Sequential Stopping for High-Throughput Experiments

DAVID ROSSELL*, PETER MÜLLER

March 3, 2013

Abstract

In high-throughput experiments sample size is typically chosen informally. Most formal sample size calculations depend critically on prior knowledge. We propose a sequential strategy which, by updating knowledge when new data is available, depends less critically on prior assumptions. Experiments are stopped or continued based on the potential benefits in obtaining additional data. The underlying decision-theoretic framework guarantees the design to proceed in a coherent fashion. We propose intuitively appealing, easy to implement utility functions. As in most sequential design problems, an exact solution is prohibitive. We propose a simulation-based approximation that uses decision boundaries. We apply the method to RNA-seq, microarray and reverse phase protein array studies and show its potential advantages. The approach has been added to the Bioconductor package `gaga`. Decision theory; Forward simulation; High-throughput

[0] To whom correspondence should be addressed.

experiments; Multiple testing; Optimal design; Sample size; Sequential design.

1 Introduction

In high-throughput studies (HTS) the sample size is usually chosen informally. The resulting experiment may either be not informative enough or unnecessarily extensive. To address this problem, we aim to develop a sequential design framework for HTS. That is, we investigate the question whether the currently available data in a typical HTS suffices and, if not, how to determine the optimal stopping strategy. We focus on experiments to perform group comparisons, although our ideas remain useful for other inferential goals. For simplicity we discuss the 2 group case, but our software allows > 2 groups. The proposal is based on Bayesian decision theory, so that decisions are coherent with respect to an underlying utility function and probability model. We emphasize ease of interpretation and use.

Several authors proposed fixed sample size calculations for HTS, *i.e.* the sample size is fixed at the beginning of the experiment (Dobbin and Simon; 2007; Lee and Whitmore; 2004; Müller et al.; 2004; Zien et al.; 2003; Pan et al.; 2002). The main limitations are the lack of flexibility to incorporate new data and the need for a good prior guess of certain features, *e.g.* effect sizes or the proportion of differentially expressed (DE) genes.

In contrast, sequential sample size designs update knowledge and make decisions as data is collected, *i.e.* they are robust with respect to prior choices. A sequential design stops or continues experimentation on the basis of all available data. A potential drawback is the need to carry out experimentation in batches. The associated increase in time or experimentation costs may outweigh the potential advantages. This should not be a major concern in many HTS, as most high-throughput technologies (*e.g.* microarrays, sequencing or mass-spectrometry) process samples in small batches. Assessing the promise of continuing experimentation after each batch seems natural. Also, samples may be costly to obtain or there may be ethical concerns, *e.g.* in human studies. These situations offer great potential for sequential strategies.

Ruppert et al. (2007), Tibshirani (2006) and Ferreira and Zwinderman (2006) proposed two-step designs focused on microarray differential expression problems. Two-step designs adapt to the observed data to a limited extent. Gibbons et al. (2005) and Durrieu and Briollais (2009) propose sequential designs that select a single/few genes and stop the trial when differences in expression can be estimated with high precision. The focus on a few genes limits the application to HTS.

Researchers typically use HTS as a screening test to identify candidates, which are then validated with more precise techniques (*e.g.* real-time PCR). The usual goal is not to estimate differential expression accurately but to find promising targets. The Durrieu and Briollais (2009) model is appropriate for paired observations, *e.g.* two-channel arrays.

We propose an approach for unpaired data that screens a large number of candidates and attempts to maximize the number of promising targets. The framework is directly applicable to many probability models and experiments, including sequencing, microarrays and reverse phase protein arrays (RPPA). With minor modifications it can be adapted to other experimental goals. The main hurdle with decision-theoretic optimal sequential designs is the prohibitive computational cost, even in single outcome experiments. Rossell et al. (2006) developed an approach based on the ideas of Müller et al.

(2006), Brockwell and Kadane (2003) and Carlin et al. (1998). They compute a summary statistic S each time that new data is observed and they use decision boundaries that partition the sample space. The experiment is terminated when S first falls in the stopping region. The sequential problem is reduced to the (non-sequential) problem of finding optimal boundaries. The choice of these boundaries accounts for all future data, which distinguishes the solution from myopic approximations. Here we extend these ideas to high dimensional data and apply them to differential expression problems.

Section 2 formalizes the problem and two convenient probability models. Section 3 describes sequential stopping and the infeasibility of an exact solution. Section 4 proposes an approximate solution. Section 5 contains several examples and Section 6 some concluding remarks. The Supplementary Material (SM) contains further theoretical and practical considerations, and an example with R code.

2 Data Format and Model

We motivate the discussion in the context of experiments that study differential gene expresion, but the proposal remains applicable to other setups. Let n be the number of outcomes (*e.g.* genes) and T be the maximum sample size. T is usually determined by budget constraints, accrual rates, or an informed guess. Let x_{ij} be the measurement for gene $i = 1, \ldots, n$ and sample $j = 1, \ldots, T$, and $z_j \in \{0, \ldots, n_z\}$ be the group of sample j. For simplicity here we assume $z_j \in \{0, 1\}$, *i.e.* we compare two groups. Generalization to $n_z > 1$ is straightforward and is implemented in the `gaga` package.

A latent variable $\delta_i = 1$ indicates that gene i is differentially expressed (DE) across groups and $\delta_i = 0$ that it is equally expressed (EE). The indicator

δ_i represents the unknown truth and is part of the parameter vector. Let $\boldsymbol{\theta}_i$ be parameters indexing a probability model for $(x_{i1}, \ldots, x_{iT})$. Optionally, let $\boldsymbol{\omega}$ be additional hyper-parameters. For example, $\boldsymbol{\omega}$ could index a regression on important covariates. Let $\mathbf{x}_t = \{x_{it}, 1 \leq i \leq n\}$ be the data obtained at time t and $\mathbf{x}_{1:t} = \{x_{ij}, 1 \leq i \leq n, 1 \leq j \leq t\}$ be all data available up to time t. Further, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$.

