



HAL
open science

On statistical inference for selective genotyping

Charles-Elie Rabier

► **To cite this version:**

| Charles-Elie Rabier. On statistical inference for selective genotyping. 2012. hal-00658583v2

HAL Id: hal-00658583

<https://hal.science/hal-00658583v2>

Preprint submitted on 30 Jun 2012 (v2), last revised 29 Mar 2020 (v3)

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

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

RESEARCH ARTICLE

*On statistical inference for selective genotyping*Charles-Elie Rabier ^{ab*}

^aUniversité de Toulouse, Institut de Mathématiques de Toulouse, U.P.S, 31062 Toulouse, France; ^bINRA UR631, Station d'Amélioration Génétique des Animaux, Chemin de Borde-Rouge, 31326 Castanet-Tolosan, France

()

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 propose different statistical tests suitable for selective genotyping and we compare their performances in a very large framework. We prove that the non extreme phenotypes (i.e. the phenotypes for which the genotypes are missing) don't bring any information for statistical inference. We also prove that we have to genotype symmetrically, that is to say the same percentage of large and small phenotypes whatever the proportions of the two genotypes in the population. Same results are obtained in the case of a selective genotyping with two correlated phenotypes.

Keywords: Hypothesis testing; Asymptotic properties of tests; Asymptotic Relative Efficiency; Selective genotyping; Quantitative Trait Locus detection

AMS Subject Classification: 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). The trait is observed on n individuals (progenies) and we denote by Y_j , $j = 1, \dots, n$, the observations, which we will assume to be independent and identically distributed (i.i.d.). In a famous article, Lander and Botstein [1] proposed, with the help of genetic markers, to scan a chromosome represented by a segment $[0, T]$, performing a likelihood ratio test (LRT) of the absence of a QTL at every location $t \in [0, T]$. It leads to a “likelihood ratio test process” $\Lambda_n(\cdot)$, and then a natural statistic is the supremum of such a process. This method is called “interval mapping”. There have been many papers related to the supremum of the LRT process. For example, we can mention Feingold and al. [2], Churchill and Doerge [3], Rebaï et al. [4], Rebaï et al. [5], Cierco [6], Piepho [7], Chang et al. [8], Azaïs and Wschebor [9], Azaïs et al. [10]. The “interval mapping” of Lander and Botstein [1] has enabled the discovery of thousands of genes in animals, humans and plants (see for instance Lynch and Walsh [11], Weller [12], Wu et al. [13]).

*Corresponding author. Email: rabier@stat.wisc.edu

In this study, contrary to the “interval mapping” where the focus is on the whole chromosome, we focus here only on a single locus which is a genetic marker. We suppose that the QTL is located on the genetic marker. We propose to study in details a famous experimental design which allows to reduce costs due to genotyping (i.e. collecting the marker information). X refers to the genetic marker with two possible genotypes : $+1$ with probability p and -1 with probability $1-p$. Note that, since we consider that the QTL is located on the genetic marker, the genotypes at the QTL and at the marker, are the same. Typically, the case $p = 1/2$ refers to a backcross population $A \times (A \times B)$, where A and B are purely homozygous lines. Indeed, for a backcross population, there are only two genotypes at the marker, each one with probability $1/2$ (under Hardy-Weinberg assumptions). The case $p \neq 1/2$ refers to a cross, between an homozygous population and an heterozygous population, and for which the Hardy-Weinberg law has been violated. According to the Hardy-Weinberg law, the heterozygous parent produces two kind of gametes in equal number. If this is not the case, the probability of the two genotypes are not equal (i.e. $p \neq 1/2$).

We assume an “analysis of variance model” for the quantitative trait (i.e. the phenotype) : $Y = \mu + qX + \varepsilon$ where ε is a Gaussian noise with mean 0 and variance σ^2 . q refers to the QTL effect. A QTL is present if and only if the QTL effect q is different from zero.

The problem is that genotyping (i.e. collecting the marker information X) is very expensive. In such a context, Lebowitz and al. [14] had a very good idea based on the observation that most of the information about the QTL is present in the extreme phenotypes (i.e. the smallest and the largest Y). So, they proposed 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 [1], formalized this approach and called it “selective genotyping”. Then, different topics have been investigated. For instance, Muranty and Goffinet [15] focused on the estimation of the QTL effect for selective genotyping. Rabbee and al. [16] studied different strategies for analyzing data in selective genotyping and gave the power associated to each strategy. Manichaikul and al. [17] focused on permutation tests for selective genotyping ... However, although there have been many papers on selective genotyping, the theory of statistical inference for selective genotyping is still missing, even in the case of only one genetic marker. In a very famous article, Darvasi and Soller [18] proposed to perform a comparison of means between the extreme individuals (i.e. 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 main goal is to propose easy and optimal statistical tests.

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 : contrary to Lander and Botstein [1], Darvasi and Soller [18], Muranty and Goffinet [15], Rabbee and al. [16], we don’t focus only on a backcross which corresponds to $p = 1/2$. On the other hand, 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 [18]), 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 ?

