 1\right\}$, it comes $I_{12}(\theta_0) = \frac{1}{\sigma^2}(2p-1)$. As a consequence :

$$I_{22}^{-1}(\theta_0) = \frac{\sigma^4}{4\ \mathcal{A}\ p(1-p)}\ .$$

$\hat{q}$, the MLE of q, can be obtained using a EM algorithm. Since the model is regular :

$$\sqrt{n}\ \hat{q}\ \ \overset{H_0}{\rightarrow}\ \ N(\ 0\ ,\ I_{22}^{-1}(\theta_0)\ )\ .$$

We can deduce the Wald test :

$$W_1 = \frac{2\sqrt{n}}{\sigma^2}\ \sqrt{\mathcal{A}\ p(1-p)}\ \ \hat{q}\ \ \overset{H_0}{\rightarrow}\ \ N(0,\ 1)\ .$$

According to Theorem 3 with $h_n = h = (0, a)$:

$$W_1 \overset{H_a}{\rightarrow} N\left(\frac{2a}{\sigma^2}\sqrt{\mathcal{A}\, p(1-p)},\ 1\right) . \tag{3}$$

8.3 Second strategy (comparison of means based on the extreme phenotypes)

8.3.1 Statistical test (μ, q, σ)

Let $\hat{\delta}$ be the following estimator :

$$\hat{\delta} = \frac{1}{p}(Y-\mu)1_{X=1} \ - \ \frac{1}{1-p}(Y-\mu)1_{X=-1} .$$

According to formula (1) in Section 7, $\mathbb{E}_{H_a}(\hat{\delta}) = 2q$ when we are in the oracle situation. So, $\hat{\delta}$ is an estimator of twice the QTL effect. If now, we consider selective genotyping, we would like to define $\hat{\delta}$ such as :

$$\hat{\delta} = \frac{1}{p}(Y-\mu)1_{\overline{X}=1} \ - \ \frac{1}{1-p}(Y-\mu)1_{\overline{X}=-1} .$$

According to Lemma 3 :

$$\begin{aligned}\mathbb{E}\left(\hat{\delta}\right) &= \frac{1}{p}\,\mathbb{E}\left(Y-\mu \mid \overline{X}=1\right)\mathbb{P}(\overline{X}=1) \ - \ \frac{1}{1-p}\,\mathbb{E}\left(Y-\mu \mid \overline{X}=-1\right)\mathbb{P}(\overline{X}=-1)\\ &= q\left(\mathbb{P}\left\{1 \mid 1\right\} + \mathbb{P}\left\{-1 \mid -1\right\}\right) \ + \ \sigma\,\varphi\left(\frac{S_+ - \mu - q}{\sigma}\right) \ - \ \sigma\,\varphi\left(\frac{S_- - \mu - q}{\sigma}\right)\\ &- \ \sigma\,\varphi\left(\frac{S_+ - \mu + q}{\sigma}\right) \ + \ \sigma\,\varphi\left(\frac{S_- - \mu + q}{\sigma}\right) .\end{aligned}$$

We remark that $\hat{\delta}$ is not a good estimator of q anymore, but we can propose a test based on $\hat{\delta}$ since the expectation depends of q. We have $\mathbb{E}_{H_0}(\hat{\delta}) = 0$ and $\mathbb{V}_{H_0}(\hat{\delta}) = \mathbb{E}_{H_0}(\hat{\delta}^2)$. Besides :

$$\hat{\delta}^2 = \frac{1}{p^2}\,(Y-\mu)^2\,1_{\overline{X}=1} \ + \ \frac{1}{(1-p)^2}\,(Y-\mu)^2\,1_{\overline{X}=-1} .$$