2.1 Probability model

Our proposal requires extensive predictive simulation and model fitting. Hence, the model must be computationally efficient. For instance, the examples in Section 5 required posterior inference in millions of simulated datasets. On the other hand, the model needs to be sufficiently flexible to capture the important features of the data.

Here we use the GaGa (Rossell; 2009) and log-normal normal with generalized variances (NN) models (Yuan and Kendziorski; 2006) models to illustrate the approach. Both offer a reasonable compromise between flexibility and computational cost. The GaGa model assumes $x_{ij} \sim \text{Ga}(\alpha_i, \alpha_i/\lambda_{iz_j})$. The NN model ueses $x_{ij} \sim N(\mu_{iz_j}, \sigma_i^2)$. The triple $\boldsymbol{\theta}_i = (\lambda_{i0}, \lambda_{i1}, \alpha_i)$ (GaGa) or $\boldsymbol{\theta}_i = (\mu_{i0}, \mu_{i1}, \sigma_i^2)$ (in the NN model) incorporates gene-specific variability and gene-by-group specific means. A hierarchical prior on $\boldsymbol{\theta}_i$ assigns positive prior probability to means being equal across groups. The GaGa hierarchical prior is

$$\begin{aligned} \lambda_{i0}^{-1} &\sim \text{Ga}(\alpha_0, \alpha_0/\nu), \quad \alpha_i|\beta, \mu \sim \text{Ga}(\beta, \beta/\mu) \\ \lambda_{i1}^{-1} \mid \lambda_{i0}, \delta_i &\sim \begin{cases} \text{Ga}(\alpha_0, \alpha_0/\nu) & \text{if } \delta_i = 1 \\ I(\lambda_{i1} = \lambda_{i0}) & \text{if } \delta_i = 0 \end{cases} \end{aligned} \tag{1}$$

and $P(\delta_i = 1) = \pi$, independently across i. The hyper-parameters are $\boldsymbol{\omega} = (\alpha_0, \nu, \beta, \mu, \pi)$. The NN hierarchical prior is $\mu_{i0} \sim N(\mu_0, \tau_0^2)$, $\sigma_i^{-2} \sim \text{Ga}(\nu_0/2, \nu_0\sigma_0^2/2)$, also with probability of ties, $P(\delta_i = 1) = \pi$, independently across i. For this model $\boldsymbol{\omega} = (\mu_0, \tau_0, \nu_0, \sigma_0, \pi)$. The GaGa sampling distribution for x_{ij} captures asymmetries that are frequently observed in HTS. The NN assumptions are similar to those of the popular limma approach (Smyth; 2004), which has been found useful in many applications. The SM (Section 2) proposes goodness-of-fit assessments to help choose the most appropriate model for a particular dataset.

In terms of computational complexity, conditional on $\boldsymbol{\omega}$ the posterior distributions are available in closed form. We treat $\boldsymbol{\omega}$ as fixed, avoiding the need for Markov Chain Monte Carlo (MCMC) simulation. This substantially increases computational speed. We estimate $\boldsymbol{\omega}$ via expectation-maximization as in Rossell (2009) and Yuan and Kendziorski (2006). The latter proposed a method of moments estimate for (ν_0, σ_0^2) which can result in over-estimating π. We illustrate this issue and outline a simple procedure to adjust $\hat{\pi}$ in the SM (Section 3).

While we use these two models in our examples, the upcoming discussion of the optimal stopping policy remains valid for any alternative probability model.

2.2 Pre-processing

We assume that the data are suitably pre-processed. This is critical for meaningful inference. For instance, ignoring batch effects may bias or add uncertainty to group comparisons. We note that some technologies such as RNA-seq may be less sensitive to batch effects, and that these can be partially mitigated by good design, *e.g.* by balancing the number of samples in each group and batch. We recommend jointly pre-processing data after every batch, as some technical biases (*e.g.* probe or GC-content biases) may be better assessed once more data is collected.

Batch effects and other sources of variability may be either addressed in the pre-processing or in the analysis by including appropriate terms in the model. Following **?** and Durrieu and Briollais (2009), we argue in favor of the former. As an illustration, let $\mathbf{y}_{ij}$ be a vector of covariates that are used in the adjustment, and assume that $E(x_{ij}|\mu_{iz_j}, \mathbf{y}_{ij}) = \mu_{iz_j} + g(\mathbf{y}_{ij})$. Here, $g(\cdot)$ captures the effect of $\mathbf{y}_{ij}$ on the outcome, and could represent a non-linear adjustment that cannot be captured by the analysis model. One could then use the partial residuals $\tilde{x}_{ij} = x_{ij} - \hat{g}(\mathbf{y_{ij}})$ as the pre-processed data, where $\hat{g}(\cdot)$ is an appropriate estimate of $g(\cdot)$.

We note that the domain of the data must match the assumptions of the model. For example, while most technologies deliver positive expression measurements, pre-processed data may sometimes present negative values which are not allowed in the GaGa model. A simple strategy to deal with negative values is to define $\tilde{x}_{iz_j} = x_{iz_j} + k$, where the offset $k > 0$ ensures that $\tilde{x}_{iz_j} > 0$. Alternatively, define $\tilde{x}_{iz_j} = e^{x_{iz_j}}$, but this option may produce outliers that decrease the model goodness-of-fit. In practice, we recommend trying several transformations and producing some goodness-of-fit (*e.g.* see SM ,Section 2).

3 Optimal Sequential Stopping

3.1 Decision criterion

We formalize sequential sample size calculation within a Bayesian decision-theoretic framework. The optimal design is chosen by maximizing the expectation of an appropriate utility function. At each decision time the expected utility is conditional on all available data (which may be no data at all) and averaged with respect to uncertainty on the model parameters and future data, assuming optimal future decisions.

