On statistical inference for selective genotyping

Charles-Elie Rabier

On statistical inference for selective genotyping☆

C.E. Rabier[a,b,c,*]

[a] *Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., F-31062 Toulouse Cedex 9, France.*
[b] *INRA UR631, Station d'Amélioration Génétique des Animaux, BP 52627-31326 Castanet-Tolosan Cedex, France.*
[c] *University of Wisconsin-Madison, Statistic Department, Medical Science Center, 1300 University Avenue, Madison, WI 53706-1532, USA.*

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 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

1. Introduction

1.1. Introducing our study

We address the problem of detecting a Quantitative Trait Locus, so-called QTL, that is to say a gene influencing a quantitative trait which is able to be measured (see for instance Wu et al. [2007] and Siegmund and Yakir [2007]). Statistical methods are crucial in QTL detection. They have enabled the discovery of thousands of genes in animals, humans and plants. For example, we can mention the work of Lander and Botstein [1989], Feingold et al. [1993], Churchill and Doerge [1994], Rebaï et al. [1995], Rebaï et al. [1994], Cierco [1998], Piepho [2001], Chang et al. [2009], Azaïs and Wschebor [2009], Azaïs et al. [2013].

*Corresponding author. Tel.:+1 608 265 9876; fax.:+1 608 262 0032
Email address: rabier@stat.wisc.edu (C.E. Rabier)

In this study, we focus only on a single locus which is a genetic marker. Indeed, in a recent study, Azaïs et al. [2013] have proved that the mathematical theory can be developed, without loss of generality, considering only one marker location. We suppose that the QTL is located on the genetic marker. 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 : $+1$ with probability p, and -1 with probability $1-p$. 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 QTL, 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. $p \neq 1/2$).

Y is the r.v. refering to the phenotype (i.e. the quantitative trait). We assume an "analysis of variance model" for the quantitative trait : $Y = \mu+qX+\varepsilon$ 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. We consider a sample of n observations (X_j, Y_j) i.i.d.

The problem is that genotyping, that is to say collecting the marker information X for all the individuals, is very expensive. In such a context, Lebowitz et al. [1987] proposed to genotype only the individuals who present an extreme phenotype (i.e. the smallest and the largest Y), since they noticed that most of the information about the QTL is present in the extreme phenotypes. This way, at a given power, a large increase of the number of individuals leads to a decrease of the number of individuals genotyped. Later, Lander and Botstein [1989] formalized this approach and called it "selective genotyping". The singularity of selective genotyping is that standard theory is not applicable : the model is not an "analysis of variance model" anymore, due to missing genotypes. In addition, the extremes phenotypes corresponding to one given genotype don't follow a classical Gaussian distribution any longer but a truncated Gaussian distribution. Until now, different topics have been investigated. Muranty and Goffinet [1997] focused on the estimation of the QTL effect for selective genotyping. Rabbee et al. [2004] studied different strategies for analyzing data in selective genotyping and gave the power associated to each strategy. Manichaikul et al. [2007] 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. In a very famous article, Darvasi and Soller [1992] 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, that is to say statistical tests easy to perform (e.g. without EM

algorithm) and with the best performances.

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 [1989], Darvasi and Soller [1992], Muranty and Goffinet [1997], Rabbee et al. [2004], we don't focus only on the case $p = 1/2$ which refers to a backcross population (see Wu et al. [2007] for some genetic background). On the other hand, we present a study as a function of the unknown parameters μ, q, σ since 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 ? Finally, we will address the question of the sample size required and the optimal proportion of individuals to genotype, in order to help geneticists in their experimental designs.

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 (see the examples given in Medugorac and Soller [2001]). In order to reduce the costs due to genotyping and due to phenotyping, 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 [1997] and Medugorac and Soller [2001], 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 easy and optimal statistical tests. To conclude, in the same way as what has been done in the first part, we will give advice to geneticists on the proportion of individuals to genotype.

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 [1992] and Rabbee et al. [2004]. 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 the same local asymptotic power as the one of the oracle test. Theorem 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 1, we have the same ARE whether we keep or not the non extreme 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 1 is a direct consequence of Theorem 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 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 $X = +1$). For $p = 1/2$, this result was expected : theory confirms what geneticists do in practice. However, when $p \neq 1/2$, this result is original : we didn't know how to analyze such data. Lemma 2 gives the sample size required in order to reach a given power. As expected, the number of individuals required increases when the QTL effect q decreases and when the variance σ^2 increases. The worst configuration corresponds to the unbalanced design, that is to say more individuals are required when $|p-1/2|$ increases. In Section 3.4, we address the question of the optimal percentage $\gamma^\star$ of individuals to genotype. We show that $\gamma^\star$ highly depends on the cost ratio (i.e. cost of genotyping divided by cost of phenotyping). However, if we assume that genotyping is at least two times more costly than phenotyping, according to our study, we should not genotype more than 30% of the individuals, even if the selective genotyping is not performed symmetrically. Section 3 ends with an illustration of the asymptotic results using extensive simulations. We show that we have a good agreement between the theoretical power and the empirical power even for small sample sizes and whatever the percentage γ of individuals genotyped. We also illustrate the fact that our test of comparison of means in selective genotyping is much more faster than a test which requires an EM algorithm.

The second part of this article, Section 4, deals 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 is the main result. According to this theorem, we still have to genotype symmetrically and the non extreme phenoypes Y don't bring any information for statistical inference on the QTL effect on Z. 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 3 presents the different test statistics, corresponding to the different tests studied. We leave the choice to geneticists between two optimal statistical tests. Note that these two tests require to perform respectively a Newton method and an EM algorithm in order to obtain the Maximum Likelihood Estimators (MLE). This way, contrary to the case of a selective genotyping with only one phenotype, we were not able to propose a test which can be performed easily (i.e. without EM algorithm or Newton method). Lemma 4 gives the sample size required in order to reach a given power. As expected, the number of individuals required increases when the QTL effect on Z decreases.

As previously, the unbalanced design corresponds to the worst configuration. Note also that when the selective genotyping is performed symmetrically, the more correlated the phenotypes Y and Z are, the less individuals are needed. In Section 4.4, we show that the optimal percentage $\tilde{\gamma}^\star$ of individuals to genotype highly depends on the correlation between the two phenotypes as well as on the cost ratio. Finally, we check the validity of our asymptotic study with the help of simulated data.

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 (i.e. no selective genotyping)

To begin with, 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 $p \in]0,1[$. Y is the r.v. refering to the phenotype :

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

where ε is a Gaussian r.v. zero-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 :

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

We will consider in particular, a local alternative $H_a : \ q = \frac{a}{\sqrt{n}}$ where a is a constant different from zero. Intuitively, when there are more and more observations, the power of detection becomes more and more trivial. As a result, we need to consider a QTL effect q which decreases when n increases, in order to make the problem harder.

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, which corresponds to a weighted estimate of the standard error times the QTL effect size. A proof is given in Appendix A. 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 written 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 answer the following questions relevant to 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 γ% individuals with the largest phenotypes? Or the γ% 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. $\overline{X}$ 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, 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 [1992], whereas strategies 1 and 3 are inspired by the simulation study of Rabbee et al. [2004]. Obviously, strategy 2 is the easiest to compute.

3.3. Results

Our main theorem is Theorem 1 :

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 Appendix B. 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, $\frac{1}{n}\sum 1_{\overline{X}_j \neq 0} \to \gamma$. 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. It comes out 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 1, that a small 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 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 : the theory confirms what geneticists do in practice. However, when $p \neq 1/2$, this result is original : we didn't know how to analyze 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,*

$\mathcal{A} = \sigma^2 \ \left\{\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\}$, $\hat{\mathcal{A}}_1 = \frac{1}{n}\sum_{j=1}^n (Y_j - \overline{Y})^2 1_{\overline{X}_j \neq 0}$
$\hat{\mathcal{A}}_3 = \frac{1}{n}\sum_{j=1}^n (Y_j - \hat{\mu}_3)^2 1_{\overline{X}_j \neq 0}$, $\hat{\sigma}^2$ *is given in Section 2.2.*

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

As previously mentioned, 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 [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, 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 $\overline{Y}$, because this estimator is $\sqrt{n}$ consistent. In the same way, we can also replace $\hat{\mathcal{A}}_3$ by $\hat{\mathcal{A}}_1$. Then, 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 analyze data, we should use this test and genotype symmetrically. In Table 1, is given the CPU time as a function of the statistical test used. As expected, the comparison of means is largely faster than the test which requires the use of an EM algorithm.

	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
QTL number	W_1	T_2	W_1	T_2	W_1	T_2	W_1	T_2
1	0.0012	0.0005	0.0020	0.0005	0.0041	0.0005	0.0090	0.0006
1000	1.8369	0.1228	2.7871	0.1267	5.1131	0.1384	9.7150	0.1535

Table 1: CPU time (in seconds), for a selective genotyping with one phenotype, as a function of the test statistic used, and as a function of the number of individuals n, and the number of QTL to analyse ($q = 0.3$, $\mu = 0$, $\sigma = 1$, $\gamma_+/\gamma = 1/2$, $p = 1/2$, EM ends when the absolute difference between old estimates and new estimates is less than 0.01).

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 whether 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 before. For strategy 2, when only q is unknown, we have to keep in mind that $\mathcal{A}$ is known. Indeed, according to the proof of Theorem 1 (see Appendix B.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 $\hat{\mu}_3$ is replaced by μ and $\hat{\mathcal{A}}_3$ by $\mathcal{A}$. Note that when we consider $(\mu,\ q)$ 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 (i.e. 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) $\forall p \in]0,1[\ \ \kappa_1$, 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 the same results as found 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 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 symmetrically for all strategies.

3.4. Sample size required and optimal percentage of individuals to genotype

In order to help geneticists in their experimental designs, we propose to focus here, on the sample size required to reach a given power β, considering a test at the α level.

Lemma 2. *If we consider a statistical model with* 3 *unknown parameters (μ, q, σ), the sample size required to reach a given power β, considering a test at the α level, is the quantity $n_{\alpha,\beta}$ which verifies :*

$$n_{\alpha,\beta} = \frac{\sigma^4\ (z_\alpha - z_\beta)^2}{4\ q^2\ \mathcal{A}\ p(1-p)}\ .$$

Note that this lemma assumes $n_{\alpha,\beta}$ large and q small. The proof is given in Appendix C.

On the other hand, another important aspect in practice, is how to choose the percentage γ of individuals to genotype. In other words, what should be the threshold for the scientist to decide whether the phenotype is extreme ? Let c_X (resp. c_Y) denote the cost of genotyping (resp. phenotyping) one individual, and let C denote the ratio c_X/c_Y. In order to give an answer to this important aspect, we have to minimize the following function :

$$F(\gamma) = n_{\alpha,\beta}\ \gamma\ c_X\ +\ n_{\alpha,\beta}\ c_Y = \frac{\sigma^4\ (z_\alpha - z_\beta)^2\ c_Y\ \ (\gamma\ C\ +\ 1)}{4\ q^2\ \mathcal{A}\ p(1-p)}\ .$$

In other words, in order to find the optimal γ, called $\gamma^\star$, we have to minimize the quantity

$$F^\star(\gamma) = \frac{\gamma\ C\ +\ 1}{\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})}$$

which is independent of p, q and σ. This quantity is the same as formula (29) of Darvasi and Soller [1992] as soon as we genotype symmetrically (i.e. $\gamma_+ = \gamma_- = \gamma/2$). We refer to the next section for differences between our study and the previous study of Darvasi and Soller [1992]. In Figure 1, is represented $\gamma^\star$ as a function of the cost ratio C and as a function of the ratio γ_+/γ. According to the figure, as mentioned in Darvasi and Soller [1992], the optimal proportion selected is a very sensitive function of C. It also depends on how the selective genotyping has been performed. Let us consider for instance a cost ratio C equal to 2. Then, if the selective genotyping is performed symmetrically, $\gamma^\star$ is equal to 0.322, that is to say we have to genotype the 16.10% individuals with the largest phenotypes and the 16.10% individuals with the smallest phenotypes (same result as in Darvasi and Soller [1992]). If for some biological reasons, the selective genotyping can not be performed symmetrically (when, for example, it is easier to genotype the individuals with the largest phenotypes), $\gamma^\star$ is equal to 0.308 (resp. 0.283) when $\gamma_+/\gamma = 3/4$ (resp. $\gamma_+/\gamma = 7/8$). On the other hand, if we want to genotype only the individuals with the largest phenotypes (or only the smallest phenotypes), $\gamma^\star$ is now equal to 1, that is to say we should genotype all the individuals. Note that, our Figure 1 was obtained assuming the ratio C independent of γ (as in Figure 4 of Darvasi and Soller [1992]), but as mentioned by these authors, we can also imagine a ratio C which depends on γ (cf. comments below their formula (29)).

