MICROARRAYS AND THEIR USE IN A COMPARATIVE EXPERIMENT

by

Bradley Efron
Robert Tibshirani
Virginia Goss
Gil Chu

Technical Report No. 213
November 2000

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California

MICROARRAYS AND THEIR USE IN A COMPARATIVE EXPERIMENT

by

BRADLEY EFRON
Department of Statistics
Stanford University

ROBERT TIBSHIRANI
Department of Statistics
Stanford University

VIRGINIA GOSS
Department of Biochemistry
Stanford University

GIL CHU
Department of Biochemistry
Stanford University

Technical Report No. 213
November 2000

This research was supported in part by National Institutes of Health
grant 2 R01 CA72028, National Science Foundation grant DMS-9971405,
National Institutes of Health grant 2R01 CA59039, and
National Sicence Foundation grant DMS-0072360

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California

Microarrays and Their Use in a Comparative Experiment

Bradley Efron [*]
Robert Tibshirani, [†]
Virginia Goss [‡]and Gil Chu [§]

October 30, 2000

Abstract

Microarrays enable genetic researchers to measure expression levels for thousands of genes simultaneously. At least that's the idea. In fact the gene expression information arrives in highly variable form, producing great quantities of data and intriguing problems of statistical analysis. This paper describes one such data analysis, involving several thousand genes and two million expression levels. The analysis is mostly at an applied level, but does offer some ideas on a more general theory for microarray data.

[*] Department of Statistics, Stanford University, Stanford CA 94305; brad@stat.stanford.edu

[†] Division of Biostatistics and Department of Statistics, Stanford University, Stanford CA 94305; tibs@stat.stanford.edu

[‡] Department of Biochemistry, Stanford University, Stanford CA 94305; goss@cmgm.stanford.edu

[§] Department of Biochemistry, Stanford University, Stanford CA 94305; chu@cmgm.stanford.edu

1 Introduction

Through the use of DNA microarrays, a new technology, it is now possible to obtain accurate quantitative measurements of the expression of thousands of genes present in a biological sample. DNA microarrays have now been used to monitor changes in gene expression during important biological processes (e.g. cellular replication and the response to changes in the environment), and to study variation in gene expression across collections of related samples (e.g. tumor samples from patients with cancer). A major statistical task is to understand the structure of the data from such studies, which often consist of measurements on thousands of genes in dozens of conditions.

This paper concerns the use of microarrays in a comparative experiment, where it is desired to compare gene expression under Treatment versus Control conditions. We wish to identify which of several thousand candidate genes have had their expression levels changed, either positively or negatively, by the Treatment. Answering this question requires an efficient data reduction strategy since microarrays deliver megabytes of information, and also statistical inference techniques that deal with the difficulties of simultaneous inference on thousands of genes. We discuss both problems here, working in the context of an experiment on radiation sensitivity discussed below.

First here is a little of the biological background. Virtually all living cells contain chromosomes, large pieces of DNA containing hundreds or thousands of genes, each of which specifies the composition and structure of a single protein. Proteins (polymers of amino acids), are the workhorse molecules of the cell, responsible, for example, for cellular structure, producing energy and important biomolecules like DNA and proteins, and for reproducing the cells chromosomes. To a first approximation every cell in an organism has the same set of chromosomes, and thus contains the same repertoire of proteins. However, cells have remarkably distinct properties, such as the differences between human eye cells, hair cells, and liver cells, , which are predominantly the result of differences in the abundance, distribution and state of the cells proteins. One of the seminal discoveries in the early days of molecular biology was that these changes in protein abundance were determined in part by changes in the levels of messenger RNA (mRNA), small and relatively unstable nucleic acid polymers that shuttle information from chromosomes to the cellular machines that synthesize new proteins. Thus, there is a logical connection between the state of a cell and the details of its protein and mRNA composition.

While it remains difficult to measure the abundance of all of a cells proteins, the recently developed DNA microarray makes it possible to quickly and efficiently measure the relative representation of each mRNA species in the total cellular mRNA population, or in more familiar terms to measure gene expression levels.

There are two major kinds of microarrays. In an oligonucleotide array, the kind studied in this paper, there are 20 probe pairs (pm, mm) for each gene. The pm (perfect match) probe is designed to match a small subsequence of the gene about 25 bases long. The mm (mismatch) probe is a control, being the same as pm except that the middle base is flipped to its complement. An experimental sample is hybridized on the microarray, and the RNA expression of the gene is estimated by the difference in signal pm-mm averaged over the 20 probe pairs. There is some belief that subtracting the mismatch numbers may actually be harmful, a question we consider in this paper.