It is convenient to distinguish sequential and terminal decisions. Sequential decisions correspond to stopping versus continuation and are made after each batch of observations. Terminal decisions are the classification of genes into EE ($\delta_i = 0$) or DE ($\delta_i = 1$), and are taken only upon stopping. Let $s_t = s(\mathbf{x}_{1:t}) = 1$ indicate the sequential decision of stopping at time t and let $s_t = 0$ indicate continuation. Equivalently, s_t can be described by the stopping time $\tau = \min\{t : \ s_t = 1\}$. We use s_t and τ interchangeably. Let $d_i(\mathbf{x}_{1:t}) = 1$ (0) indicate the terminal decision to report gene i as DE (EE). Also, let $\mathbf{d}(\mathbf{x}_{1:t}) = (d_1(\mathbf{x}_{1:t}), \ldots, d_n(\mathbf{x}_{1:t}))$. Both s_t and $\mathbf{d}(\mathbf{x}_{1:t})$ depend on all data available up to time t.

In a fully decision theoretic approach sequential and terminal decisions are chosen to jointly maximize expected utility. Instead, we assume a fixed rule for $\mathbf{d}(\mathbf{x}_{1:t})$ and focus on the optimal choice of s_t only. We take terminal decisions using the Bayes rule of Müller et al. (2004) to control the posterior expected false discovery rate (FDR) below some specified level. The posterior expected FDR is $\frac{1}{D}\sum d_i(\boldsymbol{x}_{1:t})[1 - E(\delta_i \mid \boldsymbol{x}_{1:t})]$, where $D = \sum_i d_i(\boldsymbol{x}_{1:t})$ is the number of reported positives. We use the 0.05 level throughout.

Sequential stopping decisions s_t are based on a utility function with sampling cost c and a unit reward for each correctly identified DE outcome

$$u(s_t = 1, \mathbf{d}(\boldsymbol{x}_{1:\tau}), \boldsymbol{x}_{1:\tau}, \boldsymbol{\delta}) = -c\tau + \sum_{i=1}^{n} \delta_i d_i(\boldsymbol{x}_{1:\tau}). \tag{2}$$

The second term in (2) is the number of true positives (TP). The cost c is the minimum number of TP that make it worthwhile to obtain 1 more sample. This interpretation allows for easy elicitation of c, without any reference to the formal mathematical framework. The utility function (2) focuses on statistical rather than biological significance, as the size of the effect is not considered. A simple alternative is obtained by substituting $|\mu_{i1} - \mu_{i2}|\, \delta_i d_i(\boldsymbol{x}_{1:\tau})$ in the summation in (2). See Müller et al. (2004) or Rice et al. (2008) for other interesting alternatives. The upcoming discussion is independent of the specified utility.

3.2 Optimal rule

The optimal stopping decision s_t maximizes $u(\cdot)$, in expectation over all unknowns, including parameters $(\boldsymbol{\theta}, \boldsymbol{\delta})$ and future data $\mathbf{x}_{\tau+1:T}$. An exact solution requires dynamic programming, also known as backward induction (DeGroot; 1970). At time t, the optimal decision is to stop if the posterior expected utility for $s_t = 1$, denoted $\overline{u}_t(s_t = 1, \mathbf{x}_{1:t})$, is greater than $\overline{u}_t(s_t = 0, \mathbf{x}_{1:t})$. Evaluating $\overline{u}_t(s_t = 1, \mathbf{x}_{1:t})$ is usually straightforward. For (2) we find

$$\overline{u}_t(s_t = 1, \boldsymbol{x}_{1:t}) = -ct + \sum_{i=1}^{n} P(\delta_i = 1 \mid \boldsymbol{x}_{1:t}) d_i(\boldsymbol{x}_{1:t}), \tag{3}$$

where $P(\delta_i = 1|\boldsymbol{x}_{1:t})$ is the posterior probability that outcome i is DE. The expectation is with respect to $\boldsymbol{\delta}$ only, as we fix the terminal decision $\mathbf{d}(\mathbf{x}_{1:t})$. The posterior probability $P(\delta_i = 1|\boldsymbol{x}_{1:t})$ can be computed in closed-form for some models (including GaGa and NN) or can easily be estimated from

MCMC output. Evaluating $\overline{u}_t(s_t = 0, \boldsymbol{x}_{1:t})$ is more challenging. An exact solution requires assessing the expected utility for all possible future data trajectories $\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_T$, substituting the optimal decisions $s_{t+1}, \ldots, s_T$. The computational cost is prohibitive.

4 Approximation by Optimal Decision Boundaries

Berry et al. (2001), Brockwell and Kadane (2003), DeGroot (2004), and Müller et al. (2006) discuss alternatives to an exact optimal sequential solution. Following Rossell et al. (2006) we define sequential stopping boundaries. We restrict the maximization to rules that depend on the data $\boldsymbol{x}_{1:t}$ only through a summary statistic S_t and linear boundaries that partition the sample space. We propose using $S_t = \Delta_t \mathrm{U}$, where

$$\Delta_t \mathrm{U} \equiv E_{\boldsymbol{x}_{t+1}}\left[\overline{u}_{t+1}(s_{t+1} = 1, \mathbf{x}_{1:t+1}) \mid \boldsymbol{x}_{1:t}\right] - \overline{u}_t(s_t = 1, \boldsymbol{x}_{1:t}) - c$$

is the 1-step ahead increase in expected utility and $E_{\boldsymbol{x}_{t+1}}(\cdot | \boldsymbol{x}_{1:t})$ conditions on $\boldsymbol{x}_{1:t}$ and marginalizes with respect to future data $\boldsymbol{x}_{t+1}$. For (2) we find $\Delta_t \mathrm{U} = \Delta_t(\mathrm{TP}) - c$, i.e., $\Delta_t \mathrm{U}$ is the expected increase in TP, and decision boundaries can equivalently be written in terms of $\Delta_t(\mathrm{TP})$.