In a second part of this article, we will focus on selective genotyping with two correlated phenotypes : Y and Z . Sometimes, it is difficult to measure the phenotype Z of interest : it can be expensive or it can require a lot of work. In such a situation, a second phenotype, Y , correlated to the phenotype of interest, can be measured more easily. An example given by Medugorac and Soller [19], is the “mapping of QTL to determine genetic resistance to Helminthiasis in the Red Mas-sai sheep of East Africa. Worm counts in spleen and liver are the most accurate measure of resistance. These are time consuming to perform, and require a trained professional. Faecal egg counts stand in good correlation to worm counts and are relatively easy to obtain”. In such a context, the costs due to genotyping and due to phenotyping can be reduced : a selective genotyping is performed on Y (as previously), and Z is measured only on the genotyped individuals (i.e. with extreme phenotypes Y). Obviously, in such a situation, the interest is on finding a QTL which has an effect on Z . Some theoretical results about this design are presented in Muranty and Goffinet [15] and Medugorac and Soller [19], but the theory of statistical inference is still missing. As a consequence, in our study, as in the part dealing with only one phenotype, we will focus on statistical inference and try to propose to geneticists the easiest and optimal statistical tests.

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 all the genotypes are known. 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. These strategies are inspired by Darvasi and Soller [18] and Rabbee and al. [16]. The different tests (corresponding to the different strategies) are compared in terms of Asymptotic Relative Efficiency (ARE), which determines for each test, the sample size required to obtain the same local asymptotic power as the one of the oracle test. Theorem 3.1, which gives the different ARE for the different tests, is the main result of the first part of this article which deals with only one phenotype. According to Theorem 3.1, we have the same ARE if we keep or if we don’t keep the phenotypes Y (i.e. the phenotypes for which the genotypes are 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 3.2 is a direct consequence of Theorem 3.1. We present in this lemma the different test statistics, corresponding to the different tests studied. Since the non extreme phenotypes don’t bring any information for statistical inference, an easy and optimal test is presented. It is based on the comparison of means of the extreme phenotypes.

On the other hand, a very important result of Theorem 3.1 is the following : if we want to genotype only a percentage γ of the population, we have to genotype symmetrically, 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 (i.e. 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 of this article : we deal with two correlated phenotypes Y and Z . Same kind of analysis is given, as in the first part which deals with one phenotype. Theorem 5.1 is the main result. According

to Theorem 5.1, we still have to genotype symmetrically and the non extreme phenotypes Y still don't bring any information for statistical inference. Theorem 5.1 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 5.2 presents the different test statistics, corresponding to the different tests studied. We leave the choice to geneticists between two optimal statistical 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 [20]'s work on contiguity. We refer to the book of Van der Vaart [21] 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 (i.e. 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 loss 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 (i.e. at the marker). 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. referring to the phenotype :

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

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

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 :

$$H_0 : q = 0 \text{ vs } H_1 : q \neq 0.$$

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

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

where $\hat{\sigma} = \frac{1}{\sqrt{n}} \left\{ \sum_{j=1}^n (Y_j - \bar{Y})^2 \right\}^{1/2}$ and $\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$.

The asymptotic laws are :

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

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

As said before, our main goal is to propose to geneticists the easiest statistical test. Obviously, this test has to be optimal in order to detect the QTL. As a consequence, in this section, we will have 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 (i.e. the phenotypes for which the genotypes are missing) bring any extra information for statistical inference ?
- 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 $S_- \leq S_+$. We consider that the genotype X is known if and only if the phenotype Y is extreme, i.e. if and only if $Y \leq S_-$ or $Y \geq S_+$. In order to make the reading easier, we define a new r.v. \bar{X} such as :

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

In other words, $\bar{X} = 0$ 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, i.e. the phenotypes for which the genotypes are missing) and we perform a Wald test
- 2. we keep only the extreme phenotypes (i.e. the phenotypes for which the genotypes are available) and we perform a comparison of means between the two

genotypes at the QTL

- 3. we keep only the extreme phenotypes (i.e. the phenotypes for which the genotypes are 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 the 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 is inspired by Darvasi and Soller [18], whereas strategies 1 and 3 are inspired by the simulation study of Rabbee and al. [16]. Obviously, strategy 2 is the easiest to compute.

3.3. Results

To begin, we present our main theorem :

Theorem 3.1 : *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) , $\forall p \in]0, 1[$:*

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

$$ii) \quad \kappa_1, \kappa_2 \text{ and } \kappa_3 \text{ reach their maximum, } M, \text{ when } \gamma_+ = \gamma_- = \frac{\gamma}{2}, \text{ with}$$

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

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_+ and γ_- . According to the law of large numbers, under the null hypothesis H_0 and under the local alternative H_a , $\frac{1}{n} \sum 1_{\bar{X}_j \neq 0} \rightarrow \gamma$. 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 3.1. 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 $\kappa_2 = \kappa_3$, 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 3.2, 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 (i.e. the probability that $X = +1$) and a (i.e. the constant linked to the QTL effect). It only depends on γ, γ_+ and γ_- .

According to ii) of Theorem 3.1, 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 the $\gamma/2\%$ 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 3.2, which presents explicitly, contrary to Theorem 3.1, the different tests corresponding to the different strategies.

Lemma 3.2: *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 :*

$$\begin{aligned}
 W_1 &:= \frac{2\sqrt{n}}{\hat{\sigma}^2} \sqrt{\hat{\mathcal{A}}_1 p(1-p)} \hat{q}_1 \\
 T_2 &:= \sqrt{p(1-p)} \left\{ \frac{\sum_{j=1}^n \frac{1}{p}(Y_j - \hat{\mu}_3)1_{\bar{X}_j=1} - \frac{1}{1-p}(Y_j - \hat{\mu}_3)1_{\bar{X}_j=-1}}{\sqrt{n \hat{\mathcal{A}}_3}} \right\} \\
 W_3 &:= \frac{2\sqrt{n}}{\hat{\sigma}_3^2} \sqrt{\hat{\mathcal{A}}_3 p(1-p)} \hat{q}_3
 \end{aligned}$$

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

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

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

For the proof, we refer to the proof of Theorem 3.1 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 (c.f. Rabier [22]). As a consequence, the tests corresponding to strategies one and three are difficult to perform. According to Lemma 3.2, the test based on T_2 , i.e. the comparison of means 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 \bar{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 :

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

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 symmetrically.

