

A Theoretical Analysis of Gene Selection

Sach Mukherjee* and Stephen J. Roberts
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ
U.K.
{sach,sjrob}@robots.ox.ac.uk

Abstract

A great deal of recent research has focused on the challenging task of selecting differentially expressed genes from microarray data ('gene selection'). Numerous gene selection algorithms have been proposed in the literature, but it is often unclear exactly how these algorithms respond to conditions like small sample-sizes or differing variances. Choosing an appropriate algorithm can therefore be difficult in many cases. In this paper we propose a theoretical analysis of gene selection, in which the probability of successfully selecting relevant genes, using a given gene ranking function, is explicitly calculated in terms of population parameters. The theory developed is applicable to any ranking function which has a known sampling distribution, or one which can be approximated analytically. In contrast to empirical methods, the analysis can easily be used to examine the behaviour of gene selection algorithms under a wide variety of conditions, even when the numbers of genes involved runs into the tens of thousands. The utility of our approach is illustrated by comparing three well-known gene ranking functions.

1. Introduction

The advent of microarray technology [14, 19] has meant that transcriptional responses to changes in cellular state can now be quantified for thousands of genes in a single experiment. Microarrays thus offer a window into transcriptional mechanisms underlying major events in health and disease. In recent years, an enormous amount of work has been done in this area of molecular biology, addressing questions relating to both normal cell function [6, 13] and disease [11, 12].

Perhaps the most common type of analysis involves comparing expression levels between tissues in two or more

conditions of interest, such as wild-type and mutant, or healthy and diseased. Genes relevant to the biological phenomenon under investigation are expected to be up- or down-regulated between conditions; one of the most important tasks in microarray data analysis is therefore selecting genes which are differentially expressed in this way. Although differential microarray experiments do not generally lead to a definitive understanding, they play a vital role in narrowing the field for further work. In effect, microarray studies provide geneticists with a short-list of genes worth investing hard-won funds into investigating.

However, the massive dimensionality and relatively small number of datapoints¹ in microarray datasets, coupled with the variability inherent in both experimental process and underlying biology, make their analysis a particularly challenging task. With typically many thousands of genes to choose from and perhaps a few dozen to be selected, looking for differentially expressed genes can be a little like looking for a needle in the proverbial haystack.

A large variety of algorithms, including conventional and non-parametric hypothesis tests, as well as Bayesian and information-theoretic methods [1, 2, 3, 17, 20, 22, 23] have been applied to microarray data. Yet it is often unclear how a particular algorithm will respond to specific statistical properties of the data. For example, in recent months there has been some discussion in the bioinformatics community regarding the suitability of the t-test when variances differ across genes in a systematic manner. This issue, which we will call the 't-test variance issue', is distinct from the Behrens-Fisher problem [8] of variances differing *across classes*. If relevant genes have higher variances than irrelevant ones, should the t-test be abandoned in favour of a different analysis, or does the t-statistic implicitly deal with the issue anyway? Questions such as this have serious im-

* to whom correspondence should be addressed

¹ We use *datapoint/sample* and *dimension/variable* in the following way: each gene is a dimension or variable of the data; each array or chip is an datapoint or sample. By *sample-size* we therefore mean the number of arrays.

plications for users of gene selection algorithms, yet can be difficult to resolve. Empirical studies can help, but as we shall see, the sheer number of genes involved makes it virtually impossible to get a full picture of algorithm performance by simulation alone.

This paper presents a theoretical study of gene selection in which the probability of selecting genes correctly, using a given algorithm, is derived from population parameters², sample-sizes and the number of genes under consideration. We address three key questions:

1. **Selection accuracy:** How can the probability of successfully selecting relevant genes be calculated?
2. **Multiplicity:** How can the effects of comparing thousands of genes at once be accounted for?
3. **Algorithm comparison:** How can such probabilities be used to compare gene selection algorithms?

But does it really help to calculate gene selection accuracy in terms of population parameters which are never known in practice? We will argue that it helps a great deal, in terms of obtaining a clear picture of how well a given method is likely to do under various conditions. Knowing that the performance of a particular algorithm breaks down at small sample-sizes, or when variances differ widely, makes it possible to make an informed choice of method, in accordance with prior knowledge about a specific experimental set-up.

Our initial analysis is a simple two-gene scenario: given data from two genes, only one of which is differentially expressed, how likely is a given algorithm to select the right gene? We then extend the analysis to the multi-gene case. An illustrative set of results is presented which examines the performance of three well-known gene selection methods under various conditions. We discuss our approach, highlighting connections to, and differences from, existing research, and finally look at some possibilities for future work.

We begin by briefly describing gene selection. Consider microarray slides (or chips) belonging to two classes, say, healthy and diseased. Data D for a particular gene consist of m expression levels in one class, and n in the other:

$$D = [X_1 X_2 \dots X_m \ Y_1 Y_2 \dots Y_n] \quad (1)$$

X_i are independent and identically distributed random variables with true (but unknown) mean μ_X ; Y_j are also independent and identically distributed with true mean μ_Y .

² By 'population parameters', we mean the true values of the parameters required to specify the statistical model for the data.

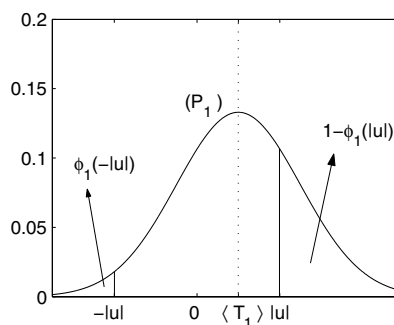


Figure 1. The relationship between the probability $P(|T_1| > |u|)$ and the cumulative distribution function ϕ_1 . The null case is similar.