3.5. 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 $Y \notin [S_-, S_+]$ (i.e. $\overline{X} \neq 0$). 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 γ% individuals with extreme phenotypes.

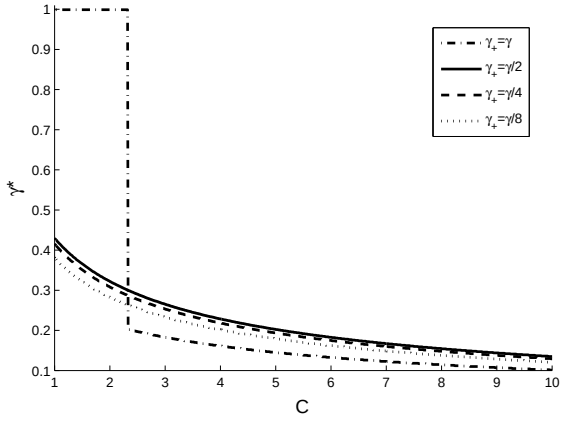


Figure 1: Optimal γ, for a selective genotyping with one phenotype, as a function of the cost ratio $C = c_X/c_Y$ and as a function of the ratio γ_+/γ.

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 using classical central limit theorem : they should use Lindeberg-Feller central limit theorem. Furthermore, they use approximations about thresholds (see their formulae (1) and (2)).

Note that in their paper, Darvasi and Soller [1992] suppose symmetry, i.e. $\mathbb{P}(Y > S_+) = \mathbb{P}(Y < S_-) = \gamma/2$. Anyway, if we consider the same configuration as Darvasi and Soller [1992] (i.e. $p = 1/2$ and symmetry), our treatment gives the same ARE as that obtained in formula (27) of Darvasi and Soller [1992]. However, it is important to recall 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 symmetry, and μ and σ unknown. In the same way, our sample size required to reach a given power (cf. Lemma 2) is more general than the one present in formula (26) of Darvasi and Soller [1992] : our formula is obtained without any symmetry, assuming σ and μ unknown, and considering the genotype frequency p.

3.6. Illustration

In this Section, we propose to illustrate our theoretical results. We consider one-sided tests at the 5% level and the situation where all the parameters are unknown. To begin with, Figure 2 represents the efficiency with respect

to the oracle test. This efficiency corresponds to Theorem 1. Note that the efficiency does not depend on the QTL effect (see Theorem 1) and the genotype frequency p. We study here the efficiency as a function of the percentage γ of individuals genotyped 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 graph, genotyping symmetrically yields to the best results. The worst is obtained when genotyping only the largest phenotypes (see $\gamma_+/\gamma = 1$) or genotyping only the smallest phenotypes (same curve as the one for $\gamma_+/\gamma = 1$). Obviously, it can be seen that when $\gamma = 1$, all the efficiencies are equal to one, since all the individuals are genotyped.

In Figure 3, the sample size required in order to reach a power β of 60%, 70%, or 80% is shown as a function of the genotype frequency p at the QTL (left-side) and as a function of the percentage γ of individuals genotyped (right-side). Note that we consider here the optimal configuration, that is to say the selective genotyping is performed symmetrically. On the left-side, we genotype only 30% of the population. We can notice that the sample size required is minimal when $p = 1/2$. For instance, for $p = 1/2$, we need respectively 116, 151 and 200 individuals in order to reach a power of 60%, 70%, and 80%. The worst configuration corresponds to the unbalanced design : the sample size required increases when $|p-1/2|$ increases. On the right-side of Figure 3, as expected, the sample size required decreases with γ. For instance, in order to reach a power of 60%, the sample sizes required are respectively 140, 116, 105, 98 and 95 if we consider respectively $\gamma = 0.2$, $\gamma = 0.3$, $\gamma = 0.4$, $\gamma = 0.5$ and $\gamma = 0.6$. This way, it seems on this example, that the gain in power is not substantial when we genotype more than 40% of the population. It is consistent with results of Figure 1 (cf. curve under symmetry). However, we have to keep in mind that results of Figure 1 are more general since they deal with a cost ratio.

In Figures 4, 5, 6 and 7, we propose to check the validity of our theoretical results using simulated data (10000 samples considered). We consider the statistical test based on the statistic T_2. It is quite 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 1). In Figure 4, we consider a QTL effect $q = 0.2$, $p = 1/2$ (backcross in genetics), and different values of n : $n = 30$, $n = 50$, $n = 100$ and $n = 200$. The selective genotyping is performed symmetrically and we consider different values of the percentage γ of individuals genotyped. According to Figure 4, the empirical power is close to the theoretical power even for small sample sizes. In Figure 5, we study different coefficients of variation q/σ. When the coefficient of variation is equal to 0.3, a slight difference occurs between the empirical power and the theoretical power, for $n = 50$ and $n = 30$. We recall that our theoretical results are established under the hypothesis of a small q and a large n values. We obtain the same kind of conclusions when we consider $p = 1/4$ (cf. Figure 6). Finally in Figure 7, the selective genotyping is not performed symmetrically

anymore. There is still a good agreement between the theoretical power and the empirical power when $\gamma_+/\gamma = 1/4$, $\gamma_+/\gamma = 1/8$ and $\gamma_+/\gamma = 1$.

To conclude, in Figure 8 and in Tables 2 and 3, we study how an estimated p affects respectively the power, the percentage of false positives and the accuracy of the QTL effect estimate. We use $\sum 1_{\overline{X}_j=1} / \sum 1_{\overline{X}_j \neq 0}$ as an estimator of p in our test statistic T_2. According to Figure 8, for large sample sizes (see $n = 100$ and $n = 200$), the power seems to be unaffected. However, for small sample sizes (see $n = 30$ and $n = 50$), we denote a loss of power, in particular when p is small (cf. $p = 0.1$). In the same way, according to Table 2, the statistical test seems to be too conservative, for small values of p and small sample sizes. In Table 3, we present a comparison between the QTL effect estimates when p is known and when p is unknown. Note that when p is known, the estimated QTL effect can be obtained according to a classical EM algorithm. When dealing with the case p unknown, we used the same EM algorithm but the proportions were estimated. The true value of q considered here is 0.3. According to Table 3, as expected, the EM algorithm presents very good performances when p is known. However, a slight bias is observed when p is unknown. In particular, a small p leads to underestimate q, whereas a large p leads to overestimate q.

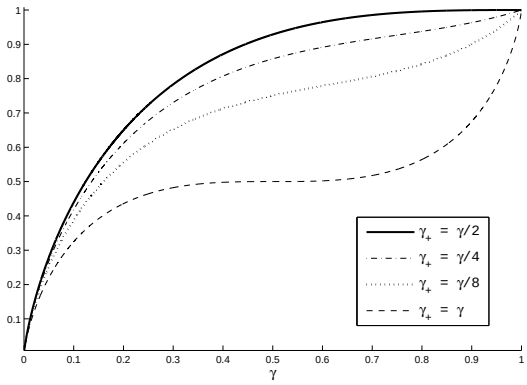


Figure 2: Efficiency, for a selective genotyping with one phenotype, as a function of the percentage γ of individuals genotyped and as a function of the ratio γ_+/γ.

4. Selective genotyping with two correlated phenotypes

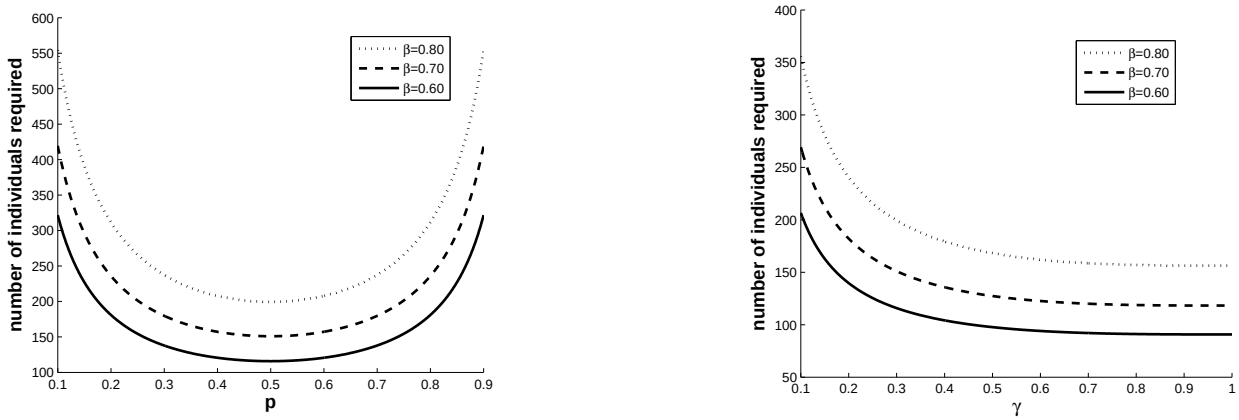
Sometimes, it is difficult to measure the phenotype of interest : it can be expensive or it can require a lot of work. In such a situation, a second phenotype correlated to the phenotype of interest, can be measured more easily (see the examples given by Medugorac and Soller [2001]). In the following, Z will denote the phenotype of interest and Y will refer to the second phenotype. In order to reduce costs due to genotyping and due to phenotyping, a selective

p	$n = 30$	$n = 50$	$n = 100$	$n = 200$
0.1	1.87%	2.67%	4.10%	5.19%
	(5.48%)	(5.02%)	(5.14%)	(5.09%)
0.2	3.50%	4.44%	4.93%	4.90%
	(5.07%)	(4.80%)	(5.12%)	(5.16%)
0.3	4.08%	5.20%	4.90%	5.26%
	(4.96%)	(5.25%)	(4.89%)	(4.87%)
0.4	5.30%	5.31%	5.14%	4.95%
	(5.21%)	(4.74%)	(4.78%)	(5.13%)
0.5	5.42%	5.16%	4.83%	4.90%
	(4.61%)	(4.67%)	(4.97%)	(5.20%)

Table 2: Percentage of false positives when p is unknown, for a selective genotyping with one phenotype, as a function of n and p (10000 samples, $q = 0$, $\mu = 0$, $\sigma = 1$, $\gamma_+/\gamma = 1/2$). In brackets is given the percentage of false positives when p is known.

	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
p	EM	EM(p?)	EM	EM(p?)	EM	EM(p?)	EM	EM(p?)
0.1	0.2885	0.28	0.2935	0.2758	0.2955	0.2718	0.2990	0.2708
	(0.3405)	(0.2108)	(0.2704)	(0.1610)	(0.1921)	(0.1105)	(0.1318)	(0.0789)
0.2	0.2964	0.3065	0.2917	0.2997	0.2967	0.3006	0.2988	0.2996
	(0.2577)	(0.2288)	(0.2024)	(0.1714)	(0.1416)	(0.1210)	(0.0997)	(0.0833)
0.3	0.2894	0.3249	0.2972	0.3215	0.2998	0.3249	0.2993	0.3207
	(0.2179)	(0.2456)	(0.1611)	(0.1851)	(0.1235)	(0.1255)	(0.0878)	(0.0893)
0.4	0.2901	0.3473	0.2964	0.3475	0.2976	0.3473	0.2999	0.3409
	(0.2067)	(0.2626)	(0.1610)	(0.1979)	(0.1140)	(0.1355)	(0.0815)	(0.0947)
0.5	0.2979	0.3620	0.2982	0.3625	0.2972	0.3620	0.2991	0.3558
	(0.2019)	(0.2831)	(0.1572)	(0.2148)	(0.1119)	(0.1457)	(0.0780)	(0.1027)

Table 3: Estimated values of the QTL effect q when p is known (EM) and when p is unknown (EM p?), for a selective genotyping with one phenotype, as a function of n and p (10000 samples, $q = 0.3$, $\mu = 0$, $\sigma = 1$, $\gamma_+/\gamma = 1/2$). In brackets is given the standard deviation.



$\gamma = 0.3$ $p = 1/2$

Figure 3: Number of individuals required in order to reach a power β of 60%, 70%, or 80%, for a selective genotyping with one phenotype, as a function of the genotype frequency p at the QTL (left side) and as a function of the percentage γ of individuals genotyped (right side). Other parameters are $\gamma_+ = \gamma_- = \gamma/2$, $q = 0.2$ and $\sigma = 1$.