Consider the example in Figure 1. The thick black line is a decision boundary. Every time we observe new data we compute $\Delta_t(\mathrm{TP})$. If $\Delta_t(\mathrm{TP})$ lies above the boundary we continue experimentation, otherwise we stop. That is, we experiment as long as enough new TP are expected.

Let $\boldsymbol{b} = (b_0, b_1)$ be the intercept and slope defining the linear boundaries, and let $U(\boldsymbol{b}, \boldsymbol{x}_{1:t})$ be the associated expected utility given data up to time t. In other words, $U(\boldsymbol{b}, \boldsymbol{x}_{1:t})$ is the expected utility conditional on $\boldsymbol{x}_{1:t}$ when the stopping decision is based on a decision boundary indexed by $\boldsymbol{b}$. Algorithm 1 details a forward simulation algorithm (Carlin et al.; 1998) to evaluate the required expectations, and a grid search to carry out the maximization of $U(\boldsymbol{b}, \boldsymbol{x}_{1:t})$ with respect to $\boldsymbol{b}$. The algorithm assumes that t samples are already available. For no data use $t = 0$ and $\boldsymbol{x}_{1:t} = \emptyset$ (see Section 5.3).

Algorithm 1: Optimal sequential boundary determination.

1. *Forward simulation.* Simulate $\boldsymbol{x}^{(j)}_{t+1:T}$ from the posterior predictive $P(\boldsymbol{x}_{t+1:T} \mid \boldsymbol{x}_{1:t})$, $j = 1, \ldots, B$. For each $\boldsymbol{x}^{(j)}_k$ compute $\Delta_t(\mathrm{TP})^{(j)}$, $k = t+1, \ldots, T$.

2. *Grid search.* For each $\boldsymbol{b}$, find the stopping times $\tau^{(j)}$ for all saved trajectories $\Delta_t(\mathrm{TP})^{(j)}$.

3. *Optimum.* Select $\boldsymbol{b}^\star \equiv \arg\max_{\boldsymbol{b}}\{\overline{U}(\boldsymbol{b}, \boldsymbol{x}_{1:t})\}$, where $\overline{U}(\boldsymbol{b}, \boldsymbol{x}_{1:t}) = \frac{1}{B}\sum_{j=1}^{B} \overline{u}(s_{\tau^{(j)}} = 1, \boldsymbol{x}^{(j)}_{1:\tau^{(j)}})$

Figure 1 shows simulated $\Delta_t(\mathrm{TP})$ as grey lines. For each boundary $\boldsymbol{b}$, we determine the stopping time for each trajectory and average the expected terminal utilities. At $t = T-1$ we do not determine stopping using $\boldsymbol{b}$ but the optimal rule $\Delta_{T-1}\mathrm{U} > 0$.

In principle $\boldsymbol{b}^\star$ can be re-computed every time that new data is observed. Re-computation can help to decide between multiple optima and update $P(\boldsymbol{x}_{t+1:T} \mid \boldsymbol{x}_{1:t})$. In our examples we determine $\boldsymbol{b}^\star$ only once, either based on a pilot dataset or prior knowledge, but we indicate the usefulness of re-computation when appropriate.

Besides the intuitive appeal, some theoretical considerations motivate our approach. First, fixed-sample designs are special cases *e.g.* $\boldsymbol{b} = c(4.5, \infty)$ results in a fixed sample size of 5. The myopic rule of continuing as long as $\Delta_t\mathrm{U} > 0$ (Berry and Fristedt; 1985, chapter 7), is the special case $\boldsymbol{b} = (c, 0)$.

We generalize the idea with an arbitrary boundary on Δ_tU. An important assurance is that Δ_tTP converges to 0 as $t \rightarrow \infty$, which guarantees eventual stopping. See SM (Section 1) for a formal statement and proof.

5 Examples

We compare our approach and the fixed sample designs of Müller et al. (2004) in several important experimental conditions. The SM discusses preprocessing and goodness-of-fit (Section 2) and an additional RNA-seq example with R code (Section 3).

5.1 Simulated Microarray Study

We plan collecting data in batches of 2 arrays per group, with a maximum of 20 per group (*i.e.* T=10 batches). Recall that c is the minimum number of new DE that compensate the cost of one more batch. We consider c =25, 50 and 100. To keep the simulation realistic we estimated the hyper-parameters based on data from a study of leukemia microarray data (Armstrong et al.; 2002). We focus on 24 acute lymphoblastic leukemia (ALL) and 18 MLL trans-location samples. The estimated proportion of DE genes is $\hat{\pi}$=0.063 under the GaGa model and 0.05 for the NN model. We find optimal strategies based on $\pi = \hat{\pi}$, but we assess performance under model miss-specification by also simulating data under $\pi = 0.5\hat{\pi}$ and $\pi = 2\hat{\pi}$, while leaving $\pi = \hat{\pi}$ unchanged in the analysis model. We obtained 250 simulations under each scenario.

