Optimal screening for promising genes in 2-stage designs

B. MOERKERKE*

*Department of Applied Mathematics and Computer Science,
Ghent University, Krijgslaan 281-S9, 9000 Gent, Belgium
beatrijs.moerkerke@ugent.be*

E. GOETGHEBEUR

*Department of Applied Mathematics and Computer Science,
Ghent University, Krijgslaan 281-S9, 9000 Gent, Belgium and Department of Biostatistics,
Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA*

SUMMARY

Detecting genetic markers with biologically relevant effects remains a challenge due to multiple testing. Standard analysis methods focus on evidence against the null and protect primarily the type I error. On the other hand, the worthwhile alternative is specified for power calculations at the design stage. The balanced test as proposed by Moerkerke *and others* (2006) and Moerkerke and Goetghebeur (2006) incorporates this alternative directly in the decision criterion to achieve better power. Genetic markers are selected and ranked in order of the balance of evidence they contain against the null and the target alternative. In this paper, we build on this guiding principle to develop 2-stage designs for screening genetic markers when the cost of measurements is high. For a given marker, a first sample may already provide sufficient evidence for or against the alternative. If not, more data are gathered at the second stage which is then followed by a binary decision based on all available data. By optimizing parameters which determine the decision process over the 2 stages (such as the area of the “gray” zone which leads to the gathering of extra data), the expected cost per marker can be reduced substantially. We also demonstrate that, compared to 1-stage designs, 2-stage designs achieve a better balance between true negatives and positives for the same cost.

Keywords: Alternative p-value; Balanced test; Cost-efficient screening; False discovery rate; Gene selection; Multiple testing; Optimal designs; Two-stage designs.

1. INTRODUCTION

The hunt for genetic markers associated with phenotype is open on such a scale that false-positive results have abounded in the literature. The scientific community is thus taking extra care before declaring significance. To guide this process, many statistical decision procedures have been developed with well-known properties under an assumed global or local null hypothesis: we know what to expect from a multiple testing strategy when the markers have no true effect. Quantifying or even ordering signals in terms of

*To whom correspondence should be addressed.

the alternative promise they truly carry has, however, proved much harder, and surprisingly, no generally accepted guiding principle has been put forward yet. To screen or propose markers for further testing, one has sometimes focused on the estimated magnitude of significant effects and ignored the imprecision at that stage (see Jin *and others*, 2001). In other instances, one has considered power issues, stressing precision (Van Steen *and others*, 2005).

The fact that both strategies survive is a reflection of the logical role both elements of evidence, magnitude of effect and precision, play when considering the strength of a signal pointing in the direction of the alternative hypothesis. With an increasing ability to experiment and relatively limited resources, formal principles of experimental design have entered our practical studies. To promote meaningful results, sample sizes are generated which yield sufficient power to detect a worthwhile alternative. This worthwhile alternative allows us to combine at the screening stage observed measures of effect size and precision from the perspective of the effect size that matters. It has led Moerkerke and Goetghebeur (2006) to introduce a new strategy, the balanced test, for selecting markers and for ranking them in order of the evidence of the target alternative they carry. Markers thus selected reflect outcomes which approximate the targeted difference as closely as possible.

The described testing/screening strategy maximizes cost-effectiveness in a simple experimental setup where m genetic markers are being tested on n observational units. When measurements are costly, it pays to look at more complex designs that can exploit extra degrees of freedom to gain efficiency. In this paper, we study 2-stage designs which carry the optimal strategy above one step further and construct optimal 2-stage designs which achieve a better balance between true negatives and true target alternatives for the same cost.

Several staged designs have been proposed to address power with relatively small samples per genetic marker. Generally, in a first stage, a large set of markers are tested on a limited number of subjects, and in a second stage, remaining resources are spent on the most promising markers that passed the first stage. Satagopan *and others* (2002, 2004) and Satagopan and Elston (2003) thus design 2-stage decision strategies specifically to maximize power. Satagopan and Elston (2003) consider statistical power keeping the global significance level constrained, while Satagopan *and others* (2002, 2004) control the probability that a given number of disease markers is among the top ranking of the markers at the end of the study. Zehetmayer *and others* (2005) develop a 2-stage test procedure based on sequential p -values which control the global false discovery rate (FDR) of the study at an α -level since control of the familywise error rate at the corresponding α -level as in Satagopan and Elston (2003) is very conservative. Thomas *and others* (2004) consider 2-stage sampling designs specifically in case-control studies where the first stage is designed to select tagging single nucleotide polymorphisms (SNPs) that are further tested in the main study. All these designs eventually combine data from both stages to draw inference, a strategy with proven efficiency despite the possibly more stringent correction for multiple testing (Skol *and others*, 2006).

In spite of the above improvements, finding biologically important effects remains a challenge due to the multiple testing problem (Shephard *and others*, 2005). Lately, one has recognized and reported an expected rate of missed findings (see, e.g. Delongchamp *and others*, 2004; De Smet *and others*, 2004; Taylor *and others*, 2005; Norris and Kahn, 2006). Moerkerke *and others* (2006) include a target alternative in the decision criterion and show how this allows to discover signals that would not be found by standard methods focusing on the null. In this paper, we combine the strength of this methodology with the capacity of cost-reducing 2-stage designs. Such an approach allows decisions at the first stage, in favor of both the null and the alternative, and gathers extra data on markers for which no binary decision is made at the first stage. The added degrees of freedom allow for further optimization.

In Section 2, 2 classes of problems are introduced that motivate balanced testing in 2-stage designs. The balanced test for classical 1-stage designs is explained in more detail in Section 3 and optimized for 2-stage designs in Section 4. Section 5 compares the performance of this testing strategy in 1- and

2-stage designs based on simulations. In Section 6, we discuss the comparison of our method with the FDR-controlling 2-stage procedure of Zehetmayer *and others* (2005).