According to Lemma 3 :

$$\begin{aligned}\mathbb{E}(\hat{\delta}^2) &= \frac{1}{p^2}\,\mathbb{E}\left\{(Y-\mu)^2 \mid \overline{X}=1\right\}\mathbb{P}(\overline{X}=1) \ + \ \frac{1}{(1-p)^2}\,\mathbb{E}\left\{(Y-\mu)^2 \mid \overline{X}=-1\right\}\mathbb{P}(\overline{X}=-1)\\ &= \frac{1}{p}\,\mathbb{E}\left\{(Y-\mu)^2 1_{Y\notin[S_-,S_+]} \mid X=1\right\} \ + \ \frac{1}{1-p}\,\mathbb{E}\left\{(Y-\mu)^2 1_{Y\notin[S_-,S_+]} \mid X=-1\right\} .\end{aligned}$$

It comes $\mathbb{E}_{H_0}(\hat{\delta}^2) = \frac{\mathcal{A}}{p(1-p)}$. So, we can define the test statistic T_2 corresponding to the second strategy. According to the Central Limit theorem,

$$T_2 = \frac{\sum_{j=1}^n \frac{1}{p}(Y_j-\mu)1_{\overline{X}_j=1} \ - \ \frac{1}{1-p}(Y_j-\mu)1_{\overline{X}_j=-1}}{\sqrt{\frac{n\ \mathcal{A}}{p(1-p)}}} \overset{H_0}{\rightarrow} N(0,\ 1) . \tag{4}$$

According to a Taylor expansion at first order :

$$\varphi\left(\frac{S_- - \mu + q}{\sigma}\right) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{S_- - \mu}{\sigma}\right)^2} \left\{1 - \frac{(S_- - \mu)\, q}{\sigma^2} + o(q)\right\} .$$

We also have (working on integrals) :

$$P\{1 \mid 1\} = \Phi\left(\frac{S_- - \mu}{\sigma}\right) - \frac{q}{\sigma}\varphi\left(\frac{S_- - \mu}{\sigma}\right) + 1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \frac{q}{\sigma}\varphi\left(\frac{S_+ - \mu}{\sigma}\right) + o(q) .$$

It comes :

$$\mathbb{E}_{H_a}\{T_2\} \to 2a\left\{\gamma - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + z_{\gamma_+}\varphi(z_{\gamma_+})\right\}\sqrt{\frac{p(1-p)}{\mathcal{A}}} .$$

We can remark that this limit is equal to $\frac{2a}{\sigma^2}\sqrt{\mathcal{A}\, p(1-p)}$.
Besides, $\mathbb{E}_{H_a}(\hat{\delta}) \to 0$.
Using Portmanteau theorem (since $\forall i \in \{-1, 1\}$, $Y \mid X = i \to N(\mu,\ \sigma^2)$):

$$\mathbb{E}_{H_a}(\hat{\delta}^2) \to \frac{\mathcal{A}}{p(1-p)} .$$

So $\mathbb{V}_{H_a}(\hat{\delta}) \to \mathbb{V}_{H_0}(\hat{\delta})$ and it comes

$$T_2 \overset{H_a}{\to} N\left(\frac{2a}{\sigma^2}\sqrt{\mathcal{A}\, p(1-p)},\ 1\right) . \tag{5}$$

Since μ and σ are unknown, we have to adapt the test statistic T_2. We can replace μ by $\hat{\mu}$, estimator which depends of the extreme phenotypes. $\hat{\mu}$ can be obtained by maximum likelihood or by the method of moments, because these two estimators are $\sqrt{n}$ consistent (same kind of proof as in Section 7). Besides, we can use the following consistent estimator of $\mathcal{A}$:

$$\hat{\mathcal{A}} = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\mu})^2 1_{\overline{X}_j \neq 0} .$$

The asymptotic laws of T_2 are unchanged.