If the true class means are distinct, the gene is said to be *differentially expressed*. In the language of hypothesis testing, the null and alternative hypotheses (H_0 and H_1 respectively) are:

$$\begin{aligned} H_0 &\rightarrow \mu_X = \mu_Y \\ H_1 &\rightarrow \mu_X \neq \mu_Y \end{aligned}$$

Gene selection algorithms *rank* genes according to some function of the data (the 'ranking function'), and then select a *set* of genes by applying a threshold to the values obtained, or by choosing a user-specified number of genes.

2. Theory

In this Section we derive probabilities of success in gene selection. Although the material presented below is necessarily somewhat technical, the question being addressed could not be more straightforward: we simply wish to calculate the probability of success using a given gene ranking function, in terms of population parameters, sample-size and numbers of genes involved. In other words we wish to understand exactly how well a given ranking function is expected to do under specified conditions. The reader less interested in the statistical analysis itself may wish to move directly to the results in Section 3.

2.1. Binary gene selection

Let us start by comparing just two genes, one of which is differentially expressed (the 'alternative gene') and the other not (the 'null gene'). The question we wish to ask is this: given a model, population parameters and sample-size, what is the probability of correctly selecting the differentially expressed gene, using a given ranking function?

We assume that data is drawn according to some model, with parameters θ_1 for the alternative gene, and θ_0 for the

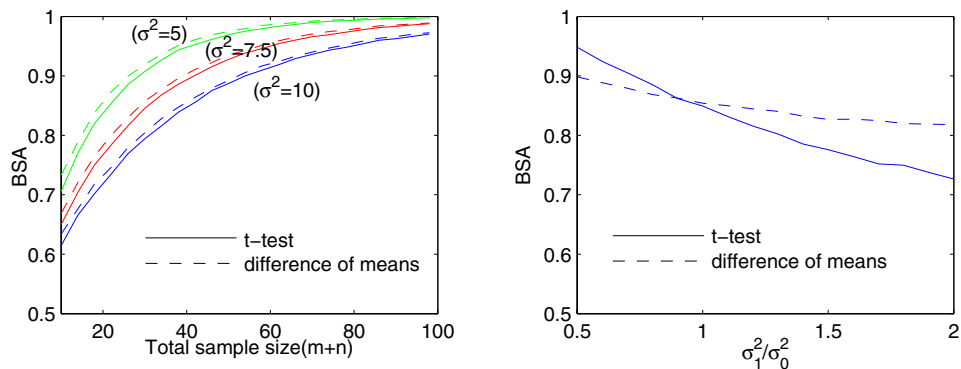


Figure 2. Binary Selection Accuracy. The left panel shows the true probability of obtaining a correct decision when comparing two genes ('Binary Selection Accuracy' or BSA) plotted against total sample-size, for two different ranking functions. The right panel shows BSA plotted against the ratio of the variance of the differentially expressed gene to the variance of the 'null' gene.

null gene (we will adopt the convention that subscript '1' refers to the alternative gene and '0' to the null gene). For example, the model might be Normal, in which case each parameter vector would contain class means and variances. Data for each gene is then a collection of random variables drawn under either the null or alternative models, denoted by D_0 or D_1 respectively. The ranking function r maps data for the null and alternative genes to ranking scores T_0 and T_1 respectively:

$$T_0 = r(D_0); \quad T_1 = r(D_1)$$

A selection decision is made in this two-gene case by simply comparing the two scores. Assuming that higher absolute ranking scores correspond to a greater chance of differential expression, the probability of choosing the correct gene is simply:

$$P(|T_0| < |T_1|) \quad (2)$$

We call the probability given by Equation 2 'Binary Selection Accuracy' or BSA. Considering absolute values has the advantage of making the analysis directly applicable to common ranking functions like the t-statistic, Fisher ratio and difference of means; however, the relevant Equations can be easily modified for the signed case.

Note that ranking scores T_0 and T_1 are functions of random data and are themselves random variables; let their sampling distributions³ be p_0 and p_1 respectively, and the

³ The *sampling distribution* of a function of random variables (i.e. a statistic) is simply the distribution of the statistic. In this context, the sampling distribution of the ranking function r can be thought of as the distribution of values that would emerge from repeatedly computing ranking scores from data sampled under the model, with population parameters and sample-sizes remaining fixed.

corresponding cumulative distribution functions be ϕ_0 and ϕ_1 . Then the probability of success given in Equation 2 can be expressed as:

$$\begin{aligned} P(|T_0| < |T_1|) &= \int_{-\infty}^{\infty} P(|T_1| > |u|) p_0(u) du \\ &= \int_{-\infty}^{\infty} (1 - \phi_1(|u|) + \phi_1(-|u|)) p_0(u) du \\ &= \langle 1 - \phi_1(|u|) + \phi_1(-|u|) \rangle_{p_0} \end{aligned} \quad (3)$$

Where $\langle \cdot \rangle_{p_0}$ denotes expectation under the density p_0 . Expressing $P(|T_1| > |u|)$ in terms of the cumulative distribution function ϕ_1 is quite straightforward, and is illustrated in Figure 1.