In a spotted cDNA microarray, the other major kind, one base sequence matching all or part of a gene is printed on a glass slide. The experimental sample is labeled with red dye and hybridized on the slide. As a control, a reference sample is labeled with green dye and hybridized on the same slide. Using a fluorescent microscope the log (red/green) intensities of RNA hybridizing at each site is measured. The red/green microarray is featured in much of the recent literature, see Newton et al. (2000), Dudoit et al. (2000), and Lee et al. (2000). Our discussion, like that in Li & Wong (2000) centers in the Affymetrix oligonucleotide microarray, but similar analysis problems arise for both types of array.

From either type of microarray we obtain several thousand expression values, one for each gene. Microarrays in current use measure anywhere from 1,000 to 25,000 genes; larger ones will soon be available. In a typical study, a number of experimental samples are each hybridized to different microarrays, in order to learn about gene expression differences across different conditions. For example (Alizadeh et al. 2000) studied gene expression patterns from tissue sample from a number of lymphoma patients, and related gene expression to patient survival. Clustering methods (Eisen et al. 1998) were the main tool used in that paper, and in a number of other similar studies. Here we will be interested in the more familiar statistical task of comparing Treatment and Control arrays, carried out though an unfamiliar setting.

The particular dataset we focus on comes from a set of 8 oligonucleotide microarrays, in an experiments to study transcriptional responses to ionizing radiation. Some cancer patents have severe (life-threatening) reactions to radiation treatment. It is important to understand the genetic basis of this sensitivity, so that such patients can be identified before the treatment is given The microarrays were labeled

$$(U1A, U1B, I1A, I1B, U2A, U2B, I2A, I2B), \tag{1.1}$$

the labels indicates the following 2 × 2 × 2 experimental design:
RNA was harvested from wild type human lymphoblastoid cell lines, designated "1" and "2", growing in an unirradiated state "U", or in an irradiated state "I", 4 hr after exposure to 5 Gy of radiation. RNA samples were labeled and divided into two identical aliquots for independent hybridizations, "A" and "B". To assess reproducibility in the data, the aliquots of an mRNA sample (for example U1A and U1B) were analyzed with two microarrays from the same manufacturing lot. Each microarray provided expression estimates for 6810 genes.

Although we focus on this specific experiment, the methods we propose are applicable to other kinds of microarrays and other experimental situations, as discussed in Section 6.

Section 2 of the paper describes the data structure of the radiation experiment, and outlines the various steps of data reduction necessary for statistical inference. A very simple model, likely to apply to any comparative experiment, is proposed in Section 3, leading to both Bayesian and frequentist inferences. We focus on empirical Bayes analysis here, but a frequentist method from Goss et al. (2000) is also discussed. Section 4 applies our methods to the radiation experiment. Different data reduction strategies are compared leading to recommendations for the radiation experiment, for example the use of $\log(pm) - .5 \cdot \log(mm)$ in pace of $pm - mm$. Section 5 employs a bootstrap resampling scheme to assess the accuracy of our recommendations. It also compares our inferences to "gold standard" Northern Blot analyses for a subset of 18 genes. Proofs and technical details are deferred to Section 6, which also suggests how our results might be used in other experimental situations.

2 Data Reduction Strategies

Microarray experiments produce enormous amounts of data, more than two million feature numbers in the relatively small experiment we are discussing here. The statistical task is to efficiently reduce this data to simple summaries of the genes' activities. Our main goal in this paper is to provide a method for comparing the statistical efficiency of different data reduction strategies.

Here is a description of the data in the radiation experiment, and the notation we will use to describe it. Expression levels were recorded for 6810 different genes,

$$\textit{genes}: \quad i = 1, 2, \ldots, n = 6810. \tag{2.1}$$

(There were actually 7129 genes, 319 of which had some missing data. For convenience this paper considers only the 6810 genes having complete data. The various analyses were also carried out on all 7129 genes, with nearly identical results.) Each gene on each plate was represented by 20 oligonucleotide "probes",

$$\textit{probes}: \quad \textit{j} = \textit{1}, \textit{2}, \ldots, J = \textit{20} \tag{2.2}$$

Finally there were 8 plates, representing the eight experimental conditions of the $2 \times 2 \times 2$ experiment described in the Introduction, (U1A, U1B, I1A, I1B, U2A, U2B, I2A, I2B),

$$\textit{plates}: \quad k = 1, 2, \ldots, K = 8 \tag{2.3}$$

Two features were recorded for each probe of each gene on each plate, a "perfect match number" pm_{ijk} and a "mismatch number" mm_{ijk}, the latter referring to a deliberately distorted version of the oligonucleotide included as a control. Table 1 shows the 20 pairs of numbers for gene $i = 2715$ on plate $k = 1$.