Until now, we have focused on the most interesting configuration : all the parameters (i.e. 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 check if we obtain the same results as previously : strategy 2 is maybe not

optimal anymore when the number of unknown parameters varies. 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 3.1 (see Section 8.2.2), we have $\mathcal{A} = \mathbb{E}_{H_0} \{(Y - \mu)^2 1_{Y \notin [S_-, S_+]}\}$. As a consequence, we will consider the test statistic T_2 of Lemma 3.2 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 3.2. Besides, in order to calculate the different ARE for the different strategies, we will obviously consider the appropriate oracle test (i.e. the oracle test with only q unknown, and the one with (μ, q) unknown).

Corollary 3.3: *If we consider a statistical model with one unknown parameter (q), then (with the previous notations) :*

- i) $\kappa_1 = \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) + (2p - 1)^2 \{1 - \gamma - z_{\gamma_+} \varphi(z_{\gamma_+}) + z_{1-\gamma_-} \varphi(z_{1-\gamma_-})\}$
- ii) $\kappa_2 = 4p(1 - p) \{\gamma - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) + z_{\gamma_+} \varphi(z_{\gamma_+})\}$
- iii) $\kappa_3 = \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) + \frac{(2p - 1)^2}{1 - \gamma} \{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\}^2 \quad \forall \gamma \neq 1$
- iv) $\kappa_1 = \kappa_2 = \kappa_3 \Leftrightarrow p = \frac{1}{2}$
- v) $\forall p \in]0, 1[\quad \kappa_1, \kappa_2 \text{ and } \kappa_3 \text{ reach their maximum for } \gamma_+ = \gamma_- = \frac{\gamma}{2}.$

Corollary 3.4: *If we consider a statistical model with two unknown parameters (μ, q) , then the results are the same as in Theorem 3.1.*

The proof of Corollary 3.3 is given in Section 2 of “Online Ressource 1”. The proof of Corollary 3.4 is obvious according to the proof of Theorem 3.1.

According to Corollary 3.4, when only the variance σ^2 is known, we have the same results as previously. So, there is no loss of generality to consider the variance known. However, according to Corollary 3.3, there is a loss of generality to consider the mean μ known. Indeed, when we consider only q unknown, the three strategies have the same ARE if and only if $p = 1/2$ (i.e. 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 [18]

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 $Y \notin [S_-, S_+]$ (i.e. $\bar{X} \neq 0$). As said previously, under H_0 and H_a , $\frac{1}{n} \sum 1_{\bar{X}_j \neq 0} \rightarrow \gamma$ 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 [18], 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 [18] suppose symmetry, 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 [18] (i.e. $p = 1/2$ and symmetry), our study gives the same ARE as presented in formula (27) of Darvasi and Soller [18]. 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 [18]. Indeed, we consider $p \in]0, 1[$, not only symmetry, 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 :

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

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) i.i.d. . Note that q_Z and q_Y are respectively the QTL effects on the 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 :

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

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 (c.f. Sections 2.2 and 7), an easy and optimal test to perform is based on the following statistic

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

The asymptotic laws are :

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

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

5. Selective genotyping with two correlated phenotypes

Sometimes, it is difficult to measure the phenotype Z of interest : it can be expensive or it can require a lot of work. In such a situation, a second phenotype, Y , correlated to the phenotype of interest, can be measured more easily. An example given by Medugorac and Soller [19], is the “mapping of QTL to determine genetic resistance to Helminthiasis in the Red Massai sheep of East Africa” (c.f. Section 1.1). In order to reduce costs due to genotyping and due to phenotyping, a selective genotyping is performed on Y , and Z is measured only on the genotyped individuals (i.e. 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 have to be correlated otherwise this selective genotyping has no sense. This way, we will focus here on statistical inference for selective genotyping with two correlated phenotypes. Note that some theoretical results about this design are already present in Muranty and Goffinet [15] and Medugorac and Soller [19]. However, the theory of statistical inference is still missing, since Muranty and Goffinet [15] focused only on the estimation of the QTL effects and Medugorac and Soller [19] focused on the power of the design using approximations.

5.1. Motivation

As previously, our main goal is to propose to geneticists the easiest statistical test which is optimal. This way, we have 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 (i.e. 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, i.e. the phenotypes for which the genotypes are missing) and we perform a Wald

test.

- 2. we keep only the extreme phenotypes Y (i.e. the phenotypes for which the genotypes are 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 the same local asymptotic power as the one of 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 analogous to strategy 3 of the first part.

5.3. Results

To begin, we present our main theorem, Theorem 5.1, which is the analogue of Corollary 3.4 for two phenotypes (the covariance matrix is known here). However, since Corollary 3.4 and Theorem 3.1 give same results, Theorem 5.1 can be also viewed as the analogue of Theorem 3.1.

Theorem 5.1: *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 $\mathbb{P}_{H_{0Y}}(Y \notin [S_-, S_+])$, $\mathbb{P}_{H_{0Y}}(Y > S_+)$ and $\mathbb{P}_{H_{0Y}}(Y < S_-)$. 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} , $\forall p \in]0, 1[$:*

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

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

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

where κ_1 and M are the quantities of Theorem 3.1.