2. PROBLEM SETTING

Problems of multiple testing appear in many forms and formats in statistical genetics. We introduce 2 classes of problems to which our general development meaningfully applies.

2.1 Problem 1: detecting differentially expressed genes

An important goal in microarray experiments is to detect genes that are differentially expressed between 2 or more conditions such as different cells or tissue samples (for a review, see Huber *and others*, 2003). Various contrasts may be of interest. For simplicity, we illustrate our approach for comparing 2 means of independent $\log_2$ expression levels for each gene j ($j = 1, \dots, m$):

$$\frac{\bar{Y}_{2j} - \bar{Y}_{1j}}{\sqrt{\frac{\sigma_{2j}^2}{n_2} + \frac{\sigma_{1j}^2}{n_1}}} \quad (2.1)$$

with $\bar{Y}_{kj}$ and σ_{kj}^2 the sample mean and population variance of the $\log_2$ expression values of gene j in condition k ($k = 1, 2$) and n_k the number of samples from condition k . When variances are unknown, σ_{kj}^2 is replaced by an estimate in (2.1) but the essence of our approach will remain unchanged. To handle variances estimated from small sample sizes and large test statistics driven by small standard errors, modifications have been proposed (see, e.g. Baldi and Long, 2001; Tusher *and others*, 2001; Lönnstedt and Speed, 2002; Smyth, 2004). In microarray studies, a 2-fold change in gene expression is typically of interest (Baldi and Long, 2001). With data on the $\log_2$ scale, one often aims to select genes with a mean difference of at least 1. At the design stage of a microarray study, one can derive the sample size from σ_{kj}^2 ($k = 1, 2$ and $j = 1, \dots, m$) to detect this alternative with the desired power, as, for example, discussed in Lee and Whitmore (2002).

2.2 Problem 2: screening SNPs in whole-genome scans in a case-control setting

SNPs are DNA variants that represent variation in a single base pair. The goal of whole-genome approaches in human studies is often to find SNPs related to disease (Burton *and others*, 2005). Typically, odds ratios of at least 1.3 are of interest (see, e.g. Shephard *and others*, 2005) whether associating outcome with genotype or allele (Sasieni, 1997). Allele frequencies capture some of the variability of the association measure per SNP.

In what follows, we use the term “marker” generically to refer to genetic markers which can be genes, SNPs, etc. In both problems, interest thus lies in selecting markers with a biologically relevant effect on the trait or outcome. Our alternative-based methodology selects markers not only based on evidence against the null (classical p -values) but also on evidence against a target alternative.

We introduce the following notations:

- Δ_j denotes the population contrast of interest for marker j ($j = 1, \dots, m$). In problems 1 and 2, Δ_j is, respectively, a difference in mean $\log_2$ expression values and a (log) odds ratio. In general, it could also be a relative risk, a log hazard ratio, etc.
- Δ^0 is the value the contrast takes under the null hypothesis of no association with the phenotype.

- Δ^1 is the target value for contrast Δ_j . Not mere non-null effects are of interest but we focus on markers with an effect of at least Δ^1 . In problems 1 and 2, we have suggested a value for Δ^1 of 1 and (log) 1.3, respectively.

3. BALANCED TESTING IN A 1-STAGE DESIGN

3.1 An alternative-based selection procedure

To develop efficient designs aimed at detecting target alternatives, we build on a decision criterion for testing H_{0j} : $\Delta_j = \Delta^0$ versus H_{Aj} : $\Delta_j = \Delta^1$ as in Moerkerke *and others* (2006). Instead of focusing exclusively on the protection of type I error rates as in classical testing procedures, the type I error rate (α_j) and type II error rate (β_j) are balanced. As a result, both the magnitude and the precision of the estimated effect determine whether evidence is judged in favor of H_{0j} or H_{Aj} . Formally, the decision criterion optimizes a gain function separately for each marker j :

$$\begin{aligned} \mathcal{G}_j &= A_j \times P(\text{Accept } H_{0j} | H_{0j}) + B_j \times P(\text{Accept } H_{Aj} | H_{Aj}) \\ &= A_j \times (1 - \alpha_j(\Delta_j = \Delta^0)) + B_j \times (1 - \beta_j(\Delta_j = \Delta^1)) \end{aligned} \quad (3.1)$$

with A_j and B_j weights given to the null and the alternative. This amounts to the selection of H_{0j} rather than H_{Aj} depending on the value of $(1 - p_{0j})/(1 - p_{1j})$, where p_{0j} is the standard p -value for testing H_{0j} versus H_{Aj} and p_{1j} its counterpart from the perspective of the alternative (testing H_{Aj} versus H_{0j}).

Remarks:

1. The optimal cut point for $(1 - p_{0j})/(1 - p_{1j})$ depends on A_j/B_j , Δ^1 , and the precision of the estimated effect which is marker-specific. Hence, even with a common $A_j/B_j = A/B$, the variance structure may impose a marker-specific optimal cutoff as explained in Section 3.2.
2. The ratio A_j/B_j can be seen as the relative cost of false positives versus false negatives. Delongchamp *and others* (2004), De Smet *and others* (2004), and Norris and Kahn (2006) also extend the weighting of error rates in single hypothesis tests to the multiple testing framework. The main difference with Moerkerke *and others* (2006) and Moerkerke and Goetghebeur (2006) is their interest in detecting any non-null effect instead of a target alternative of interest. As argued by Delongchamp *and others* (2004), De Smet *and others* (2004), and Moerkerke and Goetghebeur (2006), receiver operating characteristic curves can be used to determine the “optimal” A_j/B_j in a given context. The choice of the weight ratio can, however, also be based on an objective gain function that needs to be optimized. Moerkerke *and others* (2006) introduce such marker-specific gain functions in the context of plant breeding. The weight ratio can then be defined in terms of how many generations it takes to filter out a selected null marker as opposed to how many generations it takes to achieve the same result when selecting plants based on phenotype instead of an important marker. In general, $A_j/B_j > 1$ when only a small number of markers can be further investigated and the focus needs to be on the null. $A_j/B_j < 1$ when a more generous screening principle is handled where one first and foremost wants to protect against false negatives.
3. In general, there may be reasons to let A_j , B_j , or the target alternative, which is then denoted as Δ_j^1 , vary with the marker j . For example, stronger evidence against the null may be demanded to select an SNP in case of rare allele frequencies. This can be translated into a larger weight on the null (a larger A_j/B_j) or a larger Δ_j^1 for rare frequencies. In the context of differentially expressed genes, effects of interest are not necessarily the same for all genes. There may be biologically important genes with expected smaller effects justifying a smaller Δ_j^1 . For the sake of simplicity, we take the approach of Moerkerke and Goetghebeur (2006) and consider in the sequel a common target alternative Δ^1 for all markers.