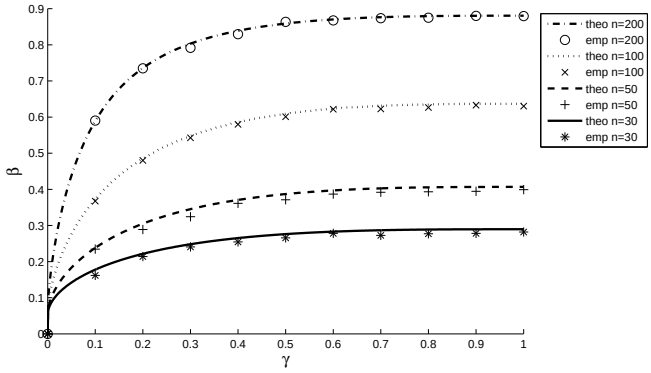
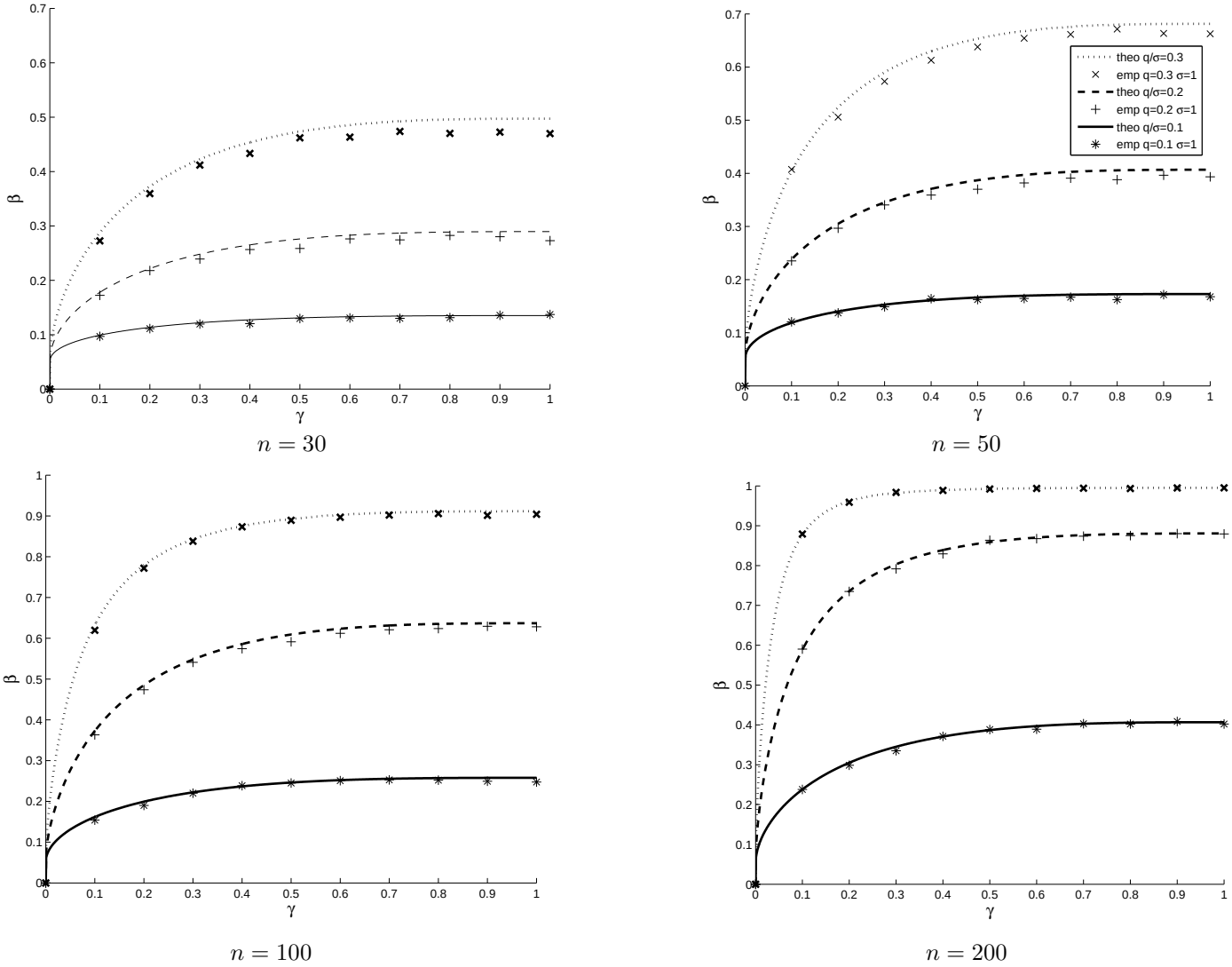


Figure 4: Theoretical power (theo) and empirical power (emp), for a selective genotyping with one phenotype, as a function of the number of individuals n and as a function of the percentage γ of individuals genotyped (10000 samples, $q = 0.2$, $\sigma = 1$, $\gamma_+ = \gamma_- = \gamma/2$, $p = 1/2$, $\mu = 0$).

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

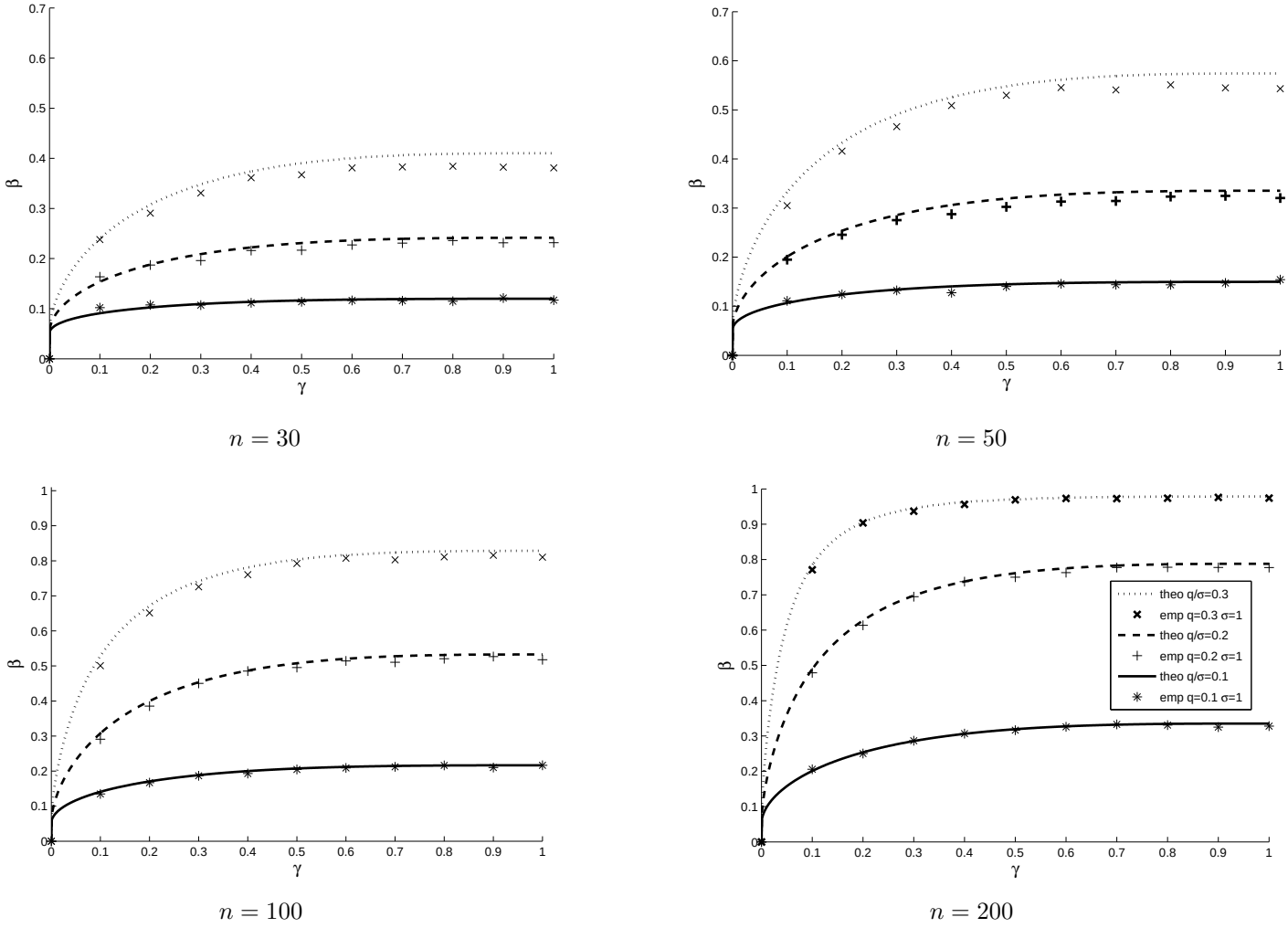


$n = 30$ $n = 50$

$n = 100$ $n = 200$

Figure 5: Theoretical power (theo) and empirical power (emp), for a selective genotyping with one phenotype, as a function of the coefficient of variation q/σ and as a function of the percentage γ of individuals genotyped (10000 samples, $\gamma_+ = \gamma_- = \gamma/2$, $p = 1/2$, $\mu = 0$).

in Muranty and Goffinet [1997] and Medugorac and Soller [2001]. However, the theory of statistical inference is still missing, since Muranty and Goffinet [1997] focused only on the estimation of the QTL effects and Medugorac and Soller [2001] focused on the power of the design using approximations.



$n = 30$ $n = 50$

$n = 100$ $n = 200$

Figure 6: Same graphs as in Figure 5, except that we consider now $p = 1/4$.

4.1. Oracle situation : model and oracle statistical test (μ_Z, q_Z)

As previously, we begin by considering the situation with no missing genotypes. We present here our model and the optimal oracle test, which will be considered as our reference test for our future study on selective genotyping. X is still the r.v. corresponding to the genotype at the QTL. We consider the following model :

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

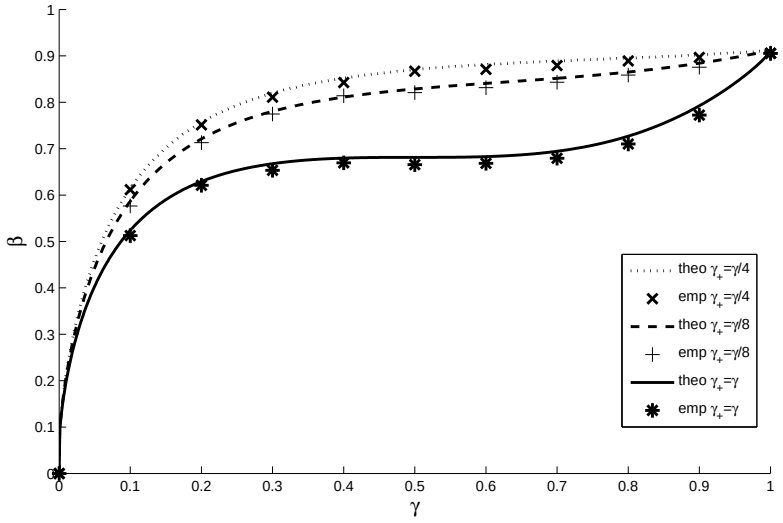


Figure 7: Theoretical power (theo) and empirical power (emp), for a selective genotyping with one phenotype, as a function of the percentage γ of individuals genotyped and as a function of the ratio γ_+/γ (10000 samples, $q = 0.3$, $\mu = 0$, $\sigma = 1$, $n = 100$, $p = 1/2$).

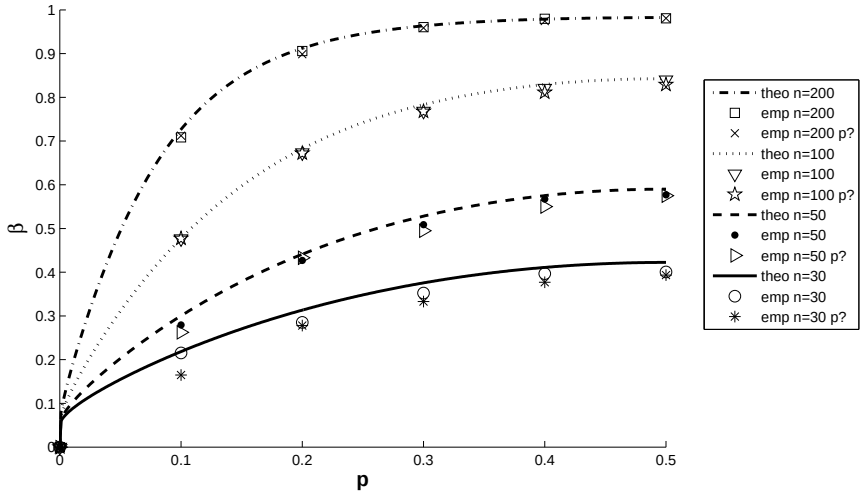


Figure 8: Theoretical power (theo), empirical power with p known (emp), and empirical power with p unknown (emp p?), for a selective genotyping with one phenotype, as a function of the genotype frequency p (10000 samples, $q = 0.3$, $\mu = 0$, $\sigma = 1$, $p = 1/2$, $\gamma = 0.3$, $\gamma_+ = \gamma_- = \gamma/2$).