The proof is given in Section 9. As expected, the ARE increase with r and γ . As previously, the non extreme phenotypes Y (i.e. for which the genotypes are missing) don't bring any extra information for statistical inference on q_Z . So, using strategy 1 instead of strategy 2 does not lead to an increase of power. Besides, we still have to genotype symmetrically for a selective genotyping with two phenotypes. Note that Theorem 5.1 establishes the relationship between the ARE of selective genotyping with one and two phenotypes. Lemma 5.2 presents the different tests corresponding to the different strategies.

Lemma 5.2: *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

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

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 3.2.

For the proof, we refer to the proof of Theorem 5.1 in Section 9. So, according to Lemma 5.2, we have two different test statistics, \tilde{W}_1 and \tilde{W}_2 , corresponding to the two different strategies. These two test statistics differ only by the MLE, \hat{q}_Z , of the QTL effect on Z . In particular, if we call \hat{q}_Z^i (resp. \hat{q}_Y^i) the MLE of q_Z (resp. q_Y) for strategy i , after some algebra (see the proof in Section 9), we obtain

$$\hat{q}_Z^i = \frac{\sigma}{2} \sqrt{1-r^2} (\hat{\mu}_{Z1}^* - \hat{\mu}_{Z-1}^*) + r \hat{q}_Y^i$$

where

$$\begin{aligned} \hat{\mu}_{Z1}^* &= \left\{ \sum_{j=1}^n \frac{(Z_j - rY_j) 1_{\bar{X}_j=1}}{\sigma \sqrt{1-r^2}} \right\} / \sum_{j=1}^n 1_{\bar{X}_j=1} \\ \hat{\mu}_{Z-1}^* &= \left\{ \sum_{j=1}^n \frac{(Z_j - rY_j) 1_{\bar{X}_j=-1}}{\sigma \sqrt{1-r^2}} \right\} / \sum_{j=1}^n 1_{\bar{X}_j=-1}. \end{aligned}$$

The key thing is that for strategy 1, \hat{q}_Y^1 can be computed by the EM algorithm, whereas for strategy 2, \hat{q}_Y^2 can be computed by a Newton method. So, although we have proved that the non extreme phenotypes don't bring any extra information, the tests suitable for selective genotyping with two correlated phenotypes, are not so simple. As said previously, it is obvious that a test of comparison of means on Z (in the same way as what has been proposed in the first part) won't be optimal. As a consequence, we leave to geneticists the choice between the two statistical tests, which are optimal and asymptotically equivalent.

We introduce now Corollary 5.3 which is the analogous of Corollary 3.3. Only q_Z and q_Y are now unknown.

Corollary 5.3: *If we consider a statistical model with two unknown parameters (q_Z, q_Y), then under H_{0Y} and under H_{aY} :*

- i) $\tilde{\kappa}_1 = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_1} \right\}^{-1}$
- ii) $\tilde{\kappa}_2 = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_3} \right\}^{-1}$
- iii) $\tilde{\kappa}_1 = \tilde{\kappa}_2 \Leftrightarrow p = \frac{1}{2}$
- iv) $\forall p \in]0, 1[\quad \tilde{\kappa}_1 \text{ and } \tilde{\kappa}_2 \text{ reach their maximum for } \gamma_+ = \gamma_- = \frac{\gamma}{2}$

where κ_1 and κ_3 are the quantities of Corollary 3.3.

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 symmetrically. In other words, we have to use strategy 1 and genotype symmetrically. Note that Corollary 5.3 establishes a link with the ARE of Corollary 3.3.

To conclude, in the following Corollary 5.4, we consider all the parameters known except q_Z .

Corollary 5.4: *If we consider a statistical model with one unknown parameter (q_Z), then $\forall p \in]0, 1[$:*

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

The proof is given in Section 4 of “Online Ressource 1”. Here, q_Y is a known constant : contrary to Theorem 5.1 and Corollary 5.3, q_Y does not depend on n . The quantity $\mathbb{P}(Y \notin [S_-, S_+])$ depends on q_Y , and is asymptotically the percentage of individuals genotyped. According to Corollary 5.4, we don’t have to genotype symmetrically anymore when q_Y is known : we can genotype only the individuals with the largest (resp. smallest) phenotypes. Besides, we can use strategy 1 or strategy 2 because the two tests have the same power. Another interesting result is that, when $\mathbb{P}(Y \notin [S_-, S_+]) > 1 - r^2$, 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 with two phenotypes (right-side). These efficiencies correspond to the two main theorems of this article : Theorem 3.1 for a selective genotyping with one phenotype, and Theorem 5.1 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 3.1 and Theorem 5.1) 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 (i.e. the percentage of individuals genotyped with large phenotypes among all the individuals genotyped). For instance, $\gamma_+/\gamma = 1/2$ means that we genotype symmetrically 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 symmetrically. The worst configuration is to genotype only the largest phenotypes (see $\gamma_+/\gamma = 1$) or to genotype only the smallest 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 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 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 easiest expression of T_2 (see the remark below Lemma 3.2). We genotype symmetrically ($\gamma_+/\gamma = 1/2$) and we consider $p = 1/2$

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, $n = 100$, $a = 2$, $q = \frac{2}{\sqrt{100}} = 0.2$, $\mu = 0$, $\sigma = 1$, $p = 1/2$, $\gamma_+/\gamma = 1/2$).

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]

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 :

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

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. We focus here on the test based on the test statistic \tilde{W}_1 of Lemma 5.2. We remind that, 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). So, we decided here to use the EM algorithm. 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 (i.e. $q_Z = 0$). We compute the percentage of false positives (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.