Although the expectation in Equation 3 will not in general be analytic, it can be calculated easily, by sampling from p_0 . The densities p_0 and p_1 thus play the role of connecting the probability of a correct decision to population parameters and sample-sizes. This becomes clear if we explicitly rewrite the probability in Equation 2 as a function of population parameters and sample-size:

$$\begin{aligned} P(|T_0| < |T_1|) &= f(\theta_0, \theta_1, m, n) \\ &= 1 - \int_{-\infty}^{\infty} \left[\int_{-\infty}^{|u|} p(T_1 = v | \theta_1, m, n) dv \right. \\ &\quad \left. - \int_{-\infty}^{-|u|} p(T_1 = w | \theta_1, m, n) dw \right] \\ &\quad p(T_0 = u | \theta_0, m, n) du \end{aligned} \quad (4)$$

Binary Selection Accuracy is thus derived directly from the sampling distributions p_0 and p_1 , and can be regarded

	Selected	Not selected	Total
Alternative genes	s_1	$g_1 - s_1$	g_1
Null genes	s_0	$g_0 - s_0$	g_0
(Total)	s	$g - s$	g

Table 1. Summary table for multiple gene selection, following [4].

as a true measure of the general ability of the ranking function to distinguish pairs of relevant and irrelevant genes, on the basis of data drawn under the model. Interestingly, it turns out that BSA is numerically equivalent to the area under a true ROC curve. ROC curves are plots of true positive against false positive rates across a range of thresholds, and are widely used in the analysis of classifiers. By ‘true ROC curve’ we mean a curve derived directly from the sampling distributions of the ranking function. Please see Appendix A for details and a short proof.

Figure 2 shows illustrative results, using Binary Selection Accuracy to compare the t-statistic and difference of means methods, under conditions detailed in Section 3.1. Of particular interest is the right panel, which shows that, at least as far as binary selection is concerned, higher variances for the differentially expressed gene really do handicap the t-test. Full results, using the multiple gene analysis developed below, are presented in Section 3.2.

2.2. Multiple gene selection

Consider a more general scenario with a total of g genes, of which g_1 are truly differentially expressed (alternative or relevant genes) and g_0 are not (null or irrelevant genes). Table 1 summarizes the relationships between the various groups of genes involved. The question we wish to ask is this: given population parameters, sample-sizes and numbers of relevant and irrelevant genes, what is the probability that one of the relevant genes is ranked in the top s places under the ranking function? Also, what is the distribution over the number of relevant genes returned among the s genes selected?

Note that we do not impose an *a priori* threshold on ranking scores, but rather treat the *number of genes to be selected* as a constant. This number tends to depend on factors like the scale of the experiment, follow-up plans and so on [15] and can reasonably be regarded as part of the experimental set-up, in much the same way as the total number of genes under study, or sample-size. In contrast, score thresholds, while statistically convenient, have no clear interpretation in the context of the experiment. A geneticist will generally have an *a priori* notion of the number of genes she wants to select, but will not care about the actual scores, be-

yond their rank order. It therefore makes sense to centre the analysis of multiple gene selection around number of genes selected, rather than threshold.

Formulating the problem: For simplicity, we assume that data for the alternative genes are identically distributed, so that their ranking scores are also identically distributed according to the density p_1 . Similarly, the ranking scores of the null genes are taken to be distributed according to p_0 .

One of the g_1 relevant genes will appear in the top s places provided its ranking score equals or exceeds the s^{th} highest score under the ranking. However, computing the probability of this event involves the distribution of the s^{th} highest score, which is a somewhat complicated order statistic. Happily, the problem can be reformulated in terms of a threshold, and thence solved relatively easily.

If we think of imposing a threshold τ on the ranking scores, such that genes having scores exceeding τ are selected as differentially expressed, there will in general be a probabilistic relationship between the threshold and number of genes selected. Now, at a given threshold τ , the probability of a relevant gene being selected by the algorithm is simply the true positive rate at τ , denoted by $\beta(\tau)$. Then, the probability of a relevant gene being selected among the top s genes is simply the expectation of the true positive rate $\beta(\tau)$, under the conditional density $p(\tau|s)$. We call this probability ‘Multiple Selection Accuracy’ or MSA:

$$\begin{aligned} MSA &= \int_0^\infty \beta(\tau)p(\tau|s) d\tau \\ &= \langle \beta(\tau) \rangle_{p(\tau|s)} \end{aligned} \quad (5)$$

Readers familiar with multiple hypothesis testing may find it useful to think of MSA as a kind of exact True Discovery Rate, but with a specified number of genes being selected (and a consequently variable threshold).

The conditional $p(\tau|s)$: But what *is* the conditional distribution of threshold τ given s , and how can the expectation in Equation 5 be computed in terms of the sampling distributions and numbers of genes? We suggest the following unnormalized density function as an approximation to $p(\tau|s)$ (the full derivation is presented in Appendix B):

$$p(\tau|s) \propto \frac{1}{\sqrt{2\pi} v(\tau)} e^{-[s-m(\tau)]^2/[2v(\tau)]} \quad (6)$$

Where, $m(\tau)$ and $v(\tau)$ are functions of threshold τ and are given by:

$$m(\tau) = g_1\beta(\tau) + g_0\alpha(\tau) \quad (7)$$

$$\begin{aligned} v(\tau) &= g_1\beta(\tau)(1 - \beta(\tau)) \\ &\quad + g_0\alpha(\tau)(1 - \alpha(\tau)) \end{aligned} \quad (8)$$