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.

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 (c.f. Section 2.2 and Appendix A), 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 :

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

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

4.2. Model and strategies under selective genotyping

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 on q_Z.
- 2. we keep only the extreme phenotypes Y (i.e. the phenotypes for which the genotypes are available) and we perform a Wald test on q_Z.

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 the same kind of questions as

for a selective genotyping with one phenotype. Note that we don't consider the comparison of means on Z : this test won't be optimal since it only uses the phenotypes Z and does not use explicitely the phenotypes Y. As a consequence, here, strategy 2 is analogous to strategy 3 of the first part.

4.3. Results

Our main theorem is Theorem 2, which is the analogue 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 analogue 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 Appendix D. According to Theorem 2, 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 2 establishes the relationship between the ARE of selective genotyping with one and two phenotypes.

On the other hand, as expected, the ARE increases with γ. Besides, the ARE increases with r^2 as soon as $z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-}) > 0$. This is true in most cases. When the selective genotyping is performed symmetrically, the ARE increases with r^2 since this quantity is equal to $2\ z_{\gamma/2}\ \varphi(z_{\gamma/2})$ which is greater than 0. However, for instance, if we genotype only the individuals with the largest phenotypes Y and if γ is greater than 50%, then the ARE does not increase with r^2. We now introduce Lemma 3 which presents the different tests corresponding to the different strategies.

Lemma 3. *If we consider a statistical model with 4 unknown parameters (μ_Z, q_Z, μ_Y, q_Y) and if we are under H_{0Y} or under 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, we refer to the proof of Theorem 2 in Appendix D. So, according to Lemma 3, 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 Appendix D), we obtain

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

where

$$\hat{\mu}_{Z1}^\star = \left\{ \sum_{j=1}^n \frac{(Z_j - rY_j)1_{\overline{X}_j=1}}{\sigma\,\sqrt{1-r^2}} \right\} / \sum_{j=1}^n 1_{\overline{X}_j=1}$$

$$\hat{\mu}_{Z-1}^\star = \left\{ \sum_{j=1}^n \frac{(Z_j - rY_j)1_{\overline{X}_j=-1}}{\sigma\,\sqrt{1-r^2}} \right\} / \sum_{j=1}^n 1_{\overline{X}_j=-1}\,.$$

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 a selective genotyping with two correlated phenotypes, are not so simple. As said previously, the test of comparison of means on Z won't be optimal. Indeed, according to the formula above, the MLE of q_Z depends on the phenotypes Y. As a consequence, we leave to geneticists the choice between the two statisticals tests, which are optimal and asymptotically equivalent.

We introduce now Corollary 3 which is the analogue 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}*:*

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

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

$$\textit{iii)} \quad \tilde{\kappa}_1 = \tilde{\kappa}_2 \Leftrightarrow p = \frac{1}{2}$$

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

where κ_1 *and* κ_3 *are the quantities of Corollary 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 symmetrically. In other words, we have to use strategy 1 and genotype symmetrically. 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 symmetrically anymore when q_Y is known : we can genotype only the individuals with the largest (or smallest) phenotypes. Note that as previously, the non-extreme phenotypes Y do not bring any information for statistical inference on q_Z. 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 a known quantity in the model studied. Indeed, the MLE $\hat{q}_Z$ of q_Z, that we have to compute in order to perform our Wald tests in selective genotyping (for both strategies), depends on q_Y which is a known quantity in the model studied (cf. Section of the 4.2 of "Online Ressource 1"). The oracle test does not use the fact that this quantity is known, and as a consequence, some information is lost.

We finally propose to explain more in details Corollary 4, focusing on the extreme cases. First, when $r = 0$, the Wald tests in selective genotyping (for the

two strategies) are both a comparison of means on Z. To prove this, just consider the expression of $\hat{q}_Z$ with $r = 0$ (cf. Section of the 4.2 of "Online Ressource 1"). Then, since $Z \mid X$ is not affected by the fact that $Y \mid X$ is extreme, the Wald tests correspond to a comparison of means on Z, but only based on an average of $n\mathbb{P}(Y \notin [S_-, S_+])$ individuals randomly sampled (contrary to the oracle test which is based on n individuals). This way, the efficiency is equal to $\mathbb{P}(Y \notin [S_-, S_+])$ when $r = 0$. Then, let us consider the extreme case $\mathbb{P}(Y \notin [S_-, S_+]) = 1$, that is to say the genotype X and the phenotype Z of each individual are collected. When $r = 0$, the efficiency is equal to 1 since the Wald test in selective genotyping is the same as the oracle test. When $r > 0$, the efficiency is equal to $\frac{1}{1-r^2}$ which is greater than 1. As a consequence, the Wald test in selective genotyping becomes more powerful than the oracle test. As previously said, it is due to the fact that q_Y is known.

4.4. Sample size required

In the same way as what has been done before for a selective genotyping with only one phenotype, we propose to focus on the sample size required to reach a given power β, considering a test at the α level.

Lemma 4. *If we consider a statistical model with* 4 *unknown parameters* (μ_Z, q_Z, μ_Y, q_Y)*, under* H_{0Y} *and under* H_{aY}*, the sample size required to reach a given power* β*, considering a test at the* α *level, is the quantity* $\tilde{n}_{\alpha,\beta}$ *which verifies :*

$$\tilde{n}_{\alpha,\beta} = \left(\frac{z_\alpha - z_\beta}{q_Z}\right)^2 \left\{\frac{\sigma^2\,(1-r^2)}{4\,p\,(1-p)\,\gamma} + \frac{\sigma^4\,r^2}{4\,p\,(1-p)\,\mathcal{A}}\right\}\,.$$

Note that this lemma assumes $\tilde{n}_{\alpha,\beta}$ large and q_Z small (q_Y has also to be small if we are under H_{aY}). The proof is given in Appendix E.

As previously, the next question is how to choose the percentage γ of individuals to genotype. We recall that c_X (resp. c_Y) denotes the cost of genotyping (resp. phenotyping) one individual. In the same way, c_Z will denote the cost of collecting the phenotype Z for one individual. We will use the notation C_{XY}, instead of C, to denote the ratio c_X/c_Y. C_{ZY} will denote the ratio c_Z/c_Y. Then, we have to minimize the following function :

$$\begin{aligned}\tilde{F}(\gamma) &= \tilde{n}_{\alpha,\beta}\,\gamma\;(c_X + c_Z)\;+\;\tilde{n}_{\alpha,\beta}\,c_Y \\ &= \left(\frac{z_\alpha - z_\beta}{q_Z}\right)^2 \left\{\frac{\sigma^2\,(1-r^2)}{4\,p\,(1-p)\,\gamma} + \frac{\sigma^4\,r^2}{4\,p\,(1-p)\,\mathcal{A}}\right\} c_Y\,\{\gamma(C_{XY} + C_{ZY}) + 1\}\,.\end{aligned}$$

In other words, in order to find the optimal γ, called $\tilde{\gamma}^\star$, we have to minimize the quantity

$$\tilde{F}^\star(\gamma) = \left\{\frac{(1-r^2)}{\gamma} + \frac{r^2}{\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})}\right\}\{\gamma(C_{XY} + C_{ZY}) + 1\}$$

which is independent of p, q_Z, q_Y, and σ.

In Figure 9, is represented $\tilde{\gamma}^\star$ as a function of the cost ratios C_{XY} and C_{ZY} and as a function of r. The selective genotyping is performed symmetrically. According to the figure, the optimal proportion selected is a very sensitive function of C_{XY}, C_{ZY} and r. We can notice that $\tilde{\gamma}^\star$ decreases when r increases. In other words, the better the proxy Y is for Z, the less individuals have to be genotyped. Note also that $\tilde{\gamma}^\star$ decreases when C_{ZY} increases : since the phenotype Z of each genotyped individual has to be collected, we do not want to genotype too many individuals when collecting Z is expensive. In the same way, $\tilde{\gamma}^\star$ decreases when C_{XY} increases : we do not want to genotype too many individuals when genotyping is expensive.

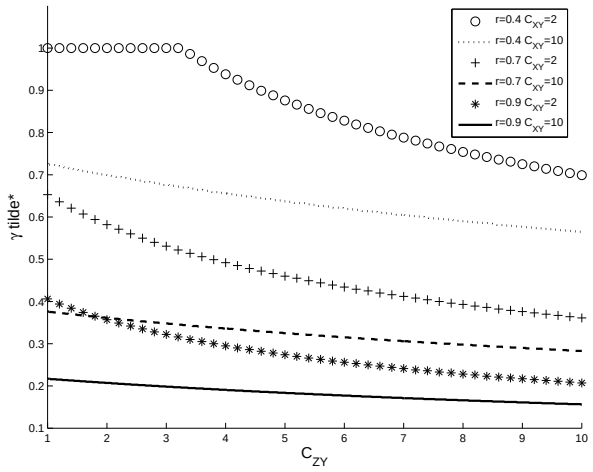


Figure 9: Optimal γ, for a selective genotyping with two phenotypes, as a function of the cost ratios C_{XY} and C_{ZY} and as a function of r ($\gamma_+ = \gamma_- = \gamma/2$).

4.5. Illustration

In this Section, we propose to illustrate our theoretical results about a selective genotyping with two phenotypes. We consider one-sided tests at the 5% level and we consider the model with 4 unknown parameters (μ_Z, q_Z, μ_Y, q_Y). Figure 10 represents the efficiency with respect to the oracle test (cf. Theorem 2). According to the graph, 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 $\gamma_+/\gamma = 1$).

In Figure 11, is represented the sample size required in order to reach a power of 60%, 70% and 80%, as a function of p, γ and r. Note that we consider here the optimal configuration, that is to say the selective genotyping is performed symmetrically. As for a selective genotyping with one phenotype, the sample size required is minimal when $p = 1/2$ and the sample size required decreases

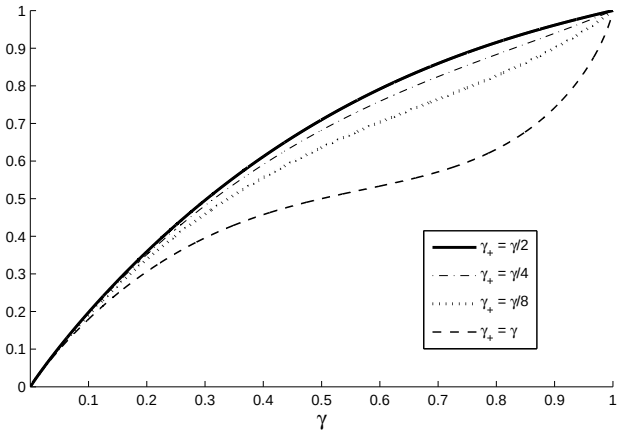
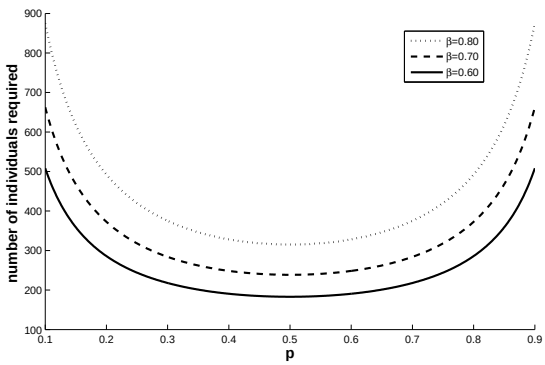
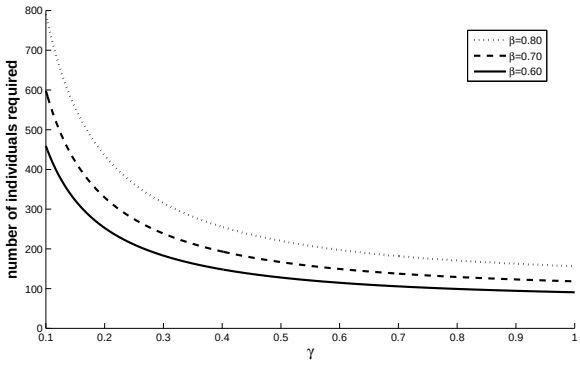


Figure 10: Efficiency, for a selective genotyping with two phenotypes, as a function of the percentage γ of individuals genotyped and as a function of the ratio γ_+/γ.