8.3.2 Asymptotic Relative Efficiency

We compute here the Asymptotic Relative Efficiency (ARE) of the test of comparison of mean based on extreme phenotypes, with respect to the oracle test (μ, q, σ) where all the genotypes are known. Until now, we have considered n individuals. Let consider now $n^\star$ individuals for a selective genotyping experiment. T_2 has to be adapted. It comes

$$T_2 = \frac{\sum_{j=1}^{n^\star} \frac{1}{p}(Y_j - \hat{\mu})1_{\overline{X}_j = 1} - \frac{1}{1-p}(Y_j - \hat{\mu})1_{\overline{X}_j = -1}}{\sqrt{\frac{n^\star\, \hat{\mathcal{A}}}{p(1-p)}}} \overset{H_0}{\to} N(0,\ 1)$$

where $\hat{\mathcal{A}}$ and $\hat{\mu}$ are the same estimators as previously but adapted for $n^\star$ individuals.
Let ζ be the quantity such as $\zeta = \frac{n^\star}{n}$, then (we remind that $q = a/\sqrt{n}$) :

$$T_2 \overset{H_a}{\to} N\left(\frac{2a}{\sigma^2}\sqrt{\zeta\, \mathcal{A}\, p(1-p)},\ 1\right) .$$

We will focus in particular on the appropriate one sided test when $a > 0$. The test based on T_2 will be more powerful than the oracle test (μ, q, σ) when (we suppose $a > 0$) :

$$z_\alpha \ - \ \frac{2a}{\sigma^2}\sqrt{\zeta\ \mathcal{A}\ p(1-p)} \ < \ z_\alpha \ - \ \frac{2a\ \sqrt{p(1-p)}}{\sigma} \ \Leftrightarrow \ \zeta \ > \ \frac{\sigma^2}{\mathcal{A}}\ .$$

As a result, the efficiency κ_2 is such as $\kappa_2 = \mathcal{A}/\sigma^2$. That is to say,

$$\kappa_2 = \gamma \ + \ z_{\gamma_+}\varphi(z_{\gamma_+}) \ - \ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\ . \tag{6}$$

8.4 Proof of i) of Theorem 1

Let $\beta_i^{(\mu,q,\sigma)}$ (resp. $\beta_i^{(\mu,q)}$) be the power of the test (μ, q, σ) (resp. (μ, q)) corresponding to strategy i. According to formulae (5) and (3) : $\beta_2^{(\mu,q,\sigma)} = \beta_1^{(\mu,q)}$. Besides, by definition : $\beta_2^{(\mu,q,\sigma)} \leqslant \beta_1^{(\mu,q,\sigma)} \leqslant \beta_1^{(\mu,q)}$. It comes $\beta_1^{(\mu,q,\sigma)} = \beta_2^{(\mu,q,\sigma)}$. As a consequence, $\kappa_1 = \kappa_2$. In the same way, by definition : $\beta_2^{(\mu,q,\sigma)} \leqslant \beta_3^{(\mu,q,\sigma)} \leqslant \beta_1^{(\mu,q,\sigma)}$. So, $\kappa_1 = \kappa_2 = \kappa_3$.

8.5 Proof of ii) of Theorem 1

We have to answer the following question : how must we choose γ_+ and γ_- to maximize the efficiency ? We remind that $\gamma_+ + \gamma_- = \gamma$. Let $g(.)$ be the function such as : $g(z_{\gamma_+}) = \Phi^{-1}\left\{\gamma - 1 + \Phi(z_{\gamma_+})\right\}$. Then, $z_{1-\gamma_-} = g(z_{\gamma_+})$.