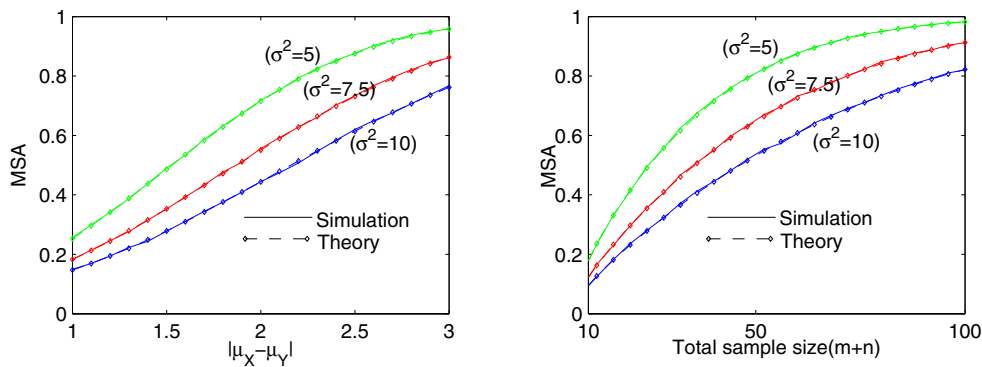


Figure 3. Theoretical and simulated Multiple Selection Accuracies. The left panel shows empirically and theoretically calculated values for the probability of correctly selecting a relevant gene ('Multiple Selection Accuracy' or MSA) using a t-statistic, plotted against the extent of differential expression (i.e. $|\mu_X - \mu_Y|$), at three levels of variance. The panel on the right shows MSA plotted against sample-size. The total number of genes is 1025, with 25 being truly differentially expressed. The number of genes selected is 50.

The true and false positive rates $\beta(\tau)$ and $\alpha(\tau)$ can be written in terms of ϕ_0 and ϕ_1 as follows:

$$\beta(\tau) = 1 - \phi_1(\tau) + \phi_1(-\tau) \quad (9)$$

$$\alpha(\tau) = 1 - \phi_0(\tau) + \phi_0(-\tau) \quad (10)$$

If the ranking function r is a valid test statistic, $\beta(\tau)$ is the power, and $\alpha(\tau)$ the P-value, at threshold τ .

Taken together, Equations 5 through 10 give Multiple Selection Accuracy in terms of p_0 , p_1 , ϕ_0 , ϕ_1 , g_0 , g_1 , and s . Now, for a given ranking function, the sampling distributions p_0 and p_1 and their corresponding cdfs depend only on population parameters θ_0 and θ_1 and sample-sizes m and n . We have thus expressed MSA in the desired form. We compute MSA by sampling from the density in Equation 6 and then computing the expectation $\langle \beta(\tau) \rangle_{p(\tau|s)}$.

The distribution over the number of relevant genes discovered is Binomial, with the number of Bernoulli trials being the smaller of s and g_1 , and the probability parameter being $MSA \times \max(g_1/s, 1)$:

$$s_1 \sim B\left(MSA \times \max\left(\frac{g_1}{s}, 1\right), \min(s, g_1)\right) \quad (11)$$

To see why this is the case, note that the greatest number of relevant genes we can possibly select is either g_1 (if s exceeds g_1) or s (otherwise). The probability parameter then follows from the definition of MSA.

Figure 3 shows comparisons of theoretically calculated MSA and brute-force simulation: the approximation used here is clearly very accurate indeed.

3. Experiments

In this Section we present the results of a case study, comparing three widely used gene ranking functions under various conditions.

3.1. Preliminaries

Model and population parameters: We assume a Normal model for (log-transformed) expression data, but emphasize that the framework developed in the previous Section is quite general, and allows selection accuracy to be calculated under any model.

The multiple selection analysis already assumed that all the relevant genes had identically distributed expression levels, as also all the irrelevant genes. Thus, the population parameters are as follows:

- *Means:* Class means for the irrelevant or null genes are by definition identical, and for the analysis of the algorithms we have chosen, need not be considered further. We therefore only explicitly refer to class means for the relevant genes, as μ_X and μ_Y respectively. Unless otherwise mentioned the extent of differential expression (i.e. $|\mu_X - \mu_Y|$) is set to 2.
- *Variances:* We assume that class variances are identical for either relevant or irrelevant genes, but may differ between them. The population variance for relevant genes is denoted by σ_1^2 , and for irrelevant genes by σ_0^2 . Unless otherwise noted, both values are set to 10.
- *Sample-size:* As before, the number of samples in each class is m (for the 'X' data) and n (for the 'Y' data).

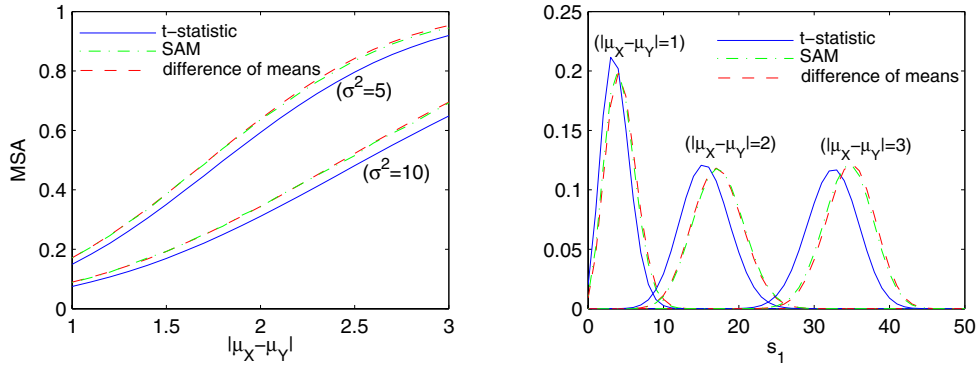


Figure 4. Multiple Selection Accuracies against extent of differential expression. The left panel shows Multiple Selection Accuracies (MSA) for three ranking functions, plotted against the extent of differential expression (i.e. $|\mu_X - \mu_Y|$), at two levels of variance. The right panel shows full distributions over the number of relevant genes selected, at three different levels of differential expression. It is assumed that variances for relevant and irrelevant genes are identical (both denoted by σ^2). The total number of genes is 6000, with 50 being truly differentially expressed; the number of genes selected is 100.