3.2 Designing optimal 1-stage designs

Let T_j denote the test statistic for testing H_{0j} versus H_{Aj} ($j = 1, \dots, m$). As in Satagopan *and others* (2004), we develop our design for a measure of marker association that is approximately normal with mean Δ_j and standard error $\sigma_{Dj}/\sqrt{n}$ where n is the sample size. In what follows, let $\Delta^0 = 0$ and $\Delta^1 > 0$. Considering standardized test statistics $T_j \sim N(\sqrt{n}\Delta_j/\sigma_{Dj}, 1)$, it is obvious that

$$T_j \stackrel{d|H_{0j}}{\approx} N(0, 1) \quad \text{and} \quad \left(T_j - \frac{\sqrt{n}\Delta^1}{\sigma_{Dj}}\right) \stackrel{d|H_{Aj}}{\approx} N(0, 1). \quad (3.2)$$

Equation (3.1) can be written as

$$\mathcal{G}_j = A_j \times P(T_j \leq c_j | H_{0j}) + B_j \times P(T_j > c_j | H_{Aj})$$

with c_j the cutoff that determines the decision based on T_j . In this case, the optimal cutoff that maximizes $\mathcal{G}_j$ has the closed form (Moerkerke *and others*, 2006)

$$c_{\text{opt},j} = \frac{\ln(A_j/B_j) + 0.5 \times (\sqrt{n}\Delta^1/\sigma_{Dj})^2}{(\sqrt{n}\Delta^1/\sigma_{Dj})}. \quad (3.3)$$

When weights A_j and B_j are common for all markers, only the different standard error $\sigma_{Dj}/\sqrt{n}$ is responsible for a marker-specific cutoff. By imposing a minimum target Δ^1 on Δ_j and not on the standardized effect as is commonly done, the magnitude and precision of the estimated signal play their distinct role in the procedure. A same observed effect smaller than Δ^1 points less in the direction of this target when it is less variable. In practice, the standard error σ_{Dj} depends on the variance of expression values and on allele frequencies which can differ dramatically over different markers.

We define cost in terms of the number of marker evaluations $n \times m$. In Section 4, we build 2-stage designs on the balanced principle where the cost is reflected by the expected number of marker evaluations.

4. BALANCED TESTING IN 2-STAGE DESIGNS

4.1 Test procedure

Two-stage designs with a screening stage preceding the second stage where final decisions are made are cost-reducing alternatives to 1-stage designs. We consider 2-stage designs where, in case of convincing evidence, decisions for markers can be made at the first stage, in favor of both the null and the alternative. First, all m markers are genotyped on n_1 individuals and tested. Then, a subset of m_2 markers for which results are inconclusive in step 1 are genotyped on n_2 additional individuals and evaluated based on the pooled data from stages 1 and 2. We are operating under the constraint that the expected cost does not exceed a given budget and the maximum sample size $n_{\max} = n_1 + n_2$ of the 2-stage design is chosen to achieve that. This expected cost equals $n_1 \times m + n_2 \times m_2^*$ with m_2^* the expected number of markers that are genotyped on extra individuals.

Let T_{j,n_k} represent the test statistic for marker j on data gathered in stage k only. As in 1-stage designs, we work with $T_{j,n_k} \sim N(\sqrt{n_k}\Delta_j/\sigma_{Dj}^{(k)}, 1)$. Assume further that data in both stages are randomly sampled from the same population. This implies that T_{j,n_1} and T_{j,n_2} are independent given the true underlying population parameters. We propose the following test statistic $T_{j,n_{\max}}$ for combining data from both stages:

$$T_{j,n_{\max}} = \frac{\sqrt{n_1}}{\sqrt{n_{\max}}} T_{j,n_1} + \frac{\sqrt{n_2}}{\sqrt{n_{\max}}} T_{j,n_2}. \quad (4.1)$$

Fig. 1. Two-stage decision strategy.

We construct a symmetric gray zone around the optimal cutoff $c_{opt,j}$ for $T_{j,n1}$ in the first stage. If the test statistic lies in the gray zone $c_{opt,j} \pm \varepsilon$, more data need to be gathered before arriving at a binary decision in favor of H_{0j} or H_{Aj} , which is then based on the optimal cutoff in the second stage ($c_{opt,j}^{(2)}$). Our strategy is graphically presented in Figure 1. When $\sigma_{Dj}^{(1)} = \sigma_{Dj}^{(2)} = \sigma_{Dj}$, (4.1) simplifies to a simple standardized test statistic that can be obtained based on the combined data and $T_{j,n_{max}} \sim N(\sqrt{n_{max}}\Delta_j/\sigma_{Dj}, 1)$ for each marker j . However, as only markers for which results are inconclusive in stage 1 are further investigated in stage 2, $T_{j,n_{max}}$ will not be normally distributed. This is accounted for in the optimization of the 2-stage designs by conditioning in stage 2 on the fact that $T_{j,n1}$ lies in the gray zone.