Let $k_1(.)$ be the following function : $k_1(z_{\gamma_+}) = z_{\gamma_+}\varphi(z_{\gamma_+}) - g(z_{\gamma_+})\ \varphi\left\{g(z_{\gamma_+})\right\}$.
In order to maximize κ_1, we have to maximize the function $k_1(.)$. Let $k_1'(.)$, $g'(.)$ and $\varphi'(.)$ be respectively the derivative of $k_1(.)$, $g(.)$ and $\varphi(.)$. We have :

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

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

Then, $k_1'(z_{\gamma/2}) = \varphi(z_{\gamma/2}) - \left\{z_{\gamma/2}\right\}^2\ \varphi(z_{\gamma/2}) - \varphi(z_{1-\gamma/2}) + \left\{z_{1-\gamma/2}\right\}^2\ \varphi(z_{1-\gamma/2}) = 0$. As a result, the efficiency κ_1 reaches its maximum when $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

9 Proof of Theorem 2

To begin, we suppose that we are in the oracle situation, ie no genotypes are missing. So, we observe Z and X whatever the value of Y. In order to perform the linear regression of $Z \mid X$ on $Y \mid X$ which will be called $\widetilde{Z} \mid X$, we define the following scalar product, for 2 r.v. U_1 and U_2 which take value in $\mathbb{R}$: $< U_1\ ,\ U_2 > = \mathbb{E}\left[U_1 U_2\right]$. We have :

$$\begin{aligned}\widetilde{Z} \mid X = & < Z \mid X\ ,\ \frac{Y \mid X - \mu_{YX}}{\sigma} > \frac{Y \mid X - \mu_{YX}}{\sigma} \quad + \quad < Z \mid X\ ,\ 1 > 1 \\ = & \ \ r\ Y \mid X \ - \ r\ \mu_{YX} \ + \ \mu_{ZX}\ .\end{aligned}$$

Let $Z^\star$ and $\mu_{ZX}^\star$ be the two following quantities :

$$Z^\star = \frac{Z \ - \ r\ Y}{\sigma\ \sqrt{1-r^2}} \quad \text{and} \quad \mu_{ZX}^\star = \frac{\mu_{ZX} \ - \ r\ \mu_{YX}}{\sigma\ \sqrt{1-r^2}}\ .$$

This way, $Z^\star \mid X \sim N(\mu^\star_{ZX}, 1)$. By construction, $(Z-\widetilde{Z}) \mid X$ and $\widetilde{Z} \mid X$ are independent. So, $Z^\star \mid X$ and $Y \mid X$ are independent.
If we consider now a selective genotyping experiment, $Z^\star$ will be available only when Y is extreme. However, since $Z^\star \mid X$ and $Y \mid X$ are independent, $Z^\star \mid X$ is not affected by the fact that Y is extreme.

9.1 First strategy (Wald test using all the phenotypes)

Notation 7 $L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y)$ is the likelihood for an observation $(\overline{X}, Y, Z^\star)$ and $L(\mu_Z, q_Z, \mu_Y, q_Y)$ is the likelihood for an observation $(\overline{X}, Y, Z)$.

Obviously, we have the relationship $L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y) = L(\mu_Z, q_Z, \mu_Y, q_Y)$.
We have :

$$L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y) = \left\{ \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right) \; + \; \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\right\} 1_{\overline{X}=0}$$
$$+ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z1})\, 1_{\overline{X}=1} \; + \; \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z-1})\, 1_{\overline{X}=-1} \; .$$

The respective MLE $\hat{\mu}_Y$ and $\hat{q}_Y$, of μ_Y and q_Y can be obtained using an EM algorithm.
Besides, since $\frac{\partial \log L^\star}{\partial \mu^\star_{Z1}} = (z^\star - \mu^\star_{Z1})1_{\overline{X}=1}$ and $\frac{\partial \log L^\star}{\partial \mu^\star_{Z-1}} = (z^\star - \mu^\star_{Z-1})1_{\overline{X}=-1}$, we easily obtain $\hat{\mu}^\star_{Z-1}$ and $\hat{\mu}^\star_{Z1}$ respective MLE of $\mu^\star_{Z-1}$ and $\mu^\star_{Z1}$ for n observations :

$$\hat{\mu}^\star_{Z1} = \frac{1}{\sum_{j=1}^n 1_{\overline{X}_j=1}} \sum_{j=1}^n z^\star_j 1_{\overline{X}_j=1} \quad \text{and} \quad \hat{\mu}^\star_{Z-1} = \frac{1}{\sum_{j=1}^n 1_{\overline{X}_j=-1}} \sum_{j=1}^n z^\star_j 1_{\overline{X}_j=-1} \; .$$

