UW Biostatistics Working Paper Series

1-27-2003

Selecting Differentially Expressed Genes from Microarray Experiments

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Gary M. Longton
Fred Hutchinson Cancer Research Center, glongton@fhcrc.org

Garnet L. Anderson
Fred Hutchinson Cancer Research Center, garnet@whi.org

Michel Schummer
Institute for Systems Biology

CONFIDENTIAL

Selecting Differentially Expressed Genes from Microarray Experiments

Margaret Sullivan Pepe,[1,2] Gary Longton,[2] Garnet L. Anderson,[2] and Michel Schummer[3]

[1]Department of Biostatistics, University of Washington

Seattle, Washington 98195-7232, U.S.A

[2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center

Seattle, Washington 98109-1024, U.S.A.

[3]Institute for Systems Biology, Seattle, Washington 98105-6099, U.S.A.

**email: mspepe@u.washington.edu*

SUMMARY. High throughput technologies, such as gene expression arrays and protein mass spectrometry, allow one to simultaneously evaluate thousands of potential biomarkers that distinguish different tissue types. Of particular interest here is cancer versus normal organ tissues. We consider statistical methods to rank genes (or proteins) in regards to differential expression between tissues. Various statistical measures are considered and we argue that two measures related to the Receiver Operating Characteristic Curve are particularly suitable for this purpose. We also propose that sampling variability in the gene rankings be quantified and suggest using the 'selection probability function,' the probability distribution of rankings for each gene. This is estimated via the bootstrap. A real data set derived from gene expression arrays of 23 normal and 30 ovarian cancer tissues are analyzed. Simulation studies are also used to assess the relative performance of different statistical gene ranking measures and our quantification of sampling variability. Our approach leads naturally to a procedure for sample size calculations appropriate for exploratory studies that seek to identify differentially expressed genes.

KEY WORDS: Classification; Discrimination; Exploratory analysis; Genomics; Prediction; Proteomics; ROC curves.

1. Introduction

The development of microarrays that provide simultaneous evaluation of mRNA expression levels for thousands of genes is one of the exciting new advances in modern medical research. It promises to identify disease at its most basic biological level, namely at that of the genes. The implications for medicine are considerable (The Chipping Forecast, 1999). Insights into genetic alterations caused by disease can lead to new therapeutic strategies. Genetic alterations that precede disease can be targets for disease prevention strategies. The research community can expect insights into the etiology of disease and pathways involved in its progression, that may well revolutionize medical practice.

There are new statistical challenges posed by data from microarray experiments, due primarily to the exploratory nature of experiments and the huge numbers of genes under investigation. It must also be recognized that different sorts of questions are addressed with microarray experiments and that the appropriate statistical approach depends, of course, on the question of interest (Dudoit et al, 2000). Categories of objectives pertaining to experiments that include multiple tissue types (e.g. cancer versus non-cancer tissue) include: i) selection of genes that are differentially expressed in different known classes of tissue; (ii) identification of a minimal combination of genes that provides discrimination between known tissue types; (iii) identification of groups of genes whose expression levels are correlated; and (iv) new classifications of tissue types defined by genes whose expression levels are related. Statistical techniques such as regression methods and discriminant analyses have been adapted for (ii) (Dudoit et al, 2000), whereas clustering techniques are more appropriate for (iii) and (iv) (Tibshirani et al, 2000, Hastie et al, 2000, Lazzeroni and Owen, 2002, Van Der Laan and Bryan, 2001). In this paper we consider statistical methods for objective (i), which, at first glance, seems to be the most straight-forward.

The particular application that motivated our work concerns the search for biomarkers of ovarian cancer that could be used in population screening. Ovarian tissue from 30 subjects with cancer and 23 subjects without cancer were analyzed for mRNA expression using glass arrays spotted for 1536 gene clones. The data, Y_{ig}, for the g^{th} gene clone in the i^{th} tissue sample, is a measure of the mRNA expression of the g^{th} gene in that tissue relative to a control tissue, with a common control employed for all experiments. We refer to Dudoit et al. (2002) and Newton et al. (2000) for a simple summary of this technology and a technical explanation for how Y_{ig} is calculated. Schummer et al. (1999) describe their data processing algorithms that are similar to those used to arrive at the relative expression values in the ovarian cancer study. Using standard terminology for these experiments, Y_{ig} is the logarithm of the ratio of the intensities of the red to green fluorescent dyes, where green dye is used for the common control and red is used for the experimental tissue.

The scientific objective is to identify genes that are differentially expressed in ovarian cancer tissue compared with normal ovarian tissue. Ovarian tissue cannot of course be used directly for population screening. However, if a gene is found that is expressed differentially in cancer tissue, then the corresponding protein product (or an antibody to it) may be detectable in blood or urine and could be the basis for a population screening test. We refer to Pepe et al (2001) for discussion of the phases of biomarker development from the initial exploratory phase that we discuss here to its application in population screening programs. In general, scientists are more interested in identifying genes that are over-expressed rather than under-expressed in cancer screening research. This is because detecting the presence of a new aberrant protein in blood is a potentially easier task than detecting the reduced level of a normal protein, particularly if that protein is also produced by normal organ tissue in the body of the patient with cancer. Therefore in this paper we focus on detection of over-expressed genes although adaptation of the methods to detection of under-expressed genes is obvious.

There are many genes over-expressed in cancer tissue that cannot lead to screening markers. For example, genes that relate simply to inflammation or growth are not candidates because those processes also occur naturally in the body. Clinical assays for some gene products may be too difficult to develop for technical reasons. Therefore we need to select a sizeable number of over-expressed genes in order to arrive at a subset that might have potential for screening. For the initial selection, we will include multiple genes that are redundant in the sense that they identify the same cancer samples so that if one gene proves useless for biomarker development we can still pursue another that could identify those same cancers.

The experimental data are used to rank candidate genes according to some statistical measure characterizing differential expression. In section 2 we discuss the choice of statistical measure. A method for quantifying the degree of confidence in the ranking of a gene provided by the data is proposed is section 3. This acknowledges the finite number of tissues examined, variability across tissues and the large number of genes investigated, all of which contribute to uncertainty in the ranking of the genes. Application to the ovarian cancer data in section 4 illustrates the approach. One approach to computing sample sizes in these exploratory studies is suggested in section 5. Some further remarks about experimental design are made in section 6. We close with some thoughts on further extensions of our proposed methods.

2. Characterizing Interesting Differential Expression

2.1 Measures of Discrimination

At each gene, data are available for n_D cancer tissues and n_C normal tissues.

$$\left\{ \begin{matrix} Y_{gi}^{D}, i = 1, \ldots, n_D \\ Y_{gj}^{C}, j = 1, \ldots, n_C \end{matrix} \right\}.$$

To say that there is differential expression at gene *g* is to say that the distribution of Y_g^D is different from that for Y_g^C. What sorts of differences are of particular interest? Figure 1 displays some hypothetical distributions that we use for discussion. Although we depict the distribution of Y_g^C as a standard normal distribution, this is a matter of convenience only and our discussion is more general in that we do not assume any particular distribution for Y_g^C. Our discussion here only concerns the *separation* between the distributions for Y_g^C and Y_g^D. Note that there always exists a transformation so that Y_g^C is standard normal and the view in Figure 1 is on this scale. Since most of the procedures we will discuss are rank based, knowledge of the specific transformation is not necessary. Moreover, our discussion about separation in this section does not require knowledge of the transformation either.

The ideal situation is represented in the top panel where there is almost complete separation between the distributions. In this case the relative expression level of gene *g* is an ideal candidate marker for cancer because the values are completely different in cancer tissue from those in normal tissue. There is a threshold value that allows one to classify cancer versus normal tissue with almost 100% accuracy.

Consider now settings where the distributions overlap. We contend that for cancer screening, the separation in panel II is of more practical interest than that in panel III. The marker clearly distinguishes a subset of cancers from normals in II, whereas in panel III marker values for cancer tissues are entirely within the range of those for non-cancer tissues. Looking ahead to population screening and assuming that gene expression translates roughly into protein expression, in panel II there is a threshold for the screening test that provides detection of about 30% of cancers while falsely identifying only 1% of non-cancers as screen positive. In screening it is important to keep false positive rates extremely low because even a small false positive rate translates into large numbers of people being subjected to diagnostic procedures that are costly and invasive. Using a similar threshold in panel III corresponding to the 1% false positive rate, detects only 2% of cancers because the distributions overlap over the whole normative range.

We suggest that statistical measures of discrimination between the distribution of Y_g^D and Y_g^C focus on separation at and beyond upper quantiles of the normative range. Figure 2 shows receiver operating characteristic (ROC) curves that characterize separations between distributions. Each point on the ROC curve, (t,ROC(t)), corresponds to a different threshold u, and by definition $t = P[Y_g^C \geq u]$, and ROC(t)=$P[Y_g^D \geq u]$. The ROC curve can be thought of as a plot of the true versus false positive rates associated with all possible thresholds for classifying a tissue as cancerous based on the relative expression level Y_g (Pepe, 2000). Because low values of t correspond to high quantiles of Y_g^C, our suggestion is to focus on the ROC curve at low values of t.