Figure 1 shows the optimal boundary for $c = 50$ and simulated Δ_t(TP) (gray lines) under the GaGa model. Table 1 reports expected utilities and stopping times. The optimal fixed sample sizes for $c = 50$ under the GaGa

		$\pi = 0.5\hat{\pi}$		$\pi = \hat{\pi}$		$\pi = 2\hat{\pi}$	
c	t_F^*	t_S^*	$U_S^* - U_F^*$	t_S^*	$U_S^* - U_F^*$	t_S^*	$U_S^* - U_F^*$
GaGa							
25	7	6.0	7.7	7.1	0.4	10.0	34.2
50	5	4.1	16.6	5.0	0.0	6.9	28.7
100	3	3.0	0.0	3.0	0.0	4.0	58.6
NN							
25	7	5.6	11.1	7.2	0	10.0	32.4
50	4	3.2	10.8	4.1	0.1	5.7	51.2
100	3	2.0	41.9	3.0	0.1	3.0	0

Table 1: Simulated data. t_F^*: fixed sample size; t_S^*: average sequential sample size; $U_S^* - U_F^*$: expected utility for sequential design minus expected utility for fixed sample.

and NN models are $t_F^* = 5$ and $t_F^* = 4$, respectively. When $\pi = \hat{\pi}$ the expected sequential sample sizes are 5.0 and 4.1 (respectively) and there is no gain in posterior expected utility. Sequential designs offer little advantages when the data matches the prior expectations. However, when prior expectations are unrealistic sequential designs adapt to the observed data. When π was overstated by the prior ($\pi = 0.5\hat{\pi}$), sequential designs stopped earlier than the fixed sample size designs. Conversely, when $\pi = 2\hat{\pi}$ they stopped later so that more DE genes could be found. For instance, for $c = 50$ the GaGa sequential design requires 4.1 data batches when $\pi = 0.5\hat{\pi}$ and 6.9 when $\pi = 2\hat{\pi}$. The fixed design always requires 5.

5.2 High-throughput Sequencing Example

We use a pilot RNA-seq dataset with 2 muscle and 1 brain human samples to design two hypothetical studies. Study 1 compares gene expression for muscle *vs.* brain. Many DE genes are expected. Hypothetical Study 2 compares the two muscle samples. No genes should be DE. In both cases we use 1 sample per group as pilot data. We consider up to $T = 5$ more samples, in batches of 1 sample. The GaGa model provided a reasonable fit to these data (SM, Section 2).

We determined the optimal boundary for sampling costs $c = 0, 1, \ldots, 100$. Figure 2(a) shows that $\Delta_t(\mathrm{TP})$ is maximal for $t = 2$ additional data batches. As suggested by Theorem 1 (SM, Section 1), the incremental reward decreases as t grows further. The dashed boundary shows that for $c > 66$ the optimal decision is to stop experimentation. For $c \leq 66$ there are multiple optimal $\boldsymbol{b}^*$. The solid black lines show two optima. In both cases, the decision at $t = 0$ is to continue. Since the simulated trajectories do not cross either boundary, we expect experimentation to continue up to $T = 5$. The future realized $\Delta_t(\mathrm{TP})$ might cross the boundary, in which case the design would adapt and stop experimentation before $T = 5$. Given that the pilot data contains 1 sample per group, we would re-determine $\boldsymbol{b}^\star$ upon observing new data.

The hypothetical muscle *vs.* muscle comparison simulation is shown in Figure 2(b). In this case $\Delta_t(\mathrm{TP})$ is negligible and the optimal design is to stop at $t = 0$ (*i.e.* not to collect any further data) for any $c > 2$. The result seems sensible as no DE genes are expected.

5.3 Microarray Example

We consider the leukemia study of Campo Dell'Orto et al. (2007) recording mRNA expression for 21 ALL and 15 MLL patients and 54,675 genes. We consider designing the study before any data was available. In such circumstances, one could estimate the hyper-parameters $\boldsymbol{\omega}$ from a similar study. We used the Armstrong et al. (2002) study (Section 5.1) as it was also carried on ALL/MLL patients and used the same microarray platform. Once fixed and sequential designs were determined, we used the historical data to compare performance. We use batches of 2 arrays per group, maximum T=7 batches and $c = 50$.

The white bars in Figure 3 (left panels) show expected utility under the GaGa and the NN priors for all fixed sample sizes. The optimal fixed sample sizes are $t_F^* = 5$ batches (GaGa) and $t_F^* = 4$ (NN). The right panels show the optimal boundaries and $\Delta_t(\mathrm{TP})$ computed from the observed data up to time $t = 1, \ldots, 7$. For both models the sequential design continues up to the time horizon. Figure 3 compares the designs by computing the posterior expected TP (gray bars) and the number of genes with limma P-values< 0.05 after the Benjamini and Yekutieli (2001) adjustment (black bars). At the time horizon both quantities increase over 2 fold compared to the recommended fixed sample size. The differences between prior and posterior expected TP show how sequential designs adapt to the observed data to correct prior miss-specifications.

5.4 Reverse Phase Protein Arrays

We design a follow-up study for the Reverse Phase Protein Array (RPPA) dataset `dataIII` that is included in the R package `RPPanalyzer` (?). The data contains expression for 75 proteins and 35 stage 2 and 25 stage 3 samples. Both models, the NN and GaGa models provide a reasonable fit (SM, Section 2). The fit under the NN model is slightly better. We find $\widehat{\pi} = 0.13$ under the NN model, and $\widehat{\pi} = 0.10$ under the GaGa mdoel. That is, the estimated number of DE proteins is 9.75 and 7.5, respectively. While we expect several DE proteins, at a posterior expected FDR < 0.05 the NN model calls 1 DE protein, and the GaGa model makes no DE calls. For comparison, only 1 protein has limma BY-adjusted P-values below 0.05.

We consider adding batches of 50 samples per group, up to a maximum of $T = 4$. We set the sampling cost to $c = 1$, reflecting that RPPA samples are relatively cheap. The study focuses on 75 carefully chosen proteins. Figure 4 shows simulated $\Delta_t(\text{TP})$ trajectories and the optimal boundaries. Inference under the NN model estimates fewer TPs. Otherwise, inference is fairly similar across models and the optimal boundaries are remarkably close. The average sample size is $t_S^* = 1.37$ under the NN model, and $t_S^* = 1.32$ under the GaGa model. The expected number of true positives at t_S^* is 9.46 (NN) and 6.11 (GaGa). That is, according to both models, most DE proteins should be detected by adding 1-2 batches, *i.e.* 50-100 samples per group. These results help assess the potential benefits of extending the experiment.