Unless otherwise mentioned, we assume $m = n$ and that total sample-size is 40.

- *Number of genes:* We assume a total of 6000 genes, of which 50 are truly differentially expressed. The number of genes selected is 100.

The parameter set-up can be chosen in accordance with the specific questions being addressed. For example, in this case, the variances are set-up to look into the ‘t-test variance issue’ mentioned previously.

Ranking functions: The three ranking functions analyzed here are briefly reviewed below:

i) t-statistic: The t-test is a canonical two-sample hypothesis test and has been widely used in the differential analysis of microarray data. The ranking statistic for a particular gene is given by:

$$r(D) = \frac{\bar{X} - \bar{Y}}{\left(\frac{1}{m} + \frac{1}{n}\right)^{\frac{1}{2}} \widehat{SD}} \quad (12)$$

Where, \bar{X} and \bar{Y} represent sample means of data in each of the two classes, and \widehat{SD} the pooled sample standard deviation.

The null sampling distribution p_0 is a t-distribution, with $(m+n-2)$ degrees of freedom. The alternative distribution p_1 is a *non-central* t-distribution, with $(m+n-2)$ degrees of freedom and non-centrality parameter ψ :

$$\psi = \frac{\mu_X - \mu_Y}{\sigma_1 \left(\frac{1}{m} + \frac{1}{n}\right)^{\frac{1}{2}}} \quad (13)$$

ii) Difference of means: The ranking function is simply the difference between sample means:

$$r(D) = \bar{X} - \bar{Y} \quad (14)$$

This function can be regarded as a log-space fold analysis [2]. Both p_0 and p_1 are Normal densities:

$$p_0 = N(0, \sigma_0^2/m + \sigma_0^2/n) \quad (15)$$

$$p_1 = N(\mu_X - \mu_Y, \sigma_1^2/m + \sigma_1^2/n) \quad (16)$$

iii) SAM statistic: The SAM statistic [23] is given by:

$$r(D) = \frac{\bar{X} - \bar{Y}}{\left(\frac{1}{m} + \frac{1}{n}\right)^{\frac{1}{2}} (\widehat{SD} + K)} \quad (17)$$

Where, \widehat{SD} is the pooled standard deviation and K a regularizing term. The value of K is determined from the entire dataset and is thus constant across genes, but data-dependent nonetheless. Thus, from the point of view of deriving a sampling distribution, K is a (somewhat complicated) random variable. The sampling distribution of the SAM statistic can be approximated as a ratio of Normals [16]; we make use of that approximation here, and refer the interested reader to the reference.

3.2. Results

Results for the three algorithms under consideration are presented below, with Multiple Selection Accuracies plotted against parameters of interest. Also shown are full distributions over the number of truly differentially expressed

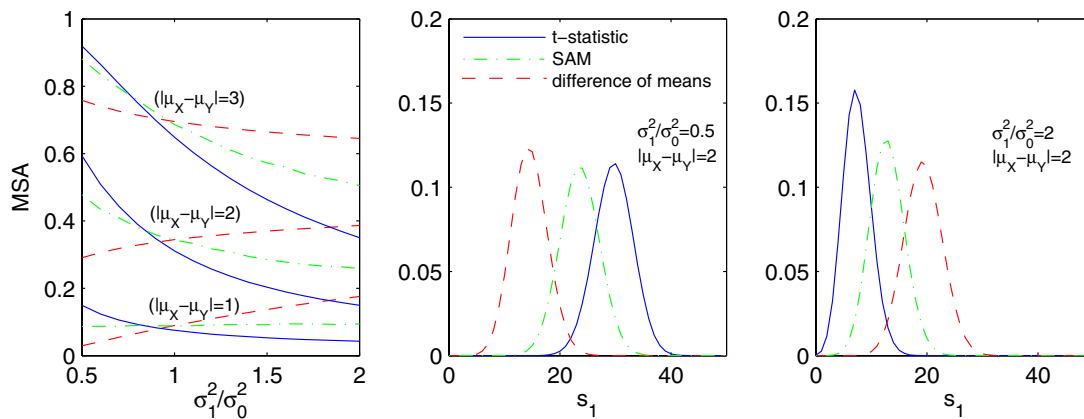


Figure 5. Unequal variances. The leftmost panel shows Multiple Selection Accuracies, plotted against the *ratio* of variances of relevant and irrelevant genes (i.e. σ_1^2/σ_0^2), at three levels of differential expression. The centre panel shows full distributions over the number of relevant genes selected, when the relevant genes' variance is lower than that of the irrelevant ones ($\sigma_0^2 = 2\sigma_1^2$); the rightmost panel shows corresponding distributions when the relevant genes' variance is higher ($\sigma_1^2 = 2\sigma_0^2$).

genes discovered. Note that these results do not themselves require significance testing, as the probabilities calculated are exact.

I. How does the extent of differential expression affect selection accuracy? The left panel of Figure 4 shows Multiple Selection Accuracy (MSA) for all three algorithms, as a function of the extent of differential expression. The right panel shows distributions over the number of relevant genes selected. Perhaps surprisingly, the simple difference of means method does slightly better than the t-test, with the effect accentuated at higher levels of differential expression. SAM performs almost exactly as well as difference of means.