Let $\theta = (\mu_Z, q_Z, \mu_Y, q_Y)$ and $\theta^\star = \left(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y\right)$. Then, θ corresponds to parameters of L and $\theta^\star$ to parameters of $L^\star$. We have :

$$q_Z = \frac{\sigma}{2}\sqrt{1-r^2}\,(\mu^\star_{Z1} - \mu^\star_{Z-1}) \; + \; r\, q_Y \; ,$$
$$\mu_Z = \frac{\sigma}{2}\sqrt{1-r^2}\,(\mu^\star_{Z1} + \mu^\star_{Z-1}) + \; r\, \mu_Y \; .$$

Let M be the matrix such as $\theta = M\theta^\star$:

$$M = \begin{pmatrix} \frac{\sigma}{2}\sqrt{1-r^2} & \frac{\sigma}{2}\sqrt{1-r^2} & r & 0 \\ -\frac{\sigma}{2}\sqrt{1-r^2} & \frac{\sigma}{2}\sqrt{1-r^2} & 0 & r \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \; .$$

The inverse of M, called M^{-1}, verifies :

$$M^{-1} = \begin{pmatrix} \frac{1}{\sigma\sqrt{1-r^2}} & -\frac{1}{\sigma\sqrt{1-r^2}} & -\frac{r}{\sigma\sqrt{1-r^2}} & \frac{r}{\sigma\sqrt{1-r^2}} \\ \frac{1}{\sigma\sqrt{1-r^2}} & \frac{1}{\sigma\sqrt{1-r^2}} & -\frac{r}{\sigma\sqrt{1-r^2}} & -\frac{r}{\sigma\sqrt{1-r^2}} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \; .$$

Let $\theta_{00} = (\mu_Z, 0, \mu_Y, 0)$ and $\theta^\star_{00} = M^{-1}\theta_{00}$. It comes :

$$\theta^\star_{00} = \left(\frac{\mu_Z}{\sigma\sqrt{1-r^2}} - \frac{r\mu_Y}{\sigma\sqrt{1-r^2}} \; , \; \frac{\mu_Z}{\sigma\sqrt{1-r^2}} - \frac{r\mu_Y}{\sigma\sqrt{1-r^2}} \; , \; \mu_Y \; , \; 0 \right) \; .$$

Notation 8 I_θ (resp. $I^\star_{\theta^\star}$) will be the Fisher information matrix corresponding to the likelihood L (resp. $L^\star$) and taken at point θ (resp. $\theta^\star$).

Let's calculate $I^\star_{\theta^\star_{00}}$:
$\frac{\partial \log L^\star}{\partial \mu_Y} \mid_{\theta^\star_{00}} = \frac{y-\mu_Y}{\sigma}$, $\frac{\partial \log L^\star}{\partial \mu^\star_{Z-1}} \mid_{\theta^\star_{00}} = (z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}})\ 1_{\overline{X}=-1}$,
$\frac{\partial \log L^\star}{\partial \mu^\star_{Z1}} \mid_{\theta^\star_{00}} = (z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}})\ 1_{\overline{X}=1}$ and
$\frac{\partial \log L^\star}{\partial q_Y} \mid_{\theta^\star_{00}} = -\left(\frac{y-\mu_Y}{\sigma^2}\right)\ 1_{\overline{X}=-1} + \left(\frac{y-\mu_Y}{\sigma^2}\right)\ 1_{\overline{X}=1} + \left(\frac{y-\mu_Y}{\sigma^2}\right)(2p-1)\ 1_{\overline{X}=0}$.
It comes

$$I^\star_{11}(\theta^\star_{00}) = (1-p)\ \gamma \ , \ I^\star_{22}(\theta^\star_{00}) = p\ \gamma \ \text{ and } \ I^\star_{33}(\theta^\star_{00}) = 1/\sigma^2 \ .$$

Let's adapt the previous notations for the configuration with two phenotypes.

