## Title
Selective Genotyping and Phenotyping Strategies in a Complex Trait Context



## Authors
Sen, Saunak
Johannes, Frank
Broman, Karl W

## Publication Date
2008-07-28

# SELECTIVE GENOTYPING AND PHENOTYPING STRATEGIES IN A COMPLEX TRAIT CONTEXT

ŚAUNAK SEN*, FRANK JOHANNES†, AND KARL W BROMAN‡

July 28, 2008

*Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143. †Gröningen Bioinformatics Centre, University of Gröningen, Gröningen, the Netherlands. ‡Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706.

## ABSTRACT

Selective genotyping and phenotyping strategies can reduce the cost of QTL (quantitative trait loci) experiments. We analyze selective genotyping and phenotyping strategies in the context of multi-locus models, and non-normal phenotypes. Our approach is based on calculations of the expected information of the experiment under different strategies. Our central conclusions are the following. (1) Selective genotyping is effective for detecting linked and epistatic QTL as long as no locus has a large effect. When one or more loci have large effects, the effectiveness of selective genotyping is unpredictable – it may be heightened or diminished relative to the small effects case. (2) Selective phenotyping efficiency decreases as the number of unlinked loci used for selection increases, and approaches random selection in the limit. However, when phenotyping is expensive, and a small fraction can be phenotyped, the efficiency of selective phenotyping is high compared to random sampling, even when over 10 loci are used for selection. (3) For time-to-event phenotypes such as lifetimes, which have a long right tail, right-tail selective genotyping is more effective than two-tail selective genotyping. For heavy-tailed phenotype distributions, such as the Cauchy distribution, the most extreme phenotypic individuals are not the most informative. (4) When the phenotype distribution is exponential, and a right-tail selective genotyping strategy is used, the optimal selection fraction (proportion genotyped) is less than 20% or 100% depending on genotyping cost. (5) For time-to-event phenotypes where followup cost increases with the lifetime of the individual, we derive the optimal followup time that maximizes the information content of the experiment relative to its cost. For example, when the cost of following up an individual for the average lifetime in the population is approximately equal to the fixed costs of genotyping and breeding, the optimal strategy is to follow up approximately 70% of the population.

## INTRODUCTION

Quantitative trait locus (QTL) experiments provide valuable clues for finding elements responsible for quantitative trait variation (LANDER AND BOTSTEIN, 1989; LYNCH AND WALSH, 1998; RAPP, 2000). For best results, QTL experiments require large numbers of individuals that need to be genotyped as well as phenotyped for the quantitative trait of interest. Because this can be a costly endeavor, investigators can employ cost-saving strategies such as selective genotyping, in

which a selected portion of the phenotyped individuals are genotyped (LEBOWITZ ET AL., 1987; LANDER AND BOTSTEIN, 1989; DARVASI AND SOLLER, 1992), and selective phenotyping, in which a selected portion of the genotyped individuals are phenotyped (JIN ET AL., 2004). The efficacy of these strategies has been evaluated in simplified settings where a single locus contributes to the phenotype, and when the phenotype (conditional on genotype) is normally distributed. It is therefore unclear how effective these strategies would be in the broader context of complex trait genetic analyses. In such settings, we suspect that multiple loci, possibly linked and epistatic, contribute to the trait, and the trait distribution may be non-normal.

SEN ET AL. (2005) examined the effectiveness of selective genotyping when two unlinked additive QTL contribute to a normally-distributed trait. Because epistasis appears to be a common and important feature of many complex traits (FRANKEL AND SCHORK, 1996), it is crucial to investigate whether epistasis can also be detected in selectively genotyped samples. Experimental studies appear to be divided over this issue, with some studies reporting epistasis in selectively genotyped samples (OHNO ET AL., 2000; ABASHT AND LAMONT, 2007) while others failed to detect it (CARR ET AL., 2006) and cited concerns about loss of power. Thus, the generality of these experimental observations require further theoretical exploration.

The potential value of selective genotyping has also been recognized in human association studies, and is currently being actively researched (CHEN ET AL., 2005; WALLACE ET AL., 2006; HUANG AND LIN, 2007). Interest in this application is primarily motivated by the fact that these studies require dense high-throughput genotyping which can be expensive. However, similar to QTL studies in experimental crosses, the theoretical results have focused primarily on normally-distributed phenotypes. In this context, GALLAIS ET AL. (2007) compared one-tail and two-tail selective genotyping and showed that the latter is superior. However, many interesting traits are non-normally distributed. Time-to-event phenotypes, such as survival times or tumor onset, are important cases when the trait is expected to be non-normally distributed, usually with a long right tail. In these situations, individuals in the right tail are likely to be genetically more informative, and it is unclear which type of selection strategy (one-tail, two-tail, or a different strategy) should be applied in this setting. Moreover, from a cost-saving perspective the additional problem arises that the most informative individuals (those in the right tail) will also be the most expensive to phenotype because of the cost of following the individuals until the event of interest has been observed. The investigator must therefore decide to either stop following up, which results in reduced cost and a loss of information due to censoring, or to follow up the entire sample until all events have been observed, which implies greater cost but a minimal loss of information. As far as we are aware, these tradeoffs have not been studied.

In applications where phenotyping is more expensive than genotyping, (JIN ET AL., 2004) proposed selective phenotyping as an approach to maximize the genetic diversity of individuals selected for phenotyping in a genomic region of interest. Their simulations showed that, for a fixed number of phenotyped individuals, this approach increases power relative to a random sample. Although this gain in power diminishes when multiple genomic regions are considered, it continues to outperform the random sample alternative. This approach is particularly attractive for genetical genomics studies (JANSEN AND NAP, 2001) where the traits of interest consist of genomewide molecular measurements (e.g. transcriptome, metabolome, proteome) obtained through microarray and mass spectrometry technology. In the specific case of genetical genomics studies, (FU AND JANSEN, 2006) proposed a related selective phenotyping strategy which seeks to

increase power by co-hybridizing the transcripts of the most genetically distant pairs of individuals in the sample onto the same array.

The above-mentioned experimental design problems have one common feature – they involve data gathering strategies (selective genotyping, selective phenotyping, choice of followup period) with a tradeoff between information and experimental cost. SEN ET AL. (2005) showed that by adopting an information perspective, one can formally study these tradeoffs in QTL studies. In that paper, the information approach was used to explain known results, and obtain some new ones. The present work extends these ideas and explores several unresolved issues in data gathering strategies in QTL studies.

Our paper is organized as follows. In the next section we briefly review the theory underlying our information perspective. In the following section we present the results obtained by applying that theory. We examine the efficacy of selective genotyping and phenotyping in the context of multilocus models when the phenotype is normally distributed. Then we study selective genotyping when the trait may not be normally distributed with a special emphasis on time-to-event (lifetime or survival) phenotypes. We conclude with a discussion of our results.

## THEORY

**Information perspective on QTL study design** Traditionally, the efficacy of QTL study designs has been investigated using power calculations. In the experimental design literature, notably industrial experimental design, study designs are evaluated using the information matrix of the design. This is because the information matrix is a fundamental statistical quantity, and is simpler to characterize. The power of a study design, the expected LOD score (which is a log likelihood ratio), and variance of estimated QTL effects, all depend on the information matrix.

The expected Fisher information is defined as the expected value of the second derivative of the log likelihood function.

$$J(\beta) = \mathsf{E}\left[-\frac{\partial^2}{\partial\beta^2}\ell(\beta|y_{obs})\right] = \mathsf{E}\left[-\frac{\partial^2}{\partial\beta^2}\log p(y_{obs}|\beta)\right],$$

where $\ell(\beta|y_{obs})$ denotes the log likelihood function for the parameter of interest $\beta$, when the observed data is $y_{obs}$. For large sample sizes, the variance of the maximum likelihood estimate, $\hat{\beta}$ is

$$V(\hat{\beta}) \simeq J(\beta)^{-1}.$$

The log likelihood ratio statistic for testing $\beta = \beta_0$ is

$$2\left(\ell(\hat{\beta}) - \ell(\beta_0)\right)$$

which, for large samples, has an approximate non-central $\chi^2$ distribution with $s$ (dimension of $\beta$) degrees of freedom and non-centrality parameter $(\beta - \beta_0)'J(\beta)(\beta - \beta_0)$. The log-likelihood ratio expressed in base 10 logarithms is the LOD score. Thus, the power of the likelihood ratio test, which depends on the non-centrality parameter, depends on the unknown state of nature ($\beta - \beta_0$),

and the information matrix, $J(\beta)$. The experimenter has no control over the state of nature, but has limited control over experimental design choices that determine the information content. Thus the information content of a study design provides us with a parsimonious description of the statistical characteristics of the study.

**Criteria for evaluating designs** We calculate the expected Fisher information (for the QTL effect parameters conditional on QTL location) for each genotyping strategy in each context. This is used to evaluate the usefulness of selective genotyping in each genetic model context. When multiple loci are involved, the information is a matrix, and we have to resort to one-dimensional summaries of the information matrix.

In the experimental design literature, a few different summary measures (optimality criteria) have been proposed for comparing alternative designs (COX AND REID, 2000). We compare the designs here using two criteria based on the information matrix $J$ , or equivalently $V = J^{-1}$:

1. D-optimality criterion: This criterion maximizes the determinant of the information matrix, $\det(J)$, or equivalently $(\det(J))^{1/k}$, where $k$ is the number of parameters. It minimizes the volume of the joint confidence ellipsoid of all the parameters. It is the most popular design criteria because it makes full use of the information matrix, and is not affected by orthogonal reparametrizations of the parameter.
2. c-optimality criterion: This criterion maximizes the inverse of the variance of a contrast $c$ between model parameters, $(c'Vc)^{-1}$.

The appropriate criterion depends on one's objective. In the present paper, we will use the c-optimality and D-optimality criteria.