As data in both stages are drawn from the same population, we assume at the design stage that σ_{Dj} remains constant over both stages. Optimal choices for the parameters involved when $\sigma_{Dj}^{(1)} = \sigma_{Dj}^{(2)} = \sigma_{Dj}$ are given in Section 4.2.

4.2 Optimal designs

Acknowledging different variance structures over the markers leads to a different sample size need per marker, which may pose practical complications. We derive the sample sizes first and then discuss implications.

As each test has its own variance-dependent decision cut point, the probability that a marker goes to the second stage differs and hence the cost that results from the same maximum number of tests. A fixed budget per marker translates into a different maximum sample size per marker. Assuming that the expected number to test per marker stays below n_E , the maximum sample size $n_{max,j} = n_{1j} + n_{2j}$ follows from

$$n_E = P(\text{Accept } H_{0j} \text{ or } H_{Aj} \text{ in stage 1}) \times n_{1j} + P(\text{Go to stage 2 for marker } j) \times n_{max,j}, \quad (4.2)$$

where $P(\text{Accept } H_{0j} \text{ or } H_{Aj} \text{ in stage 1}) = 1 - P(\text{Go to stage 2 for marker } j)$ equals

$$\begin{aligned} & \left(P(T_{j,n1j} < c_{opt,j} - \varepsilon | H_{0j}) + P(T_{j,n1j} > c_{opt,j} + \varepsilon | H_{0j}) \right) P(H_{0j}) \\ & + \left(P(T_{j,n1j} < c_{opt,j} - \varepsilon | H_{Aj}) + P(T_{j,n1j} > c_{opt,j} + \varepsilon | H_{Aj}) \right) P(H_{Aj}). \end{aligned}$$

Constructing the gray zone in the first stage around the cutoff (3.3), which depends on σ_{Dj} (assumed constant over both stages), leaves parameters $c_{\text{opt},j}^{(2)}$ and ε to optimize the gain function subject to the fixed budget. The relation between ε and $n_{\max,j}$ leads to ε_j and marker-specific zones. Different detection probabilities and different sample sizes thus naturally follow from the different amounts of information the markers contain. In practice, it may, however, be difficult to specify marker-specific σ_{Dj} -values at the design stage, unless specific structures impose themselves. When different sample sizes per marker are feasible and marker-specific variances can be obtained, the added degree of freedom allows to gain efficiency. For example, consider studies where the goal is to compare a (quantitative) trait between different genotypes for each marker. The standard error of the observed measure of association differs over markers due to different proportions of genotypes. As these proportions follow simple Mendelian rules in studies where offspring is investigated, prior knowledge may be available to enable efficient designs.

When marker-specific design rules are not attainable, even though marker-specific variance structures have been recognized, one can choose to

1. take the minimum over all $n_{\max,j}$ ($j = 1, \dots, m$) values to guarantee a total cost that respects the budget (The consequence is that not all possible resources are used and valuable information may be lost.);
2. prioritize some markers and divide the budget unevenly over the markers such that the maximum sample size and detection probabilities are the same for all markers.

In what follows, we develop an optimal design for a single marker j with a fixed budget n_E . When no marker-specific criteria are of interest, this is the design for all m markers with a total expected cost of $m \times n_E$.

Assume for now that the proportion of the maximum sample size possibly investigated over both stages is fixed, for example, $n_{1j} = q \times n_{\max,j}$ and $n_{2j} = (1 - q) \times n_{\max,j}$, and choose $q = 0.5$: $n_{1j} = n_{2j} = n_{\max,j}/2$. We then seek $c_{\text{opt},j}^{(2)}$ that maximizes the expected gain (3.1) for a given value of $c_{\text{opt},j}$ and ε_j :

$$\begin{aligned} P(\text{Accept } H_{0j} | H_{0j}) &= P(\text{Accept } H_{0j} \text{ in stage 1 or 2} | H_{0j}) \\ &= P(\text{Accept } H_{0j} \text{ in stage 1} | H_{0j}) + P(\text{Accept } H_{0j} \text{ in stage 2} | H_{0j}) \\ &= P(T_{j,n_{1j}} < c_{\text{opt},j} - \varepsilon_j | H_{0j}) \\ &\quad + P(c_{\text{opt},j} - \varepsilon_j \leq T_{j,n_{1j}} \leq c_{\text{opt},j} + \varepsilon_j, T_{j,n_{\max,j}} \leq c_{\text{opt},j}^{(2)} | H_{0j}) \end{aligned}$$

and

$$\begin{aligned} P(\text{Accept } H_{Aj} | H_{Aj}) &= P(\text{Accept } H_{Aj} \text{ in stage 1 or 2} | H_{Aj}) \\ &= P(\text{Accept } H_{Aj} \text{ in stage 1} | H_{Aj}) + P(\text{Accept } H_{Aj} \text{ in stage 2} | H_{Aj}) \\ &= P(T_{j,n_{1j}} > c_{\text{opt},j} + \varepsilon_j | H_{Aj}) \\ &\quad + P(c_{\text{opt},j} - \varepsilon_j \leq T_{j,n_{1j}} \leq c_{\text{opt},j} + \varepsilon_j, T_{j,n_{\max,j}} > c_{\text{opt},j}^{(2)} | H_{Aj}). \end{aligned}$$

Under (3.2), it can be shown that $c_{\text{opt},j}^{(2)}$ that maximizes (3.1) given $c_{\text{opt},j}$ and ε_j equals

$$c_{\text{opt},j}^{(2)} = \frac{\ln(A_j/B_j) + (\sqrt{(n_{\max,j}/2)}\Delta^1/\sigma_{Dj})^2}{\sqrt{2}(\sqrt{(n_{\max,j}/2)}\Delta^1/\sigma_{Dj})}. \quad (4.3)$$

This is essentially the same optimal cutoff as in stage 1 but now based on data of both stages 1 and 2. See Section 1 of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>) for more details.