II. How do the algorithms respond to unequal variances between relevant and irrelevant genes? Figure 5 shows selection accuracy as a function of the *ratio* of variances between relevant and irrelevant genes (i.e. σ_1^2/σ_0^2), as well as distributions over the number of truly differentially expressed genes discovered. When the variance of relevant genes is lower than that of the others, the t-statistic does better than difference of means; but when the situation is reversed, difference of means clearly outperforms the t-statistic. SAM sits between these two extremes. The t-statistic clearly does *not* automatically take account of this heterogeneity in variances across genes, even though the familiar assumptions of normality and identical variances in both classes hold for each gene. These results clearly resolve the 't-test variance issue', and verify the binary selection results presented earlier. The curves also illustrate a major strength of the SAM statistic: it does relatively well

across the whole range of relative variances. This makes SAM an excellent choice when little is known about the distribution of variances between relevant and irrelevant genes.

III. How successfully can the algorithms cope with small sample-sizes? The left panel of Figure 6 shows selection accuracy as a function of total sample-size. The right panel shows distributions over the number of relevant genes selected. Difference of means and SAM do noticeably better than the t-test, especially at small sample-sizes. Previous work [7] has drawn attention to the low power of the t-test at small sample sizes - our results clearly quantify the impact of this effect in the multiple gene context.

4. Related work

The central problem in applying classical hypothesis tests to microarray data is the issue of multiple comparisons. Hypothesis tests were developed, for the most part, with the aim of making inferences about a single variable, rather than comparing thousands of variables against each other. From the outset, when we considered binary selection, our analysis has therefore explicitly dealt with the fact that gene selection involves the *comparison* of ranking scores.

Statistical power analysis can be thought of as relating the probability of false negatives (or Type II error) in a hypothesis test, to population parameters. The analysis presented here deals explicitly with the issue of comparisons, leading to a definition of Binary Selection Accuracy, for example, that is effectively an integral over the power func-

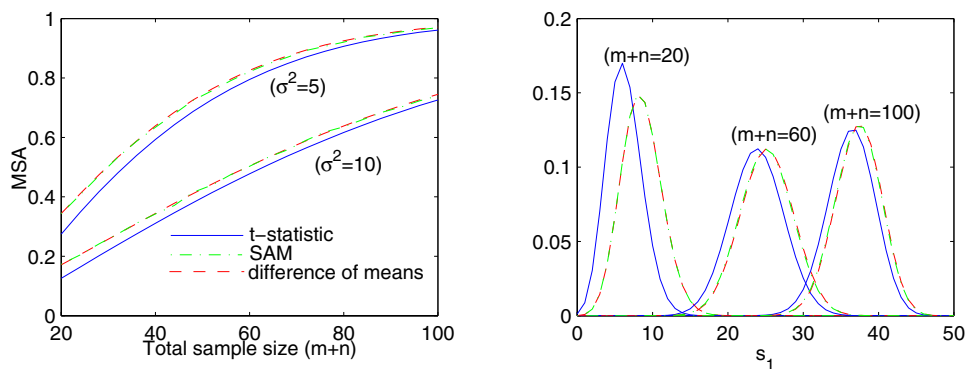


Figure 6. Sample-size. The left panel shows Multiple Selection Accuracies plotted against sample-size, at two levels of variance. The right panel shows full distributions over the number of relevant genes selected, for three sample-sizes. The number of datapoints in each class is taken to be equal, i.e. $m = n$.

tion⁴. This already makes our approach quite distinct from power analysis, even at the level of binary selection. Our treatment of errors and approach to algorithm comparison is quite different too. Classical testing theory first fixes an acceptable level of Type I error (via the P-value) and then seeks to maximize power. In contrast, starting out with a definition of selection accuracy which was motivated by the way in which gene selection algorithms are actually used, we found that the balance between the two types of error emerged naturally and did not need to be explicitly specified.

An aspect of the multiple comparisons issue which has been widely addressed in the literature is the question of P-value adjustments. Methods for adjusting P-values [4, 10, 18, 21] include the Bonferroni correction and the less conservative False Discovery Rate (FDR). Dudoit *et al.* present an excellent review [9] of current work in this important area. As noted earlier, Multiple Selection Accuracy can be thought of as similar to a True Discovery Rate⁵. However, MSA is calculated exactly, and in terms of the number of genes selected rather than threshold. Moreover, FDR methods and MSA address related but distinct problems. Our aim has been to understand analytically how algorithms behave - and differ in their behaviour - in various regions of parameter space. Consequently, we have derived selection accuracy in terms of population parameters. FDR methods, in contrast, aim to calibrate P-values for a given dataset

and algorithm. The ‘true’ FDR value is generally approximated, without using population parameters, and subsequently used to determine an appropriate selection threshold.

Recent work [5] has compared gene selection algorithms by using ROC curves computed by simulation. However, no attempt is made to derive selection accuracy as a function of population parameters. Simulation studies are necessarily limited, as only a small subset of conditions can possibly be explored, particularly if multiplicity is to be taken into account. Consider the multiple gene experiments presented above: a brute-force simulation would have required 6000 sets of samples to be drawn and ranked for *each* point on *each* curve.