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

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

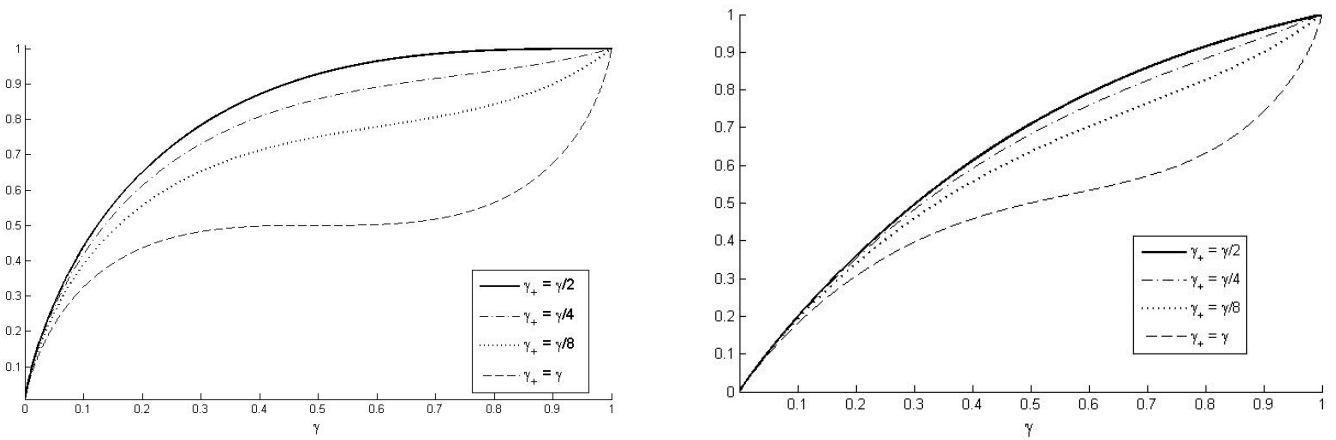
$$\frac{1}{2} \left\{ \frac{\sum_{j=1}^n Y_j 1_{X_j=1}}{\sum_{j=1}^n 1_{X_j=1}} - \frac{\sum_{j=1}^n Y_j 1_{X_j=-1}}{\sum_{j=1}^n 1_{X_j=-1}} \right\}$$

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	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 3. Study of strategy 1 for a 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).

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]



(a) Selective genotyping with one phenotype

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

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

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 :

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

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

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

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 \tilde{T} be the following test statistic :

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

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

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

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

We have

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

Notation 7.1: $o_P(1)$ 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, $\bar{\eta} = O_P\left(\frac{1}{\sqrt{n}}\right)$ 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,

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

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

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

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

statistic T . T is now such as :

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

The asymptotic laws are unchanged : $T \xrightarrow{H_0} N(0, 1)$ and $T \xrightarrow{H_1} N\left(\frac{2a \sqrt{p(1-p)}}{\sigma}, 1\right)$.

This test has the same asymptotic laws as the Wald test (proof given in Section 1 of “Online Ressource 1”).

8. Proof of Theorem 3.1

Notation 8.1: 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 8.2: Let C_1, \dots, 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 [21] 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):

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

Notation 8.3: $Q_n \triangleleft P_n$ 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 [21] :

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

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 [21] :

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

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

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

Then, according to Le Cam's third lemma :

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{Q_v} N(h, I^{-1}(\theta_0)).$$

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. \bar{X} is such as :

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

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

Notation 8.4: $\forall i \in \{-1, 1\}$ and $\forall k \in \{-1, 0, 1\}$, $\bar{\mathbb{P}}\{i | k\}$ and $\mathbb{P}\{k | i\}$ are the quantities such as :

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

Notation 8.5: q_{-1} , q_1 and q_0 are the quantities such as :

$$q_{-1} = \mathbb{P}(\bar{X} = -1), \quad q_1 = \mathbb{P}(\bar{X} = 1) \quad \text{and} \quad q_0 = \mathbb{P}(\bar{X} = 0).$$

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

As a consequence :

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

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

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

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

where $\varphi(\cdot)$ denotes the density of a standard normal distribution.

It comes :

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

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

Besides,

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

Then,

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

Finally, the likelihood L for an observation (\bar{X}, Y) is such as :

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

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 8.6: Let $V \sim N(\mu, \sigma^2)$, then :

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

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

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

$$- \sigma (S_- - \mu) \varphi \left(\frac{S_- - \mu}{\sigma} \right)$$

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

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

Notation 8.7: γ , γ_+ and γ_- are respectively the quantities $\mathbb{P}_{H_0}(Y \notin [S_-, S_+])$, $\mathbb{P}_{H_0}(Y > S_+)$ and $\mathbb{P}_{H_0}(Y < S_-)$. z_α denote the quantile of order $1 - \alpha$ of a standard normal distribution. \mathcal{A} is the quantity such as $\mathcal{A} = \sigma^2 \{ \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) \}$.