Per design, the population standard deviation σ_{Dj} stays constant over both stages. If it appears that within the samples $\sigma_{Dj}^{(1)} \neq \sigma_{Dj}^{(2)}$, for instance due to group membership proportions that differ in both stages, this can be incorporated at the stage of data analysis. σ_{Dj} in (4.3) can then be replaced by the correctly estimated standard deviation.

We find that (4.3) only depends on ε_j through $n_{\max,j}$. The question arises whether an optimal choice for ε_j is possible. When ε_j becomes very small or large, the width of the gray zone goes to zero or infinity, and we have the extreme scenarios:

$$\begin{aligned}\varepsilon_j \rightarrow 0 &: \text{Never go to second stage: } n_{\max,j}/2 = n_E, \\ \varepsilon_j \rightarrow +\infty &: \text{Always go to second stage: } n_{\max,j} = n_E.\end{aligned}$$

Both situations result in a 1-stage design with sample size n_E .

The problem is that $P(\text{Accept } H_{0j} \text{ or } H_{Aj} \text{ in stage 1})$ is in fact unknown and depends on the underlying distribution of the population effects. Given our philosophy and the role of the sharp null and alternative in the optimization procedure and ensuing decision criterion, we are contrasting the sharp null with the sharp alternative also here (see also Satagopan and Elston, 2003) and approximate $P(\text{Accept } H_{0j} \text{ or } H_{Aj} \text{ in stage 1})$ by

$$\begin{aligned} &(\Phi(c_{\text{opt},j} - \varepsilon_j) + 1 - \Phi(c_{\text{opt},j} + \varepsilon_j))P(H_{0j}) + (\Phi(c_{\text{opt},j} - \varepsilon_j - \sqrt{(n_{\max,j}/2)\Delta^1/\sigma_{Dj}}) \\ &+ 1 - \Phi(c_{\text{opt},j} + \varepsilon_j - \sqrt{(n_{\max,j}/2)\Delta^1/\sigma_{Dj}}))P(H_{Aj}) \end{aligned}$$

with $P(H_{0j}) + P(H_{Aj}) = 1$. In Sections 4.3 and 4.4, we present an algorithm that allows us to optimize the 2-stage designs when $P(H_{0j})$ and $P(H_{Aj})$ are unknown. Alternatively, the choice for $P(H_{0j})$ and $P(H_{Aj})$ can be based on prior information about the true underlying effects and $P(H_{0j})$ can be defined as the proportion of true effects smaller than Δ^1 .

In Section 4.3, we find an optimal ε_j when $P(H_{0j})$ and $P(H_{Aj})$ are known and compare the 2-stage design with corresponding 1-stage designs. In Section 4.4, we present an algorithm to obtain an optimal ε_j through numerical optimization when $P(H_{0j})$ and $P(H_{Aj})$ are unknown.

4.3 Two-stage designs versus 1-stage designs

This section illustrates optimization of a 2-stage design for a given marker j . Remember that when no marker-specific criteria are of interest, the design is the same for all m markers. Let the target effect Δ^1 equal 0.3 and assume Δ_j is either 0 or 0.3 with probability $P(H_{0j}) = 0.9$ or $P(H_{Aj}) = 0.1$, respectively, and let σ_{Dj} equal 1. The weights A_j and B_j given to the null and alternative in the gain function can be chosen to reflect financial gains following a correct decision under the null or the alternative or they can more generally reflect the relative importance of the null and the alternative in the given context. We chose the latter here and scale the weights to sum to 1. The null is considered 4 times more important than the alternative, hence $A_j = 0.8$ and $B_j = 0.2$. The available budget or expected number to test for marker j is $n_E = 80$.

Determining an optimal ε_j happens numerically. We let ε_j vary from 0 to 5 and, for each ε_j , the maximum sample size $n_{\max,j}$ of the 2-stage design is obtained under the constraint of the fixed n_E . This is an iterative process: the $(\varepsilon_j, n_{\max,j})$ combination determines the optimal cutoff in the first stage which, together with ε_j , determines n_E (see (4.2)). For each ε_j and corresponding $n_{\max,j}$, the expected gain (3.1) of the procedure is obtained. Note that the available data $n_{\max,j}$ are equally divided over both stages. The results are graphically displayed in Figure 2.

Fig. 2. Expected gain and maximum sample size in function of ε_j .

We find that the optimal ε_j for the 2-stage design is approximately equal to 0.904 with corresponding $n_{\max,j} = 132$. The horizontal dashed line on the left-hand side plot is the expected gain of a 1-stage design with sample size 80. When ε_j becomes very small, $n_{\max,j} \rightarrow 160$, and when ε_j becomes large, $n_{\max,j} \rightarrow 80$. Both limiting situations result in a 1-stage design with sample size 80, and hence the expected gain is indeed the same as for this 1-stage design.

It is clear that 2-stage designs can be optimized with respect to ε_j and that the corresponding expected gain is larger than that for the 1-stage designs with the same cost. The increase in gain may seem small but it should be noted that the definition of the gain in (3.1) is heavily scale dependent and depends on how the weights A_j and B_j are chosen. We therefore judge the increase in expected gain as a percentage of the distance between the expected gain $\mathcal{G}_j^P$ obtained by a perfectly informed decision and the expected gain $\mathcal{G}_j^N$ corresponding to a non-informed decision. More specifically,

$$\begin{aligned}\mathcal{G}_j^P &= A_j + B_j, \\ \mathcal{G}_j^N &= A_j \times P(H_{0j}) + B_j \times P(H_{Aj}).\end{aligned}\tag{4.4}$$

In this example, $\mathcal{G}_j^P - \mathcal{G}_j^N = 0.26$. The relative increase in expected gain that follows from working with a 2-stage design with $n_E = 80$ instead of a 1-stage design with the same cost is 11.2%. In contrast, the relative increase when working with a 1-stage design with sample size 132 compared to our 2-stage design with $n_E = 80$ and $n_{\max,j} = 132$ is only 1.9%.