Notation 9 γ, γ_+ and γ_- are respectively the quantities
$\mathbb{P}_{H_{0Y}}(Y \notin [S_-,\ S_+])$, $\mathbb{P}_{H_{0Y}}(Y > S_+)$ and $\mathbb{P}_{H_{0Y}}(Y < S_-)$.

We remind that $\mathcal{A} = \sigma^2\ \left\{\gamma\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\right\}$. According to Section 8.2.2, we have

$$I^\star_{44}(\theta^\star_{00}) = \frac{\mathcal{A}}{\sigma^4} + \frac{(2p-1)^2}{\sigma^4}(\sigma^2 - \mathcal{A}) \quad \text{and} \quad I^\star_{34}(\theta^\star_{00}) = \frac{2p-1}{\sigma^2} \ .$$

Besides, all the other terms of $I^\star_{\theta^\star_{00}}$ are equal to zero.
Let $\hat\theta$ and $\hat\theta^\star$ be the respective MLE of θ and $\theta^\star$, then we have $\hat\theta = M\hat\theta^\star$. Since the model is regular :

$$\mathbb{V}\left\{\ \sqrt{n}\ (\hat\theta^\star - \theta^\star_{00})\ \right\} \overset{H_{0Y}H_{0Z}}{\to}\ I^{\star\,-1}_{\theta^\star_{00}} \ .$$

Besides, $\sqrt{n}\ (\hat\theta - \theta_{00}) = \sqrt{n}\ M\ (\hat\theta^\star - \theta^\star_{00})$, it comes :

$$\mathbb{V}\left\{\ \sqrt{n}\ (\hat\theta - \theta_{00})\ \right\} \overset{H_{0Y}H_{0Z}}{\to}\ M\ I^{\star\,-1}_{\theta^\star_{00}}\ M^t \quad \text{and} \quad I^{-1}_{\theta_{00}} = M\ I^{\star\,-1}_{\theta^\star_{00}}\ M^t \ .$$

After some calculations, we obtain :

$$I^{-1}_{22}(\theta_{00}) = \frac{\sigma^2\ (1-r^2)}{4\ p\ (1-p)\ \gamma} \ + \ \frac{\sigma^4\ r^2}{4\ p\ (1-p)\ \mathcal{A}} \ .$$

Let define the Wald statistic W_1 :

$$W_1 = \sqrt{n}\ \ \hat q_Z / \sqrt{I^{-1}_{22}(\theta_{00})} \ .$$

The MLE $\hat q_Z$ can easily be obtained using the MLE $\hat\mu^\star_{Z-1}$, $\hat\mu^\star_{Z1}$, and $\hat q_Y$ ($\hat q_Y$ can be obtained by EM). Since the model is regular :

$$W_1 \quad \overset{H_{0Z}H_{0Y}}{\to} \quad N(0,\ 1) \ .$$

We apply Theorem 3 respectively with $h_n = h = (0,0,0,a)$, $h_n = h = (0,b,0,0)$, $h_n = h = (0,b,0,a)$. Then, we have :

$$\begin{aligned} W_1 &\overset{H_{0Z}H_{aY}}{\rightarrow} N(0,\ 1) \\ W_1 &\overset{H_{bZ}H_{0Y}}{\rightarrow} N\left(b/\sqrt{I_{22}^{-1}(\theta_{00})}\ ,\ 1\right) \\ W_1 &\overset{H_{bZ}H_{aY}}{\rightarrow} N\left(b/\sqrt{I_{22}^{-1}(\theta_{00})}\ ,\ 1\right)\ . \end{aligned}$$