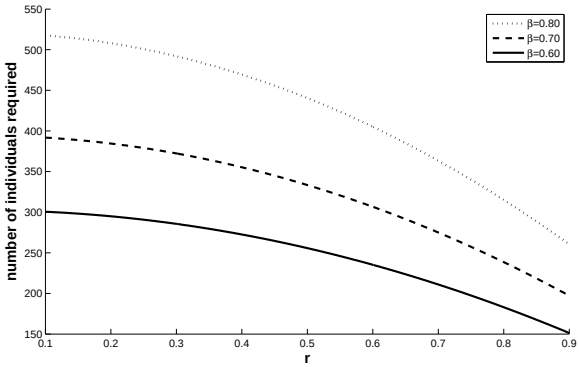
when γ increases. Note also that the sample size required decreases when r increases : this design is interesting when the proxy Y is good for Z. In Figures 12, 13 and 14, we propose to compare the theoretical power and the empirical power (10000 samples). We focus here on the test based on the test statistic $\tilde{W}_1$ of Lemma 3. It has to be reminded that, in order to obtain 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 Appendix D for details). So, we decided here to use the EM algorithm. According to Figure 12 (obtained under symmetry), there is a good agreement between the theoretical power and the empirical power even for small values of n and whatever the values of r and q_Y. We also observe a good agreement between the theoretical power and the empirical power when γ varies (see Figure 13). However, when γ and n are very small (see $\gamma = 0.1$, $\gamma = 0.2$ for $n = 30$ and $n = 50$), there are some differences. Finally, in Figure 14, the selective genotyping is not performed symmetrically anymore. The empirical power ($n = 100$) and the theoretical power are similar whatever the ratio γ_+/γ is. To conclude, in Table 4, we focus on the null hypothesis, that is to say the situation where the QTL has no effect on the phenotype Z (i.e. $q_Z = 0$). We compute the percentage of false positives for different values of r, n and q_Y, under symmetry and assuming $\gamma = 0.3$. We can see that, for $n = 100$ and $n = 200$, the percentage of false positives is always close to 5%. However, when $n = 30$ (or $n = 50$) and $q_Y = 0$, the percentage of false positives is found to be overestimated.



$\gamma = 0.3$ and $r = 0.8$

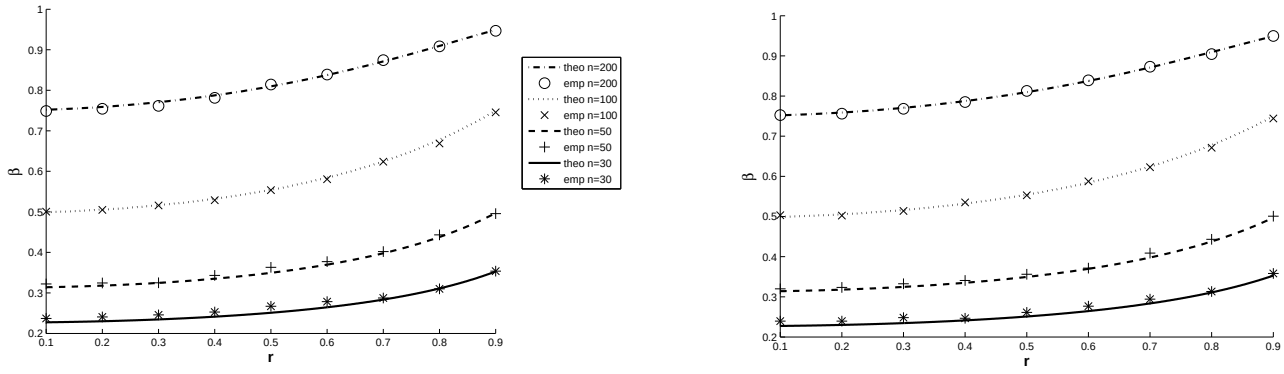


$p = 1/2$ and $r = 0.8$



$\gamma = 0.3$ and $p = 1/2$

Figure 11: Number of individuals required in order to reach a power β of 60%, 70%, or 80%, for a selective genotyping with two phenotypes, as a function of the genotype frequency p (upper left-side), as a function of the percentage γ of individuals genotyped (upper right-side), and as a function of r (bottom). Other parameters are $\gamma_+ = \gamma_- = \gamma/2$, $q_Z = 0.2$ and $\sigma = 1$.

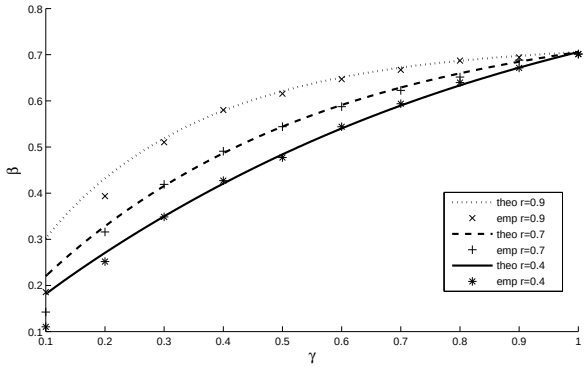


$q_Y = 0$ $q_Y = 0.2$

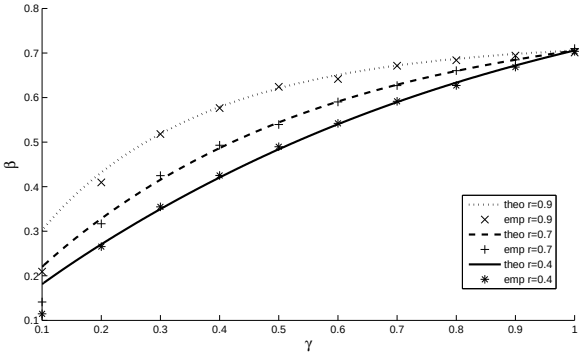
Figure 12: Theoretical power (theo) and empirical power (emp), for a selective genotyping with two phenotypes, as a function of the number of individuals n and as function of r (10000 samples, $q_Z = 0.3$, $\mu_Z = 0$, $\mu_Y = 0$, $\sigma = 1$, $p = 1/2$, $\gamma = 0.3$, $\gamma_+ = \gamma_- = \gamma/2$).

	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
r	$q_Y = 0$	$q_Y = 0.3$	$q_Y = 0$	$q_Y = 0.3$	$q_Y = 0$	$q_Y = 0.3$	$q_Y = 0$	$q_Y = 0.3$
0.1	6.31%	5.81%	5.56%	5.41%	5.26%	4.98%	5.33%	4.58%
0.2	6.85%	5.85%	6.06%	5.61%	5.46%	4.66%	4.70%	4.85%
0.3	7.21%	6.02%	5.88%	5.23%	5.22%	4.74%	4.97%	4.34%
0.4	6.77%	5.80%	5.87%	4.95%	5.25%	5.01%	5.16%	4.58%
0.5	6.87%	5.10%	5.87%	4.78%	5.32%	4.66%	4.89%	4.88%
0.6	6.37%	5.02%	5.78%	4.96%	5.45%	4.51%	5.12%	4.38%
0.7	6.35%	4.80%	5.61%	4.87%	5.31%	4.62%	5.13%	4.65%
0.8	6.03%	5.08%	5.36%	4.66%	5.32%	4.84%	5.32%	4.73%
0.9	5.41%	5.11%	5.60%	4.29%	5.21%	4.65%	4.66%	4.64%

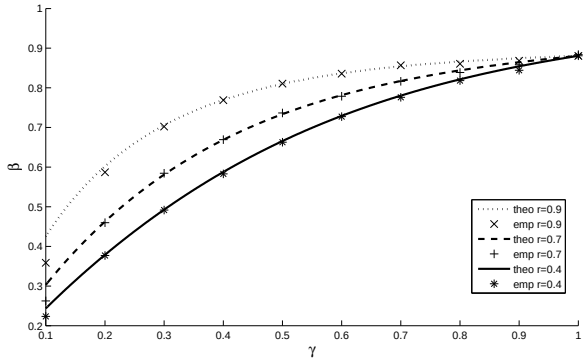
Table 4: Percentage of false positives as a function of q_Y, r and n (10000 samples, $q_Z = 0$, $\mu_Z = 0$, $\mu_Y = 0$, $\sigma = 1$, $p = 1/2$, $\gamma = 0.3$, $\gamma_+/\gamma = 1/2$).



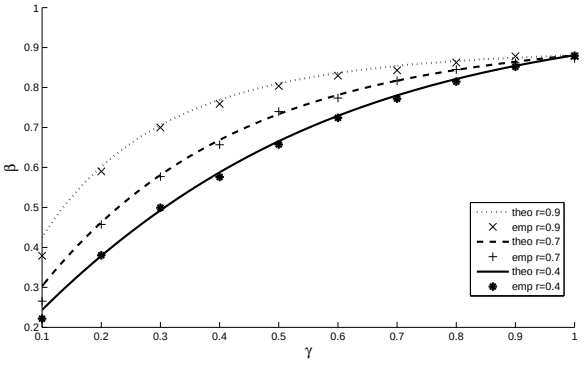
$q_Y = 0$ and $n = 30$



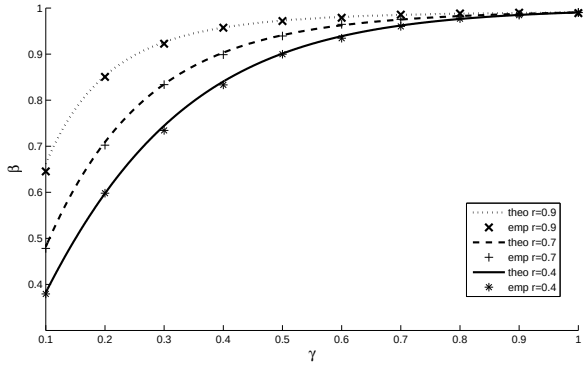
$q_Y = 0.2$ and $n = 30$



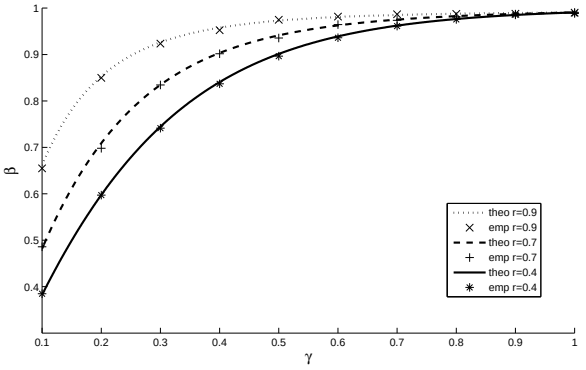
$q_Y = 0$ and $n = 50$



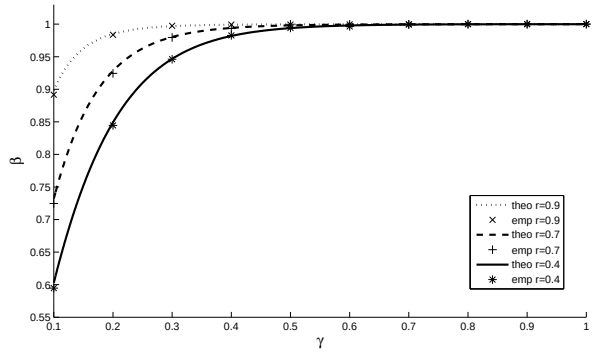
$q_Y = 0.2$ and $n = 50$



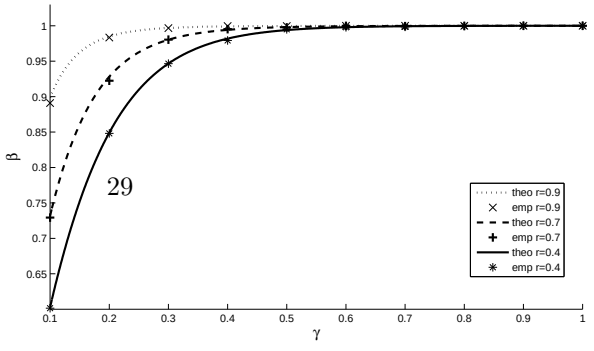
$q_Y = 0$ and $n = 100$



$q_Y = 0.2$ and $n = 100$



$q_Y = 0$ and $n = 200$



$q_Y = 0.2$ and $n = 200$

Figure 13: Theoretical power (theo) and empirical power (emp), for a selective genotyping with two phenotypes, as a function of the percentage γ of individuals genotyped (10000 samples, $q_Z = 0.4$, $\mu_Z = 0$, $\mu_Y = 0$, $\sigma = 1$, $p = 1/2$, $\gamma_+ = \gamma_- = \gamma/2$).