Another type of comparison between 1- and 2-stage designs determines n_E in a 2-stage design achieving the same expected gain of a 1-stage design with sample size 80. This happens numerically: for a range of n_E s, we determine the optimal ε_j and corresponding expected gain. We select n_E , the expected number to test, that yields the expected gain of the 1-stage design. We find that the 2-stage design with $\varepsilon_j = 0.79$ and $n_E = 54$ ($n_{\max,j} = 90$) yields the expected gain of the 1-stage design with sample size 80.

Fig. 3. Expected gain in function of expected number to test for the 2-stage design.

This is graphically shown in Figure 3 and underlines how the 2-stage designs are indeed cost-reducing alternatives to 1-stage designs.

4.4 An algorithm to optimize 2-stage designs

In practice, it is unlikely that the true underlying effect Δ_j of marker j is exactly Δ^0 or exactly the target Δ^1 . However, for simplicity and in line with classical design decisions which focus on the significance level and the power from the perspective of an effect of exactly Δ^0 and Δ^1 , respectively, we prefer to optimize (3.1) and the 2-stage designs accordingly. In the study population, one may view $P(H_{0j})$ as the proportion of markers with an effect smaller than Δ^1 and $P(H_{Aj})$ as the proportion with an effect of at least Δ^1 from which follows that $P(H_{0j}) + P(H_{Aj}) = 1$. When the true $P(H_{Aj})$ is unknown, the following steps provide an algorithm that leads to an optimal 2-stage design for a marker j :

1. Fix n_E .
2. Let $\pi_A = P(H_{Aj})$ vary from 0 to 1; $P(H_{0j}) = 1 - P(H_{Aj})$.
3. For each π_A , let ε_j vary between 0 and a given maximum (in our previous example this was 5).
4. Derive from π_A and ε_j , the $n_{\max,j}$ that generates the n_E from step 1.
5. Calculate the expected gain $\mathcal{G}_j$ in the corresponding 2-stage design, choose the optimal ε_j and go to next value of π_A .

The result is a range of optimal ε_j s corresponding to every possible value of $P(H_{Aj})$. Ignoring the true $P(H_{Aj})$, one has the following options:

- Choose the minimum $n_{\max,j}$ to guarantee that the budget is respected.
- Choose the average $n_{\max,j}$, accepting that the average sample size may tend to be larger or smaller than n_E .
- Choose the maximum $n_{\max,j}$, knowing that n_E will be a lower bound for the average sample size.

The different approaches are discussed in more detail following the simulations in Section 5 of this paper and Section 4 of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>).

5. RESULTS

The aim of this section is to evaluate the achieved gain in (3.1) for 1- and 2-stage balanced testing designs under several scenarios. This gain reflects the obtained balance between correct decisions under the null and correct decisions under the alternative which is the target for optimization of the balanced test. We define the average α -level of m tests as the expected proportion of wrong rejections in the set of markers with an effect smaller than Δ^1 . The average β -level is then the expected proportion of non-rejections in the set with an effect of at least Δ^1 . A power of 1 minus the average β -level follows. For a fixed A_j/B_j , Δ^1 , and sample size, a higher gain implies a smaller α - and/or a smaller β -level. The study of the α - and β -levels separately and other error rates that may be of interest is not covered in this section. Moerkerke and Goetghebeur (2006) show that a better trade-off between α -levels and power can be obtained with the balanced test than with methods based solely on classical p -values. We investigate through simulations if this balance is further improved with 2-stage designs. In this section, we present the conclusions of these simulations. The results are provided in detail in Section 4 of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>).

One- and 2-stage designs are evaluated for $m = 3000$ markers and a budget per marker is $n_E = 50$ and 80 (the overall budget is then $m \times n_E$). The true underlying effects of the markers are generated as follows: $m \times P(\Delta_j = a)$ markers have an effect $\Delta_j = a$ with $a = 0, 0.25$, and 0.5 . The underlying distribution of the effect is thus different than what we assume in the “sharp” optimization of the algorithm in Section 4.4 ($\Delta_j = 0$ or $\Delta_j = \Delta^1$). It is of interest to see how the designs we select cope with this. We consider a target alternative Δ^1 of 0.4 and 0.5 . When $\Delta^1 = 0.4$, this target effect is not present among the markers. We consider all markers with $\Delta_j \geq \Delta^1$ of interest, and the achieved gain in the simulations is evaluated accordingly.

For the 2-stage designs, the algorithm presented in Section 4.4 is used, no prior knowledge about the distribution of the true underlying effects is incorporated. This implies that per 2-stage design, we look at 3 possible scenarios corresponding to the minimum, maximum, and average $n_{\max,j}$ over all possible configurations of $P(H_{0j})$ and $P(H_{Aj})$.

Three different series are simulated:

1. Series A: independent markers with the same variance structure over all markers.
2. Series B: 2 sets of independent markers. A first set with marker prevalence $\ell = 0.5$ and a second set with marker prevalence $\ell = 0.75$. These different prevalences result in a different variance structure.
3. Series C: correlated tests.

Section 4 of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>) contains more details on this.