According to this lemma, we have $\mathcal{A} = \mathbb{E}_{H_0} \{ (Y - \mu)^2 1_{Y \notin [S_-, S_+]} \}$. 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

$$\begin{aligned} \frac{\partial \log L}{\partial q} \Big|_{\theta_0} &= - \left(\frac{y - \mu}{\sigma^2} \right) 1_{\bar{X} = -1} + \left(\frac{y - \mu}{\sigma^2} \right) 1_{\bar{X} = 1} + \left(\frac{y - \mu}{\sigma^2} \right) (2p - 1) 1_{\bar{X} = 0}, \\ \left(\frac{\partial \log L}{\partial q} \Big|_{\theta_0} \right)^2 &= \frac{(y - \mu)^2}{\sigma^4} 1_{\bar{X} = -1} + \frac{(y - \mu)^2}{\sigma^4} 1_{\bar{X} = 1} + \frac{(y - \mu)^2}{\sigma^4} (2p - 1)^2 1_{\bar{X} = 0}. \end{aligned}$$

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

$$\frac{\partial \log L}{\partial q \partial \mu} \Big|_{\theta_0} = \frac{1}{\sigma^2} 1_{\bar{X} = -1} - \frac{1}{\sigma^2} 1_{\bar{X} = 1} - \frac{1}{\sigma^2} (2p - 1) 1_{\bar{X} = 0}.$$

Since we are under H_0 , $P_{H_0} \{-1 | -1\} = P_{H_0} \{1 | 1\}$, 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 :

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

We can deduce the Wald test :

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

According to Theorem 8.2 with $h_n = h = (0, a)$:

$$W_1 \xrightarrow{H_0} N \left(\frac{2a}{\sigma^2} \sqrt{\mathcal{A} p(1 - p)}, 1 \right). \quad (3)$$

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 a selective genotyping, we would like to define $\hat{\delta}$ such as :

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

According to Lemma 8.6 :

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

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 :

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

According to Lemma 8.6 :

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

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

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

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) :

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

It comes :

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

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

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

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

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

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} :

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

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's consider now n^* individuals for a selective genotyping experiment. T_2 has to be adapted. It comes

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

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

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

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 3.1

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 : $\beta_2^{(\mu,q,\sigma)} \leq \beta_1^{(\mu,q,\sigma)} \leq \beta_1^{(\mu,q)}$. 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 : $\beta_2^{(\mu,q,\sigma)} \leq \beta_3^{(\mu,q,\sigma)} \leq \beta_1^{(\mu,q,\sigma)}$. So, $\kappa_1 = \kappa_2 = \kappa_3$.

8.5. Proof of ii) of Theorem 3.1

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 $g(\cdot)$ be the function such as : $g(z_{\gamma_+}) = \Phi^{-1} \{ \gamma - 1 + \Phi(z_{\gamma_+}) \}$. Then, $z_{1-\gamma_-} = g(z_{\gamma_+})$.

Let $k_1(\cdot)$ be the following function : $k_1(z_{\gamma_+}) = z_{\gamma_+} \varphi(z_{\gamma_+}) - g(z_{\gamma_+}) \varphi \{ g(z_{\gamma_+}) \}$. In order to maximize κ_1 , we have to maximize the function $k_1(\cdot)$. Let $k'_1(\cdot)$, $g'(\cdot)$ and $\varphi'(\cdot)$ be respectively the derivative of $k_1(\cdot)$, $g(\cdot)$ and $\varphi(\cdot)$. We have :

$$k'_1(z_{\gamma_+}) = \varphi(z_{\gamma_+}) + z_{\gamma_+} \varphi'(z_{\gamma_+}) - g'(z_{\gamma_+}) \varphi \{ g(z_{\gamma_+}) \} - g(z_{\gamma_+}) g'(z_{\gamma_+}) \varphi' \{ g(z_{\gamma_+}) \} ,$$

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

Then, $k'_1(z_{\gamma/2}) = \varphi(z_{\gamma/2}) - \{ z_{\gamma/2} \}^2 \varphi(z_{\gamma/2}) - \varphi(z_{1-\gamma/2}) + \{ z_{1-\gamma/2} \}^2 \varphi(z_{1-\gamma/2}) = 0$. As a result, the efficiency κ_1 reaches its maximum when $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

9. Proof of Theorem 5.1

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

$$\tilde{Z} | X = \langle Z | X, \frac{Y | X - \mu_{YX}}{\sigma} \rangle \frac{Y | X - \mu_{YX}}{\sigma} + \langle Z | X, 1 \rangle 1$$

$$= r Y | X - r \mu_{YX} + \mu_{ZX} .$$

Let Z^* and μ_{ZX}^* be the two following quantities :

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

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

9.1. First strategy (Wald test using all the phenotypes)

Notation 9.1: $L^*(\mu_{Z-1}^*, \mu_{Z1}^*, \mu_Y, q_Y)$ is the likelihood for an observation (\bar{X}, Y, Z^*) and $L(\mu_Z, q_Z, \mu_Y, q_Y)$ is the likelihood for an observation (\bar{X}, Y, Z) .

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

$$L^*(\mu_{Z-1}^*, \mu_{Z1}^*, \mu_Y, q_Y) = \left\{ \frac{1-p}{\sigma} \varphi\left(\frac{y - \mu_Y + q_Y}{\sigma}\right) + \frac{p}{\sigma} \varphi\left(\frac{y - \mu_Y - q_Y}{\sigma}\right) \right\} 1_{\bar{X}=0} + \frac{p}{\sigma} \varphi\left(\frac{y - \mu_Y - q_Y}{\sigma}\right) \varphi(z^* - \mu_{Z1}^*) 1_{\bar{X}=1} + \frac{1-p}{\sigma} \varphi\left(\frac{y - \mu_Y + q_Y}{\sigma}\right) \varphi(z^* - \mu_{Z-1}^*) 1_{\bar{X}=-1} .$$