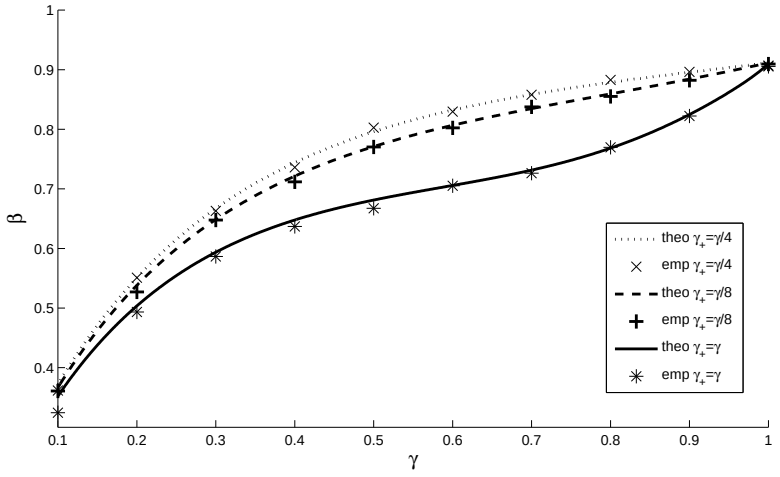


Figure 14: Theoretical power (theo) and empirical power (emp), for a selective genotyping with two phenotypes, as a function of the percentage γ of individuals genotyped and as a function of the ratio γ_+/γ (10000 samples, $q_Z = 0.3$, $\mu_Z = 0$, $q_Y = 0.2$, $\mu_Y = 0$, $r = 0.8$, $\sigma = 1$, $n = 100$, $p = 1/2$).

5. Acknowledgements

I am very grateful to my PhD advisor Jean-Marc Azaïs for valuable guidance and advice in completing this work. I thank Laurent Bordes, Céline Delmas, Jean-Michel Elsen, Bruno Goffinet 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.

Appendix A. 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

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

As a consequence, $\mathbb{V}_{H_0}\left(\frac{\eta}{p}\,1_{X=1}\;-\;\frac{\eta}{1-p}\,1_{X=-1}\right)=\frac{\sigma^2}{p(1-p)}$.
Besides, under the local alternative H_a :

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

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

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

We remark that $\mathbb{V}_{H_a}\left(\frac{\eta}{p}\,1_{X=1}\;-\;\frac{\eta}{1-p}\,1_{X=-1}\right)\to\mathbb{V}_{H_0}\left(\frac{\eta}{p}\,1_{X=1}\;-\;\frac{\eta}{1-p}\,1_{X=-1}\right)$.
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 : $\widetilde{T}\overset{H_0}{\to}N\,(0,\;1)$ and $\widetilde{T}\overset{H_a}{\to}N\left(\frac{2a\;\sqrt{p\;(1-p)}}{\sigma},\;1\right)$.
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 :

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

We have

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

Notations 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 $\sum_{j=1}^{n}\;\frac{1}{1-p}1_{X_j=-1}-\frac{1}{p}\,1_{X_j=1}=O_P(\sqrt{n})$.
As a result,

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

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

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

So, 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 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 :

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

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

Appendix B. Proof of Theorem 1

Notations 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.*

Appendix B.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 identically 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 $h_n \to h$, as $n \to +\infty$, we have :*

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

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

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):

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

Notations 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 [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 :

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

Then, according to Le Cam's third lemma :

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

This gives the result.

Appendix B.2. First strategy (Wald test using all the phenotypes)

Appendix B.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 counting measure $\times$ the Lebesgue measure.

Notations 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 :*

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

Notations 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* $q_0 = \mathbb{P}(\overline{X} = 0)$.

As a result, $\mathbb{P}\{i \mid i\} = \Phi\left(\frac{S_- - \mu - iq}{\sigma}\right) + 1 - \Phi\left(\frac{S_+ - \mu - iq}{\sigma}\right)$ 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.
So,

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

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

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 :

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

Appendix B.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 5. *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) $\mathbb{E}\left\{(V-\mu)^2 1_{V\in[S_-,\ S_+]}\right\} = \sigma^2 - \sigma^2\mathbb{P}(V \notin [S_-,\ S_+]) - \sigma(S_+ - \mu)\ \varphi\left(\frac{S_+-\mu}{\sigma}\right)$
$+\ \sigma\ (S_- - \mu)\ \varphi\left(\frac{S_--\mu}{\sigma}\right)$.

Notations 6. *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* $\mathcal{A} = \sigma^2\ \left\{\gamma\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\right\}$.

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

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

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

As a consequence $I_{22}(\theta_0) = \frac{\mathcal{A}}{\sigma^4}\ +\ \frac{(2p-1)^2}{\sigma^4}\ (\sigma^2 - \mathcal{A})$. 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, $P_{H_0}\{-1 \mid -1\} = P_{H_0}\{1 \mid 1\}$. So, we have $I_{12}(\theta_0) = \frac{1}{\sigma^2}(2p-1)$. 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 :

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

We can deduce the Wald test :

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

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

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

Appendix B.3. Second strategy (comparison of means based on the extreme phenotypes)

Appendix B.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 (A.1), $\mathbb{E}_{H_a}(\hat{\delta}) = 2q$ 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 :

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

According to Lemma 5 :

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

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 5 :

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

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

$$T_2 = \frac{\sum_{j=1}^{n} \frac{1}{p}(Y_j - \mu)1_{\overline{X}_j=1} \; - \; \frac{1}{1-p}(Y_j - \mu)1_{\overline{X}_j=-1}}{\sqrt{\frac{n\ \mathcal{A}}{p(1-p)}}} \overset{H_0}{\to} N(0,\ 1)\ . \qquad \text{(B.2)}$$

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) :

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

As a consequence :

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

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

$$T_2 \overset{H_a}{\to} N\left(\frac{2a}{\sigma^2}\ \sqrt{\mathcal{A}\ p(1-p)},\ 1\right)\ . \qquad \text{(B.3)}$$

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 Appendix A). 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.

Appendix B.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 $n^\star$ individuals for a selective genotyping experiment. T_2 has to be adapted. We now have

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

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 $q = a/\sqrt{n}$) :

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

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 $a > 0$) :

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

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

$$\kappa_2 = \gamma \; + \; z_{\gamma_+} \varphi(z_{\gamma_+}) \; - \; z_{1-\gamma_-} \; \varphi(z_{1-\gamma_-}) . \tag{B.4}$$

Appendix B.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 (B.3) and (B.1) : $\beta_2^{(\mu,q,\sigma)} = \beta_1^{(\mu,q)}$. Besides, by definition : $\beta_2^{(\mu,q,\sigma)} \leq \beta_1^{(\mu,q,\sigma)} \leq \beta_1^{(\mu,q)}$. So, $\beta_1^{(\mu,q,\sigma)} = \beta_2^{(\mu,q,\sigma)}$. 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, $\kappa_1 = \kappa_2 = \kappa_3$.

Appendix B.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 $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

Appendix C. Proof of Lemma 2

We recall that we consider here three unknown parameters (μ, q, σ). Since the powers are exactly the same for all the strategies, $\beta^{(\mu,q,\sigma)}$ will denote the corresponding power. According to Lemma 1, $\beta^{(\mu,q,\sigma)} = 1-\Phi\left(z_\alpha - 2a\sqrt{\mathcal{A}p(1-p)}/\sigma^2\right)$. As a consequence, for n large and q small, we have to find the value of n which verifies the following relationship :

$$z_\alpha - \frac{2\,q\,\sqrt{n\,\mathcal{A}\,p\,(1-p)}}{\sigma^2} = \Phi^{-1}(1-\beta)\;.$$

As a consequence,

$$\sqrt{n} = \frac{\sigma^2\;\left\{z_\alpha - \Phi^{-1}(1-\beta)\right\}}{2\,q\,\sqrt{\mathcal{A}\,p\,(1-p)}} \quad \text{and} \quad n = \frac{\sigma^4\,(z_\alpha - z_\beta)^2}{4\,q^2\,\mathcal{A}\,p\,(1-p)}\;.$$

Appendix D. Proof of Theorem 2

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 $Z \mid X$ on $Y \mid X$ which will be called $\underline{Z} \mid X$, we define the following scalar product, for 2 r.v. U_1 and U_2 which take value in $\mathbb{R}$: $< U_1\;,\; U_2 > = \mathbb{E}\,[U_1 U_2]$. We have :

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

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

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

This way, $Z^\star \mid X \sim N(\mu^\star_{ZX}, 1)$. By construction, $(Z - \underline{Z}) \mid X$ and $\underline{Z} \mid X$ 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.

Appendix D.1. First strategy (Wald test using all the phenotypes)

Notations 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 :

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

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 :

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

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 $\theta^\star_{00} = M^{-1}\theta_{00}$. As a result :

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

Notations 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 $I^\star_{\theta^\star_{00}}$:

$\frac{\partial \log L^\star}{\partial \mu_Y} \, |_{\theta^\star_{00}} = \frac{y-\mu_Y}{\sigma}$, $\frac{\partial \log L^\star}{\partial \mu^\star_{Z-1}} \, |_{\theta^\star_{00}} = (z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}}) \ 1_{\overline{X}=-1}$,

$\frac{\partial \log L^\star}{\partial \mu^\star_{Z1}} \, |_{\theta^\star_{00}} = (z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}}) \ 1_{\overline{X}=1}$ and

$\frac{\partial \log L^\star}{\partial q_Y} \, |_{\theta^\star_{00}} = -\left(\frac{y-\mu_Y}{\sigma^2}\right) \ 1_{\overline{X}=-1} \ + \ \left(\frac{y-\mu_Y}{\sigma^2}\right) \ 1_{\overline{X}=1} \ + \ \left(\frac{y-\mu_Y}{\sigma^2}\right) (2p-1) \ 1_{\overline{X}=0}$.

We finally obtain

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

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

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

We remind that $\mathcal{A} = \sigma^2\ \left\{\gamma\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\right\}$. According to Appendix B.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 respectively the 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, $\sqrt{n}\ (\hat{\theta} - \theta_{00}) = \sqrt{n}\ M\ (\hat{\theta}^\star - \theta^\star_{00})$. We have :

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

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's define the Wald statistic $\tilde{W}_1$:

$$\tilde{W}_1 = \sqrt{n}\ \ \hat{q}_Z / \sqrt{I^{-1}_{22}(\theta_{00})}\ .$$

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 :

$$\tilde{W}_1\ \ \overset{H_{0Z}H_{0Y}}{\to}\ \ N(0,\ 1)\ .$$

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 :

$$\begin{aligned}
\tilde{W}_1\ &\overset{H_{0Z}H_{aY}}{\to}\ N\,(0,\ 1) \\
\tilde{W}_1\ &\overset{H_{bZ}H_{0Y}}{\to}\ N\left(b/\sqrt{I^{-1}_{22}(\theta_{00})}\ ,\ 1\right) \\
\tilde{W}_1\ &\overset{H_{bZ}H_{aY}}{\to}\ N\left(b/\sqrt{I^{-1}_{22}(\theta_{00})}\ ,\ 1\right)\ .
\end{aligned}$$

As a consequence, under either hypothesis on Y (i.e. the null hypothesis or the local alternative), we always have :

$$\tilde{W}_1\ \overset{H_{0Z}}{\to}\ N(0,\ 1) \quad \text{and} \quad \tilde{W}_1\ \overset{H_{bZ}}{\to}\ N\left(b/\sqrt{I^{-1}_{22}(\theta_{00})}\ ,\ 1\right)\ .$$

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 $\tilde{\kappa}_1$.

Appendix D.2. Second strategy (Wald test using only the extreme phenotypes Y)

In this case, the likelihood is :

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

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,

$$\frac{\partial \log L^\star}{\partial \mu_Y}\ |_{\theta^\star_{00}} = \frac{y-\mu_Y}{\sigma^2}\ \left\{1_{\overline{X}=-1} + 1_{\overline{X}=1}\right\} \ + \ \frac{\varphi\left(z_{1-\gamma_-}\right) - \varphi\left(z_{\gamma_+}\right)}{\sigma\ (1-\gamma)}\ 1_{\overline{X}=0} \ ,$$

$$I^\star_{33}(\theta^\star_{00}) = \frac{\mathcal{A}}{\sigma^4} \ + \ \frac{\left\{\varphi\left(z_{1-\gamma_-}\right) - \varphi\left(z_{\gamma_+}\right)\right\}^2}{\sigma^2\ (1-\gamma)} \ .$$

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,

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

As a result :

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

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's 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 :

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

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

In the same way as previously :

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

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 $\tilde{W}_2$ and its asymptotic laws (same proof as for the first strategy)

$$\tilde{W}_2 = \sqrt{n}\;\; \hat{q}_Z / \sqrt{I^{-1}_{22}(\theta_{00})} \;\; \stackrel{H_{0Z}}{\rightarrow} \;\; N(0,\,1)$$

$$\tilde{W}_2 \;\; \stackrel{H_{bZ}}{\rightarrow} \;\; N\left(b/\sqrt{I^{-1}_{22}(\theta_{00})}\;\;,\;\; 1\right) .$$

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.