For Series A and B, results show that 2-stage designs perform convincingly better than 1-stage designs. The achieved gain is higher in 2-stage designs that are as expensive as 1-stage designs. When comparing 1- and 2-stage designs with the same gain, we find that there is a remarkable reduction in cost when working with 2-stage designs. As expected, the balanced test performs better when Δ^1 is among the true effect of the markers. For the designs with the maximum $n_{\max,j}$, the expected cost is on average close to (slightly higher than) the budget, and given the small variability on the expected cost, this reflects that the expected cost for some designs is slightly above and for other designs slightly below the budget. The minimal $n_{\max,j}$ provides the smallest cost which is far below the budget, but naturally this coincides with

a smaller average gain as not all available resources are exploited. The “safest” solution in all cases is to choose the average $n_{\max,j}$, but again, we see that in general we go below the budget.

The largest improvements in gain are seen for $n_E = 50$ as the room for improvement is higher there. However, in terms of cost, we gain more when $n_E = 80$. This is also a logical result: in 1-stage designs, more extra samples are typically needed to improve a gain that is already very high than to improve a smaller gain. Two-stage designs appear to reduce this extra cost considerably.

When introducing correlation (Series C), results are comparable with those of Series A and B. However, we find that the correlation strongly affects the variability of the expected number to test per marker (expected cost per marker). This variability is now much larger reflecting the instability (Qiu *and others*, 2006) of the expected cost over all markers. Practically, this implies that it is very hard to accurately predict the cost of a study which is unfavorable. Choosing the minimal $n_{\max,j}$ keeps the cost within acceptable boundaries when taking the variability into account. Variability on the achieved gain for the simulations is very low for uncorrelated markers but also increases in the case of correlation (results not shown). The variance on this gain is in general higher for 1-stage designs than for 2-stage designs. Variability on the achieved gain reflects the variability on the average α - and β -levels, and as could be expected, correlation also introduces more instability at that level.

6. COMPARISON WITH AN FDR-CONTROLLING 2-STAGE DESIGN

Our procedure is fundamentally different from classical procedures because its decision criterion is not solely based on classical p -values. To explore the difference in strategy and resulting outcome of the 2-stage balanced test with such a classical design, we compare our approach with the procedure of Zehetmayer *and others* (2005) through simulations. Like ours, the 2-stage designs of Zehetmayer *and others* (2005) require specification of a (biologically relevant) alternative of interest Δ^1 , but they aim to control the global FDR while maximizing the power. In this section, we summarize this comparison. For more technical details and all results, we refer the reader to Section 5 of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>).

We focus on expected gain as a natural target for comparing balanced 1- and 2-stage designs, which aim to balance type I and type II error rates. When comparing the balanced and FDR-controlling procedure, a complication arises because the classical test is interested in any non-null effect while we seek to only detect effects larger than Δ^1 . To reconcile both approaches, we start by obtaining the empirical FDR of our procedure for different scenarios through simulation and use this as the controlled FDR for the corresponding optimal 2-stage designs of Zehetmayer *and others* (2005). The same Δ^1 is used for both procedures. This allows us to look at other important differences (which result from the different philosophy of both procedures), while keeping key parameters at the same level.

In Section 5, results are presented for scenarios that differ from those assumed in the “sharp” optimization of the balanced test and the procedure of Zehetmayer *and others* (2005). We return to such scenarios as they reflect more realistic situations using a different distribution than a sharp null and alternative for the true underlying effects. This means that these scenarios are not optimal for either procedure which makes comparison more fair.

In summary, we find that on average, more markers need to be selected with the FDR-controlling procedure to select the same amount of biologically relevant markers. The variability on the number of selected markers for the FDR-based method is also larger. The higher average number of selected genes can be explained by the focus on null versus non-null effects. FDR-based methods select a larger number of non-null markers for which the effect is smaller than the alternative of interest Δ^1 , our procedure avoids this by incorporating Δ^1 in the decision criterion. In the long run, this means that the procedure of Zehetmayer *and others* (2005) is more expensive as more markers that are not of scientific interest need follow-up. In addition, the higher variability on the number of selected genes in the FDR-controlling

2-stage design reflects the higher instability of FDR-controlling procedures (Qiu *and others*, 2006) which makes such procedures often unfavorable.

When introducing correlation, we find that the instability of both procedures increases as the variability on the number of selected markers increases. This is in line with the findings of Qiu *and others* (2006). We find that the effect of correlation is larger for the FDR-controlling procedure as the variability on the number of selected markers is larger. However, the distribution of the number of selected markers becomes heavily skewed for both procedures. For the balanced test, the expected cost per marker does not change on average when introducing correlation, only the standard deviation on the expected cost increases as mentioned before. Remarkably, for the procedure of Zehetmayer *and others* (2005), the expected cost also decreases on average which implies that not all available resources are fully used.

7. DISCUSSION

We found that 2-stage designs lead to a considerably lower expected cost than corresponding 1-stage designs for the balanced test. Theoretically, we minimize the expected cost function in settings where either the null or the sharp alternative holds, where it depends on the number of alternative or disease markers as in Satagopan and Elston (2003). They search for 2-stage designs with a power close to the corresponding 1-stage designs. Zehetmayer *and others* (2005) optimize power for their FDR-controlling 2-stage designs, also under the assumption that the true underlying effect of a marker is either zero or a prespecified alternative. We, however, optimize an expected gain instead of power. We take this also one step further and investigate through simulations what happens in more realistic situations where settings do not necessarily correspond to parameters used in the theoretical calculations. In all cases, 2-stage designs prove to be superior and design parameters can be chosen to keep costs within predefined boundaries. Note, however, that the expected gain for the 1-stage designs we considered is already quite high. Consequently, the space for improvement is low. Further optimizing near-optimal designs typically costs more than the same relative improvement starting from a smaller expected gain. Nevertheless, 2-stage designs need considerably less extra samples to achieve similar results.