Two summary measures of discrimination that are commonly used in ROC analysis are:

$$ROC(t_0) = P[Y_g^D \geq y^C(1-t_0)]$$

and

$$pAUC(t_0) = \int_0^{t_0} ROC(t)dt.$$

where t_0 is some small false positive rate and $y^C(1-t_0)$ is the quantile in the upper tail of the normative range corresponding to t_0. The measure $\mathrm{ROC}(t_0)$ is easily conceived of by non-statisticians, as the proportion of cancer tissues with expression levels above the $(1-t_0)$ quantile of normal tissues. The partial area under the curve, $\mathrm{pAUC}(t_0)$, in effect averages this proportion across values of $t < t_0$ (McClish, 1989). If two curves have the same value of $\mathrm{ROC}(t_0)$, the curve with larger $\mathrm{pAUC}(t_0)$ would indicate better separation at that gene because for some values of $t < t_0$, $\mathrm{ROC}(t)$ must be higher for that gene.

The $\mathrm{ROC}(t_0)$ or $\mathrm{pAUC}(t_0)$ statistic calculated for the three settings of Figure 1, ranks biomarker II better than biomarker III for small values of t_0 ($t_0 \leq 0.10$). On the other hand, other classic measures of discrimination such as the two-sample t-statistic or the Mann-Whitney U statistic(equivalently the Wilcoxon statistic) rank biomarker III better than biomarker II. We regard this as a serious weakness of those statistics for our application. We also see from Figure 2 that all of these statistics rank biomarker I as the best, regardless of t_0, and indeed any reasonable statistic should because biomarker I is almost perfect.

How should one choose t_0? Ideally the choice of t_0 will depend on false positive rates that are acceptable in practice, and t_0 could be chosen as the maximally acceptable one. The magnitudes of false positive rates that are acceptable will vary with the application since they depend on the costs and consequences of the errors. Very small t_0 are in general required for cancer screening. However, with small numbers of tissue samples, estimation of $\text{pAUC}(t_0)$ or $\text{ROC}(t_0)$ at very small t_0 will not be possible. Thus in our application we chose t_0 to be small, but large enough that the estimates of $\text{ROC}(t_0)$ and $\text{pAUC}(t_0)$ were reasonably precise for our purposes. Further research into appropriate choices for t_0 in large and small sample studies would be worthwhile.

We suggest that empirical estimates of $\text{ROC}(t_0)$ and $\text{pAUC}(t_0)$ be used to rank genes for differential expression in cancer versus normal tissue. Other measures of discrimination that we calculate for comparison are: (i) Zstat, the standardized difference in means, i.e., the two-sample t-statistic and (ii) AUC, the area under the entire ROC curve

$$AUC = \int_0^1 ROC(t)dt.$$

Interestingly the empirical AUC is equal to the numerator of the Mann-Whitney U-statistic, $\sum_i \sum_j I[Y_{gi}^D \geq Y_{gj}^C]/n_D n_C$, and hence equivalent to the Wilcoxon ranksum statistic for comparing the distribution of Y_g^D and Y_g^C. It can be interpreted as an estimate of $P\left[Y_g^D \geq Y_g^C\right]$ (Bamber, 1975). Each of $\text{ROC}(t_0)$, $\text{pAUC}(t_0)$ and AUC are distribution free rank statistics whereas Zstat depends on the underlying probability distributions for Y_g^D and Y_g^C.

2.2 Illustration

To illustrate our ideas we consider a small dataset comprising the first 100 genes in our ovarian cancer dataset. Table 1 displays the top 10 ranking genes in order when ranked according to the different statistical measures. Later we will return to the larger dataset. For illustration purposes, we chose a smaller set here because this provided substantial variation in the discrimination capacities of the top 10 genes while the top 10 genes from the larger pool of genes were less varied.

Turning to Table 1 we see that to a large extent the same genes were identified by all discrimination measures, although the order of ranking differed. Consider, however, genes 5 and 97 for which raw data and ROC curves are displayed in Figure 3. The Mann-Whitney U-statistic (AUC) ranked these genes very similarly, as the 6th and 8th, respectively. On the other hand, the pAUC statistic ranked them quite differently as the 3rd and 31st ranking genes, respectively. The raw data and the ROC curves indicate that indeed for gene 5 more of the cancer tissues are above the bulk of the normative range than is the case for gene 97. The pAUC statistic picks up on this fact and gives it a far higher rank than it gives gene 97. It suggests to these authors that gene 5 should receive higher priority for biomarker development than gene 97.

Insert Table 1

2.3 Additional steps for selection.

The main point we wish to make in this section is that careful consideration of the statistical measure used to rank genes in regards to differential expression is warranted in applications. In disease screening, ROC or pAUC measures are proposed. The ranking is of course only one step in the process of selecting genes for further study. One will investigate the actual separation achieved between the distributions of Y_g^D and Y_g^C for genes that rank well. ROC curves such as shown in Figure 3b should be considered in this evaluation because they display the separation achieved on a scale that is relevant to the problem and that allows for direct comparisons between genes. It is more difficult to compare genes using frequency distributions of the raw expression data (Figure 3a).

The next step towards selecting genes for further experimental work is to investigate what is already known about the function of the genes that appear to have promising differential expression. Libraries of information are available from the public and private domains. Genes may be eliminated from further investigation for a variety of reasons related to their known function or prior experience with assay development. Investigators then select some set of genes for further investigation. The number depends on multiple factors, not the least of which is the resources available for experimental work.

In this paper we focus on the initial step in the gene selection process, namely the ranking step that orders genes in regards to a statistical measure of differential expression. In section 3 we discuss the sampling variability in the rank ordering. Related to this, in section 5 we propose that the achievement of adequate rank ordering be a basis for choosing sample sizes in these studies. In particular we propose that the study design should ensure that genes with promising differential expression should rank high and therefore be drawn to the attention of the investigators.

3. Assessing Variability

3.1 The probability of gene selection.

The relative rankings of genes is the primary outcome of the study. However, the rankings are subject to sampling variability. How should this variability be acknowledged? Standard errors or p-values don't seem to be directly relevant to the task because the stated objective is neither to estimate parameters nor to test hypotheses. Rather, the task is to rank genes and to select the top genes for further study. Therefore we propose the following quantity to quantify our degree of confidence in choosing the g^{th} gene among the top k.

$$P_g(k) = P[\text{gene } g \text{ ranked in the top } k]$$

$$= P[\text{Rank } (g) \leq k]$$

The value of $P_g(k)$ may be of particular interest for *k* equal to a predetermined number of genes to be selected (10 in the small illustration). However, the whole survivor function can be considered, $\{P_g(k),\ k \geq 1\}$, and this gives a more full description of sampling variability in the ranking. Various factors contribute to the variability in Rank(g): (i) the number of cancer tissues and normal tissues studied, n_D and n_C; (ii) the extent and type of differential expression of the g^{th} gene; (iii) the number of genes in the selection pool, which we denote by N; (iv) the differential expression of genes other than the g^{th} gene; and not least, (v) the algorithm or statistical measure used to rank genes. The quantity, $P_g(k)$, will be affected by all of these factors.

Intuitively, as sample sizes increase, the $P_g(k)$ function will tend to 0 or 1 for differentially expressed genes according to whether the true asymptotic discriminating measure for the g^{th} gene ranks below k or not. Genes that in truth are very highly discriminatory will certainly have high ranks even in experiments with small sample sizes and $P_g(k)$ will be close to 1. This may be reduced by chance if there is a large number of competing genes and in particular if a substantial number of competing genes also exhibit differential expression. Observe that at the opposite extreme, if no genes are differentially expressed, then $P_g(k) = k / N$.

The selection probabilities, $P_g(k)$, as we call them, can be estimated by the bootstrap with the resampling unit being at the tissue level. Thus, when a tissue is included in the bootstrap sample, the entire vector of data relating to all genes for that tissue is entered into the bootstrapped dataset, and genes are ranked within the dataset according to the statistical measure chosen. The bootstrapping therefore acknowledges the complex correlations between genes.

All of our statistical measures but Zstat are rank statistics. Tied data points influence the distribution of rank statistics and we note that tied data points ensue with simple resampling of observed data. However, real data, such as the original dataset, do not have ties because Y_g is measured on a continuous scale. Thus, we modified the bootstrapping to randomly break ties by adding miniscule random noise (jitter) to the expression levels. This was done in an effort to make the bootstrap distribution of the rank statistics more reflective of the actual distribution across different realizations of the experiment.