It comes, whatever that we consider the null hypothesis or the local alternative for Y, we always have :

$$W_1 \overset{H_{0Z}}{\rightarrow} N(0,\ 1) \quad \text{and} \quad W_1 \overset{H_{bZ}}{\rightarrow} N\left(b/\sqrt{I_{22}^{-1}(\theta_{00})}\ ,\ 1\right)\ .$$

The efficiency $\tilde{\kappa}_1$ of this test, with respect to the oracle test (μ_Z, q_Z) is obtained easily :

$$\tilde{\kappa}_1 = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\gamma\ +\ z_{\gamma_+}\varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})} \right\}^{-1}\ .$$

We remark that :

$$\tilde{\kappa}_1 = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_1} \right\}^{-1}$$

where κ_1 is given in Theorem 1. According to Theorem 1, κ_1 reaches its maximum for $\gamma_+ = \gamma_- = \gamma/2$. So, it is the same for $\tilde{\kappa}_1$.

9.2 Second strategy (Wald test using only the extreme phenotypes Y)

In this case, the likelihood is :

$$\begin{aligned} L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y) = \mathbb{P}(\overline{X}=0)\ 1_{\overline{X}=0}\ +\ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z1})\ 1_{\overline{X}=1} \\ +\ \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z-1})\ 1_{\overline{X}=-1}\ . \end{aligned}$$

Let's calculate the Fisher Information matrix. $I^\star_{11}(\theta^\star_{00})$ and $I^\star_{22}(\theta^\star_{00})$ are the same as previously :

$$I^\star_{11}(\theta^\star_{00}) = (1-p)\ \gamma\ \ ,\ \ I^\star_{22}(\theta^\star_{00}) = p\ \gamma\ .$$

Besides,

$$\begin{aligned} \frac{\partial \log L^\star}{\partial \mu_Y}\ |_{\theta^\star_{00}} &= \frac{y-\mu_Y}{\sigma^2}\ \left\{1_{\overline{X}=-1} + 1_{\overline{X}=1}\right\}\ +\ \frac{\varphi\left(z_{1-\gamma_-}\right)-\varphi\left(z_{\gamma_+}\right)}{\sigma\ (1-\gamma)}\ 1_{\overline{X}=0}\ , \\ I^\star_{33}(\theta^\star_{00}) &= \frac{\mathcal{A}}{\sigma^4}\ +\ \frac{\left\{\varphi\left(z_{1-\gamma_-}\right)-\varphi\left(z_{\gamma_+}\right)\right\}^2}{\sigma^2\ (1-\gamma)}\ . \end{aligned}$$

According to formula (1) of Section 2.4 of "Online Ressource 1" :

$$I^{\star}_{44}(\theta^{\star}_{00}) = \frac{\mathcal{A}}{\sigma^4} \ + \ (2p-1)^2 \ \frac{\left\{\varphi\left(z_{1-\gamma_-}\right) - \varphi\left(z_{\gamma_+}\right)\right\}^2}{\sigma^2 \ (1-\gamma)} \ .$$

Besides,

$$\begin{aligned}\frac{\partial \log L^{\star}}{\partial \mu_Y \partial q_Y} \ |_{\theta^{\star}_{00}} &= \frac{1}{\sigma^2}(1_{\overline{X}=-1} - 1_{\overline{X}=1}) \ + \ \frac{2p-1}{\sigma^2(1-\gamma)} \left\{z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) - z_{\gamma_+}\varphi(z_{\gamma_+})\right\} 1_{\overline{X}=0} \\ &\quad - \frac{2p-1}{\sigma^2(1-\gamma)^2} \left\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\right\}^2 \ 1_{\overline{X}=0} \ .\end{aligned}$$

It comes :

$$I^{\star}_{34}(\theta^{\star}_{00}) = (1-2p)\left[\frac{\mathcal{A}}{\sigma^4} + \frac{\left\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\right\}^2}{\sigma^2(1-\gamma)}\right] \ .$$