**Calculation of the information content** We use missing data methods to calculated the expected information content under selective genotyping. When we use selective genotyping we deliberately choose not to collect genotyped data on certain individuals based on their phenotype. This data is *missing data.* We use the missing data principle to calculate the expected information content of any genotyping strategy. The missing information principle (ORCHARD AND WOODBURY, 1972; MCLACHLAN AND KRISHNAN, 1996) states that the observed information, $I_o$, may be calculated as

$$I_o = I_c - I_m, .$$

where the observed information is

$$I_o(\beta) = -\frac{\partial^2}{\partial \beta^2} \log p(y_{obs}|\beta),$$

the missing information is

$$I_m(\beta) = \mathsf{E}\left[-\frac{\partial^2}{\partial \beta^2} \log p(y_{mis}|y_{obs},\beta)\Big|y_{obs},\beta\right],$$

and the complete information is

$$I_c(\beta) = \mathsf{E}\left[-\frac{\partial^2}{\partial \beta^2} \log p(y_{mis},y_{obs},\beta)\Big|y_{obs},\beta\right]$$

In the selective genotyping context, $y_{obs}$ consists of the phenotypes, and the genotypes of genotyped individuals. The missing data, $y_{mis}$ consists of the genotype data at all ungenotyped locations. Since the expected information satisfies, $J(\beta) = \mathsf{E}(I_o(\beta))$, we can use the missing information principle to calculate the expected information content of genotyping designs.

**Backcross population, single locus model** Assume we have a population of $n$ individuals. Let $y$ denote the phenotype of an individual and let $g$ the genotype at a particular locus. Let $\beta$ denote the genetic model parameters. In general, $\beta$ is a vector. The phenotype is assumed to be normally distributed given the QTL genotypes with mean depending on $\beta$ and variance 1.

Assume that a single locus contributes to the trait variation, and consider a single individual with phenotype $y$, and with $q$ denoting the conditional probability that the individual is homozygous at a locus, given the available marker data. SEN ET AL. (2005) showed that, in this case, the contribution of the individual to the observed information is

$$1 - 4y^2 q^*(1-q^*),$$

where $q^*/(1-q^*) = e^{2\beta}q/(1-q)$. At a locus with no nearby markers genotyped, $q$=1/2, so that the observed information is

$$1 - 4y^2 \frac{e^{2\beta y}}{1+e^{2\beta y}} \frac{1}{1+e^{2\beta y}} = 1 - y^2 \text{sech}^2(\beta y) = 1 - y^2 + y^4\beta^2 - \frac{(2y^6)}{3}\beta^4 + \ldots$$

If the individual is genotyped, then the observed information is simply 1. When selectively genotyping with selection fraction $\alpha$, we genotype an $\alpha$ fraction of the most extreme phenotypic individuals. Thus, if $z(\alpha, \beta)$ is the upper $\alpha$ point of the phenotype distribution when the QTL effect is $\beta$, then the expected information using the two-tail selective genotyping strategy is

$$J(\alpha, \beta) = \alpha + \int_{-z_{\alpha/2,\beta}}^{+z_{\alpha/2,\beta}} (1 - y^2 \text{sech}^2(\beta y))dy = \alpha + 2\, z_{\alpha/2,\beta}\phi(z_{\alpha/2,\beta}) + O(\beta^2).$$

For small $\beta$,

$$J(\alpha, \beta) \simeq J(\alpha) = \alpha + 2\, z_\alpha \phi(z_\alpha),$$

where $z_\alpha$ is the upper alpha point of the standard normal distribution.

The observed information corresponding to an individual phenotype $y$ gives an indication of the value of genotyping that individual. Integrating over the observed information corresponding to a genotyping strategy, we can get the expected information resulting from that genotyping strategy. Thus, we can devise and evaluate strategies by examining the observed information and expected information.

**Backcross population, two unlinked loci** Let $g_1$ and $g_2$ denote the QTL genotypes at two unlinked loci. Assuming that the two loci, are additive, we can write the genetic model for the phenotype as:

$$y = \beta_0 + \beta_1(2g_1 - 1) + \beta_2(2g_2 - 1) + \epsilon, \tag{1}$$

where $\beta_0$ is the overall mean, $\beta_1$ and $\beta_2$ are the effects of the first and second QTL respectively, and $\epsilon$ is the random error which is normally distributed. For simplicity, we will assume $\beta_0$=0 for the

rest of this paper. SEN ET AL. (2005) used the above-mentioned approach to calculate the expected information for $\underline{\beta} = (\beta_1, \beta_2)$, when $\beta_1$=0, and $\beta_2$=$\beta$. This gives us the missing information when the first QTL has small effect as a function of the effect of the second QTL. In this setting the missing information matrix for an ungenotyped individual with phenotype $y$ was shown to be

$$I_m = \begin{pmatrix} (y^2+\beta^2) + 2\beta y \tanh(\beta y) & 0 \\ 0 & y^2\text{sech}^2(\beta y) \end{pmatrix}.$$

Note that when the second QTL has a small effect the expected information under selective genotyping behaves similarly as with a single QTL. As the strength of the second QTL the information content for the first QTL progressively decreases. The worst scenario is when the second QTL has a really dramatic effect. In this setting, when half the extreme individuals are genotyped, only half the information is obtained – this is the same as genotyping randomly selected individuals (random genotyping). When more than half of the extreme individuals are genotyped, selective genotyping performs worse than random genotyping.

## RESULTS

**Two linked loci** Let $g_1$ and $g_2$ denote the QTL genotypes at two loci separated by a recombination fraction $\theta$. Our objective is to evaluate the expected information content of a selective genotyping design where $\alpha$ fraction of the extreme phenotypic individuals are genotyped. Assume that the QTL act additively, i.e. the genetic model for the phenotypes is the same as (1). First note that with complete genotyping, the expected information matrix per observation is

$$J = \begin{pmatrix} 1 & 1-2\theta \\ 1-2\theta & 1 \end{pmatrix}.$$

Thus, $\det(J) = 4\theta(1-\theta)$. The variance of the parameter estimates is thus

$$V = J^{-1} = \frac{1}{4\theta(1-\theta)} \begin{pmatrix} 1 & 2\theta-1 \\ 2\theta-1 & 1 \end{pmatrix}.$$

Using either a D-optimality criterion (the determinant of the information matrix) or the inverse of the variance of the first locus effect, $\hat{\beta}_1$, we see that the informativeness for two linked loci is a function of $\theta(1-\theta)$ which is maximum when the two loci are unlinked, and gets progressively smaller as theta approaches 0.

Our goal is to examine how selective genotyping affects the information to detect two linked loci. To do this, we have to calculate the information matrix using the missing information principle. The missing information matrix is

$$I_m = \begin{pmatrix} y^2\text{sech}^2(\beta y) + 4\theta(1-\theta)\big(y\tanh(\beta y)-\beta\big)^2 & (1-2\theta)y^2\text{sech}^2(\beta y) \\ (1-2\theta)y^2\text{sech}^2(\beta y) & y^2\text{sech}^2(\beta y) \end{pmatrix}.$$

Note that when the QTL are unlinked, i.e. $\theta$=$\frac{1}{2}$, the missing information for $\beta_1$ is

$$\begin{aligned} y^2\text{sech}^2(\beta y) + \big(y\tanh(\beta y)-\beta\big)^2 &= y^2\big(\text{sech}^2(\beta y) + \tanh^2(\beta y)\big) + \beta^2 + 2\beta y\tanh(\beta y) \\ &= y^2 + \beta^2 + 2\beta y\tanh(\beta y), \end{aligned}$$

which coincides with the result for unlinked loci derived earlier. Further,

$$I_m = \begin{pmatrix} A + 4\theta(1-\theta)B & (1-2\theta)A \\ (1-2\theta)A & A \end{pmatrix},$$

where $A{=}y^2\text{sech}^2(\beta y)$, and $B{=}\big(y\tanh(\beta y){-}\beta\big)^2$. Thus,

$$I_o = I_c - I_m = \begin{pmatrix} 1 & 1-2\theta \\ 1-2\theta & 1 \end{pmatrix} - \begin{pmatrix} A + 4\theta(1-\theta)B & (1-2\theta)A \\ (1-2\theta)A & A \end{pmatrix}$$

Therefore the expected information has the form

$$J(\alpha,\beta) = \begin{pmatrix} A^* + 4\theta(1-\theta)B^* & (1-2\theta)A^* \\ (1-2\theta)A^* & A^* \end{pmatrix},$$

where $A^*{=}(1{-}A)$, and $B^*{=}{-}B$. Note that $A$, $A^*$, $B$, and $B^*$ are independent of $\theta$. Now we will determine how the two quantities of interest, depend on how close the two loci are to each other (as measured by $\theta$). Since the information matrix is not scalar, we will use our two scalar summaries, the determinant and the inverse of the variance of the $\hat{\beta}_1$. First, note that

$$\det(J(\alpha,\beta)) = 4\theta(1-\theta)(1{-}A^*)(1{-}A^*{-}B^*),$$

which means that by the D-optimality criterion, the effect of selective genotyping and the closeness of the two loci are independent. This implies, that beyond the loss of information due to linked loci, the effect of selective genotyping is exactly as for unlinked loci. Next, note that the variance matrix is

$$V(\alpha,\beta) = J(\alpha,\beta)^{-1} = \frac{1}{4\theta(1-\theta)(1{-}A^*{-}B^*)} \begin{pmatrix} 1 & 2\theta{-}1 \\ 2\theta{-}1 & 1 - 4\theta(1{-}\theta)B^*/\big(1{-}A^*\big) \end{pmatrix}.$$

Here also, the variance of $\hat{\beta}_1$ is the product of two terms, one that depends on how linked the loci are, and another that depends on the selective genotyping scheme. This also implies that the relative change in information to detect a locus with small effect in the presence of a linked locus, does not depend on the extent of linkage.

**Two epistatic loci** We will analyze the case of two epistatic loci with the same approach as for two linked loci. Consider the following linear model for the phenotype.

$$y = \beta_0 + \beta_1(2g_1{-}1) + \beta_2(2g_2{-}1) + \beta_3(2g_1{-}1)(2g_2{-}1) + \epsilon, \tag{2}$$

where $\beta_3$ is the epistatic effect of the two QTL. We will consider two important special cases when the epistatic effect is small: (a) when there is one major main effect and the other locus has a small effect ($\beta_1{=}\beta, \beta_2{=}0, \beta_3{=}0$), and (b) when both loci have equal but non-zero main effects ($\beta_1{=}\beta, \beta_2{=}\beta, \beta_3{=}0$). The analytic expressions for the observed information matrix are included in the supplementary information. We graph the functions in Figures 1 and 2.