3.2 Illustration

Returning to the small illustration described earlier, Table 1 shows $P_g(10)$ based on 200 bootstrapped samples for each gene ranked in the top 10. Thus if the strategy of the experiment is to select the top 10 genes ranked on the basis of ROC(0.1), we are highly confident $(P_g(10) > 90\%)$ about the selection of genes 93, 76, 65 and 42. However, we estimate that, due to sampling variability, genes 35, 23 and 52 have $\leq 60\%$ chance of ranking in the top 10 if the experiment were repeated.

A comparison of the two estimators of $P_g(10)$, with and without jitter in the bootstrap sample, suggested that they are quite similar. That is, we arrive at the same conclusions about $P_g(10)$ for the rank-based measures if the data are jittered or not. Thus tied datapoints in the bootstrap samples do not appear to affect the $P_g(10)$ estimates substantially.

3.3 Simulation

As a simple example we simulated data on 2000 genes for equal numbers of cancers and normal tissues. Of the 2000 genes, 95% were configured to be non-informative in the sense that Y_g^D and Y_g^C had the same distributions, namely standard normal (without loss of generality). For 100 genes the distributions were normal with mean 1 and standard deviation 2 for cancer tissues and standard normal for non-cancer tissues. For an informative gene therefore, the area under the corresponding ROC curve was $\Phi((1-0)/\sqrt{2^2+1^2}) = 0.67$ (Reiser and Guttman, 1986). Data for different genes were generated independently. We set the number of genes to be selected at $k = 100$. Table 2 panel A shows the proportions of informative markers selected averaged across 100 simulation studies. That is, it shows P[Rank(g)≤$k \mid g$ is an informative gene].

Insert table 2

The results suggest that the top 100 genes consist primarily of informative genes even with relatively small sample sizes. An informative gene has a 68% chance of being in the top 100 ranked on the basis of the pAUC(0.2) statistic when 30 samples are analyzed, 15 cancer and 15 normal tissues. The chance reaches 91% when a total of 35 cancer and 35 normals are evaluated.

In this particular example, (setting A of Table 2), the pAUC statistic was most effective at selecting informative genes. Interestingly it outperformed the full area under the curve the AUC statistic. That is, focusing on differences between the normal and cancer tissues only in the upper end of the normative range, yielded a better selection algorithm. This will not always be the case. In another set of simulations (also shown in Table 2) where informative genes had a mean 1 and standard deviation 1 in cancer tissues compared with standard normal in non-cancer tissues, the AUC statistic performed better.

These simulations assume statistical independence of genes and hence are unlikely to reflect real data. In practice one might find that subsets of informative genes are correlated statistically, and likewise subsets of uninformative genes are correlated. Intuition suggests that this would lead to higher selection probabilities for informative genes than those in Table 2 because correlated genes will behave as a unit and reduce variability compared with the setting where all genes are independent. We will return to simulation studies later when we consider sample size calculations.

4. Analysis of the Full Ovarian Cancer Dataset

The 1536 genes spotted on the glass arrays were ranked according to each of the discriminatory statistics defined above. Sixty-five genes were ranked in the top 100 by all 4 ranking statistics while 16 genes were selected among the top 100 by only one of the statistics (7 by ROC (0.10), 9 by pAUC (0.10), 0 by AUC and 0 by Zstat only).

The stability of their selection, quantified by $P_g(100)$, was estimated with 200 bootstrap samples. Figure 4 displays the results. The selection probabilities for the AUC and Zstat statistics are overall higher than those for the pAUC (0.10) and ROC (0.10) statistics. This presumably indicates less variability in the statistics that use more of data, namely AUC and Zstat. The selection algorithms based on them therefore are less variable and more reproducible across experiments. However, we saw earlier (Table 2) that this does not necessarily induce higher sensitivity to differential expression and in particular to the sorts of differential expression of most interest to biologists.

Another display of the resampling results, specifically for the pAUC (0.1) statistic, is shown in Figure 5. For each gene selected we calculated its ranking in each bootstrap sample. Its 80th and 90th percentile across the bootstrap samples is shown. We observe for example that gene 1483, which ranked best in the original dataset, ranked at or above 14 in 90% of the resampled datasets and at or above 8 in 80% of the resampled datasets. Gene 65, which ranked 50th in the original dataset, had ranks of 148 and 115 at its 90th and 80th bootstrap percentiles, respectively. We see that for all genes ranked in the top 24, their rankings were better than 100 in at least 90% of the bootstrap samples. Thus, we have high confidence in the good ranking of these genes, in the sense that it is unlikely to be attributable to sampling variability. On the other hand, all of the genes that ranked worse than 63rd in the original data were at the 150th rank or worse in at least 10% of bootstrap samples, and 15/37 (41%) had 90th percentiles above 200.

Let's briefly consider the biological relevance of the highest ranking genes. The top 10 ranking clones for the pAUC(0.1) statistic are SPINT2 (2 clones), TACSTD1, HE4, Oviductal glycoprotein, Keratin 8, Argininosuccinate synthetase (ASS), 2 ESTs, and a novel gene. Of the six genes with known function, five are tumor-related: SPINT2 is expressed in colorectal cancer (Kataoka et al., 2000); TACSTD1, an adenocarcinoma-associated antigen, is currently being used in a clinical trial as a target in the treatment of gastro-intestinal carcinomas (Staib et al., 2001); HE4 is a potential ovarian cancer marker (Schummer et al., 1999), which is currently being evaluated in a serum assay (unpublished results); oviductal glycoprotein has a role in fertilization (Verhage et al., 1997) and was found to be expressed at higher levels in ovarian carcinomas (unpublished results); and Keratin 8 expression is associated with cervical cancer progression (Smedts et al., 1990). Moreover, one of the two EST-related clones is homologous to a putative integral membrane transporter protein discovered in hepatocellular carcinoma (NCBI website http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=retrieve&db=Nucleotide &list_uids=7320864&dopt=GenBank).

With five of six top-ranking genes known to be related to cancer, our biologist colleagues are motivated to study further the remaining 3 genes with unknown function. They suspect that those genes may be tumor-related as well. This of course remains to be seen.

Of the top 10 genes selected according to the pAUC(.1) statistic, 6 were also ranked in the top 10 by the AUC statistic. The 4 additional genes ranked in the top 10 by AUC included one with unknown function, one that relates to a brain protein not found in normal ovary, one “housekeeping genes” involved in glycolysis and one (IFI27) that has been found to be overexpressed in breast carcinomas. The last two were ranked 28 and 30, respectively, with the pAUC(.1) statistic having values that were approximately 64% of the ideal value of 0.1. In contrast the 9th and 10th ranking genes according to the pAUC(.1) algorithm were 80% of the ideal value.

5. Sample Size Calculations

Gene expression microarray experiments are expensive. Therefore in practice sample sizes tend to be small. Our simulation study and analysis of the ovarian cancer dataset nevertheless, suggested that an informative analysis, properly accounting for sampling variability, can be based on experiments with relatively small sample sizes. This initially seemed surprising to us, but in retrospect it is intuitively reasonable. Informative genes will show themselves to be so even with small sample sizes.

What advice can the statistician offer for choices of sample sizes in exploratory gene expression studies? Since the task is to select informative genes from the pool of genes studied, the criterion for choosing sample sizes should be that they be large enough to ensure that informative genes have a high chance of being selected for further study on the basis of data from the experiment. Again, traditional notions of basing sample size calculations on hypothesis tests or on precision of estimators seem inappropriate.

Suppose that resources exist such that the top ranked k_0 genes will be considered for further study, $k_0 = 100$, say. The sample size might be driven by the requirement that an informative gene, ranking in truth in the top k_1 genes ($k_1 = 30$ say), has a probability of at least β of being ranked in the top k_0 in the experiment. That is, the investigator suspects that a gene that ranks in the top k_1 will be of interest and wants to be assured that such a gene will be identified in the experiment. Therefore, one might choose n_D and n_c so that

$$P_g(k_0 \vert \in k_1) = P[\text{Rank}(g) \le k_0 \vert \text{TrueRank}(g) \le k_1] = \beta$$

where True Rank*(g)* is the ranking of gene *g* according to the ranking statistic chosen, if an infinite number of tissue samples were studied, $n_D = n_c = \infty$. This probability is related to the selection probability $P_g(k)$ defined earlier. However, here instead of quantifying confidence in observed results, it now quantifies the power of the experiment to select a gene that in truth is a high ranking gene. Like $P_g(k)$ it depends on the size and contents of the pool of genes considered, the ranking statistic used and importantly on n_D and n_c .

To calculate $P_g\,(k_0\,|\!\in k_1)$ we suggest that a simulation study be performed. In fact the simulation study described in section 3.3 was our first attempt at this. In that setting we calculated $P_g(100\,|\!\in 100)$ for various sample sizes and showed that even with a total sample size of 30, $n_D = n_c = 15$, the study design had a power β = 68% assuming that the pAUC (0.1) ranking statistic was used for analysis. The data generating mechanism in that simulation, however, is very simple and is not based on a theoretically justifiable model. Sample size calculations cannot be used for practical application without such justification. Unfortunately, it is extremely unlikely that one can ever stipulate a simulation model for gene expression array data that is based on adequate biological theory and knowledge of laboratory processes.