6 Discussion

We proposed a sequential strategy for massive multiple hypothesis testing. An important advantage lies in the generality of the proposed design. We discussed three RNA-seq, one microarray and one RPPA experiment. Sequential designs are robust with respect to to inaccurate prior guesses and provide substantial advantages over fixed sample designs.

The proposal is formulated in a decision-theoretic framework and emphasizes interpretability. We monitor the one-step ahead expected increment in utility and stop the experiment when it falls below a boundary. The approach includes fixed sample size and myopic designs as special cases. We use terminal decisions that control the posterior expected FDR. While inconsistent with a strict decision-theoretic setup where all decisions are taken to maximize the expectation of a single utility, we feel that our choice offers a pragmatic compromise.

The method allows stopping when only 1 or 2 samples are available, which requires making strong parametric assumptions. For instance, in Figure 3 the posterior expected TP and $\Delta_t(\text{TP})$ based on 2 samples per group differ widely between the GaGa and NN models. Nevertheless, both models correctly indicate to continue and show good agreement for later samples. Whenever possible, we recommend using a minimum burn-in (*e.g.* ≥ 3 samples) before starting sequential stopping. When not feasible, we recommend assessing goodness-of-fit carefully and updating the forward simulation when more data is available.

We focused on group comparison experiments, but the framework can serve as the basis for other HTS. Interesting extensions include classification, clustering or network discovery studies. These would require adjusting the utility function and possibly the probability model, *e.g.* to capture strong dependencies between outcomes.

Sequential designs are most appealing in moderate to large studies, where technical limitations require gathering data in batches. They should also prove valuable when samples are costly to obtain or there are ethical considerations, *e.g.* in human studies. Overall, they help save valuable resources and guarantee that sufficient data is collected to answer the scientific questions.

7 Software

An implementation of the proposed approach was added to the Bioconductor package gaga.

8 Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank Antonio Zorzano and Deborah Naon for providing the RNA-seq data. *Conflict of Interest*: None declared.

References

Armstrong, S., Staunton, J., Silverman, L., Pieters, R., Boer, M., Minden, M., Sallan, E., Lander, E., Golub, T. and Korsmeyer, S. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia, *Nature Genetics* **30**: 41–47.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* **29**: 1165–1188.

Berry, D. and Fristedt, B. (1985). *Bandit problems*, Chapman & Hall.

Berry, D., Müller, P., Grieve, A., Smith, M., Parke, T., Blazed, R., Mitchard, N. and Krams, M. (2001). *Adaptive Bayesian Designs for Dose-Ranging Drug Trials*, Vol. V, Springer-Verlag, New York.

Brockwell, A. and Kadane, J. (2003). A gridding method for Bayesian sequential decision problems, *Journal of Computational and Graphical Statistics* **12**: 566–584.

Campo Dell'Orto, M., Zangrando, A., Trentin, L., Li, R., Liu, W., te Kronnie, G., Basso, G. and Kohlmann, A. (2007). New data on robustness of gene expression signatures in leukemia: comparison of three distinct total rna preparation procedures, *BMC Genomics* **8:188**.

Carlin, B., Kadane, J. and Gelfand, A. (1998). Approaches for optimal sequential decision analysis in clinical trials, *Biometrics* **54**: 964–975.

DeGroot, M. (1970). *Optimal Statistical Decisions*, McGraw Hill, New York.

DeGroot, M. (2004). *Optimal Statistical Decisions*, Wiley-Interscience.

Dobbin, K. and Simon, R. (2007). Sample size planning for developing classifiers using high-dimensional dna microarray data, *Biostatistics* **8**: 101–117.

Durrieu, G. and Briollais, L. (2009). Sequential design for microarray experiments, *Journal of the American Statistical Association* **104**: 650–660.

Ferreira, J. and Zwinderman, A. (2006). Approximate power and sample size calculations with the benjamini-hochberg method, *The International Journal of Biostatistics* **2(1)**.

Gibbons, R., Bhaumik, D., Cox, D., Grayson, D., Davis, J. and Sharma, R. (2005). Sequential prediction bounds for identifying differentially expressed genes in replicated microarray experiments, *Journal of Statistical Planning and Inference* **129**: 19–37.

Lee, M. and Whitmore, G. (2004). Power and sample size for microarray studies, *Statistics in medicine* **11**: 3543–3570.

Müller, P., Berry, D., Grieve, A., Smith, M. and Krams, M. (2006). Simulation-based sequential bayesian design, *Journal of Statistical Planning and Inference* **137**: 3140–50.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays, *Journal of the American Statistical Association* **99**: 990–1001.

Pan, W., Lin, J. and C.T., L. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments?, *Genome Biology* **3**: research0022.1–0022.10.

Rice, K. M., Lumley, T. and Szpiro, A. A. (2008). Trading bias for precision: Decision theory for intervals and sets, *Technical Report Working Paper 336*, UW Biostatistics, http://www.bepress.com/uwbiostat/paper336.

Rossell, D. (2009). GaGa: a simple and flexible hierarchical model for differential expression analysis, *Annals of Applied Statistics (to Appear)* .

Rossell, D., Müller, P. and Rosner, G. (2006). Screening designs for drug development, *Biostatistics* **8**: 595–608.

Ruppert, D., Nettleton, D. and Hwang, J. (2007). Exploring the information in p-values for the analysis and planning of multiple-test experiments, *Biometrics* **63**: 483–95.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology* **3**.

Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments, *BMC Bioinformatics* **7**: 106.

Yuan, M. and Kendziorski, C. (2006). A unified approach for simultaneous gene clustering and differential expression identification, *Biometrics* **62**: 1089–1098.

Zien, A., Fluck, J., Zimmer, R. and Lengauer, T. (2003). Microarrays: how many do you need?, *Journal of Computational Biology* **10**: 653–667.