5. Conclusions

To conclude, let us return to the three questions posed at the beginning of this paper, and examine the extent to which they have been addressed:

1. **Selection accuracy:** We have shown how sampling distributions can be used to link gene selection accuracy to population parameters. ‘Binary Selection Accuracy’ was defined and derived in an intuitive but principled manner, taking full account of the comparative nature of gene selection.
2. **Multiplicity:** The analysis was extended to deal with an arbitrary number of relevant and irrelevant genes. For simplicity it was assumed that all relevant genes are identically distributed, as also irrelevant ones. Although this is certainly a strong assumption, it nonetheless allowed us to usefully ex-

⁴ Appendix A makes this clear, by looking the relationship between BSA, and the area under a true ROC curve.

⁵ True Discovery Rate is usually defined as the expected proportion of relevant genes among those selected at a given threshold. If $TDR(s)$ is taken to represent the True Discovery Rate for s genes selected, it can be shown that $MSA \approx TDR(s) \times \frac{s}{g_1}$.

amine the effects of multiplicity on gene selection.

3. Algorithm comparison: We presented a case-study which compared three widely-used selection methods across a range of parameter settings. Three particularly interesting results emerged:

- i) The t-statistic can be an inappropriate choice even when data is normally distributed and class variances identical,
- ii) A simple fold method can outdo the t-statistic in a number of plausible situations, and
- iii) The SAM statistic performs well across a range of conditions.

Although the results presented illustrate the utility of the proposed analysis, it is worth emphasizing that practically *any* aspect of algorithm performance can be queried within the framework put forward here.

Three major directions for future work will be:

i) *Extending the comparative analysis to other algorithms.* This will require relevant sampling distributions to be derived, or approximated analytically. While this may not always be easy, it could be argued that without the information provided by sampling distributions, it is difficult to get a true picture of the behaviour of an algorithm under various conditions, beyond necessarily limited empirical studies. Microarray experiments have come to play an important role in guiding further exploration; as a consequence, errors in gene selection can be costly in terms of wasted resources and missed opportunities. A full understanding of the properties of selection algorithms is therefore of more than merely theoretical interest!

ii) *Examining other selection problems.* This paper focused on the topical problem of gene selection, but it is worth emphasizing the generic nature of the questions that were addressed. An increasing number of problems in scientific data analysis involve selecting variables from high-dimensional data; in principle, any task of this type can be analyzed within the framework developed here. The selection of biologically relevant genes from time-varying microarray data is one example, which despite being a quite different problem to static gene selection, raises many of the same questions about selection accuracy. It would certainly be interesting to extend our approach to algorithms for time-course analysis.

iii) *Generalizing the analysis by placing priors over population parameters, sample-sizes and numbers of genes.* For compactness, let π represent *all* these quantities. Treating everything but the model and form of the ranking function as ‘nuisance parameters’ to be integrated out, would leave

us with a single probability of selection accuracy:

$$\int f_{MSA}(\pi)p(\pi) d\pi \quad (18)$$

Where $f_{MSA}(\pi)$ represents Multiple Selection Accuracy as a function of the parameter vector π .

The prior $p(\pi)$ would be chosen to capture knowledge about the experimental set-up. This approach could also be used to relax some of the simplifying assumptions made here, such as the assumption that all relevant genes are identically distributed. The probability of selection accuracy given by Equation 18 would in effect represent an *objective function* for gene selection algorithms. An optimal algorithm, for the given model and priors, would simply maximize this quantity. This approach could therefore lead to a method for learning an optimal ranking function from a family of candidate functions.

In conclusion, we have presented a theoretical analysis of gene selection, which can be used to thoroughly examine the behaviour of selection algorithms, even when the number of genes runs into the tens of thousands. The analysis can highlight strengths and weaknesses of algorithms, and thus help bioinformaticians choose methods appropriate to particular experimental conditions.

Acknowledgements

SM gratefully acknowledges the support of the U.K. Biotechnology and Biological Sciences Research Council for financial support; thanks also to Sarah Gurr, Nick Hughes and Peter Sykacek for helpful discussions.

References

- [1] P. Baldi and G. W. Hatfield. *DNA Microarrays and Gene Expression*. Cambridge University Press, 2002.
- [2] P. Baldi and A. D. Long. A Bayesian Framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–19, 2001.
- [3] A. Ben-Dor, N. Friedman, and Z. Yakhini. Scoring genes for relevance. Technical Report 2000-38, School of Computer Science and Engineering, Hebrew University, 2000.
- [4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, 57:289–300, 1995.
- [5] P. Broberg. Statistical methods for ranking differentially expressed genes. *Genome Biology*, 4(6), 2003.
- [6] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian,

- D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2:65–73, 1998.
- [7] X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(210), 2003.
- [8] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison Wesley, 3rd edition, 2002.
- [9] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.
- [10] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. Technical Report 633, Department of Statistics, University of California, Berkeley, 2003.
- [11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–37, 1999.
- [12] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bitner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene-Expression Profiles in Hereditary Breast Cancer. *N. Engl. J. Med.*, 344(8):539–48, 2001.
- [13] F. Holstege, E. Jennings, J. Wyrick, T. Lee, C. Hengartner, M. Green, T. Golub, E. Lander, and R. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–28, 1998.
- [14] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–80, 1996.
- [15] I. Lonnstedt and T. P. Speed. Replicated microarray data. *Stat Sinica*, 12:31–46, 2002.
- [16] S. Mukherjee. Sampling Distribution of the SAM-statistic. Technical Report PARG-04-01, Department of Engineering Science, University of Oxford, 2004. Available at <http://www.robots.ox.ac.uk/~parg/publications.html>.
- [17] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002.
- [18] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures. *Bioinformatics*, 19:368–75, 2003.
- [19] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70, 1995.
- [20] P. Sebastiani, E. Gussoni, I. S. Kohane, and M. Ramoni. Statistical Challenges in Functional Genomics. *Statistical Science*, 18(1):33–70, 2003.
- [21] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, 100(16):9440–45, 2003.
- [22] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–61, 2002.
- [23] V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98(9):5116–21, 2001.