Ideally a set of pilot data would be available upon which to base a simulation. To illustrate, suppose that the ovarian cancer data represents a dataset from a pilot study. We based a second simulation study on these data. Specifically we resampled with replacement the entire data vector of gene expressions for n_D^* cancer tissues and n_c^* normal tissues from the original dataset and determined $P_g(k_0\,\big|\in k_1)$. Various sample sizes, n_D^* and n_c^* were considered. That is, the distributions of observed data were regarded as the population distributions for cancers and normals and we randomly selected from those (infinite) populations in order to simulate data for the planned experiments. (In this sense bootstrapping can be considered as a simulation.) Table 3 displays $P_g(100 \,|\!\in k_1)$ for various choices of sample sizes. We see that a gene that in truth ranks in the top 10 according to the pAUC (0.1) measure is almost certainly selected with data from a study involving as few as 30 tissues, if the selection criterion is that its pAUC (0.1) statistic ranks in the top 100 in the study. A gene, truly in the top 50 is likely to be selected (*β*=91%) from a study using 25 cancer and 25 non-cancer tissues.

The power , $P_g(k_0 \,|\!\in k_1)$, quantifies how likely a gene randomly selected from the top k_1 is likely to be ranked in the top k_0. Table 3 also displays $P_g(k_0 \,| \cup k_1)$ which is the probability that *all* k_1 truly top-ranking genes will rank in the top k_0 when the experiment involving n_D^* and n_C^* tissues is performed. These probabilities are much lower because the criterion to be met is more stringent. In order that all top 30 genes be likely to be selected it appears that at least 100 tissues $n_D^* = n_C^* = 50$ should be studied ($P_g(100 \,| \cup 30)$=84%).

In these simulations we only considered equal numbers of cases and controls. Unequal sample sizes could be chosen. It would be interesting to see if, in general, relatively more cases than controls are desirable and how this should in general relate to the relative variability of gene expressions in cases versus controls. Another aspect that we feel should be explored further relates to the likely over-optimism of the pilot data that we use for simulation. Efron and Tibshirani (1994, section 25.5) suggest some caution about plugging in parameters from a pilot study for power calculations and their concerns apply here too. One could add noise to the observed data in the simulations for more conservative sample size calculations.

6. Additional Design and Data Analysis Considerations

We have considered only the comparative design where assays for both the normal tissues and cancer tissues are performed, each using a common control tissue. Thus a sample of relative expression values are obtained for both the normals and the cancers, represented by $\left\{Y_{ig}^{D}, i=1,\ldots,n_D\right\}$, and $\left\{Y_{jg}^{C}, j=1,\ldots,n_C\right\}$, respectively. In this design the distribution of Y_g^D can be compared with that of Y_g^C , the latter being the appropriate reference distribution.

An alternative design frequently cited in the statistical literature (Van Der Laan and Bryan (2001)) entails using a non-cancer tissue as a control within the assay for a cancer tissue. The data at gene *g* from such an experiment can be represented as $\{Z_{ig}, i=1,\ldots,n_D\}$ where Z_{ig} is the expression in the cancer tissue relative to the normal control transformed to a log scale. Typically the mean of the distribution of Z_g is compared with 0, the null value if expression at the gene g is the same on average in cancer and normal tissue.

We have argued in section 2 that the mean difference $E\left\{Y_g^D - Y_g^C\right\} = E\left\{Z_g\right\}$ is only one summary of the separation between the distributions of Y_g^D and Y_g^C, and that in many cases alternative summary measures are more relevant. Unfortunately summary measures, such as pAUC, are not identifiable from the distribution of Z_g. Indeed we believe that the two distributions for Y_g^D and Y_g^C, respectively, or at least their ROC curve should be generated by an experiment in order to adequately assess differential expression. Unfortunately they simply cannot be reconstructed from the single distribution of the composite variable Z_g. Clearly many different pairs of random variables (Y_g^D, Y_g^C) can give rise to a single composite $Z_g = Y_g^D - Y_g^C$.

In summary, for the type of application we consider in this paper, we prefer the design that yields relative expression levels for both normals and cases instead of just the composite Z_g. This design allows a full and flexible comparison of the two distributions, that for normal tissues yielding a reference distribution against which the cancer tissue distribution can be compared. Such is not achieved with the design that evaluates normals only within the assay for the cancer tissue.

7. Concluding Remarks

In this paper we have considered the identification of a subset of genes that are differentially expressed between two tissue types from a large pool of candidate genes. The same statistical problem arises in experiments involving other recently developed high throughput technologies. For example, protein mass spectrometry can be used to identify a set of proteins differentially expressed from amongst a large set of candidate proteins. Large arrays of tumor antigens are used to select a subset to which antibodies are differentially present in subjects with and without cancer. The concept of ranking genes using a statistical measure of discrimination between tissues, applies equally well to proteins in protein spectrometry and to antigens in tumor immunogenicity experiments. Thus, our methods will also be useful in these settings.

We have emphasized that investigators must carefully choose the statistical measure for ranking the genes so that it fits the purpose of the experiment. For disease screening we have argued that biomarkers must be highly specific. This could be argued for other applications too, such as in the identification of treatment targets. Statistical measures such as the $\text{pAUC}(t_0)$ or $\text{ROC}(t_0)$ statistics are appealing when specificity is important. Dudoit et al (2002) use Zstat, the standardized difference in means, to rank genes. Efron et al (2000) also use a difference in means with a somewhat different standardization. Their rationale for using these measures over others was not discussed. One feature of those measures is that they depend on the absolute values of Y_g, whereas the empirical ROC statistics do not since they are rank statistics. This presumably infers robustness on the ROC statistics but at the expense of disallowing the magnitudes of relative expression to influence the relative ordering of genes. Whether or not the magnitude of Y_g should influence the gene rankings over and above the separation between the probability distributions of Y_g^D and Y_g^C, is a debatable point since magnitude of expression does not translate directly into biological effect in the body. Another feature of the $\text{pAUC}(t_0)$ and

ROC(t_0) statistics is that they are not influenced by variability in the measurement of Y_g at the lower end of the scale, at values below the (1-t_0) quantile of Y_g^C .

We have suggested the selection probability, $P_g(k)$, to quantify sampling variability and confidence in the gene ranking, and as the basis for sample size calculations. Dudoit et al (2000) use *p*-values for a related purpose. However, we find the interpretation of $P_g(k)$ more compelling given the exploratory nature of the study. The purpose is not to test a null hypothesis about equal distributions of gene expression versus unequal distributions. More importantly, the objective is to rank genes according to the extent of differential expression. Although the measures used by Dudoit et al (2000) and by Newton et al (2000) are statistics for testing a null hypothesis, they are used more in the same spirit as we use statistics, namely to rank genes according to the extent of differential expression. Efron et al (2000) consider two probabilities: a *p*-value, *Prob* {data at gene *g* | null hypothesis of equal expression}, and a Bayesian probability, *Prob* {gene *g* affected | data at gene *g*} = 1-*P* {equal expression | data at gene *g*}. Again, since many genes will be differentially expressed, probabilities relating to the null state of equal expression seem less compelling than ranking the extent of differential expression amongst the genes. Moreover, Efron et al (2000) use the probabilities to rank the genes, whereas we use the selection probabilities only to quantify sampling variability in the rankings.

Our selection probabilities are more closely related in this regard to the 'single gene probabilities' proposed by Van der Laan and Bryan (2001). The single gene probabilities are used to quantify sampling variability in a gene clustering algorithm, and are estimated by a parametric bootstrap approach. Kerr and Churchill (2001) also assess reliability of clustering algorithms with the bootstrap. Our selection probabilities quantify sampling variability in a gene ranking algorithm, and are estimated with a non-parametric bootstrap procedure. Since the bootstrap provides consistent estimates of the distributions of the vectors $\{Y_1^D,\ldots,Y_N^B\}$ and $\{Y_1^C,\ldots,Y_N^C\}$, and for any given g, $P_g(k)$ is a function of these distributions it seems intuitive that the bootstrap estimate of $P_g(k)$ will be consistent. However, it is likely that bootstrap or any data-based estimates of $P_g(k)$ will be correlated with the data-based ordering of the gene. This correlation implies that if attention is restricted to a subset of genes that are observed to rank high say, then as a group their estimated selection probabilities will tend to be biased upwards. Efforts to reduce this bias would be worthwhile.

The initial motivation for our research was to develop a strategy for sample size calculations. The strategy we propose is based on selection probabilities for informative genes, and is implemented with bootstrap simulation studies using pilot data. A similar strategy could be used for calculating sample sizes in studies that have the determination of gene clusters as the ultimate purpose. The single gene probabilities of Van der Laan and Bryan (2001) or some related construct could take the place of the selection probabilities in that sort of application.