We find that as long as the proportion of variance explained by the main effect QTLs remains less than 20% the effectiveness of selective genotyping is approximately the same as that for the case when a single locus with a main effect is segregating in the cross. When the proportion of variance explained by the main effect QTLs is larger, the efficiency of selective genotyping for detecting epistasis varies. In some cases, it can be less efficient than random sampling (Figure 1); in other cases it may have more information than that for the main effect loci (Figure 2).

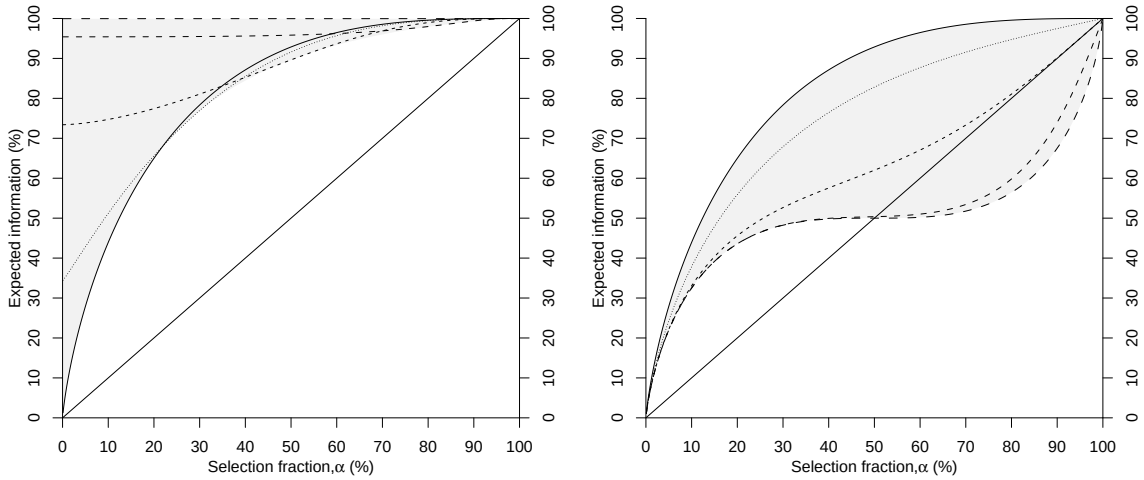


**Figure 1:** Expected information in a two-QTL model with epistasis, as a function of the selection fraction $\alpha$ for main effects ($\beta_1$, left panel) and epistatic effects ($\beta_3$, right panel). The information is plotted as we vary the size of the main effect of the first QTL, while the second QTL and epistatic effect size is assumed to be zero. The shaded region is the space of variation as $\beta_1$ varies from 0 to $\infty$. The solid line corresponds to $\beta_1$=0, and successive dashed lines (as the size of the dash increases) correspond to the proportion of variance explained by the first QTL equal to 20%, 50%, 75% and 90%. If the proportion of variance explained by the main effect QTL is less than 20%, the expected information is approximately equal to that when the proportion variance explained is 0%. Information for the main effect increases as the size of the effect increases. The information for the epistatic effect decreases as the size of the main effect increases. For selection fractions greater than 50% selective genotyping may be less efficient than even random sampling (solid diagonal line), for which the expected information is equal to the selection fraction. This is specially so when the variance explained by the main effect exceeds 50%.

**Selective phenotyping with multiple regions of interest** JIN ET AL. (2004) proposed selective phenotyping as a cost-savings measure when phenotyping is substantially more expensive than genotyping. Here we analyze the effect of selection based on multiple unlinked regions on the information content of the experiment.

The fundamental idea of selective phenotyping is to pick a subset of individuals who are as genotypically diverse as possible at a set of candidate regions. The efficiency of this approach decreases as the number of unlinked loci considered increases. To motivate the general result we first consider a single locus, then two unlinked loci, and then the general case. Throughout we consider selective phenotyping in an $F_2$ population where genotypes at any given locus are coded 0, 1, and 2 corresponding to the number of alleles from a particular inbred strain. We focus on detecting the additive effect of a locus.

Single locus The most efficient strategy is to first pick equal numbers of the two homozygotes (corresponding to genotypes 0 and 2) until they are exhausted. Then we pick the heterozygotes (corresponding to the genotype 1). Note that for detecting additive effects, heterozygotes are not

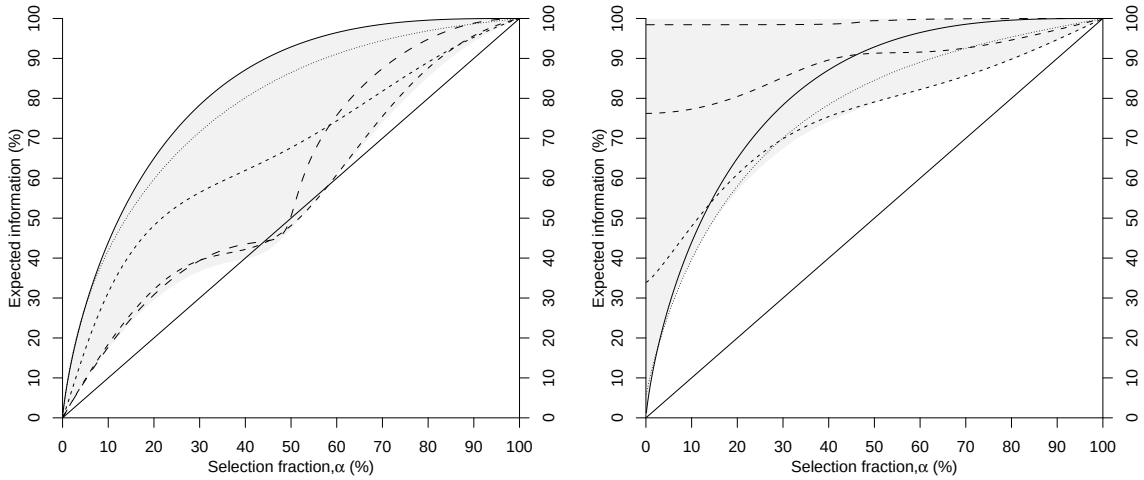


**Figure 2:** Expected information in a two-QTL model with epistasis, as a function of the selection fraction $\alpha$ main effects ($\beta_1$, left panel) and epistatic effects ($\beta_3$, right panel). The information is plotted as we vary the size of the main effects of the both QTL, assumed to be of equal effect, while the epistatic effect size is assumed to be zero. The shaded region is the space of variation as $\beta_1$ varies from 0 to $\infty$. The solid line corresponds to $\beta_1$=$\beta_2$=0, and successive dashed lines (as the size of the dash increases) correspond to the proportion of variance explained by the main effect QTLs equal to 20%, 50%, 75% and 90%. The diagonal solid line is the efficiency of genotyping a random subset. If the proportion of variance explained by the main effect QTL is less than 20%, the expected information is slightly less than that when the proportion variance explained is 0%. Information for the main effects decrease as the size of the effects increases, but the pattern is not monotonic with the effect size. The information for the epistatic effect approaches 100% as the size of the main effects increases. This "hyper-efficiency" relative to when the main effect size is zero is most pronounced when the proportion of variance explained by the main effects exceeds 75%.

informative, so on average, just studying half the population is as effective as studying all of it. This is reflected in Figure 3.

Suppose we select an $\alpha$ proportion of the sample for phenotyping, and of those a proportion $\tau$ are homozygotes . Then it is easily seen that the information content of the sample relative to studying the full sample is $2\tau\alpha$. We will use this result for proving the general result for an arbitrary number of loci.

Two loci When selective phenotyping is performed using two loci, the genotypes can be represented as in Figure 4. There are three genotype classes depending on the number of homozygous loci (0, 1, or 2). These correspond to the center point, the inner circle, and the outer circle respectively. The outer circle genotypes are the most different, and represent the greatest genetic diversity, followed by the inner circle, and finally the center point. Thus, the optimal strategy is to first select equal numbers of individuals from the outer circle (2 homozygous loci), then the inner circle (1 homozygous locus), and finally the center point (0 homozygous loci). The outer circle

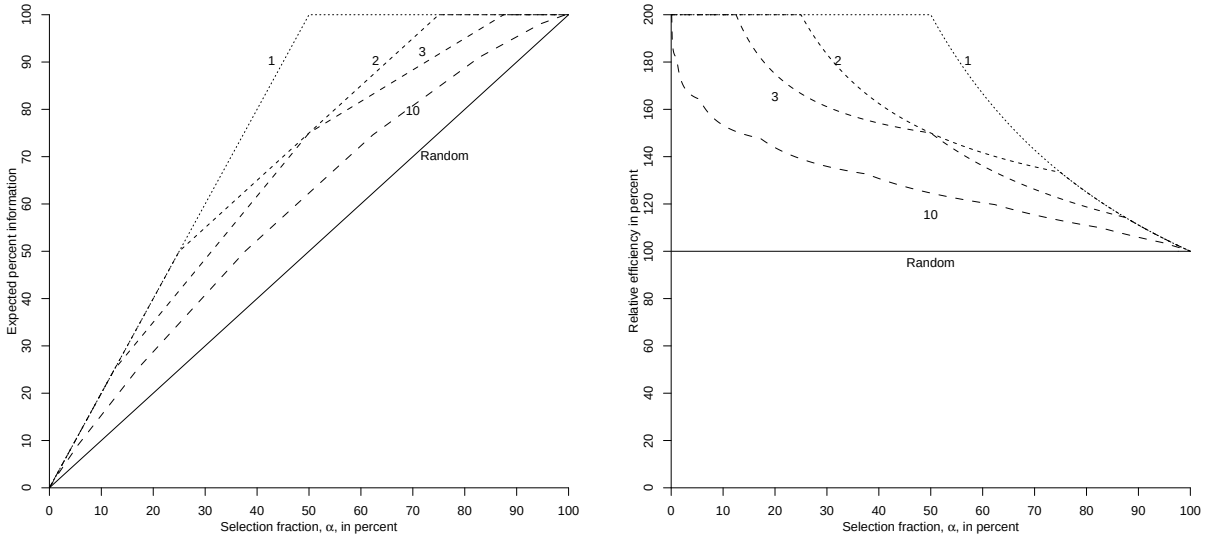


**Figure 3:** Information of selective phenotyping as a function of the selection fraction and the number of unlinked genetic loci used for selection. The left panel shows the expected information from selected phenotyping as a function of the selection fraction when one, two, three, or ten loci are used for selection. The solid line shows the expected information from random sampling. The right panel shows the information from selective phenotyping relative to random sampling as a function of the selection fraction. We see that as the number of loci increases, the efficiency of selective phenotyping approaches random selection. However, the relative efficiency for small selection fractions can be quite high even when ten loci are used for selection.

covers $1/4$ of the sample, the inner circle $1/2$ and the center point $1/4$.