Appendix E. Proof of Lemma 4

We recall that we consider here four unknown parameters (μ_Z, q_Z, μ_Y, q_Y). Since the powers are exactly the same for all the strategies, $\beta^{(\mu_Z,q_Z,\mu_Y,q_Y)}$ will

denote the corresponding power. According to Lemma 3, under H_{0Y} and H_{aY}, $\beta^{(\mu_Z,q_Z,\mu_Y,q_Y)} = 1 - \Phi\left(z_\alpha - b\left\{\frac{\sigma^2\ (1-r^2)}{4\ p\ (1-p)\ \gamma} + \frac{\sigma^4\ r^2}{4\ p\ (1-p)\ \mathcal{A}}\right\}^{-1/2}\right)$. As a consequence, for n large and q_Z small (q_Y has also to be small if we are under under H_{aY}), we have to find the value of n which verifies the following relationship :

$$z_\alpha - q_Z\sqrt{n}\left\{\frac{\sigma^2\ (1-r^2)}{4\ p\ (1-p)\ \gamma} + \frac{\sigma^4\ r^2}{4\ p\ (1-p)\ \mathcal{A}}\right\}^{-1/2} = \Phi^{-1}(1-\beta)\ .$$

As a consequence,

$$n = \left(\frac{z_\alpha - z_\beta}{q_Z}\right)^2\ \left\{\frac{\sigma^2\ (1-r^2)}{4\ p\ (1-p)\ \gamma} + \frac{\sigma^4\ r^2}{4\ p\ (1-p)\ \mathcal{A}}\right\}.$$

References

Azaïs, J.M., Wschebor, M., 2009. Level sets and extrema of random processes and fields, Wiley, New-York.

Azaïs, J.M., Delmas, C., Rabier, C.E., 2013. Likelihood Ratio Test process for Quantitative Trait Locus detection. To appear in Statistics.

Chang, M. N., Wu, R., Wu, S. S., Casella, G., 2009. Score statistics for mapping quantitative trait loci. Stat. Appl. Genet. Mol. Biol. 8, 1, 16.

Churchill, G.A., Doerge, R, 1994. Empirical threshold values for quantitative trait mapping. Genetics. 138, 963-971.

Cierco, C., 1998. Asymptotic distribution of the maximum likelihood ratio test for gene detection. Statistics. 31, 261-285.

Darvasi, D., Soller, M., 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. 85, 353-359.

Feingold, E., Brown, P.O., Siegmund, D., 1993. Gaussian models for genetic linkage analysis using complete high-resolution maps of idendity by descent. Am. J. Hum. Genet. 53, 234-251.

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. Theor. Appl. Genet. 73, 556-562.

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

Manichaikul, A., Palmer, A., Sen, S., Broman, K., 2007. Significance thresholds for Quantitative Trait Locus mapping under selective genotyping. Genetics. 177, 1963-1966.

Medugorac, I., Soller, M., 2001. Selective genotyping with a main and a correlated trait. J. Anim. Breed. Genet. 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.

Piepho, H.P., 2001. A quick method for computing approximate thresholds fo quantitative trait loci detection. Genetics. 157, 425-432.

Rabbee, N., Speca, D., Armstrong, N., Speed, T., 2004. Power calculations for selective genotyping in QTL mapping in backcross mice. Genet. Res. Camb. 84, 103-108.

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

Rebaï, A., Goffinet, B., Mangin, B., 1994. Approximate thresholds of interval mapping tests for QTL detection. Genetics. 138, 235-240.

Rebaï, A., Goffinet, B., Mangin, B., 1995. Comparing power of different methods for QTL detection. Biometrics. 51, 87-99.

Siegmund, D., Yakir, B., 2007. The statistics of gene mapping, Springer, New York.

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

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

Online Resource 1 for the article , “On statistical inference for selective genotyping"

Charles-Elie Rabier

Institut de Mathématiques de Toulouse, Toulouse, France.
INRA UR631, Auzeville, France.
Statistic Department, University of Wisconsin-Madison, USA.
rabier@stat.wisc.edu

1. The oracle statistical test (μ, q, σ) is the most powerful test we can perform

First, we focus on the Wald test for a statistical model with two unknown parameters (μ, q). Let $\theta = (\mu, q)$ and $\theta_0 = (\mu, 0)$.

The likelihood L for an observation (X, Y) is :

$$L = \frac{p}{\sigma}\,\varphi\left(\frac{y-\mu-q}{\sigma}\right)1_{X=1} \;+\; \frac{1-p}{\sigma}\,\varphi\left(\frac{y-\mu+q}{\sigma}\right)1_{X=-1}$$

where $\varphi(x)$ denotes the density of a standard normal distribution taken at point x.
Let's compute the Fisher information matrix.

Notations : 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.

We have

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

$$\frac{\partial^2 logL}{\partial \mu^2} \;=\; -\frac{1}{\sigma^2} \quad \text{, so} \quad I_{11}(\theta_0) \;=\; \frac{1}{\sigma^2}\,.$$

Besides,

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

$$\frac{\partial^2 logL}{\partial q^2} \;=\; -\frac{1}{\sigma^2} \quad \text{, so} \quad I_{22}(\theta_0) \;=\; \frac{1}{\sigma^2}\,.$$

Finally,

$$\frac{\partial logL}{\partial \mu\,\partial q} = -\frac{1}{\sigma^2}1_{X=1} \;+\; \frac{1}{\sigma^2}1_{X=-1}\,,$$

$$I_{12}(\theta_0) = \frac{1}{\sigma^2}\,(2p-1)\,.$$

As a consequence :

$$I_{\theta_0} = \begin{pmatrix} \frac{1}{\sigma^2} & \frac{1}{\sigma^2}\,(2p-1) \\ \frac{1}{\sigma^2}\,(2p-1) & \frac{1}{\sigma^2} \end{pmatrix}\,.$$

The Maximum Likelihood Estimator (MLE) $\hat{\theta}$ is such $\hat{\theta} = (\hat{\mu}, \hat{q})$ where $\hat{\mu}$ and $\hat{q}$ are the respective MLE of μ and q.
We want to test :

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

Since the model is regular, we use Theorem 3 (cf. Appendix B.1 of the article) :

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

We have :

$$I_{\theta_0}^{-1} = \frac{\sigma^2}{4(1-p)p}\begin{pmatrix} 1 & 1-2p \\ 1-2p & 1 \end{pmatrix}\,.$$

So, the test statistic W is such as :

$$W = \sqrt{n}\,\frac{2\,\sqrt{(1-p)p}}{\sigma}\,\hat{q} \ \overset{H_0}{\to}\ N(0,\,1)\,.$$

We apply Theorem 3 (cf. Appendix B.1 of the article) with $h_n = h = (0, a)$:

$$W \overset{H_a}{\to} N\left(\frac{2a\,\sqrt{p(1-p)}}{\sigma},\,1\right)\,.$$

By definition, the Wald test (μ, q, σ) has a greater or the same power as the oracle test (μ, q, σ). Besides, we remark that the Wald test (μ, q) has the same asymptotic laws as the oracle test (μ, q, σ) (see Section 2.2 of the article). As a consequence, the Wald test (μ, q, σ) and the oracle test (μ, q, σ) have the same asymptotic laws.

2. Proof of Corollary 1

2.1. *Oracle statistical test* (q)

We suppose here that μ and σ are known. So, we consider a statistical model with one unknown parameter (q). If we have a look at the oracle statistical test (μ, q, σ) (cf Section 2.2 of the article), we would like to consider the same test statistic T but with μ and σ known. As a consequence, we have :

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

$$T \overset{H_0}{\to} N(0,\ 1)\ ,$$

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

However, this test is not the best to perform. If we consider a Wald test, and using the calculations of Section 1 of this document, the MLE $\hat{q}$ of q :

$$\hat{q} = \frac{1}{n}\ \sum_{j=1}^{n} (Y_j - \mu) 1_{X_j=1}\ -\ (Y_j - \mu) 1_{X_j=-1}\ .$$

Using $I_{22}(\theta_0)$ of Section 1, the Wald statistic is :

$$W = \frac{\sum_{j=1}^{n} (Y_j - \mu) 1_{X_j=1}\ -\ (Y_j - \mu) 1_{X_j=-1}}{\sigma\ \sqrt{n}}\ .$$

Let $\tilde{q}$ be the MLE of q if we consider only one observation. We have

$$\tilde{q} = (Y - \mu) 1_{X=1}\ -\ (Y - \mu) 1_{X=-1}\ ,$$

$$\mathbb{E}_{H_0}\left(\tilde{q}\right) = 0 \quad \text{and} \quad \mathbb{E}_{H_0}\left(\tilde{q}^2\right) = \sigma^2\ .$$

According to the central limit theorem :

$$W \overset{H_0}{\to} N(0,\ 1)\ .$$

Besides :

$$\mathbb{E}_{H_a}\left(\tilde{q}\right) = q \quad \text{and} \quad \mathbb{V}_{H_a}\left(\tilde{q}\right) \to \sigma^2\ .$$

According to the central limit theorem :

$$W \overset{H_a}{\to} N\left(\frac{a}{\sigma},\ 1\right)\ .$$

This Wald test will be our oracle test (q).
Note that we have :

$$\{W = T\}\ \Leftrightarrow\ \left\{p = \frac{1}{2}\right\}\ .$$

2.2. First strategy

According to the Fisher information matrix in the proof of Theorem 1 of the article (cf. Appendix B), the Wald test for a model (q) and corresponding to the first strategy is :

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

Using Theorem 3 of the article with $h_n = h = a$, we obtain :

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

Considering the oracle test (q), we obtain easily the efficiency κ_1 for the first strategy :

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

2.3. Second strategy

According to formula (B.2) of the proof of Theorem 1 of the article, the test statistic T_2 for a model (q) is :

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

The asymptotic laws were :

$$T_2 \overset{H_a}{\to} N(0,\ 1) \quad , \quad T_2 \overset{H_a}{\to} N\left(\frac{2a}{\sigma^2}\ \sqrt{\mathcal{A}\ p(1-p)},\ 1\right)\ .$$

As a result, the efficiency is :

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

2.4. Third strategy

We focus here on the Wald test under a statistical model (q). $\theta = (q)$ and $\theta_0 = (0)$. We suppose $\gamma \neq 1$. The likelihood for an observation is :

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

Besides,

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

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

As a result,

$$I_{\theta_0} = \frac{\mathcal{A}}{\sigma^4}\ +\ \frac{\left\{(1-2p)\ \varphi\left(z_{\gamma_+}\right)\ +\ (2p-1)\ \varphi\left(z_{1-\gamma_-}\right)\right\}^2}{\sigma^2(1\ -\gamma)}\ . \tag{1}$$

Let $\hat{q}$ the MLE of q. It can be obtained using a Newton method.
We have :

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

The Wald test statistic W_3 corresponding to strategy 3 is :

$$W_3 = \sqrt{n}\ \left[\frac{\mathcal{A}}{\sigma^4}\ +\ \frac{\left\{(1-2p)\ \varphi\left(z_{\gamma_+}\right)\ +\ (2p-1)\ \varphi\left(z_{1-\gamma_-}\right)\right\}^2}{\sigma^2(1\ -\gamma)}\right]^{1/2}\ \hat{q}\ \ \overset{H_0}{\to}\ \ N(0,\ 1)\ .$$

According to Theorem 3 of the article with $h_n = h = a$, we have :

$$W_3 \ \stackrel{H_a}{\rightarrow} \ N(a\sqrt{I_{\theta_0}},\ 1)\ .$$

We easily obtain the efficiency correponding to strategy 3 :

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

2.5. How to genotype in order to maximize the efficiencies ?

We want to answer the following question : how to choose γ_+ and γ_- to maximize the efficiencies ? We remind that $\gamma_+ + \gamma_- = \gamma$.
We study here the different strategies. In order to make the calculation easier, we first consider strategy 2. We remind that :

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

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

In order to maximize κ_2, we have to maximize the following function called $k_2(.)$:

$$k_2(z_{\gamma_+}) = z_{\gamma_+}\varphi(z_{\gamma_+}) - g(z_{\gamma_+})\ \varphi\left\{g(z_{\gamma_+})\right\}\ .$$

Let $k_1'(.)$, $k_2'(.)$, $k_3'(.)$ and $g'(.)$ be respectively the derivative of $k_1(.)$, $k_2(.)$, $k_3(.)$ and $g(.)$. We have :

$$k_2'(z_{\gamma_+}) = \varphi(z_{\gamma_+}) \ + \ z_{\gamma_+}\varphi'(z_{\gamma_+}) \ - \ g'(z_{\gamma_+})\ \varphi\left\{g(z_{\gamma_+})\right\} \ - \ g(z_{\gamma_+})\ g'(z_{\gamma_+})\ \varphi'\left\{g(z_{\gamma_+})\right\}\ ,$$

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

Then :

$$k_2'(z_{\gamma/2}) = \varphi(z_{\gamma/2}) \ - \ \left\{z_{\gamma/2}\right\}^2\ \varphi(z_{\gamma/2}) \ - \ \varphi(z_{1-\gamma/2}) \ + \ \left\{z_{1-\gamma/2}\right\}^2\ \varphi(z_{1-\gamma/2}) = 0\ .$$

As a consequence, the efficiency κ_2 of strategy 2 is maximum when $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

Let's focus now on κ_1. We remind that :

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

We have to maximize the fuction $k_1(.)$ defined such as :

$$k_1(z_{\gamma_+}) = k_2(z_{\gamma_+}) \ - \ k_2(z_{\gamma_+})\ (2p-1)^2\ .$$

We have :

$$k_1'(z_{\gamma_+}) = k_2'(z_{\gamma_+}) \ - \ k_2'(z_{\gamma_+})\ (2p-1)^2\ .$$

Since $k_2'(z_{\gamma/2}) = 0$, we have $k_1'(z_{\gamma/2}) = 0$.