Laboratory techniques for measuring gene expression with microarrays are certainly imperfect. Moreover, data processing procedures for calculating the relative expression values from the raw images data are evolving. The variability and biases in the derived values will surely impact on the gene rankings observed. It would be interesting to determine the extent of these impacts and if the use of ROC statistics can mitigate some of these problems because they are rank based. Hopefully, laboratory and data processing procedures will improve in the future, to alleviate related statistical concerns.

Although the identification of differentially expressed genes is the first objective, it is not the only objective of an exploratory gene expression study. In cancer research, it is recognized that cancer is a heterogeneous disease and that different unidentified subtypes may be characterized by unique sets of overexpressed genes. Thus, if a single gene doesn't completely discriminate cancer from non-cancer it may be possible that a small set of genes each flagging one subtype will. Statistical methods to identify such minimal subsets are needed. Ranking of different subsets of genes might draw on ideas presented here. In addition, the identifications of clusters of genes, that is, genes that are over- or underexpressed in the same cancer tissues would be of interest. Biological insights into the pathways and pathogenesis of cancer would likely result. Some modifications of the plaid models (Lazzeroni and Owen, 2002) to include a baseline reference group of tissues (non-cancer in our case) might be useful for this purpose.

Acknowledgements

Support for this research was provided by NIH grants GM 54438 and CA 86368 and by the Fred Hutchinson Cancer Research Center Ovarian SPORE grant. We thank Lu Chen who performed the initial simulation studies, and colleagues in Seattle, particularly Nicole Urban, Ziding Feng, and David Haynor for valuable discussions.

References

Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12:** 387–415.

“The Chipping Forecast,” (1999). *Nature Genetics*, **21 supplement**.

Dudoit S., Fridlyand J., and Speed T.P. (2000). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**(457):77–87.

Dudoit S., Yang Y. H., Speed T. P., Callow, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, Vol. 12, No. 1, p. 111–139.

Efron B. and Tibshirani RJ. (1994). An Introduction to the Bootstrap. Chapman and Hall, CRC Press.

Efron B., Tibshirani R., Goss V., and Chu G. (2000). Microarrays and their use in a comparative experiment. *Technical Report #213*, Division of Biostatistics, Stanford University. http://www-stat.stanford.edu/~tibs/ftp/microarrays.pdf

Hastie T., Tibshirani R., Eisen M., Brown P., Ross D., Scherf U., Weinstein J., Alizadeh A., Staudt L., and Botstein D. (2000). Gene Shaving: a new class of clustering methods for expression arrays. *Technical Report*, Department of Statistics, Stanford University. http://www-stat.stanford.edu/~hastie/Papers/

Hintze J.L. and Nelson R.D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician* **52**(2):181–184.

Kataoka H., Itoh H., Uchino H., Hamasura N., Nabeshima K., Kono M. (2000). Conserved expression of hepatocyte growth factor activator inhibitor type-2/placental bikunin in human colorectal carcinomas. *Cancer Letter* **148**(2):127–134.

Kerr MK, Churchill GA. (2001).Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci* U S A. **98**(16):8961–5.

Lazzeroni L. and Owen A. (2002). Plaid models for gene expression data. *Statistica Sinica* Vol. **12**(1):61–86.

McClish D.K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9:** 190–195.

Newton M.A., Kendziorski C.M., Richmond C.S., Blattner F.R. and Tsui K.W. (2000). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computer Biology* **8**1:37–52.

Pepe M.S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association* **95:** 308–311.

Pepe M.S., Etzioni R., Feng Z., Potter J., Thompson M.L., Thornquist M., Winget M., and Yasui Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*. **93**(14):1054–61

Reiser B. and Guttman I. (1986). Statistical inference for *Pr(Y<X)*: The normal case. *Technometrics* **28:** 253-257.

Schummer M., Ng W.V., Bumgarner R.E., Nelson P.S., Schummer B., Bednarski D.W., Hassell L., Baldwin R.L., Karlan B.Y., Hood L. (1999). Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Genetics* **238**(2):375–385.

Smedts F., Ramaekers F., Robben H., Pruszcynski M., van Mujien G., Lave B. Leigh I., Vooijs P. (1990). Changing patterns of keratin expression during progression of cervical epithelial neoplasia. *American Journal of Pathology* **136**(3):657–668.

Staib L., Birebent B., Somasundaram R., Purev E., Braumuller H., Leeser C., Kuttner N., Li W., Zhu D., Diao J., Wunner W., Speicher D., Beger H.G., Song H., Herlyn D. (2001). Immunogenicity of recombinant GA733-2E antigen (CO17-1A, EGP, KS1-4 Ep-CAM) in gastro-intestinal carcinoma patients. *International Journal of Cancer* **92**(1):79–87.

Tibshirani R., Hastie T., Eisen M., Ross D., Botstein D., and Brown P. (2000). Clustering methods for the analysis of DNA microarray data, *Technical Report*, Department of Statistics, Stanford University. http://www-stat.stanford.edu/~tibs/lab/publications.html

Van der Laan M. and Bryan J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* **2**(3):1–17.

Verhage H.G., Fazleabas A.T., Mayrogianis P.A., O'Day-Bowman M.B., Schmidt A., Arias E.B., Jaffe R.C. (1997). Characteristics of an oviductal glycoprotein and its potential role in fertility. *Journal of Reproduction and Fertility* (Supplement) **51**:217–226.

Figure Legends

Figure 1. Hypothetical distributions for gene expression data showing different sorts of separations between cancer tissue and normal tissue.

Figure 2. Receiver operating characteristic curves corresponding to pairs of distributions shown in Figure 1.

Figure 3(a). Frequency distributions and (b)ROC curves corresponding to gene 5 and 97 in the ovarian cancer data set.

Figure 4. (a) Violin plots (Hintze and Nelson, 1998) of selection probabilities for the top 100 ranked genes in the ovarian cancer datasets using 4 different ranking statistics. Probability estimates are based on 200 bootstrap samples. The median is indicated by a short horizontal line, the first to third interquartile range by the narrow shaded box, and a vertical line extends to the upper and lower adjacent values. The surrounding violin shell consists of mirrored local kernal density estimates of the distribution. The y-axis is labeled at the minimum, median, and maximum values.

(b) A comparison of the selection probabilities for the AUC statistic and the pAUC (0.1) statistic.

Figure 5. Gene rank percentiles (90th and 80th) in the bootstrap distribution for the ovarian cancer data set. Shown are results for the top ranked 100 genes. 200 bootstrap samples were drawn with the sampling unit being tissue.

Table 1.

Gene number, selection probability $P_g(10)$, and value of the discriminatory measure for the top 10 ranking genes among the first 100 genes in the ovarian cancer dataset.

ROC(.10)				pAUC(.10)				AUC				Z-stat			
rank	gene	$P_g(10)$	statistic	rank	gene	$P_g(10)$	statistic	rank	gene	$P_g(10)$	statistic	rank	gene	$P_g(10)$	statistic
1	93	1.00	0.900	1	93	1.00	0.090	1	93	1.00	0.971	1	93	1.00	6.149
2	76	0.81	0.767	2	65	0.99	0.059	2	42	1.00	0.870	2	65	1.00	5.090
3	65	0.98	0.733	3	5	0.94	0.051	3	76	1.00	0.864	3	42	0.96	4.238
4	42	0.92	0.667	4	23	0.83	0.044	4	65	0.98	0.854	4	97	0.74	3.543
5	5	0.89	0.600	5	42	0.60	0.041	5	16	0.82	0.804	5	39	0.71	3.321
6.5	16	0.71	0.533	6	51	0.68	0.040	6	5	0.74	0.789	6	23	0.60	3.032
6.5	39	0.61	0.533	7	52	0.63	0.040	7	52	0.74	0.784	7	35	0.55	3.011
8	35	0.58	0.500	8	35	0.38	0.033	8	97	0.71	0.780	8	76	0.50	2.664
9.5	23	0.54	0.467	9	73	0.38	0.032	9	39	0.52	0.752	9	63	0.40	2.567
9.5	52	0.43	0.467	10	76	0.48	0.032	10	75	0.43	0.736	10	5	0.47	2.554