If the loci considered are unlinked, the effect estimates corresponding to the loci are uncorrelated with each other, and hence, orthogonal. Thus, using symmetry, the information content of the whole sample can be evaluated through the information of any single locus.

Let us consider the information content corresponding to three key $\alpha$ values, $1/4$, when the outer circle points are included, $3/4$, when the outer and inner circle points are included, and $1$, when all points are included. When $\alpha$ is $1/4$, at any given locus all individuals are homozygous. Thus, the information content is $2 \times 1 \times \frac{1}{4} = \frac{1}{2}$. When $\alpha$ is $3/4$, all homozygous individuals are in the sample; they comprise $2/3$ of the selected sample. Thus, the information content is $2 \times \frac{2}{3} \times \frac{3}{4} = 1$. When $\alpha$ is 1, all individuals are in the sample, and thus the information content is 1. The information content of all other $\alpha$ values can be calculated by linear interpolation as in Figure 3.

Arbitrary number of loci We can now tackle the general case with $m$ unlinked loci, where the genotypes can be represented as points on a lattice in an $m$-dimensional space. There are $m+1$ classes of points corresponding to their distance from the center point representing an individual heterozygous at all loci. The classes are defined by the number of homozygous loci, $0$ through $m$. The proportion of the sample in each of these classes is given by the probability mass function of

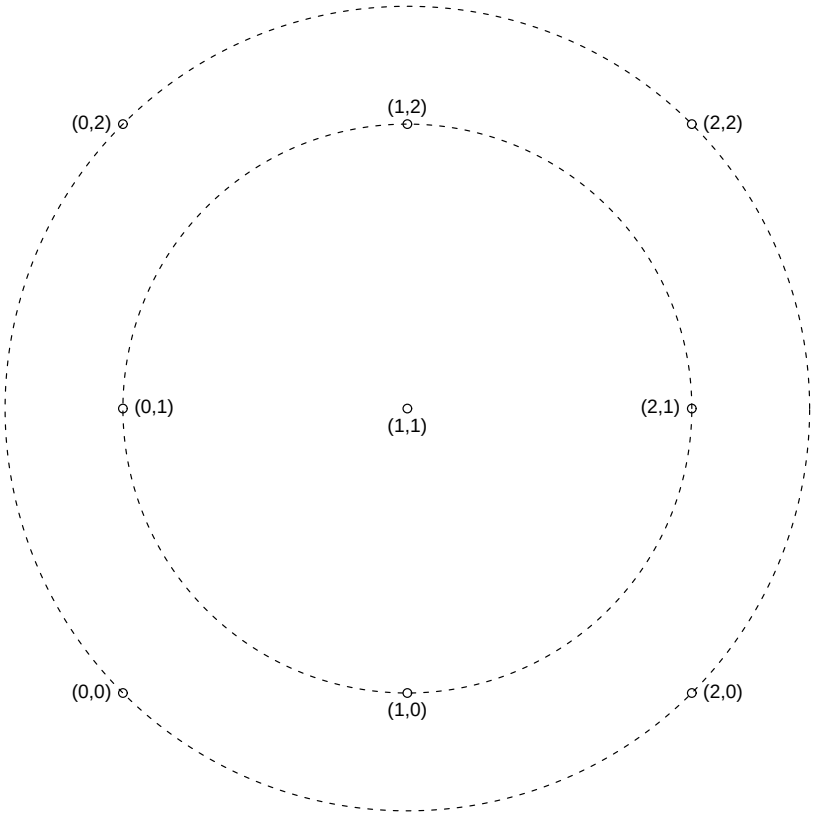


**Figure 4:** Distances of genotypes from the average genotype for two-locus genotypes in an $F_2$ intercross. We code genotypes at each locus as 0, 1, or 2. The x-axis and y-axis are used to plot the genotypes at the first and second locus respectively. The average genotype is the (1,1) genotype (double heterozygote) at the center of the figure. Two concentric circles are drawn to depict two sets of equidistant point from the center. The outermost circle consists of the homozygous genotypes, the points (0,0), (0,2), (2,2), and (2,0). These are the points most distant from the center. The inner circle consists of genotypes homozygous at one locus, and heterozygous at the other, the points (0,1), (1,2), (2,1), and (1,0). These are the next most distant from the center. To pick the most genotypically diverse individuals, one would first pick individuals with genotypes in the outermost circle, then the inner circle, and finally the center.

a binomial distribution with parameters $m$ and $1/2$. The expected information content of the class with $k$ homozygous loci is

$$2 \times \frac{k}{n} \times \binom{m}{k} \times \left(\frac{1}{2}\right)^m.$$

Thus, the information content of a sample that has chosen the classes $m, m-1, \ldots, k$ is

$$\sum_{i=k}^{m} 2\frac{i}{n}\binom{m}{i}\left(\frac{1}{2}\right)^m.$$

The information content corresponding to intermediate selection fractions can be found by linear interpolation. The function `info.pheno` in the R/qtlDesign package (SEN ET AL., 2007) calculates the information content of selective phenotyping.

We find that as the number of loci used for selective phenotyping increases, the efficiency of selective phenotyping decreases. In the limit, it reduces to random selection. However, it is notable that the gain in efficiency relative to random selection is higher for small selection fractions (the fraction of individuals selected for selective phenotyping). In other words, if phenotyping is very expensive relative to genotyping and rearing, then even if a large number of loci (or the whole genome) is used for selective phenotyping, it will be effective. These findings are consistent with the simulation studies of JIN ET AL. (2004), and provide a theoretical justification for their observations.

**Non-normal phenotypes** The logic of selective genotyping is that extreme phenotypic individuals provide the most information. This may not hold for all situations. For example, if the phenotype is heavy-tailed, the most extreme individuals are less informative. In other words, we expect individuals with moderately high, but not the most extreme phenotypes, to be the most informative. This argument implicitly assumed that both extremes of the phenotype are equally important. For lifetime distributions, it is reasonable to expect that the right tail is more important than the left tail, but this asymmetry is not reflected in two-tailed selective genotyping strategies. To help us choose a genotyping strategy based on the nature of the phenotype distribution, we develop the idea of the information gain function below.

Information gain function We develop our ideas in the context of a backcross. Let $y$ be the phenotype of an individual, $g{=}0, 1$ be the QTL genotype at a locus of interest, and let $q{=}\mathsf{P}(g{=}1|m)$ be the probability of the 1 genotype given the marker genotype information, $m$. Let the distribution of the phenotype given the QTL genotype be $p(y|g)$. The observed data consists of $(y, m)$, while the missing data is $g$. We want to know, based on an individual's phenotype, how informative that individual will be. Let $p(y|g{=}0) = f(y, -\delta)$, and $p(y|g{=}1) = f(y, +\delta)$ where $f$ is the phenotype density. In our context, the missing data are the unobserved QTL genotypes and the observed data consist of the marker genotypes and the phenotypes. The parameter of interest is $\delta$. Thus the distribution of the missing data conditional on the observed data is $q^{*g}(1{-}q^*)^{1-g}$, where $q^* = P(g{=}1|y, m, \delta)$. Since $q = P(g{=}1|m)$, by Bayes theorem it is easy to see that

$$q^* = \frac{q\ f(y, +\delta)}{q\ f(y, -\delta) + (1{-}q)\ f(y, +\delta)}.$$

Hence the missing data log likelihood is

$$\ell^* = \Big(g\ \log(q^*) + (1{-}g)\ \log(1{-}q^*)\Big)$$

Differentiating twice, we get

$$\frac{\partial^2 \ell^*}{\partial \delta^2} \quad = \quad \left(\frac{\partial^2 q^*}{\partial \delta^2}\right)\left[\frac{g}{q^*} - \frac{(1{-}g)}{(1{-}q^*)}\right] + \left(\frac{\partial q^*}{\partial \delta}\right)^2\left[-\frac{g}{q^{*2}} - \frac{1{-}g}{(1{-}q^*)^2}\right]$$

Hence,

$$-E\left(\frac{\partial^2 \ell^*}{\partial \delta^2}\middle| y, m, \delta\right) \;=\; \left(\frac{\partial q^*}{\partial \delta}\right)^2 \left[\frac{1}{q^*(1-q^*)}\right] = h(y,q,\delta)\, q^*(1-q^*),$$

where

$$h(y,q,\delta) = \left(\frac{\partial q^*}{\partial \delta}\right)^2 \left[\frac{1}{q^*(1-q^*)}\right]^2,$$

is a function depending on the phenotype density $f$. We expect this function to change with the shape of the phenotype distribution. By calculating this function, which we call the *information gain function,* for different functional forms of $f$, we can identify the individuals that are best to genotype. We use Taylor expansions for small $\delta$, the most interesting scenario. For the normal distribution, $h(y,q,\delta){=}y^2$ and captures the fact that most information is to be gained from the extremes of the distribution. Information gain functions for selected distributions is shown in Table 1.

Location shift – symmetric distributions We first examine, symmetric distributions with a location shift depending on genotype. Our calculations show very different information gain functions for the normal and Cauchy distributions (Figure 5). This suggests that the most extreme phenotypic individuals are not as informative when the phenotype distribution is Cauchy, as it is when the phenotype distribution is normal.

To study this further, we conducted a simulation study as follows. We simulated 10,000,000 individuals from a backcross. Conditional on the genotype, the phenotype in the two genotype groups was location-shifted by 0.1 times the IQR (inter-quartile range). Then we examined the genotype ratios conditional on the percentile of the phenotype distribution. Uninformative percentiles would be those where the genotype ratio is 0.5. The further the deviation from 0.5, the more informative the percentile. Assuming that the two genotypes are coded 0 and 1, let $p_q$ be the proportion of 1 genotypes conditional on the phenotype $y$ being in the $q$-th percentile. We plot $(p_q{-}0.5)^2$ as a function of $q$ to see which percentiles most discriminate between the two genotypes (Figure 5).

The simulation study confirmed what the information gain function suggests – the most extreme individuals are most informative when the phenotype distribution is normal or logistic; however, they are not the most informative if the phenotype has a Cauchy distribution. This shows that the best selective genotyping strategy depends on the shape of the phenotype distribution, and that the traditional two-tail selective genotyping strategy is not always the best. We explore this further by examining the information gain function for typical survival distributions.