A. Area under the ROC curve

The true ROC curve is simply a parametric curve $(\alpha(\tau), \beta(\tau))$, parameterized by a ranking threshold τ , where $\alpha(\tau)$ and $\beta(\tau)$ are respectively the false positive and true positive rates given threshold τ . Note that in gene selection true and false positives refer to the genes being selected, not samples being classified. The threshold is taken to be always non-negative. Then, given sampling distributions p_0 and p_1 , each of β and α is a function of threshold τ only, as shown in the main text (Equations 9 and 10 respectively). The area under the ROC curve (*AUC*) is given by:

$$AUC = \int_0^1 \beta d\alpha$$

The bounds on the integral can be thought of as going from ‘strict’ to ‘lenient’ - thus, a false positive rate of zero (the lower bound in terms of α) corresponds to a threshold of $+\infty$, while a false positive rate of one, corresponds to a threshold of zero. Bearing in mind that $d\alpha = \alpha'(\tau)d\tau$ we can change variable to τ as follows:

$$AUC = \int_{\infty}^0 \beta(\tau)\alpha'(\tau)d\tau$$

Now, $\alpha'(\tau)$ and $\beta(\tau)$ can be expressed in terms of p_0 and ϕ_1 :

$$\begin{aligned}\alpha'(\tau) &= -p_0(\tau) - p_0(-\tau) \\ \beta(\tau) &= 1 - \phi_1(\tau) + \phi_1(-\tau)\end{aligned}$$

Substituting these expressions into the integral:

$$\begin{aligned}AUC &= \int_{\infty}^0 (1 - \phi_1(\tau) + \phi_1(-\tau))(-p_0(\tau) - p_0(-\tau))d\tau \\ &= -\int_{\infty}^0 (1 - \phi_1(\tau) + \phi_1(-\tau))p_0(\tau)d\tau \\ &\quad - \int_{\infty}^0 (1 - \phi_1(\tau) + \phi_1(-\tau))p_0(-\tau)d\tau \\ &= \int_0^{\infty} (1 - \phi_1(\tau) + \phi_1(-\tau))p_0(\tau)d\tau \\ &\quad + \int_0^{\infty} (1 - \phi_1(\tau) + \phi_1(-\tau))p_0(-\tau)d\tau\end{aligned}$$

If p_0 is symmetric about 0 (as is the case for most common ranking functions):

$$AUC = 2 \int_0^{\infty} (1 - \phi_1(\tau) + \phi_1(-\tau))p_0(\tau)d\tau$$

Now, Binary Selection Accuracy was expressed in Equation 3 in the following form:

$$BSA = \int_{-\infty}^{\infty} (1 - \phi_1(|u|) + \phi_1(-|u|))p_0(u)du$$

Since $(1 - \phi_1(|u|) + \phi_1(-|u|))$ is an even function of u :

$$\begin{aligned}BSA &= 2 \int_0^{\infty} (1 - \phi_1(|u|) + \phi_1(-|u|))p_0(u)du \\ &= 2 \int_0^{\infty} (1 - \phi_1(u) + \phi_1(-u))p_0(u)du\end{aligned}$$

This latter expression is clearly identical to *AUC*.

B. Approximation to the conditional $p(\tau|s)$

As noted in the main text, the general relationship between threshold τ and number of genes s is probabilistic. Applying Bayes theorem:

$$p(\tau|s) = \frac{P(s|\tau)p(\tau)}{P(s)}$$

The likelihood term $P(s|\tau)$ represents the distribution over the number of genes selected at a given threshold τ . The number of genes selected is simply the sum of the number of relevant and irrelevant genes selected (s_1 and s_0 respectively) at threshold τ . Given τ , s_1 and s_0 are Binomial:

$$\begin{aligned}s|\tau &= s_1|\tau + s_0|\tau \\ s_1|\tau &\sim B(\beta(\tau), g_1) \\ s_0|\tau &\sim B(\alpha(\tau), g_0)\end{aligned}$$

Where, as before, $\beta(\tau)$ and $\alpha(\tau)$ are respectively the true and false positive rates at threshold τ . Twice making use of the Normal approximation to the Binomial and adding the two resulting Normals we can therefore approximate the distribution over the number s of genes selected, given threshold τ , as follows:

$$\begin{aligned}s|\tau &\sim N(m(\tau), v(\tau)) \\ m(\tau) &= g_1\beta(\tau) + g_0\alpha(\tau) \\ v(\tau) &= g_1\beta(\tau)(1 - \beta(\tau)) \\ &\quad + g_0\alpha(\tau)(1 - \alpha(\tau))\end{aligned}$$

Further assuming uniform priors for τ and s , we get the unnormalized density function for $p(\tau|s)$ suggested in the main text.

Principally due to the properties of the Normal approximation to the Binomial, the approximation is very accurate for moderately large values of g_0 , g_1 and s ; this partly explains the empirically verified accuracy of the approximation for realistic gene selection scenarios.

An alternative approach to working out $p(\tau|s)$ would be to find the distribution of the order statistic $T^{(s')}$, where T is a single random variable representing the complete set of g_1 draws from p_1 and g_0 draws from p_0 , and $s' = g - s + 1$. The distribution we have suggested is thus effectively an approximation to $p(T^{(s')})$.