Table 2

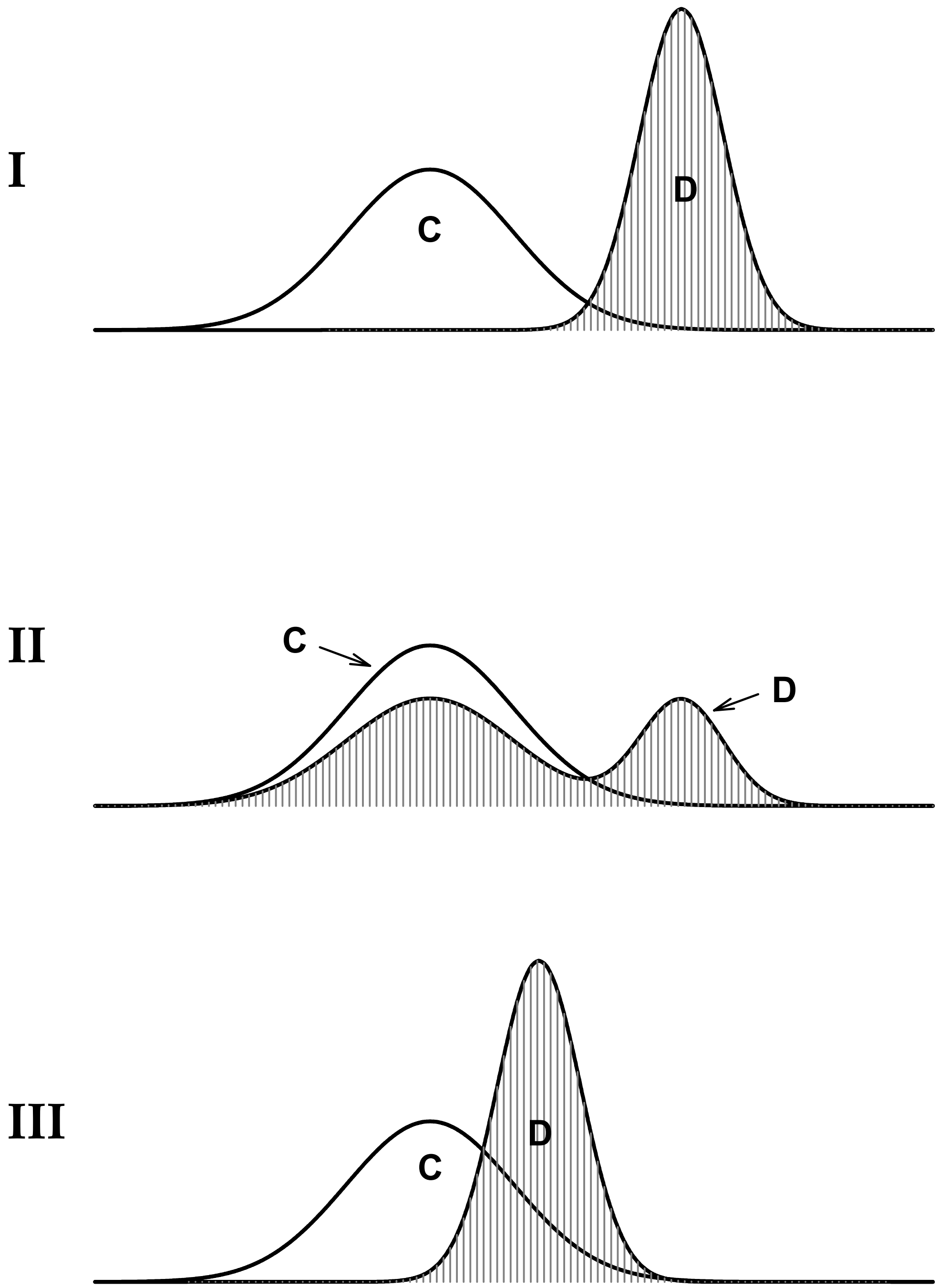
Results of a simulation study with N=2000 genes of which 100 are informative about disease status. Shown are P*[Rank(g)≤100] for informative genes. In all simulations* Y_g *has a standard normal distribution among controls and for non-informative genes among cases. The distribution of* Y_g^D *for informative genes is N(1,2) in setting A and N(1,1) in setting B.*

	A			B		
	n=#cases=#controls			**n=#cases=#controls**		
Statistic	*15*	*25*	*35*	*15*	*25*	*35*
ROC(.10)	.69	.82	.89	.57	.69	.77
pAUC(.10)	.68	.83	.92	.50	.62	.72
ROC(.20)	.59	.75	.83	.65	.76	.84
pAUC(.20)	.68	.83	.91	.58	.71	.81
AUC	.42	.56	.66	.68	.84	.92
T statistic	.42	.58	.68	.69	.84	.93

Table 3

Study power $\mathrm{P_g}$ $\{100| \in \mathrm{k}_1\}$ *as a function of sample size using the ovarian cancer data as a simulation model. Also shown is the power for the more stringent criterion* $\mathrm{P_g}$ $\{100| \cup \mathrm{k}_1\}$.

	Pg $\{100| \in \mathrm{k}_1\}$				
True Ranking (k_1)	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50
(n_D, n_c)					
(15, 15)	.997	.982	.934	.893	.850
(25, 25)	1.000	.996	.973	.949	.914
(50, 50)	1.000	1.000	.994	.987	.968
(100, 100)	1.000	1.000	.999	.998	.990
	$\mathrm{P_g}$ $\{100| \cup \mathrm{k}_1\}$.				
(15, 15)	.960	.654	.120	.016	.000
(25, 25)	1.000	.928	.486	.202	.024
(50, 50)	1.000	1.000	.836	.638	.206
(100, 100)	1.000	1.000	.984	.928	.608