Scale shift – lifetime distributions For lifetime distributions we focus on the exponential distribution, and two families extending it – the Gamma and Weibull distributions. Calculating the information gain function for a scale shift (Figure 6) we find that the upper tail, containing individuals with the longest lifetimes (top 15%), is more informative than the shortest lived individuals. This suggests that for phenotypes with a long right tail we should selectively genotype by oversampling the right tail. Although the information gain function for Weibull and exponential distributions appear different in functional form (Table 1), they are identical as a function of phenotype percentile, .

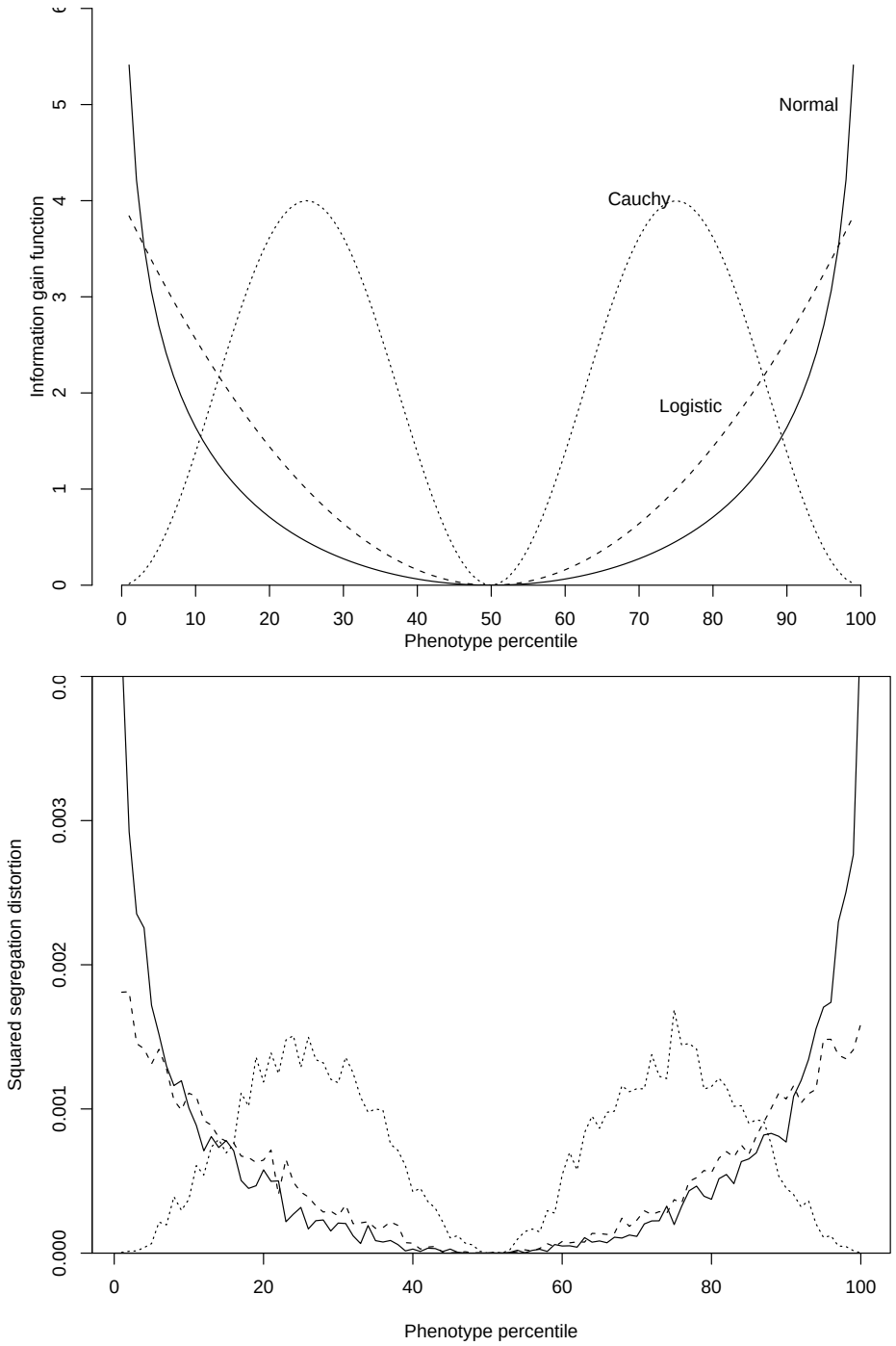


**Figure 5:** Top panel: Plot of the information gain function against the phenotype percentile for normal, Cauchy, and logistic distributions. We see that the extreme phenotypic individuals are very informative if the phenotype has a normal or logistic distribution. However, if the phenotype follows a Cauchy distribution, the extreme phenotypic individuals are not very informative. The most informative individuals are those near the first and third quartiles. Bottom panel: Plot of the squared deviation of the segregation ratio from the expected 50% by percentile of phenotype distribution from 10,000,000 simulations. The squared segregation ratios conditional on phenotype have shapes similar to the information gain function.

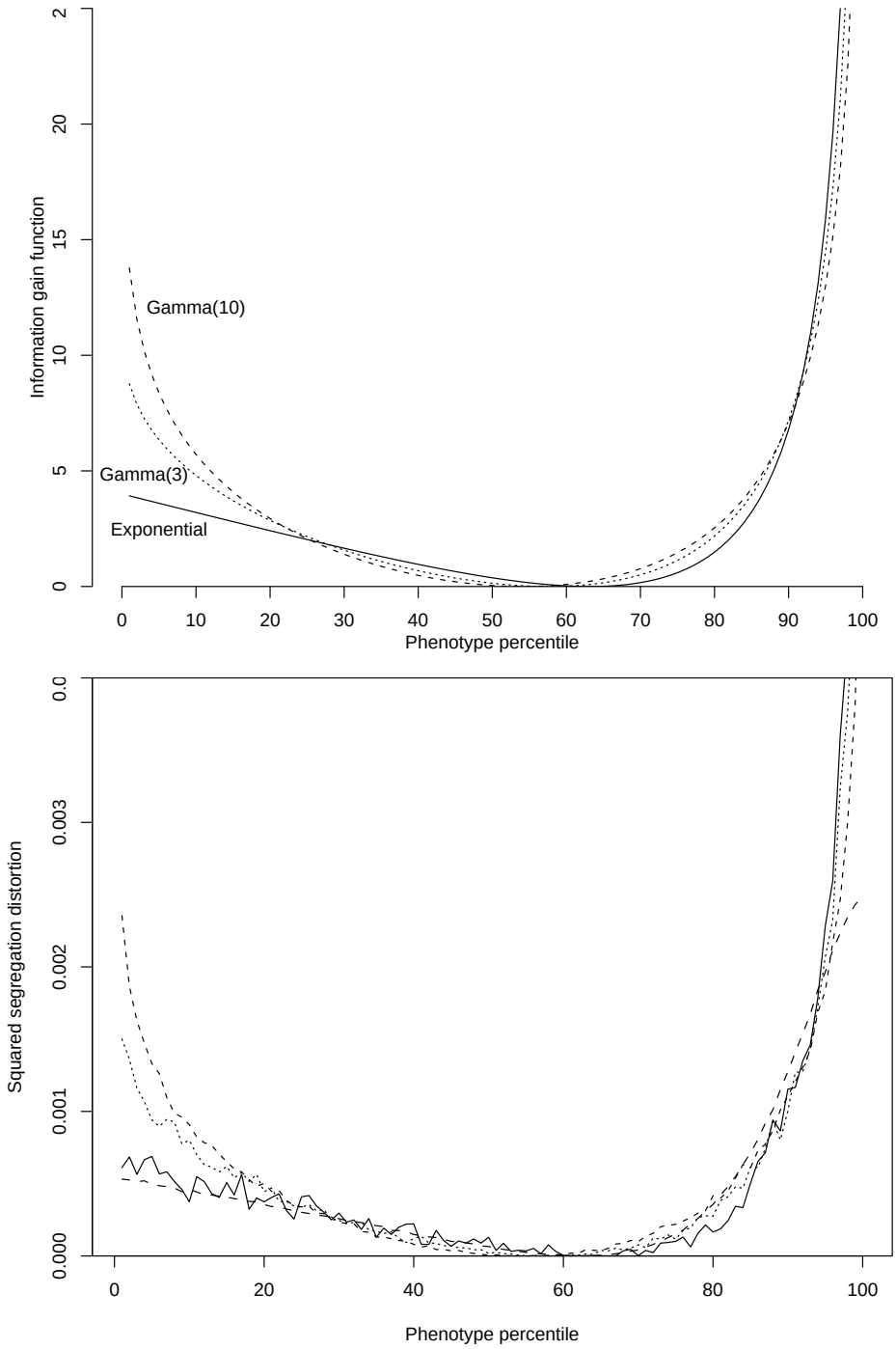


**Figure 6:** Top panel: Information gain function against the phenotype percentile for lifetime distributions (Exponential, and Gamma). The upper (right) tail is more informative than the lower (left) tail. The importance of the right tail decreases as the shape parameter of the Gamma distribution increases (and approaches a normal distribution). Bottom panel: Plot of the squared deviation of the segregation ratio by percentile of phenotype distribution from 10,000,000 simulations. The solid line corresponds to the exponential distribution. The dashed lines (in order of increasing width of dashes) are Gamma distribution shape parameter 3, Gamma distribution with shape parameter 10, and Weibull distribution with shape parameter 10. The shape of the squared segregation ratios resembles that of the information gain function.

As with the symmetric distributions we simulated 10,000,000 individuals from a backcross. Conditional on the genotype, the scale parameter of the phenotype in the two genotype groups was shifted by 10%. We examined the genotype ratios conditional on the phenotype percentile (Figure 6). As with the symmetric distributions, the shape of the information gain function parallels that of the squared deviation of the segegation ratios from $1/2$. This further demonstrates the fundamental role of the information gain function for prioritizing individuals for genotyping.