We will investigate three separate stages of data reduction: probe reduction, the mapping which takes the 20 probe pair numbers into an expression value "M_{ik}" for gene i on plate k,

$$\textit{probe reduction}: \quad \{(pm_{ijk}, mm_{ijk}), j = \textit{1}, \textit{2}, \ldots, \textit{20}\} \rightarrow M_{ik}; \tag{2.4}$$

gene reduction, the mapping that takes the $K = 8$ expression values M_{ik} for gene i into a single expression score "Z_i",

$$\textit{gene reduction}: \quad \{M_{ik}, k = 1, 2, \ldots, K = 8\} \rightarrow Z_i; \tag{2.5}$$

Table 1: *The 20 pairs of perfect match and mismatch feature numbers for gene $i = 2715$ on plate $k = 1$ (U1A).*

probe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
pm	1054	3242	1470	4050	1356	1476	561	606	1307	1057	974	1584	802	1399	1670	2514	2096	6592	5662	2244
mm	793	2333	826	1912	561	558	942	526	699	1060	829	1771	601	569	840	950	700	8717	1484	668

and finally an *inference mapping* that re-expresses Z_i in terms of a statistical inference concerning gene i's activity. Two inference mappings will be discussed, a Bayesian version of the form $\text{Prob}\{\text{Event}_i|Z_i\}$, where Event_i is an event of interest such as "gene i's activity was affected by radiation", and a frequentist version that assigns a significance to Event_i.

Here is an example of our results, using a form of data reduction that will be shown to be quite efficient. For the probe reduction let

$$M_{ik} = \text{mean}\{\log(pm_{ijk}) - .5 \cdot \log(mm_{ijk}), j = 1, 2, \ldots, 20\}. \tag{2.6}$$

For the gene reduction, first compute the 4 differences $(D_{i1}, D_{i2}, D_{i3}, D_{i4})$ between the irradiated and unirradiated values within the same wildtype sample and aliquot, e.g.

$$D_{i1} = M_{i3} - M_{i1}, \tag{2.7}$$

the difference between the I1A and U1A values M_{ik}. Then take

$$Z_i = \bar{D}_i/(a_0 + S_i) \tag{2.8}$$

where $\bar{D}_i$ is the average of the 4 differences, S_i is their sample standard deviation, and a_0 is the 90th percentile of the 6810 S values. Specifications (2.6)-2.8 will be used in all of our numerical examples unless stated otherwise.

Figure 1 displays the Bayesian inference mapping $\text{Prob}\{\text{Event}|Z\}$ based on the probe and gene reductions (2.6), (2.8). The methodology behind Figure 2, and the reasons for choosing (2.6)-(2.8), are explained in Section 4. The plotted Z values are actually a monotonic transformation of (2.8) that makes the empirical distribution of the 6810 Z's nearly a standard normal, $N(0,1)$. [The algorithm producing Figure 1 includes two additional steps between (2.6) and (2.8), designed to remove systematic differences between the eight plates, and to better condition the inference mapping. Remark A (Section 6) describes these two steps, which were included in all the results of this paper.]

Eighteen of the 6810 genes were independently assessed by a Northern Blot analysis. Seven of these, indicated by "+" in Figure 1, were deemed "affected positively by radiation", five indicated by "-" were "affected negatively", and six indicated by "o" were "not affected". (Full results are given in Section 5.) There is reasonably good agreement between the Northern Blot assessments and the probabilities assigned in Figure 1. In particular, gene #2715 is, appropriately, assigned probability .998 of being affected.

3 A Simple Model for Inference

Besides analyzing the radiation data, our goal here is to provide data analysis techniques that could be useful in a variety of microarray situations. With generality in mind we will avoid highly specified models, relying instead on a simple inference model that is likely to apply to most comparative experiments: that a gene is either affected on unaffected by the treatment of interest, radiation in our case, giving two possible distributions for the expression score "Z", (2.5). Lee et al. (2000) use a normal theory version of this idea, as, less directly, Li & Wong (2000), Newton et al. (2000) use a Gamma model. Here we will avoid such parametric assumptions.