I
D
C
II
C
D
III
D
C

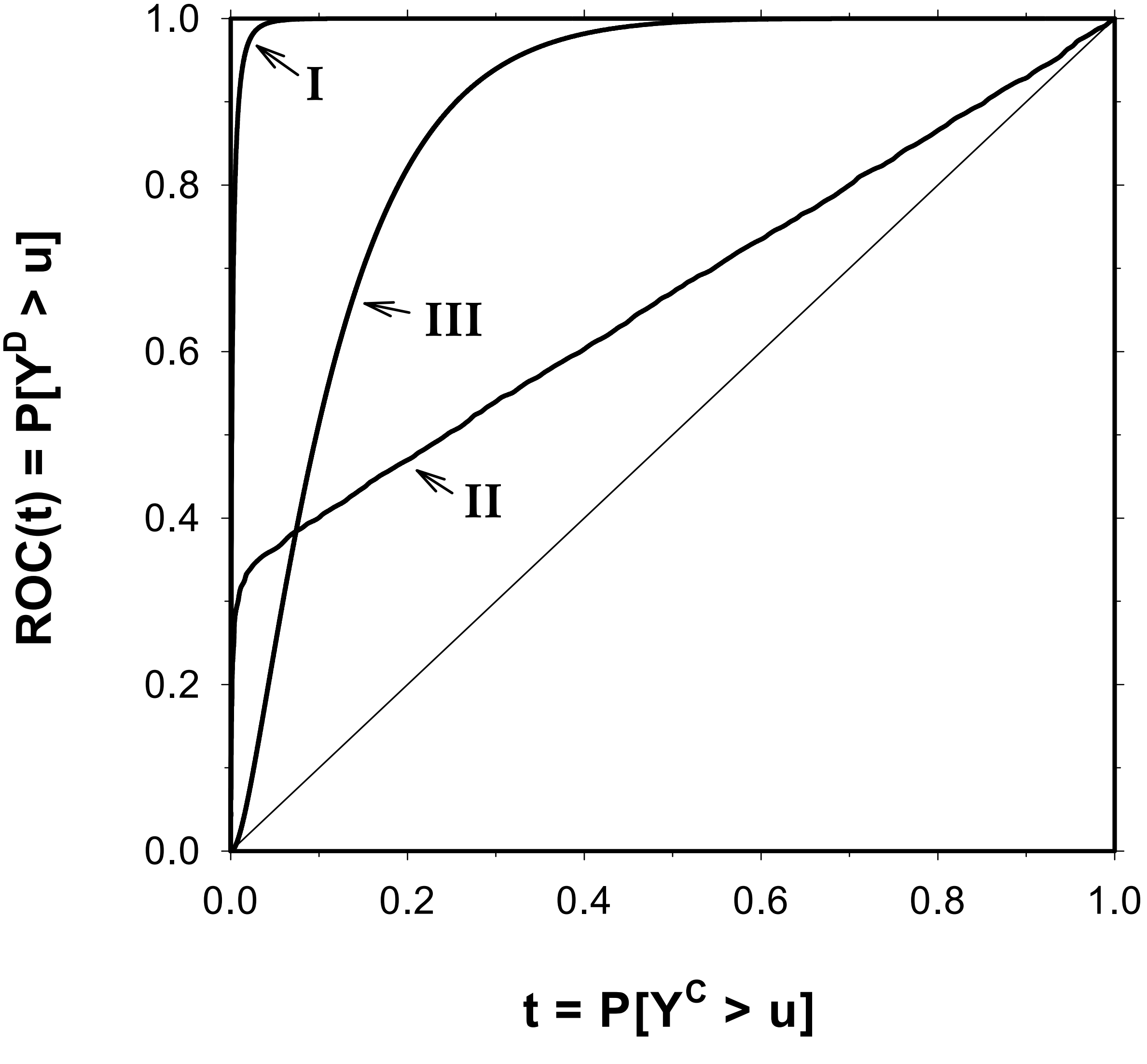


Figure 2

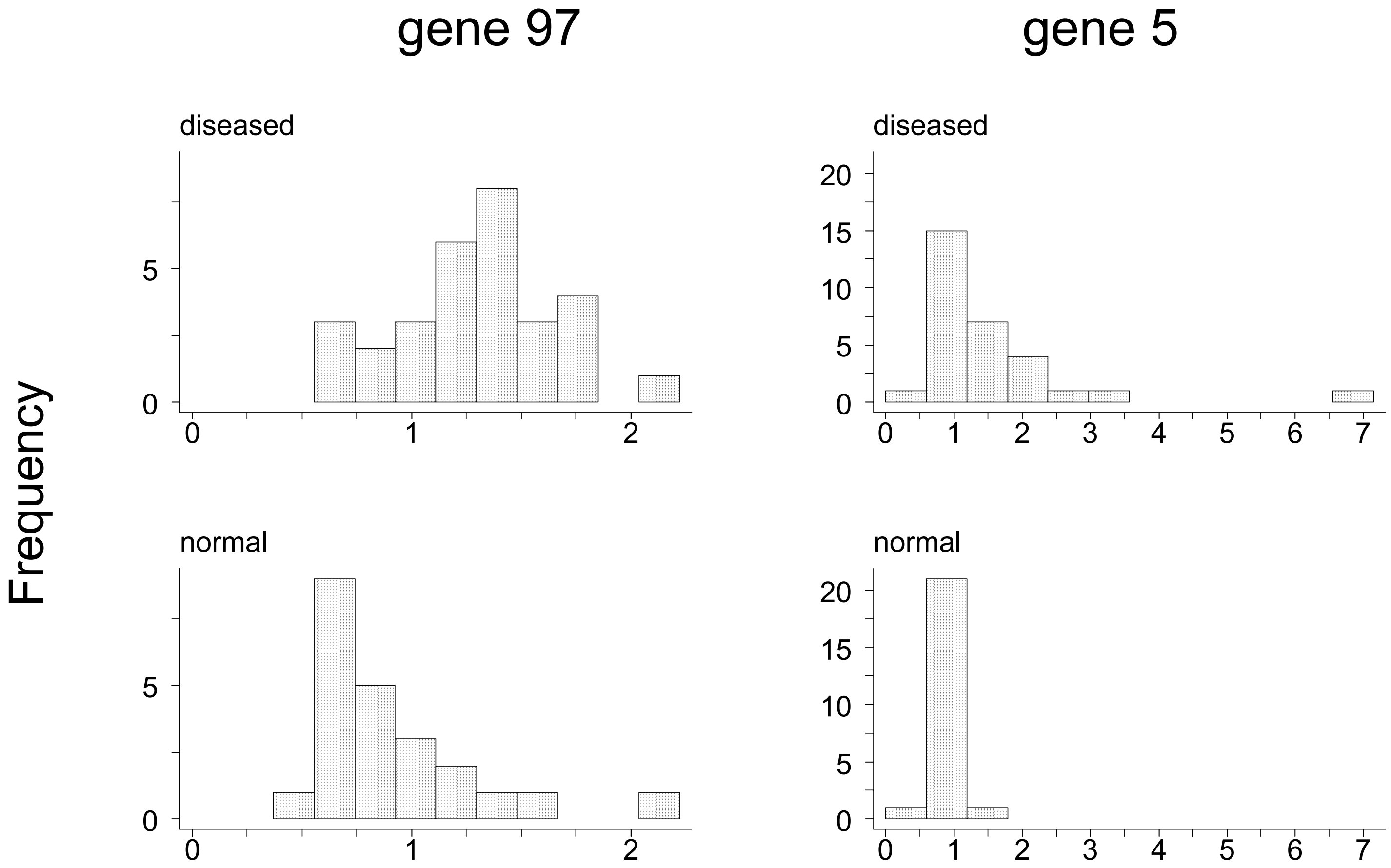
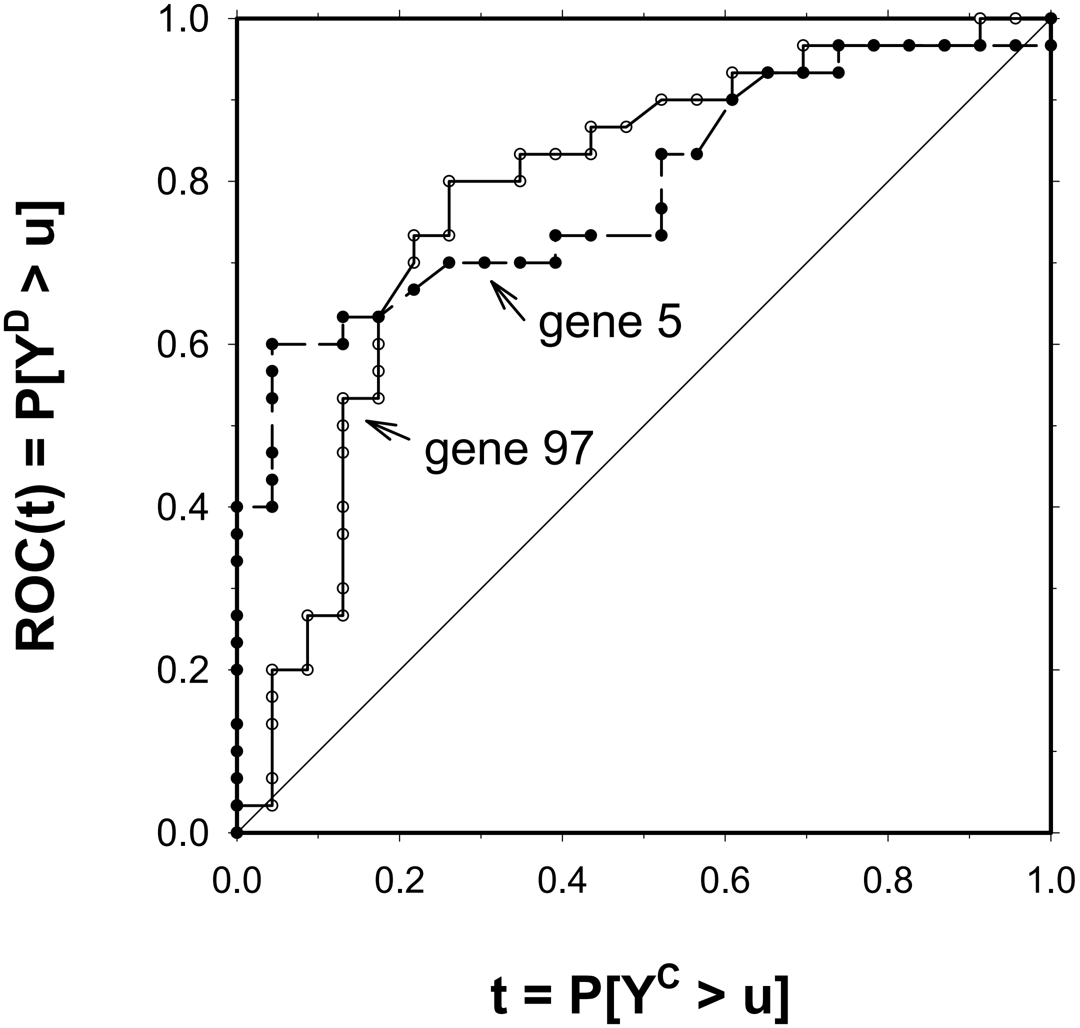

gene 97
gene 5
Frequency
diseased
5
0
0
1
2
diseased
20
15
10
5
0
0
1
2
3
4
5
6
7
normal
5
0
0
1
2
normal
20
15
10
5
0
0
1
2
3
4
5
6
7

Figure 3a


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC(t) = P[Y^D > u]
t = P[Y^C > u]
gene 5
gene 97