**Selective genotyping for lifetimes** Since the right tail is more informative for phenotypes with a long right tail such as lifetimes, we investigate single-tail selective genotyping, where individuals with the longest lifetimes are genotyped. We concentrate on the exponential distribution which has a central role in the analysis of lifetimes (and time-to-event data). The expected information for small effect sizes as a function of the selection fraction, $\alpha$ has a simple form:

$$J(\alpha) = \alpha + \alpha \log(\alpha)^2.$$

Comparing this with the expected information from traditional two-tail selective genotyping for normally distributed phenotypes (Figure 7) reveals important differences. Although the expected information rises more steeply for small $\alpha$, it flattens out for $\alpha$ between 20% and 70%. This is because after about 20% of the individuals have been genotyped, one-tail genotyping is no longer the most efficient strategy (as indicated by the information gain function); the best strategy is to genotype both tails after that point. Nevertheless, a one-tail genotyping strategy is simpler to implement in practice.

Next we consider the impact of genotyping cost on selective genotyping. As in SEN ET AL. (2005) we consider a simple linear cost function. Let $c$ be the cost of genotyping relative to raising and phenotyping an individual. Our goal is to maximize information relative to cost by focusing on the information-cost ratio:

$$\frac{J(\alpha)}{1 + c\alpha}.$$

The optimal selection fraction is the value of $\alpha$ that maximizes this ratio (Figure 8). We observe a "phase transition" in the optimal selection fraction when the genotyping cost is approximately half that of raising and phenotyping an individual. If genotyping is very expensive then we should genotype a small fraction of the population. As genotyping costs get smaller, the best strategy is to progressively genotype more individuals. If genotyping is cheaper than half the cost of phenotyping and rearing, the best strategy is to genotype everyone.

**Followup time for lifetimes** For many lifetimes (time-to-event phenotypes), such as time to tumorogenesis (in animals), flowering time (in plants) and lifespan, an investigator may have to decide how long to wait for the event of interest (tumorogenesis, flowering, or death, in the examples above) to occur. Individuals for whom the event has not occurred in the followup period are considered "censored" in the language of survival analysis. For these individuals we do not know the time to event exactly, but we know that it is greater than the followup time.

We consider the problem of choosing the followup duration when measuring lifetimes. We consider the tradeoffs between loss of information due to incomplete followup, and the greater cost of full followup. We develop our ideas in the context of a backcross population.

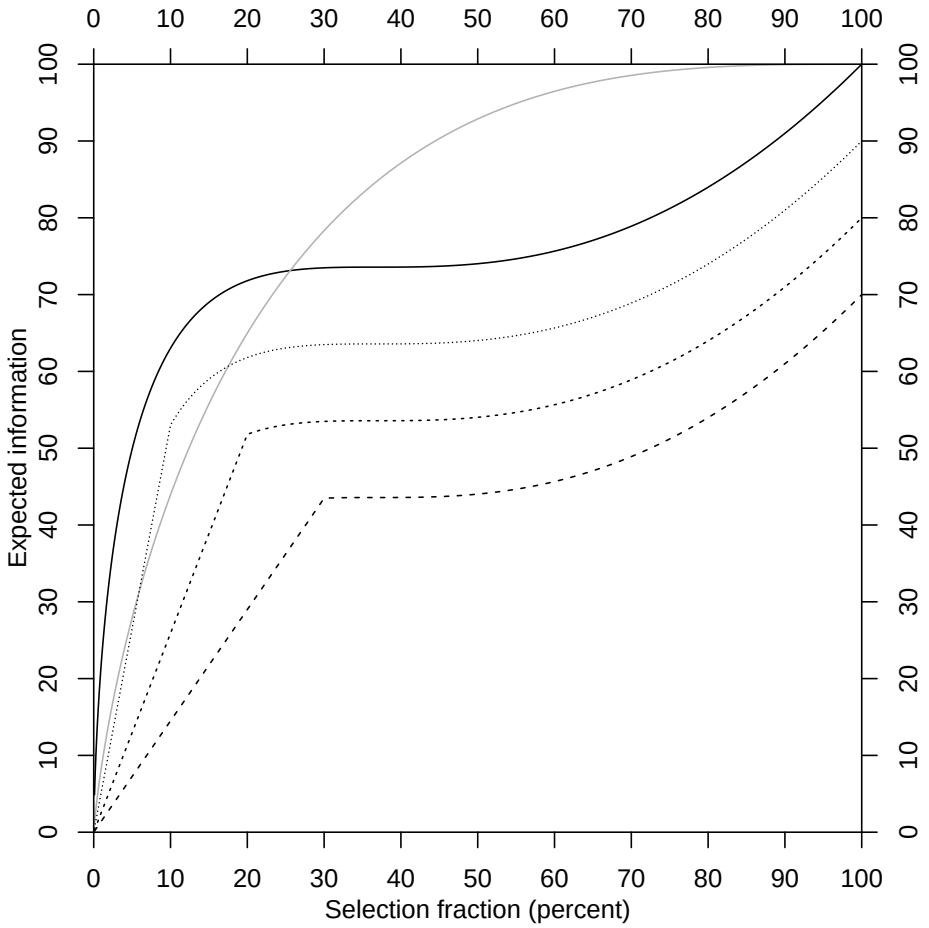


**Figure 7:** Expected information for single-tail selective genotyping as a function of the selection fraction, and proportion censored. We assume that the trait distribution is exponential, and that the effect size affecting the scale parameter is small. We assume that all individuals until a certain time are followed up, and the rest are censored. The solid black line shows the expected information for an exponential information with no censoring (100% followup). The dashed lines with increasing dash size show, respectively, the expected information with 10%, 20% and 30% censoring. The solid grey line shows, for reference, the expected information for a normal distribution with two-tail selective genotyping. With no censoring, genotyping 20% of the longest lived individuals gives us almost 75% of the information. However, the gains from selective genotyping more individuals are modest thereafter.

Let $y$ denote the time to an event, and $g$ denote the (0/1) genotype of an individual in a backcross when the event time distribution conditional on the genotype is exponential. Assume that the followup period for all individuals is $T$, $0 < T < \infty$. Then, we can write

$$p(y|g,\delta) \quad = \quad \begin{cases} f(y|g,\delta), & \text{if } y \leq T \\ \bar{F}(T|g,\delta), & \text{if } y > T, \end{cases}$$

where $f(\cdot)$ is the density function of the event times, and $\bar{F}(\cdot)$ is the survival function (the complement of the cumulative distribution function). Without loss of generality we rescale time so that

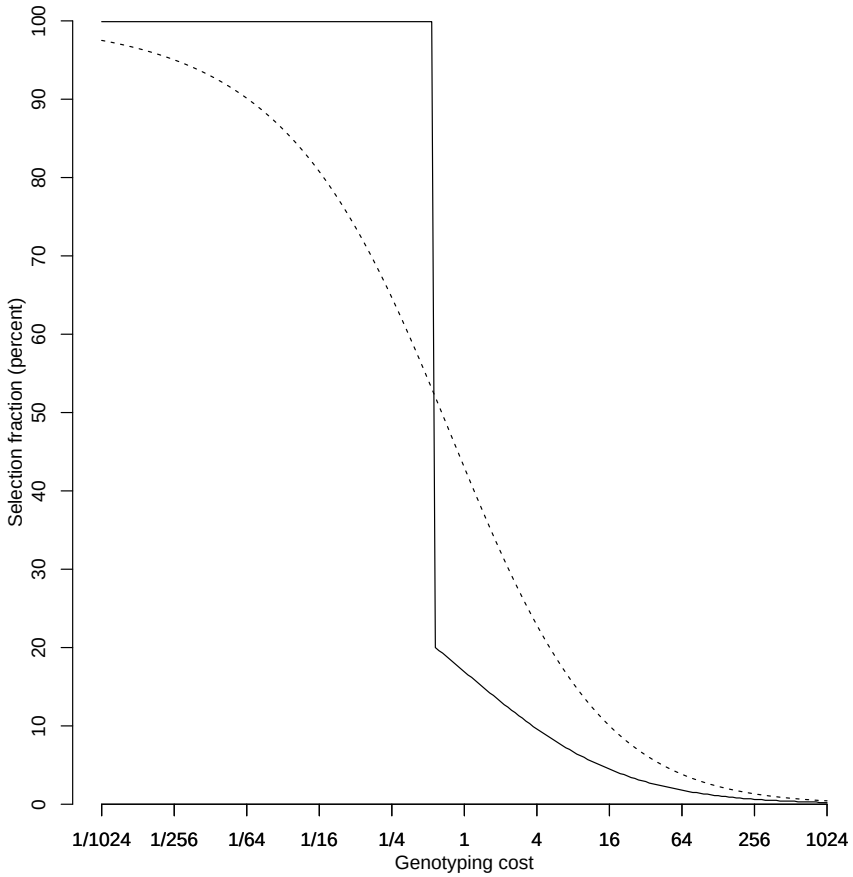


**Figure 8:** Optimal selection fraction as a function of genotyping cost for exponentially distributed waiting time phenotypes. We assume a one-tail selective genotyping scheme is being used. The cost of genotyping is measured relative to the cost of raising and phenotyping, assuming everyone is followed up with no censoring. The dotted line gives the optimal selection fraction for two-tail selective genotyping when the trait is normally distributed. As expected, it is more efficient to genotype less as the cost of genotyping increases relative to raising and phenotyping. However, for one-tail selective genotyping there is a "phase transition" or a sudden change in the optimal fraction when the cost of genotyping is comparable to the cost of phenotyping and raising (by contrast the change is gradual with traditional two-tail selective genotyping). The best strategy is to genotype everyone, or less than 20% of the individuals depending on genotyping cost.

the average waiting time for the "0" genotype is $\exp(-\delta)$ and that of the "1" genotype is $\exp(\delta)$. Then we obtain,

$$p(y|g,\delta) \quad = \quad \begin{cases} \exp(-(2g-1)\delta)\exp(-y/\exp(2(g-1)\delta)), & \text{if } y \leq T \\ \exp(-T/\exp((2g-1)\delta)), & \text{if } y > T \end{cases}$$

We use this to construct the log-likelihood function and to derive the expected Fisher information

for $\delta$,

$$\begin{aligned} I(\delta) &= 1 - \frac{\exp(-\exp(\delta)T) + \exp(-\exp(-\delta)T)}{2} \\ &= 1 - \exp(-T) + \frac{\delta^2}{2}(T^2 - T)\exp(-T) + O(\delta^4) \end{aligned} \tag{3}$$

If $C_f$ is the fixed cost per individual (for rearing, and genotyping, for example), and $C_w$ is the cost of waiting per unit time, then for followup period $T$ the information-cost ratio is

$$\frac{I(\delta)}{C_f + TC_w} \simeq \frac{1}{C_f}\frac{1-\exp(-T)}{1+TC_w/C_f} \propto \frac{1-\exp(-T)}{1+TC},$$

where $C = C_w/C_f$. Thus, if we are willing to assume that the genetic effect, $\delta$ is small, we only need to maximize the ratio $(1-\exp(-T))/(1+TC)$. Elementary calculus shows that maximizing that ratio is equivalent to solving, for $T$, the equation

$$\exp(T) - 1 - T = 1/C.$$

The solution of the equation, $T^*$, the optimal time, has a one-to-one relationship with the optimal proportion of uncensored individuals, $1-\exp(-T^*)$. Figure 9 shows the optimal proportion of uncensored individuals as a function of $C$, the ratio of the cost of followup and the fixed costs per individual. The function, `opt.wait` in the R/qtlDesign package (SEN ET AL., 2007) calculates the optimal waiting time and the optimal proportion of uncensored individuals given the cost ratio, $C$.