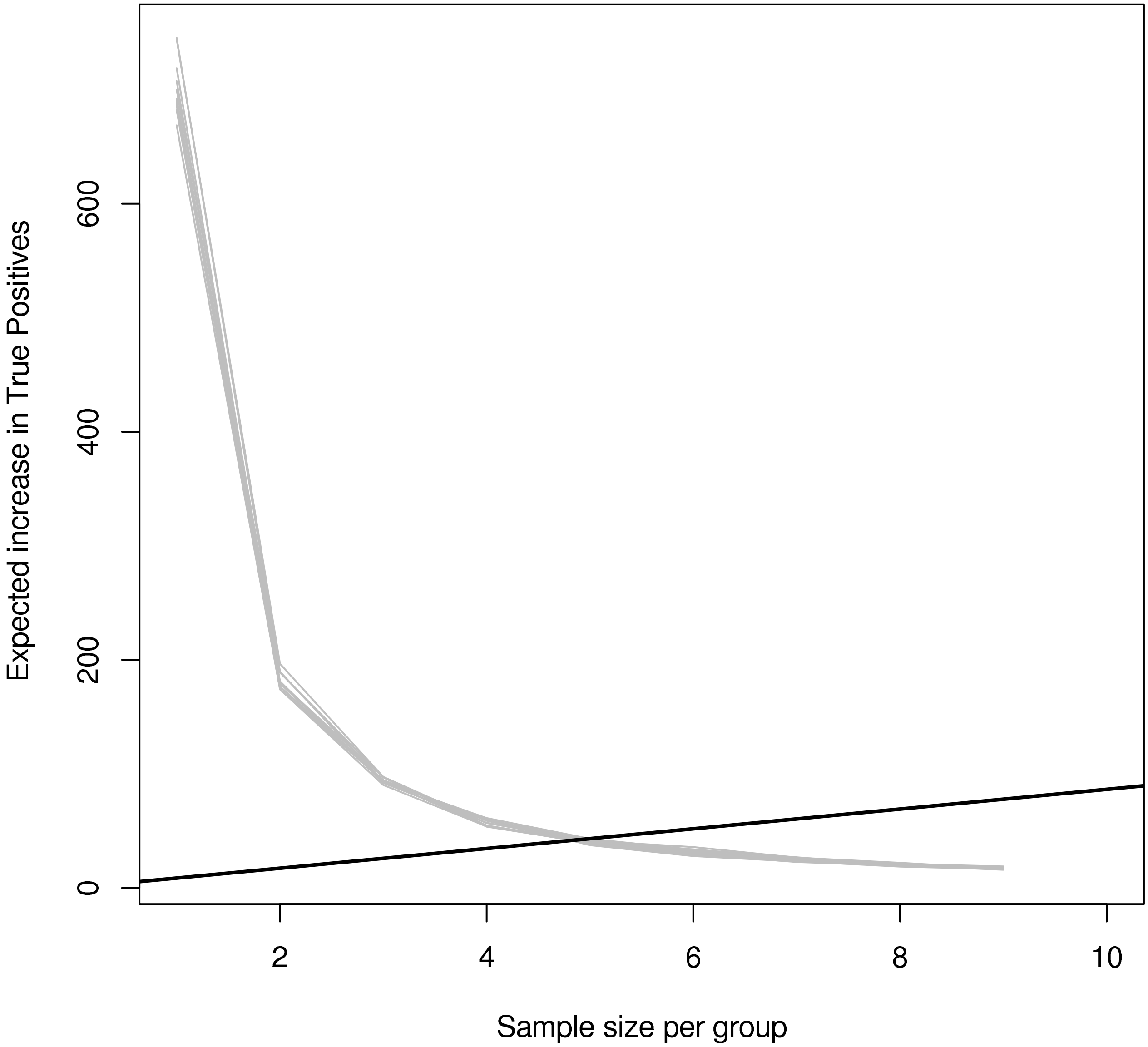

Figure 1: GaGa model based optimal sequential boundary for $c = 50$ (thick black line) and forward simulation trajectories (light grey lines) for example in Section 5.1.

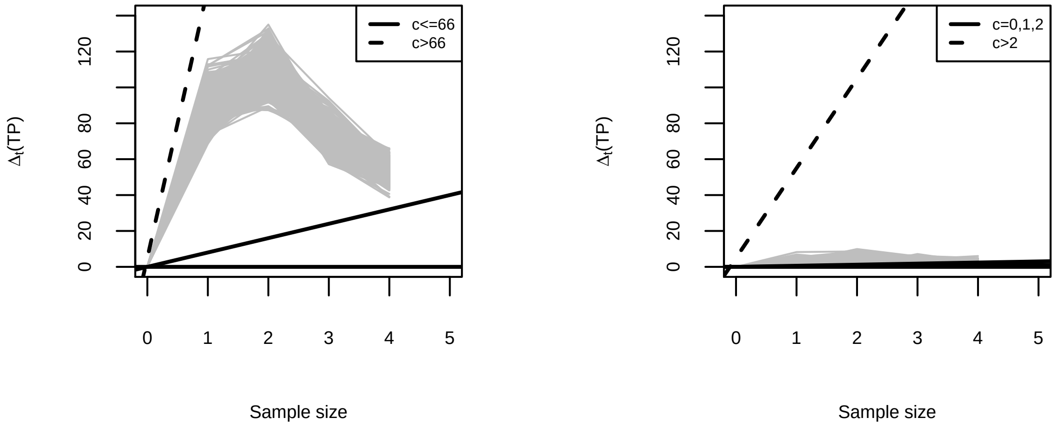

Figure 2: Simulated one-step expected increase in true discoveries Δ_t(TP) (gray lines) and optimal boundaries for several sampling costs c (black lines). Left: brain *vs.* muscle (two multiple optimal boundaries shown for $c \leq 66$) Right: muscle *vs.* muscle.