The other components of the Fisher Information matrix ar equal to zeros. Using block matrix inversion, we obtain :

$$I^{\star\,-1}_{11}(\theta^{\star}_{00}) = \frac{1}{(1-p)\ \gamma} \quad , \quad I^{\star\,-1}_{22}(\theta^{\star}_{00}) = \frac{1}{p\ \gamma} \quad , \quad I^{\star\,-1}_{44}(\theta^{\star}_{00}) = \frac{\sigma^4}{4\ \mathcal{A}\ p(1-p)} \ .$$

Let define Λ such as :

$$\Lambda = \left\{\frac{4\ \mathcal{A}\ p(1-p)}{\sigma^4}\left[\frac{\mathcal{A}}{\sigma^4} + \frac{\left\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\right\}^2}{\sigma^2\ (1-\gamma)}\right]\right\}^{-1} \ .$$

Then :

$$\begin{aligned} I^{\star\,-1}_{33}(\theta^{\star}_{00}) &= \frac{\Lambda}{\sigma^4}\left[\mathcal{A} + (2p-1)^2 \frac{\left\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\right\}^2}{1-\gamma}\right] \\ I^{\star\,-1}_{34}(\theta^{\star}_{00}) &= \Lambda\ (2p-1)\ \left[\frac{\mathcal{A}}{\sigma^4} + \frac{\left\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\right\}^2}{\sigma^2\ (1-\gamma)}\right] \ .\end{aligned}$$

In the same way as previously :

$$I^{-1}_{\theta_{00}} = M\ I^{\star\,-1}_{\theta^{\star}_{00}}\ M^t \ .$$

After calculations, we obtain :

$$I^{-1}_{22}(\theta_{00}) = \frac{\sigma^2(1-r^2)}{4\ \gamma\ p(1-p)} \ + \ \frac{r^2\sigma^4}{4\ \mathcal{A}\ p(1-p)} \ .$$

We deduce the Wald test statistic W_2 and its asymptotic law (same proof as for the first strategy)

$$\begin{aligned} W_2 &= \sqrt{n}\ \ \hat{q}_Z / \sqrt{I^{-1}_{22}(\theta_{00})} \ \ \overset{H_{0Z}}{\rightarrow} \ \ N(0,\ 1) \\ W_2 \ &\overset{H_{bZ}}{\rightarrow} \ N\left(b/\sqrt{I^{-1}_{22}(\theta_{00})}\ ,\ 1\right) \ .\end{aligned}$$

The MLE $\hat{q}_Z$ can be obtained using $\hat{\mu}^{\star}_{Z-1}$, $\hat{\mu}^{\star}_{Z1}$ and $\hat{q}_Y$ ($\hat{q}_Y$ can be obtained using a Newton method). This test has the same power as the test corresponding to the first strategy. It concludes the proof.

Acknowledgements I am very grateful to my PhD advisor Jean-Marc Azaïs. I thank Laurent Bordes, Céline Delmas, Jean-Michel Elsen and Bernard Prum for fruitful discussions. This work has been supported by the French National Center for Scientific Research (CNRS) and the Animal Genetic Department of the French National Institute for Agricultural Research and SABRE.

References

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

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

Lebowitz, R.J., Soller, M., Beckmann, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics*, **73**, 556-562.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer.

Lynch, M., Walsh, B. (2007). *Genetics and Analysis of Quantitative Traits*, Sinauer.

Medugorac, I., Soller, M. (2001). Selective genotyping with a main and a correlated trait. *Journal of Animal Breeding and Genetics*, **118**, 285-295.

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

Rabier, C-E. (2010). *PhD thesis*, Université Toulouse 3, Paul Sabatier.

Van der Vaart, A.W. (1998). *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

Weller, J.I. (2001). *Quantitative Trait Loci in Animals*, Cabi Publishing.

Wu, R., MA, C.X., Casella, G. (2007). *Statistical Genetics of Quantitative Traits*, Springer.