**Selective genotyping and followup time for lifetimes** Selectively genotyping the longest lived individuals is a good strategy for lifetime phenotypes. On the other hand, the longest lived are the most expensive to follow up, and may be censored to save cost. What is the best strategy when a fraction of the longest lived individuals are censored by design? We investigate this question when the lifetimes are exponentially distributed.

Suppose the individuals are followed up until time $T$. Treating $T$ as a parameter, we can calculate the information gain function, and the expected information, as with the previously considered distributions. The expected information for small $\delta$ is

$$\begin{aligned} J_\beta(\alpha) &= \begin{cases} \alpha - \beta + \alpha\log(\alpha)^2, & \text{if } \alpha \geq \beta \\ \alpha\log(\beta)^2, & \text{if } \alpha < \beta \end{cases} \\ &= ((\alpha-\beta)\wedge 0) + \alpha\log(\alpha\wedge\beta)^2, \end{aligned}$$

where $\beta = \exp(-T)$ is the proportion of censored individuals and $\wedge$ is the maximum operator (Figure 7). Notice that the upper bound for information with $\beta$ proportion censored is $\beta$.

The information gain function for the censored exponential distribution is

$$\begin{cases} 4(y-1)^2, & \text{if } y \leq T \\ 4T^2, & \text{if } y > T. \end{cases}$$

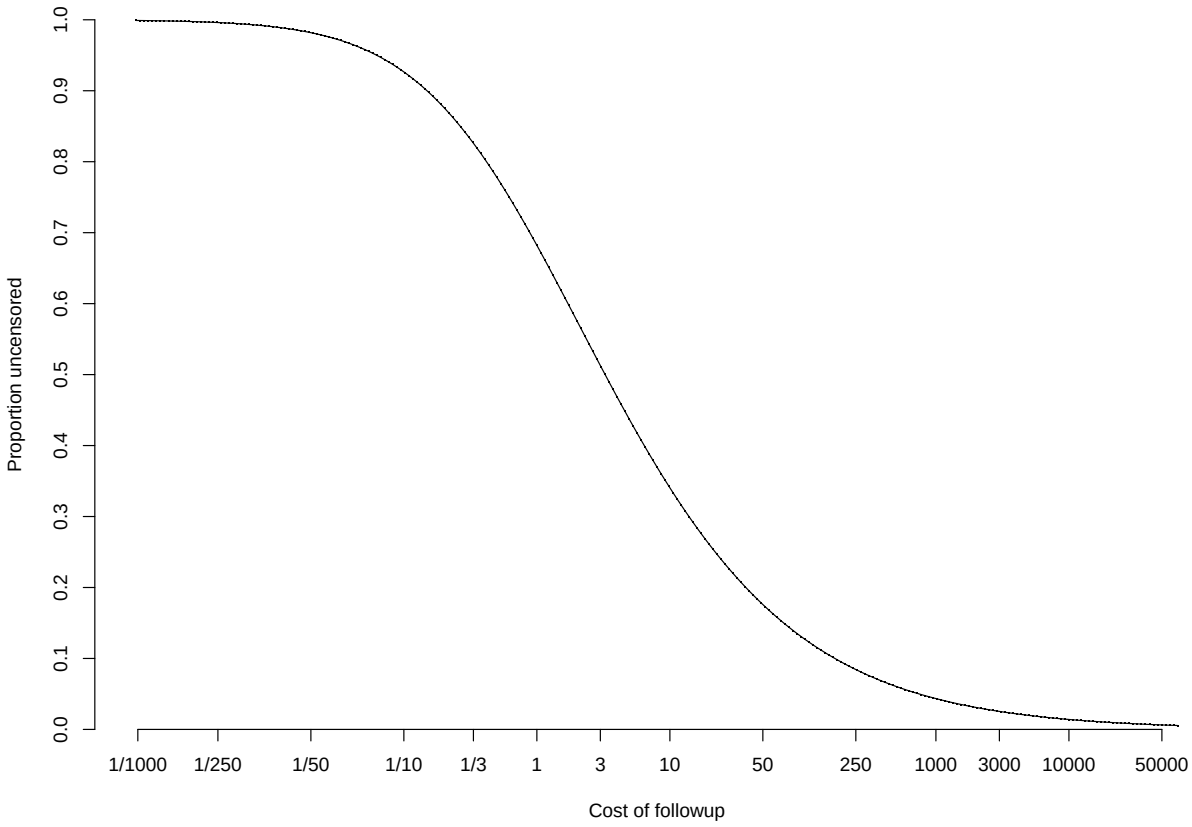


**Figure 9:** Optimal followup strategy in a backcross with exponential survival distributions. We plot the optimal proportion of uncensored individuals as a function of the cost of followup assuming that the difference in mean waiting times between two groups to be compared is small. The cost of followup is measured as the cost of following up an individual until the mean event time divided by the fixed costs per individual (genotyping, raising). The optimal proportion of individuals to be followed up decreases as the followup costs increase. If the cost of following until the mean event time is approximately the same as the fixed costs for that individual, we should follow up until approximately 70% of the events have been observed.

Figure 10 shows this function when the proportion censored is 15%. We can see that a one-tail selective genotyping strategy would be a good one, even in the presence of censoring.

As with the case with no censoring, we investigated the effect of followup cost and genotyping cost on the genotyping/followup strategy. Let $c_F$ be the cost of following up an individual for an average lifetime, and let $c_G$ be the cost of genotyping an individual. Both costs are measured relative to the fixed cost of rearing an individual. Then the cost per individual of a study that genotypes $\alpha$ proportion, and censors $\beta$ proportion of the population is

$$1 + \alpha c_G + (1 - \beta c_F),$$

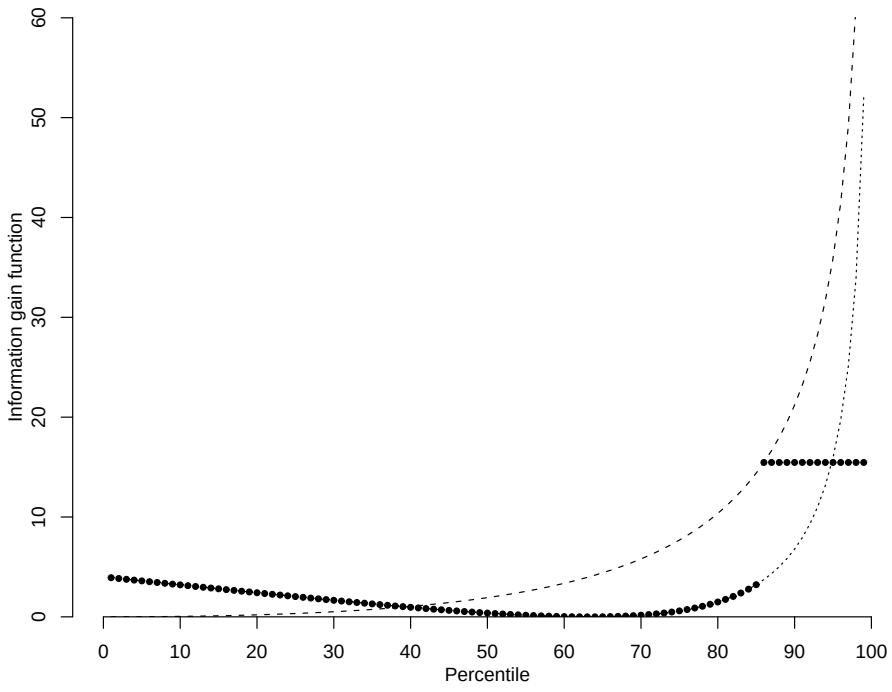


**Figure 10:** Information gain function for exponential phenotypes in the presence of censoring. The dotted line shows the information gain function for uncensored individuals as a function of the percentile of their phenotype. The dashed line shows the level of the information gain function for censored individuals as a function of the percentile of individuals who are followed up. As an example, the solid dots show the information gain function for the case when all individuals above the 85th percentile are censored. The information gain function for the first 85% of individuals follows the usual pattern for exponential phenotypes. The information gain for the censored 15% individuals is horizontal level indicated by dashed line. We see that one-tail selective genotyping is a good strategy even in the presence of censoring.

and the information cost ratio is

$$\frac{J_\beta(\alpha)}{1+\alpha c_G+(1-\beta c_F)}.$$

Given the cost structure, $(c_F, c_G)$, we can find $(\alpha, \beta)$ that minimize the information cost ratio (Figure 11). The optimal selection fraction, $\alpha$, shows an abrupt change, while the censoring proportion, $\beta$, does not. This is consistent with the optimal $\alpha$ when there is no censoring ($\beta$=0), and the optimal $\beta$ when everyone is genotyped ($\alpha$=1).