So, the efficiency κ_1 is maximum for $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

Let's focus now on the efficiency κ_3 corresponding to the third strategy. We remind that :

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

We have to maximize the function $k_3(.)$ such as :

$$k_3(z_{\gamma_+}) = k_2(z_{\gamma_+}) \ + \ \frac{(2p-1)^2}{1 \ - \gamma} \ \left[\ \varphi\left\{g(z_{\gamma_+})\right\} \ - \ \varphi(z_{\gamma_+}) \ \right] \ .$$

We have :

$$k_3'(z_{\gamma_+}) = k_2'(z_{\gamma_+}) \ + \ \frac{(2p-1)^2}{1 \ - \gamma} \ 2 \ \left[\ g'(z_{\gamma_+})\varphi'\left\{g(z_{\gamma_+})\right\} \ - \ \varphi'(z_{\gamma_+}) \ \right] \ \left[\ \varphi\left\{g(z_{\gamma_+})\right\} - \varphi(z_{\gamma_+}) \ \right] \ .$$

Then $k_3'(z_{\gamma/2}) = 0$. We deduce that the efficiency κ_3 is maximum for $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

3. Proof of Corollary 3

3.1. *Oracle statistical test* (q_Z)

We consider a statistical model with one parameter (q_Z). The oracle statistical test is easy to obtain : we just have to consider what has been done in Section 2.1 of this document. So, the oracle test is based on the following statistic :

$$W = \frac{\sum_{j=1}^{n}(Z_j - \mu_Z)1_{X_j=1} \ - \ (Z_j - \mu_Z)1_{X_j=-1}}{\sigma \ \sqrt{n}} \ .$$

And we have

$$W \overset{H_{0Z}}{\rightarrow} N\left(0, \ 1\right) \ \text{ and } \ W \overset{H_{bZ}}{\rightarrow} N\left(\frac{b}{\sigma}, \ 1\right) \ .$$

3.2. *First strategy*

We remind that we study a statistical model with two parameters (q_Z, q_Y). The calculations are largely inspired of the proof of Theorem 2 of the article.
Let $\hat{q}_Z$ be the MLE of q_Z. The Wald statistic is :

$$\tilde{W}_1 = \sqrt{\frac{n}{\Delta}} \ \hat{q}_Z \quad \text{with } \ \Delta = \frac{\sigma^2 \ (1 - r^2)}{\mathbb{P}_{H_{0Y}}(Y \notin [S_-, S_+])} \ + \ \frac{\sigma^4 \ r^2}{\mathcal{A} + (2p-1)^2(\sigma^2 - \mathcal{A})} \ .$$

According to Theorem 3 of the article :

$$\begin{aligned} \tilde{W}_1 \ &\overset{H_{0Z}}{\rightarrow} \ N(0, \ 1) \\ \tilde{W}_1 \ &\overset{H_{bZ}}{\rightarrow} \ N\left(\frac{b}{\sqrt{\Delta}}, \ 1\right) \ . \end{aligned}$$

We can calculate the efficiency $\tilde{\kappa}_1$ of this test, with respect to the oracle test (q_Z). We have :

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

where κ_1 is the efficiency introduced in Corollary 1 of the article.

3.3. Second strategy

We use the same notations as in the proof of Theorem 2 of the article. Besides, let's define the following quantities :

$$\mu_Z^\star = \frac{\mu_Z - r\,\mu_Y}{\sigma\,\sqrt{1-r^2}} \qquad \text{and} \qquad q_Z^\star = \frac{q_Z - r\,q_Y}{\sigma\,\sqrt{1-r^2}}.$$

As a consequence, $\mu_{ZX}^\star = \mu_Z^\star + q_Z^\star X$ and $Z^\star \mid X \sim N(\mu_Z^\star + q_Z^\star X,\ 1)$.
Since we keep only the extreme phenotypes Y, the likelihood for an observation $(\overline{X}, Y, Z^\star)$ is :

$$L^\star(q_Z^\star, q_Y) = \mathbb{P}(\overline{X}=0)\ 1_{\overline{X}=0}$$
$$+\ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\varphi(z^\star-\mu_Z^\star-q_Z^\star)\ 1_{\overline{X}=1}\ +\ \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\varphi(z^\star-\mu_Z^\star+q_Z^\star)\ 1_{\overline{X}=-1}.$$

Let $\theta = (q_Z, q_Y)$ and $\theta^\star = (q_Z^\star, q_Y)$. Let M be the matrix such as $\theta = M\theta^\star$:

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

As in the proof of Theorem 2 of the article, I_θ is the Fisher information matrix taken at point θ and $I_{\theta^\star}^\star$ is the Fisher information matrix taken at point $\theta^\star$.
Besides, let's define $\theta_{00} = (0,0)$ and $\theta_{00}^\star = M^{-1}\theta_{00}$. So, $\theta_{00}^\star = (0,0)$. We have :

$$\frac{\partial \log L^\star}{\partial q_Z^\star}\ |_{\theta_{00}^\star} = (z^\star - \mu_Z^\star)\ 1_{\overline{X}=1}\ -\ (z^\star-\mu_Z^\star)\ 1_{\overline{X}=-1}\ ,$$
$$\frac{\partial^2 \log L^\star}{\partial q_Z^\star \partial q_Y}\ |_{\theta_{00}^\star} = 0\ ,$$

$$I_{11}^\star(\theta_{00}^\star) = \gamma \quad \text{and} \quad I_{12}^\star(\theta_{00}^\star) = 0\ .$$

Besides, according to formula (1) of Section 2 of this document :

$$I_{22}^\star(\theta_{00}^\star) = \frac{\mathcal{A}}{\sigma^4}\ +\ \frac{\left\{(1-2p)\ \varphi\left(z_{\gamma_+}\right)\ +\ (2p-1)\ \varphi\left(z_{1-\gamma_-}\right)\right\}^2}{\sigma^2\ (1-\gamma)}\ .$$

We have :

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

After some calculations, we obtain :

$$I_{11}^{-1}(\theta_{00}) = \frac{\mathcal{A}}{\sigma^4} + \frac{(2p-1)^2\left\{\varphi\left(z_{\gamma_+}\right) - \varphi\left(z_{1-\gamma_-}\right)\right\}^2}{\sigma^2\ (1-\gamma)}\ .$$

We define the Wald test statistic $\tilde{W}_2$:

$$\tilde{W}_2 = \sqrt{\frac{n}{I_{11}^{-1}(\theta_{00})}}\ \hat{q}_Z\ .$$

Since the model is regular, we have :

$$\tilde{W}_2 \overset{H_{0Z}}{\rightarrow} N(0,\ 1)\ ,$$

According to Theorem 3 of the article, we have :

$$\tilde{W}_2 \overset{H_{bZ}}{\rightarrow} N\left(\frac{b}{\sqrt{I_{11}^{-1}(\theta_{00})}},\ 1\right)\ .$$

The efficiency $\tilde{\kappa}_2$ of this test, with respect to the oracle test (q_Z), is such as :

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

where κ_3 is the efficiency introduced in Corollary 1 of the article.

4. Proof of Corollary 4

4.1. *Oracle statistical test* (q_Z)

We consider a statistical model with one parameter (q_Z). The oracle statistical test is the same as in Section 3.1 of this document.

4.2. *First strategy*

We remind that we deal here with a statistical model with only one parameter (q_Z). We use same notations as in the proof of Theorem 2 of the article. Let $\theta = (q_Z)$ and $\theta_0 = (0)$.

In order to make the calculations easier for the score function, we write the likelihood L using $z^\star$. We have :

$$\begin{aligned}
L(q_Z) &= \left\{\frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\ +\ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\right\}\ 1_{\overline{X}=0} \\
&+ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\varphi\left(z^\star - \frac{\mu_Z}{\sigma\ \sqrt{1-r^2}} + \frac{r\ \mu_Y}{\sigma\ \sqrt{1-r^2}} - \frac{q_Z}{\sigma\ \sqrt{1-r^2}} + \frac{rq_Y}{\sigma\ \sqrt{1-r^2}}\right)\ 1_{\overline{X}=1} \\
&+ \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\varphi\left(z^\star - \frac{\mu_Z}{\sigma\ \sqrt{1-r^2}} + \frac{r\ \mu_Y}{\sigma\ \sqrt{1-r^2}} + \frac{q_Z}{\sigma\ \sqrt{1-r^2}} - \frac{rq_Y}{\sigma\ \sqrt{1-r^2}}\right)\ 1_{\overline{X}=-1}\ .
\end{aligned}$$

The score function is :

$$\begin{aligned}
\frac{\partial \log L}{\partial q_z} &= \left\{\frac{1}{\sigma\sqrt{1-r^2}}\ \left(z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} - \frac{q_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}} + \frac{rq_Y}{\sigma\sqrt{1-r^2}}\right)\right\} 1_{\overline{X}=1} \\
&- \left\{\frac{1}{\sigma\sqrt{1-r^2}}\ \left(z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{q_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}} - \frac{rq_Y}{\sigma\sqrt{1-r^2}}\right)\right\} 1_{\overline{X}=-1}\ .
\end{aligned}$$

The MLE $\hat{q}_Z$ of q_Z for n observations is such as :

$$\begin{aligned}
\hat{q}_Z = \Bigg\{ &\sum_{j=1}^{n}\left(\sigma\ \sqrt{1-r^2}\ z_j^\star - \mu_Z + r\mu_Y + rq_Y\right) 1_{\overline{X}_j=1} \\
&- \left(\sigma\ \sqrt{1-r^2}\ z_j^\star - \mu_Z + r\mu_Y - rq_Y\right) 1_{\overline{X}_j=-1}\Bigg\} \Big/ \sum_{j=1}^{n} 1_{\overline{X}_j\neq 0}\ .
\end{aligned}$$

And :

$$I_{\theta_0} = \frac{\mathbb{P}(Y \notin [S_-, S_+])}{\sigma^2(1-r^2)} \ .$$

Since the model is regular, the Wald test is :

$$\tilde{W}_1 = \frac{\sqrt{n}\ \hat{q}_Z}{\sigma\sqrt{1-r^2}}\ \sqrt{\mathbb{P}(Y \notin [S_-, S_+])} \ \overset{H_{0Z}}{\to}\ N(0,\ 1)\ .$$

We apply Theorem 3 (cf. Appendix B.1 of the article) with $h_n = h = b$:

$$\tilde{W}_1 \ \overset{H_{bz}}{\to}\ N\left(\frac{b\ \sqrt{\mathbb{P}(Y \notin [S_-, S_+])}}{\sigma\ \sqrt{1-r^2}},\ 1\right)\ .$$

Then, the efficiency $\tilde{\kappa}_1$ of the first strategy with respect to the oracle test (q_Z), is :

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

4.3. Second strategy

The likelihood is :

$$\begin{aligned} L = {} & \mathbb{P}(\overline{X}=0)\ 1_{\overline{X}=0} \\ & + \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\varphi\left(z^\star - \frac{\mu_Z}{\sigma\ \sqrt{1-r^2}} + \frac{r\ \mu_Y}{\sigma\ \sqrt{1-r^2}} - \frac{q_Z}{\sigma\ \sqrt{1-r^2}} + \frac{rq_Y}{\sigma\ \sqrt{1-r^2}}\right)\ 1_{\overline{X}=1} \\ & + \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\varphi\left(z^\star - \frac{\mu_Z}{\sigma\ \sqrt{1-r^2}} + \frac{r\ \mu_Y}{\sigma\ \sqrt{1-r^2}} + \frac{q_Z}{\sigma\ \sqrt{1-r^2}} - \frac{rq_Y}{\sigma\ \sqrt{1-r^2}}\right)\ 1_{\overline{X}=-1}\ . \end{aligned}$$

We remark that we have the same term which depends on q_Z as for the first strategy. As a consequence, the test will be exactly the same as the test for the first strategy. And naturally, the efficiency for the second strategy will be the same as for the first strategy, so $\tilde{\kappa}_2 = \tilde{\kappa}_1$.