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 $\frac{\partial \log L^*}{\partial \mu_{Z1}^*} = (z^* - \mu_{Z1}^*) 1_{\bar{X}=1}$ and $\frac{\partial \log L^*}{\partial \mu_{Z-1}^*} = (z^* - \mu_{Z-1}^*) 1_{\bar{X}=-1}$, we easily obtain $\hat{\mu}_{Z-1}^*$ and $\hat{\mu}_{Z1}^*$ respective MLE of μ_{Z-1}^* and μ_{Z1}^* for n observations :

$$\hat{\mu}_{Z1}^* = \frac{1}{\sum_{j=1}^n 1_{\bar{X}_j=1}} \sum_{j=1}^n z_j^* 1_{\bar{X}_j=1} \quad \text{and} \quad \hat{\mu}_{Z-1}^* = \frac{1}{\sum_{j=1}^n 1_{\bar{X}_j=-1}} \sum_{j=1}^n z_j^* 1_{\bar{X}_j=-1} .$$

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

$$q_Z = \frac{\sigma}{2} \sqrt{1-r^2} (\mu_{Z1}^* - \mu_{Z-1}^*) + r q_Y ,$$

$$\mu_Z = \frac{\sigma}{2} \sqrt{1-r^2} (\mu_{Z1}^* + \mu_{Z-1}^*) + r \mu_Y .$$

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

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

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 $\theta_{00}^* = M^{-1}\theta_{00}$. It comes :

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

Notation 9.2: I_θ (resp. I_{θ^*}) will be the Fisher information matrix corresponding to the likelihood L (resp. L^*) and taken at point θ (resp. θ^*).

Let's calculate $I_{\theta_{00}^*}^*$:

$$\frac{\partial \log L^*}{\partial \mu_Y} \Big|_{\theta_{00}^*} = \frac{y-\mu_Y}{\sigma}, \quad \frac{\partial \log L^*}{\partial \mu_{Z-1}^*} \Big|_{\theta_{00}^*} = \left(z^* - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}} \right) 1_{\bar{X}=-1},$$

$$\frac{\partial \log L^*}{\partial \mu_{Z1}^*} \Big|_{\theta_{00}^*} = \left(z^* - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}} \right) 1_{\bar{X}=1} \quad \text{and}$$

$$\frac{\partial \log L^*}{\partial q_Y} \Big|_{\theta_{00}^*} = - \left(\frac{y-\mu_Y}{\sigma^2} \right) 1_{\bar{X}=-1} + \left(\frac{y-\mu_Y}{\sigma^2} \right) 1_{\bar{X}=1} + \left(\frac{y-\mu_Y}{\sigma^2} \right) (2p-1) 1_{\bar{X}=0}.$$

It comes

$$I_{11}^*(\theta_{00}^*) = (1-p)\gamma, \quad I_{22}^*(\theta_{00}^*) = p\gamma \quad \text{and} \quad I_{33}^*(\theta_{00}^*) = 1/\sigma^2.$$

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

Notation 9.3: 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 $\mathcal{A} = \sigma^2 \{ \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) \}$. According to Section 8.2.2, we have

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

Besides, all the other terms of $I_{\theta_{00}^*}^*$ are equal to zero.

Let $\hat{\theta}$ and $\hat{\theta}^*$ be the respective MLE of θ and θ^* , then we have $\hat{\theta} = M\hat{\theta}^*$. Since the model is regular :

$$\mathbb{V} \left\{ \sqrt{n} (\hat{\theta}^* - \theta_{00}^*) \right\} \xrightarrow{H_{0Y}H_{0Z}} I_{\theta_{00}^*}^{*-1}.$$

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

$$\mathbb{V} \left\{ \sqrt{n} (\hat{\theta} - \theta_{00}) \right\} \xrightarrow{H_{0Y}H_{0Z}} M I_{\theta_{00}^*}^{*-1} M^t \quad \text{and} \quad I_{\theta_{00}}^{-1} = M I_{\theta_{00}^*}^{*-1} M^t.$$

After some calculations, we obtain :

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

Let's define the Wald statistic W_1 :

$$W_1 = \sqrt{n} \hat{q}_Z / \sqrt{I_{22}^{-1}(\theta_{00})}.$$

The MLE \hat{q}_Z can easily be obtained using the MLE $\hat{\mu}_{Z-1}^*$, $\hat{\mu}_{Z1}^*$, and \hat{q}_Y (\hat{q}_Y can be

obtained by EM). Since the model is regular :

$$W_1 \xrightarrow{H_{0z}H_{0Y}} N(0, 1).$$

We apply Theorem 8.2 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 :

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

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

$$W_1 \xrightarrow{H_{0z}} N(0, 1) \quad \text{and} \quad W_1 \xrightarrow{H_{bz}} N\left(b/\sqrt{I_{22}^{-1}(\theta_{00})}, 1\right).$$

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 3.1. According to Theorem 3.1, 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 :

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

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

$$I_{11}^*(\theta_{00}^*) = (1-p)\gamma, \quad I_{22}^*(\theta_{00}^*) = p\gamma.$$

Besides,

$$\frac{\partial \log L^*}{\partial \mu_Y} \Big|_{\theta_{00}^*} = \frac{y - \mu_Y}{\sigma^2} \{1_{\bar{X}=-1} + 1_{\bar{X}=1}\} + \frac{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})}{\sigma(1-\gamma)} 1_{\bar{X}=0},$$

$$I_{33}^*(\theta_{00}^*) = \frac{\mathcal{A}}{\sigma^4} + \frac{\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\}^2}{\sigma^2(1-\gamma)}.$$