## DISCUSSION

In this paper we have analyzed data gathering strategies in QTL experimental design in the context of non-normal phenotype distributions and multi-locus models. Our approach to analyzing QTL study design is based on the information content of design choices. Genotyping and phenotyping strategies can be analyzed using this framework. This approach can provide useful

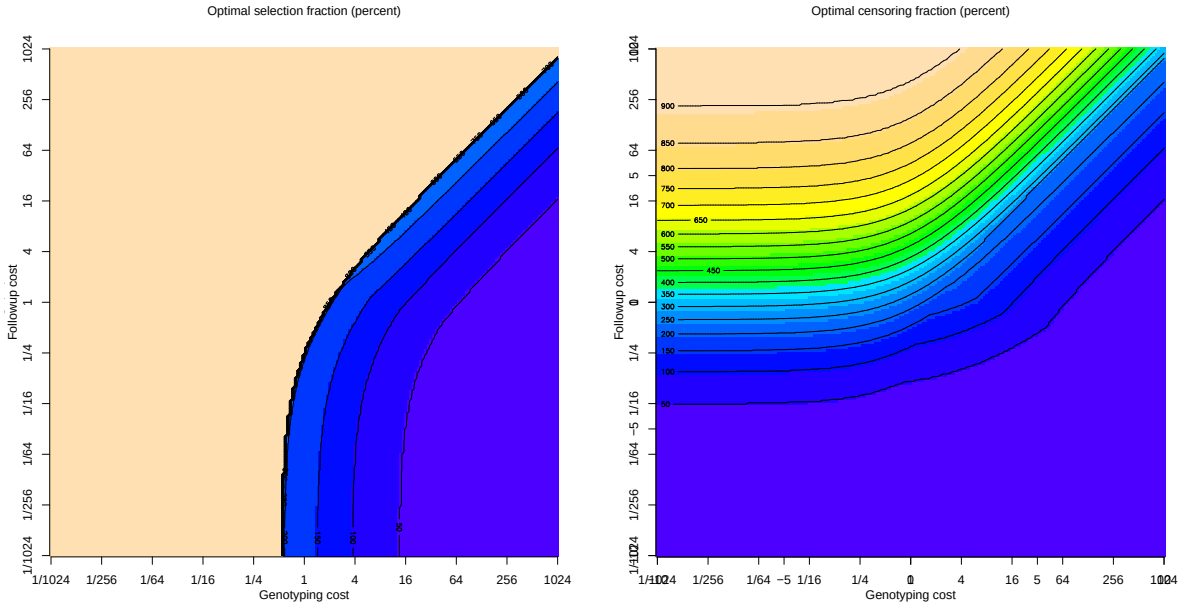


**Figure 11:** Optimal selection fraction, and censoring fraction as a function of genotyping and followup cost. The left panel shows the optimal selection fraction in percent, and the right panel shows the optimal followup proportion. A "phase transition" is observed with the selection fraction, but the followup proportion changes more gradually.

guidance for a wide range of scenarios.

A limitation of our approach is that it is model-based, asymptotic, and makes assumptions about the nature of the phenotype and genotype distributions. Our information analysis does necessarily reflect how the data will be analyzed; a sample may be more informative, but the analysis method may not take full advantage of it. However, making design choices necessitates making assumptions about yet unseen data. From that perspective, our methods allow contemplation of a range of choices.

Our information approach requires that we know (or guess) the *conditional* distribution of the phenotype given genotype. In practice, only the *marginal* distribution of the phenotype is known, thus posing difficulties for our analytic approach. However, for the most interesting and useful scenarios, when the effect size is small, the marginal and conditional distributions are approximately the same. Thus, for the purposes of selecting a genotyping scheme, it is reasonable to use the marginal phenotype distribution as a guide.

Our analysis of selective genotyping was performed in the context of a backcross population. However, we expect the conclusions to apply to more general settings including $F_2$ intercrosses, and human association studies with multiple haplotypes as SEN ET AL. (2005) showed that expected information for any contrast between haplotypes behaves the same way as in a backcross. Our results indicate that selective genotyping in genome-wide association studies may be effective since the effect sizes are expected to be small. Additionally, our results indicate how followup time for time-to-event phenotypes can be optimized in conjunction with selective genotyping.

The discontinuity in the optimal selection fraction for one-tail selective genotyping with exponen-

tially distributed phenotypes is surprising. However, as we have noted, it is not the most efficient genotyping strategy, but a "good" one that is easily implemented. A strategy devised using the information gain function will have an expected information function that will equal or exceed that for one-tail selective genotyping. That strategy will not exhibit a discontinuity in the optimal genotyping fraction as a function of cost.

Our analysis of selective genotyping for non-normal trait distributions has led us to conclusions similar to those of PARK (1996) and ZHENG AND GASTWIRTH (2000) who used more involved analytical techniques calculating the exact distribution of order statistics. Our results are asymptotic, but easily calculated.

Computer code used for symbolic algebra using Maxima (`http://maxima.sourceforge.net`), and for numerical calculation using R (`http://www.r-project.org`) will be made available at `http://www.biostat.ucsf.edu/sen`.

## ACKNOWLEDGMENTS

This research was supported by grants from the National Institutes of Health (GM074244 and GM078338).

**Table 1:** Table showing the density function and the information gain function for select distributions. The parameter of interest for the first three distribution is the location parameter. For the last three, the para-mater of interest is the scale parameter, the respective *shape* parameters being fixed. $\psi((\cdot,\cdot)$ is the incomplete Gamma function.

| Distribution | Density function | Information Gain | Expected information | Parameter |
|---|---|---|---|---|
| Normal | $\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(y-\theta)^2)$ | $y^2$ | $\alpha + 2x\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x-\theta)^2)$ | $\theta$ |
| Cauchy | $\frac{1}{\pi}\frac{1}{1+(y-\theta)^2}$ | $16\,\frac{y^2}{(y^2+1)^2}$ | $1-4\frac{\tan^{-1}(x)}{2\pi}+4\frac{(x-x^3)}{2\pi(x^2+1)^2}$ | $\theta$ |
| Logistic | $\frac{\exp((y-\theta))}{1+\exp((y-\theta))}$ | $4\,\frac{(\exp(y)-1)^2}{(\exp(y)+1)^2}$ | $2\frac{3\exp(2x)+1}{(\exp(x)+1)^3}$ | $\theta$ |
| Exponential | $(1/\sigma)\,\exp(-y/\sigma),\ \ y>0$ | $4\,(y-1)^2$ | $\alpha+\alpha\log(\alpha)^2$ | $\theta=\log(\sigma)$ |
| Gamma | $\frac{(y/\sigma)^{\nu-1}\,\exp(-y/\sigma)}{\sigma\,\Gamma(\nu)},\ \ y>0$ | $4\,(y-\nu)^2$ | $\psi(x,m+1)+(x-m)\frac{\exp(-x)x^m}{\Gamma(m+1)}$ | $\theta=\log(\sigma)$ |
| Weibull | $(\nu/\sigma)\,(y/\sigma)^{\nu-1}\,\exp(-(y/\sigma)^{\nu}),\ \ y>0$ | $4\,\nu^2\,(y^{\nu}-1)^2$ | $\alpha+\alpha\log(\alpha)^2$ | $\theta=\log(\sigma)$ |

## LITERATURE CITED


ABASHT, B. AND S. J. LAMONT (2007) Genome-wide association analysis reveals cryptic alleles as an important factor in heterosis for fatness in chicken F-2 population. *Animal Genetics*, **38**:491–498.

CARR, L., K. HABEGGER, J. SPENCE, L. LIU, L. LUMENG, AND T. FOROUD (2006) Development of congenic rat strains for alcohol consumption derived from the alcohol-preferring and non-preferring rats. *Behav. Genet.*, **36**:285–290.

CHEN, Z., G. ZHENG, K. GHOSH, AND Z. LI (2005) Linkage disequilibrium mapping of quantitative-trait Loci by selective genotyping. *Am. J. Hum. Genet.*, **77**:661–669.

COX, D. AND N. REID (2000) *The theory of the design of experiments*. Chapman and Hall/CRC.

DARVASI, A. AND M. SOLLER (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics*, **85**:353–359.

FRANKEL, W. AND N. SCHORK (1996) Who's afraid of epistasis? *Nature Genetics*, **14**:371–373.

FU, J. AND R. JANSEN (2006) Optimal design and analysis of genetic studies on gene expression. *Genetics*, **172**:1993–1999.

GALLAIS, A., L. MOREAU, AND A. CHARCOSSET (2007) Detection of marker-QTL associations by studying change in marker frequencies with selection. *Theor. Appl. Genet.*, **114**:669–681.

HUANG, B. AND D. LIN (2007) Efficient association mapping of quantitative trait loci with selective genotyping. *Am. J. Hum. Genet.*, **80**:567–576.

JANSEN, R. C. AND J. P. NAP (2001) Genetical genomics: the added value from segregation. *Trends in Genetics*, **17**:388–391.

JIN, C., H. LAN, A. D. ATTIE, G. A. CHURCHILL, AND B. S. YANDELL (2004) Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics*, **168**:2285–2293.

LANDER, E. S. AND D. BOTSTEIN (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**:185–199.

LEBOWITZ, R., M. SOLLER, AND J. BECKMANN (1987) Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics*, **73**:556–562.

LYNCH, M. AND B. WALSH (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates Inc.

MCLACHLAN, G. J. AND T. KRISHNAN (1996) *The EM algorithm and extensions*. John Wiley and Sons.

OHNO, Y., H. TANASE, T. NABIKA, K. OTSUKA, T. SASAKI, T. SUZAWA, T. MORII, Y. YAMORI, AND T. SARUTA (2000) Selective genotyping with epistasis can be utilized for a major quantitative trait locus mapping in hypertension in rats. *Genetics*, **155**:785–792.

ORCHARD, T. AND M. WOODBURY (1972) A missing information principle:theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics*, volume 1, pages 697–715.

PARK, S. (1996) Fisher information in order statistics. *Journal of the American Statistical Association*, **91**:385–390.

RAPP, J. P. (2000) Genetic analysis of inherited hypertension in the rat. *Physiological Reviews*, **80**:131–172.

SEN, S., J. M. SATAGOPAN, K. W. BROMAN, AND G. A. CHURCHILL (2007) R/qtlDesign: Inbred line cross experimental design. *Mammalian Genome*, **18**:87–93, in press.

SEN, S., J. M. SATAGOPAN, AND G. A. CHURCHILL (2005) Quantitative trait loci study design from an information perspective. *Genetics*, **170**:447–464.

WALLACE, C., J. CHAPMAN, AND D. CLAYTON (2006) Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *Am. J. Hum. Genet.*, **78**:498–504.

ZHENG, G. AND J. L. GASTWIRTH (2000) Where is the Fisher information in an ordered sample? *Statistica Sinica*, **10**:1267–1280.