In the algorithm presented in Section 4.4, we assume no prior information about the prevalence of the worthwhile alternative and optimize 2-stage designs for a range of possible prevalences. In reality, however, some knowledge about a plausible range of prevalences could be obtained from previous studies with estimates of the proportion of “null” and “alternative” markers. To optimally include this information, more research is needed about the relation between the optimal ε and the prevalence, together with the interplay with the weight ratio. Again, we need to keep in mind that the optimization procedure is developed for scenarios where either the null or the sharp alternative is true, while the simulations applied it to analyze data generated under more realistic alternative scenarios.

In the calculations and simulations, sample sizes are too small to be realistic for studies involving SNPs. However, the sought-after alternatives and true underlying effects (on the log scale) are in those instances also substantially smaller, resulting in a comparable standardized alternative. Correlation between different markers is for simplicity simulated as in Qiu *and others* (2006) and will typically produce results that may heavily depend on the seed of the random number generator. This need not reflect the nature of biologically induced correlations.

We have only considered 2-stage designs where the maximum sample size is equally divided over both stages. The proportion of data used in the first stage is in fact a design parameter which could also be allowed to change in order to optimize 2-stage designs. In Section 2 of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>), we incorporate this extra parameter in the 2-stage design of Section 4.3 and show that 2-stage designs can indeed be further optimized using this parameter. However, the extra level of complexity provides only a small relative increase in expected gain for this example.

If prior knowledge about the true distribution of the underlying effects is available, this should be incorporated. In Section 3 of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>), we use a normal prior distribution on the effects and show that 2-stage designs can be optimized exactly in that case. This is because the probability that more data need to be gathered for a marker after the first stage can be determined and this enables better budget control.

In summary, we find the 2-stage balanced test to provide a cost-efficient flexible approach with a clear target and good properties.

ACKNOWLEDGMENTS

The authors wish to thank Professor Xiaole Liu from the Harvard School of Public Health for helpful discussions. *Conflict of Interest*: None declared.

FUNDING

Bijzonder Onderzoeksfonds Universiteit Gent (01J16607 to B.M.), National Institutes of Health (U54 LM008748 to E.G.), Interuniversity Attraction Pole research network from the Belgian government (Belgian Science Policy) (P06/03 to E.G.).

REFERENCES

- BALDI, P. AND LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.
- BURTON, P. R., TOBIN, M. D. AND HOPPER, J. L. (2005). Key concepts in genetic epidemiology. *The Lancet* **366**, 941–951.
- DELONGCHAMP, R. R., BOWYER, J. F., CHEN, J. J. AND KODELL, R. L. (2004). Multiple-testing strategy for analyzing cDNA array data on gene expression. *Biometrics* **60**, 774–782.
- DE SMET, F., MOREAU, Y., ENGELEN, K., TIMMERMAN, D., VERGOTE, I. AND DE MOOR, B. (2004). Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer* **91**, 1160–1165.
- HUBER, W., VON HEYDEBRECK, A. AND VINGRON, M. (2003). Analysis of microarray gene expression data. In: Balding, D. J., Bishop, M. and Cannings, C. (editors), *Handbook of Statistical Genetics*, 2nd edition. Chichester, United Kingdom: John Wiley & Sons, pp. 162–187.
- JIN, W., RILEY, R. M., WOLFINGER, R. D., WHITE, K. P., PASSADOR-GURGEL, G. AND GIBSON, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29**, 389–395.
- LEE, M. L. T. AND WHITMORE, G. A. (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine* **21**, 3543–3570.
- LÖNNSTEDT, I. AND SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.
- MOERKERKE, B. AND GOETGHEBEUR, E. (2006). Selecting ‘significant’ differentially expressed genes from the combined perspective of the null and the alternative. *Journal of Computational Biology* **13**, 1513–1531.
- MOERKERKE, B., GOETGHEBEUR, E., DE RIEK, J. AND ROLDAN-RUIZ, I. (2006). Significance and impotence: towards a balanced view of the null and the alternative hypotheses in marker selection for plant breeding. *Journal of the Royal Statistical Society, Series A, Statistics in Society* **169**, 61–79.
- NORRIS, A. W. AND KAHN, C. R. (2006). Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 649–653.

- QIU, X., XIAO, Y. H., GORDON, A. AND YAKOVLEV, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics* **7**, 50.
- SASIENI, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.
- SATAGOPAN, J. M. AND ELSTON, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology* **25**, 149–157.
- SATAGOPAN, J. M., VENKATRAMAN, E. S. AND BEGG, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60**, 589–597.
- SATAGOPAN, J. M., VERBEL, D. A., VENKATRAMAN, E. S., OFFIT, K. E. AND BEGG, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* **58**, 163–170.
- SHEPHARD, N., JOHN, S., CARDON, L., MCCARTHY, M. I. AND ZEGGINI, E. (2005). Will the real disease gene please stand up? *BMC Genetics* **6** (Suppl 1), S66.
- SKOL, A. D., SCOTT, L. J., ABECASIS, G. R. AND BOEHNKE, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* **38**, 209–213.
- SMYTH, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3.
- TAYLOR, J., TIBSHIRANI, R. AND EFRON, B. (2005). The ‘miss rate’ for the analysis of gene expression data. *Biostatistics* **6**, 111–117.
- THOMAS, D., XIE, R. R. AND GEBREGZIABHER, M. (2004). Two-stage sampling designs for gene association studies. *Genetic Epidemiology* **27**, 401–414.
- TUSHER, V. G., TIBSHIRANI, R. AND CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121.
- VAN STEEN, K., MCQUEEN, M. B., HERBERT, A., RABY, B., LYON, H., DEMEO, D. L., MURPHY, A., SU, J., DATTA, S., ROSENOW, C. and others (2005). Genomic screening and replication using the same data set in family-based association testing. *Nature Genetics* **37**, 683–691.
- ZEHEMAYER, S., BAUER, P. AND POSCH, M. (2005). Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* **21**, 3771–3777.

[Received September 5, 2006; revised July 3, 2007; second revision December 31, 2007;
accepted for publication January 22, 2008]