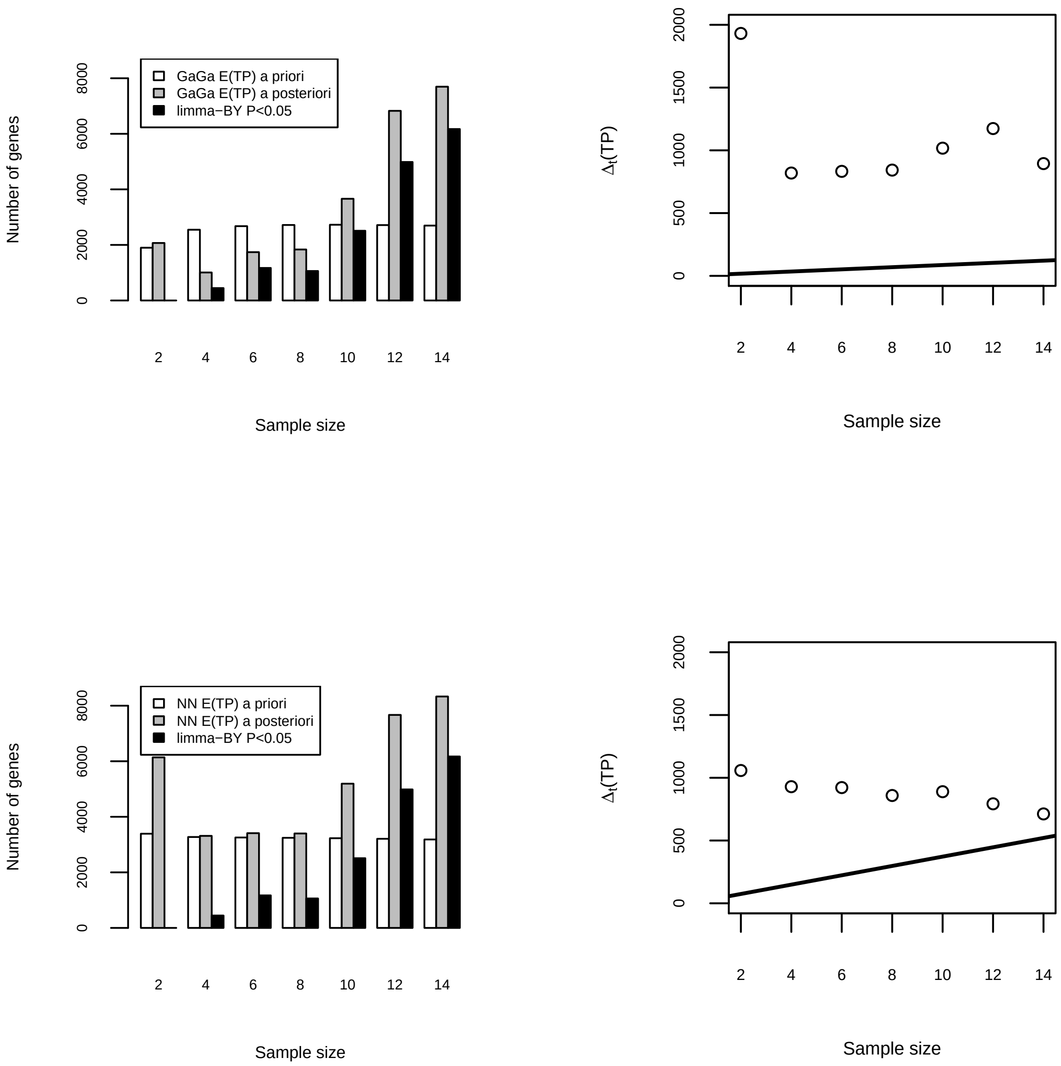

Figure 3: Sequential Analysis of Campo Dell'Orto's data based on GaGa (top panels) and NN models (bottom panels). Left panels: expected number of true positives (*a priori* & *a posteriori*) *vs.* sample size. Black bars indicate the number of genes with Benjanimi-Yekutieli adjusted limma P-values $< .05$. Right panels: $\Delta_t(\text{TP})$ *vs.* sample size and optimal sequential boundary. $\Delta_t(\text{TP})$ being above the boundary for all t indicates experimentation to continue up to the maximum sample size.

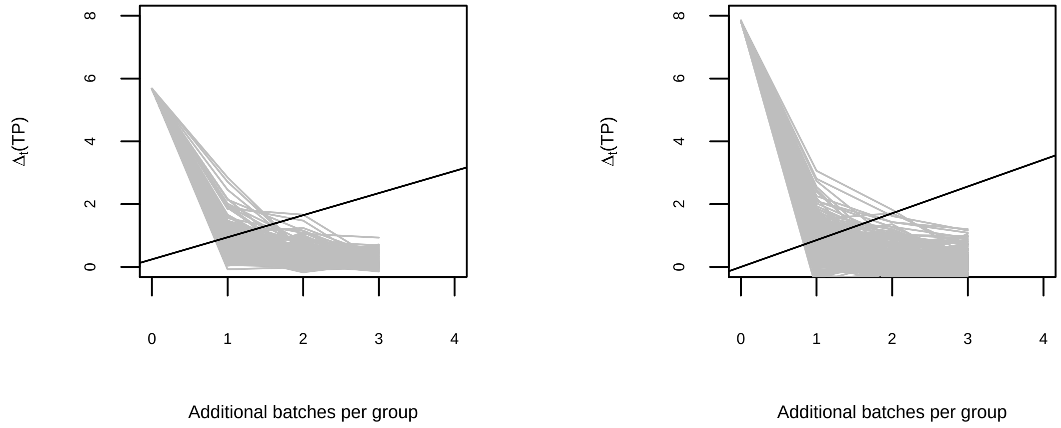

Figure 4: Simulated Δ_t(TP) and optimal boundaries for $c = 1$ in RPPA data using GaGa (left) and NN (right) models