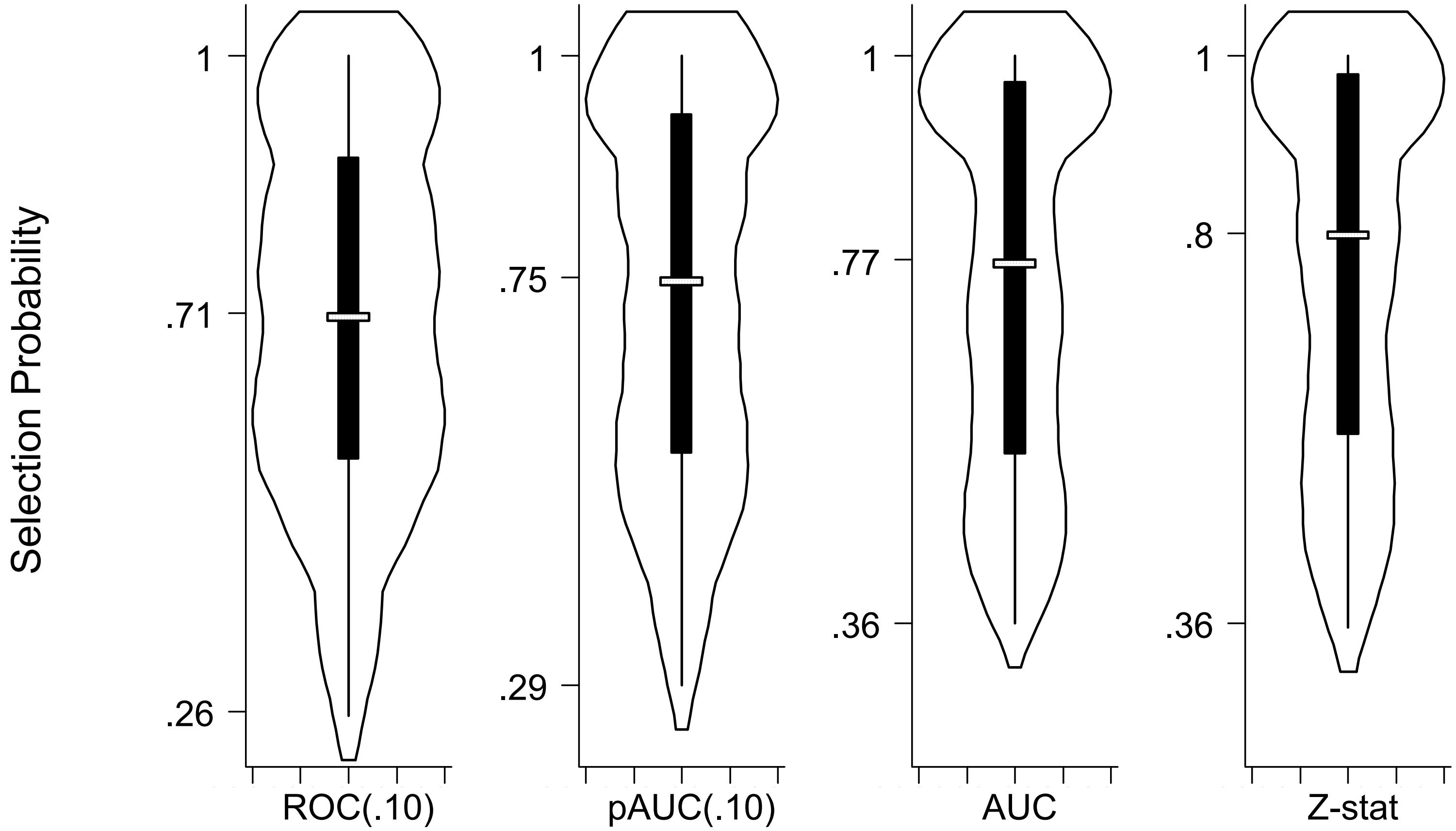
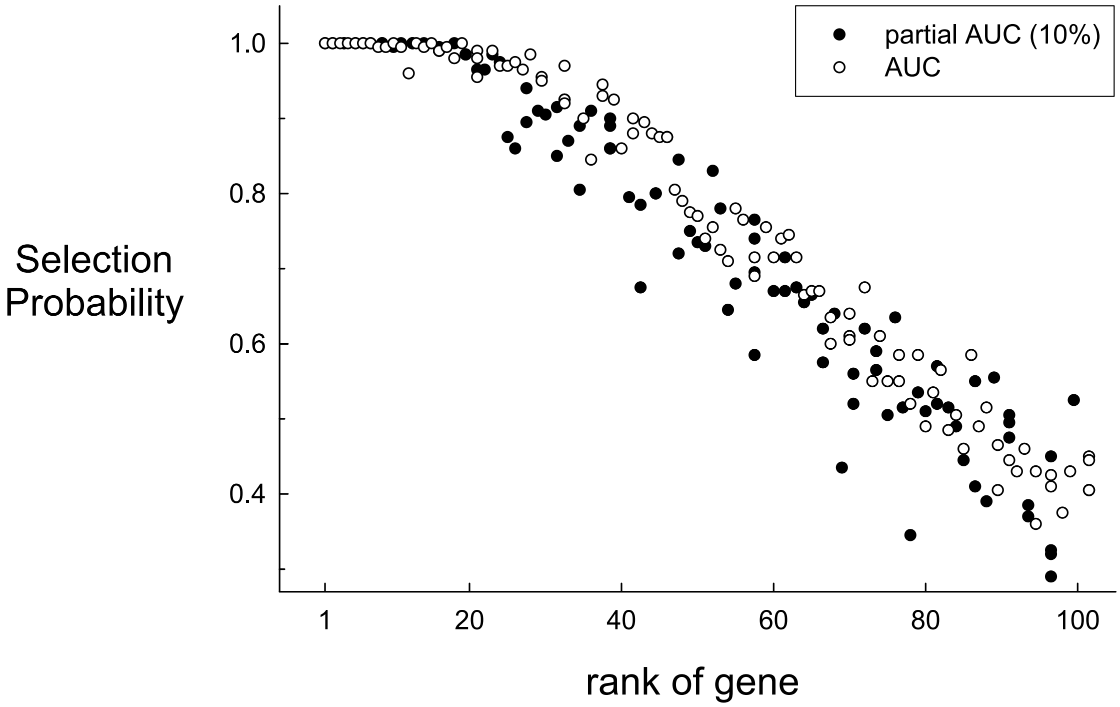

Selection Probability
1
.71
.26
ROC(.10)
1
.75
.29
pAUC(.10)
1
.77
.36
AUC
1
.8
.36
Z-stat

Figure 4a



partial AUC (10%)
AUC
Selection Probability
1.0
0.8
0.6
0.4
1
20
40
60
80
100
rank of gene

Figure 4b

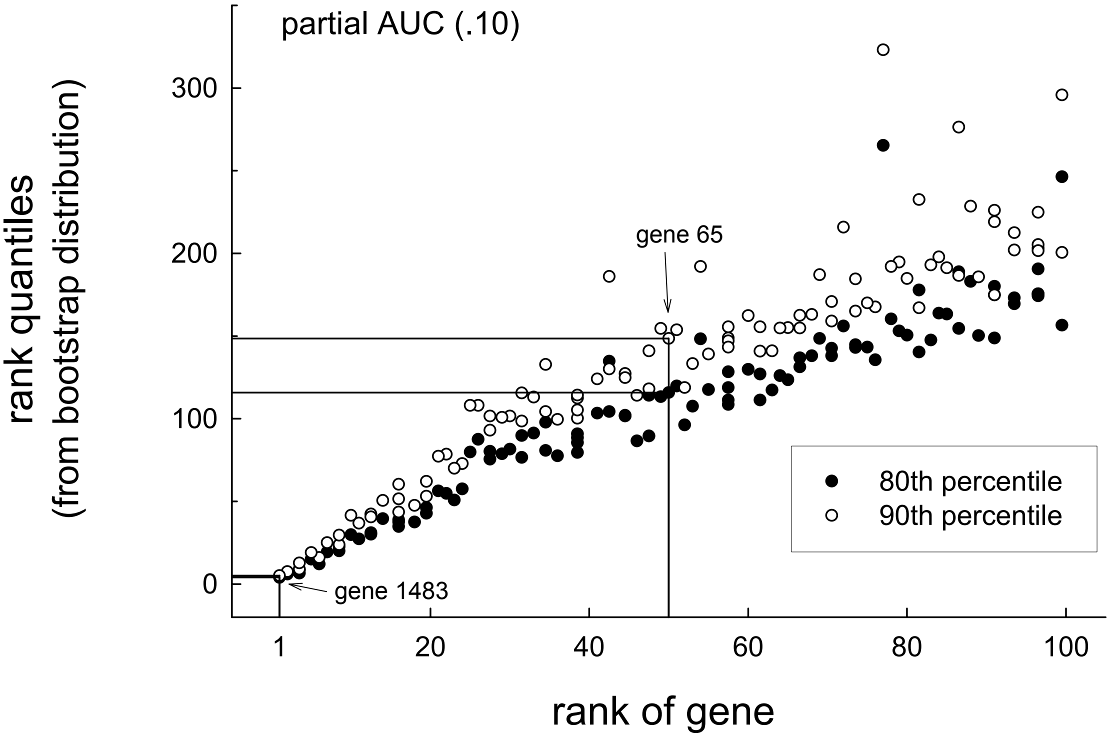

partial AUC (.10)
rank quantiles
(from bootstrap distribution)
300
200
100
0
gene 65
gene 1483
80th percentile
90th percentile
1
20
40
60
80
100
rank of gene

Figure 5