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

$$I_{44}^*(\theta_{00}^*) = \frac{\mathcal{A}}{\sigma^4} + (2p-1)^2 \frac{\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\}^2}{\sigma^2(1-\gamma)}.$$

Besides,

$$\frac{\partial \log L^*}{\partial \mu_Y \partial q_Y} \Big|_{\theta_{00}^*} = \frac{1}{\sigma^2} (1_{\bar{X}=-1} - 1_{\bar{X}=1}) + \frac{2p-1}{\sigma^2(1-\gamma)} \{z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) - z_{\gamma_+} \varphi(z_{\gamma_+})\} 1_{\bar{X}=0}$$

$$- \frac{2p-1}{\sigma^2(1-\gamma)^2} \{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\}^2 1_{\bar{X}=0}.$$

It comes :

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

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

$$I_{11}^{*-1}(\theta_{00}^*) = \frac{1}{(1-p)\gamma}, \quad I_{22}^{*-1}(\theta_{00}^*) = \frac{1}{p\gamma}, \quad I_{44}^{*-1}(\theta_{00}^*) = \frac{\sigma^4}{4\mathcal{A}p(1-p)}.$$

Let's define Λ such as :

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

Then :

$$I_{33}^{*-1}(\theta_{00}^*) = \frac{\Lambda}{\sigma^4} \left[\mathcal{A} + (2p-1)^2 \frac{\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\}^2}{1-\gamma} \right]$$

$$I_{34}^{*-1}(\theta_{00}^*) = \Lambda (2p-1) \left[\frac{\mathcal{A}}{\sigma^4} + \frac{\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\}^2}{\sigma^2(1-\gamma)} \right].$$

In the same way as previously :

$$I_{\theta_{00}}^{-1} = M I_{\theta_{00}^*}^{-1} M^t.$$

We obtain :

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

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

$$W_2 = \sqrt{n} \hat{q}_Z / \sqrt{I_{22}^{-1}(\theta_{00})} \xrightarrow{H_{0Z}} N(0, 1)$$

$$W_2 \xrightarrow{H_{bZ}} N\left(b / \sqrt{I_{22}^{-1}(\theta_{00})}, 1\right).$$

The MLE \hat{q}_Z can be obtained using $\hat{\mu}_{Z-1}^*$, $\hat{\mu}_{Z1}^*$ 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 Céline Delmas, Jean-Michel Elsen, Bruno Goffinet and the two referees of my PhD 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

- [1] E.S. Lander and D. Botstein, *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps*, Genetics 121 (1989), pp. 235–240.
- [2] E. Feingold, P.O. Brown, D. Siegmund, *Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent*, Am. J. Human. Genet. 53 (1993), pp. 234–251.
- [3] G.A. Churchill and R.W. Doerge, *Empirical threshold values for quantitative trait mapping*, Genetics 138 (1994), pp. 963–971.
- [4] A. Rebaï, B. Goffinet, B. Mangin, *Comparing power of different methods for QTL detection*, Biometrics 51 (1995), pp. 87–99.
- [5] ———, *Approximate thresholds of interval mapping tests for QTL detection*, Genetics 138 (1994), pp. 235–240.
- [6] C. Cierco, *Asymptotic distribution of the maximum likelihood ratio test for gene detection*, Statistics 31 (1998), pp. 261–285.
- [7] H.P. Piepho, *A quick method for computing approximate thresholds for quantitative trait loci detection*, Genetics 157 (2001), pp. 425–432.
- [8] M.N. Chang, R. Wu, S.S. Wu, G. Casella, *Score statistics for mapping quantitative trait loci*, Statistical Application in Genetics and Molecular Biology 8(1) 16 (2009).
- [9] J.M. Azaïs and M. Wschebor, *Level sets and extrema of random processes and fields*, Wiley, New-York (2009).
- [10] J.M. Azaïs, C. Delmas, C.E. Rabier, *Likelihood ratio test process for quantitative trait locus detection*, hal 00483171 (2012).
- [11] M. Lynch and B. Walsh, *Genetics and Analysis of Quantitative Traits*, Sinauer (1997).
- [12] J.I. Weller, *Quantitative Trait Loci in Animals*, Cabi Publishing (2001).
- [13] R. Wu, C.X. Ma, G. Casella, *Statistical Genetics of Quantitative Traits*, Springer (2007).

- [14] 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*, Theoretical and Applied Genetics, 73 (1987), pp. 556–562.
- [15] H. Muranty and B. Goffinet, *Selective genotyping for location and estimation of the effect of the effect of a quantitative trait locus*, Biometrics 53 (1997), pp. 629–643.
- [16] N. Rabbee, D. Speca, N. Armstrong, T. Speed, *Power calculations for selective genotyping in QTL mapping in backcross mice*, Genet. Res. Camb. 84 (2004), pp. 103–108.
- [17] A. Manichaikul, A. Palmer, S. Sen, K. Broman, *Significance thresholds for Quantitative Trait Locus mapping under selective genotyping*, Genetics 177 (2007), pp. 1963–1966.
- [18] D. Darvasi and M. Soller, *Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus*, Theoretical and Applied Genetics 85 (1992), pp. 353–359.
- [19] I. Medugorac and M. Soller, *Selective genotyping with a main and a correlated trait*, Journal of Animal Breeding and Genetics 118 (2001), pp. 285–295.
- [20] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Springer (1986).
- [21] A.W. Van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (1998).
- [22] C.E. Rabier, *PhD thesis*, Université Toulouse 3 Paul Sabatier (2010).