Let

$$\begin{aligned} p_1 &= \text{probability that a gene is affected} \\ p_0 &= 1 - p_1 = \text{probability unaffected,} \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} f_1(z) &= \text{the density of } Z \text{ for affected genes} \\ f_0(z) &= \text{the density of } Z \text{ for unaffected genes.} \end{aligned} \tag{3.2}$$

Then

$$f(z) = p_0 f_0(z) + p_1 f_2(z) \tag{3.3}$$

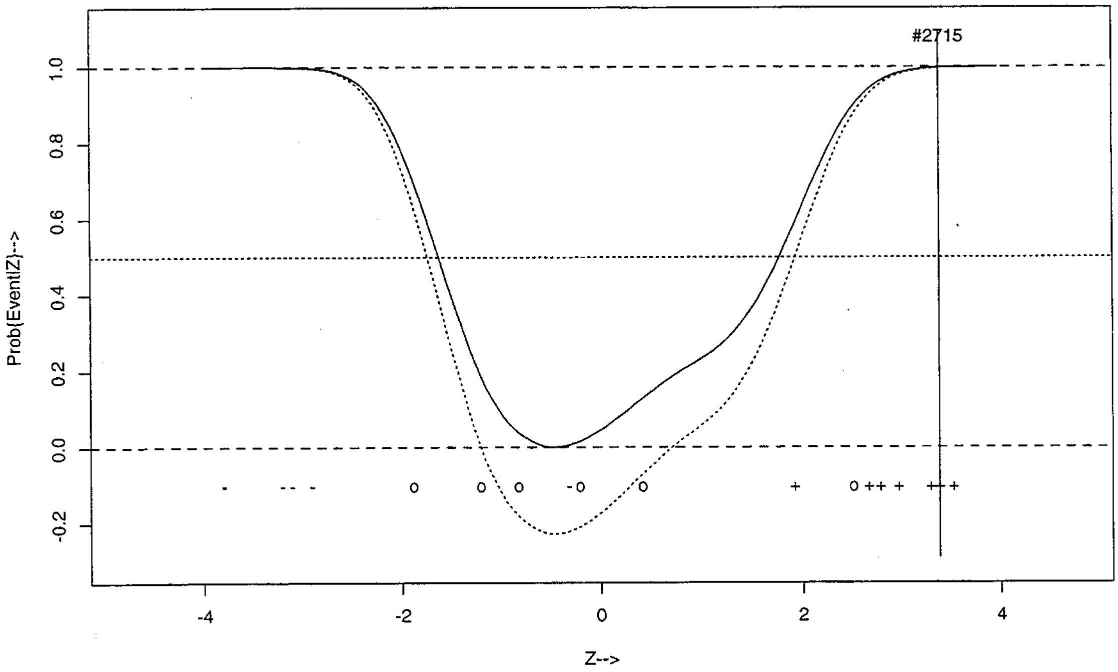

Figure 1: *Solid curve: Bayesian inference mapping* $\text{Prob}\{\text{Event}_i|Z_i\}$ *from data reductions (2.6), (2.8);* Event_i *is "gene i affected by radiation". Symbols show Z values for 18 genes separately analyzed by Northern Blot: "+" positively affected, "-" negatively affected", "o" not affected.*

is the mixture density of the two populations. In our situation we can estimate $f(z)$ directly from the 6810 expression scores Z_i obtained from the data reduction (2.4), (2.5).

In the absence of strong parametric assumptions such as normality, model (3.3) is useless without an estimate of the "null density" $f_0(z)$. Fortunately it is easy to obtain such estimates. What follows is the method we used to estimate $f_0(z)$ in the radiation experiment. Remark C of Section 6 discusses variants of this method applicable more generally.

The 6810×8 matrix $\mathbf{M}$ of expression values (2.4), one value for each gene on each plate, gives a 6810×4 matrix $\mathbf{D}$ of differences between the irradiated and unirradiated expression values, as in (2.7). Let $\mathbf{M}_k$ indicate the kth column of $\mathbf{M}$, a 6810 vector. With the plates ordered as before, (U1A, U1B, I1A, I1B, U2A, U2B, I2A, I2B), the "difference matrix" $\mathbf{D}$ is

$$\mathbf{D} = (\mathbf{M}_3 - \mathbf{M}_1, \mathbf{M}_4 - \mathbf{M}_2, \mathbf{M}_7 - \mathbf{M}_5, \mathbf{M}_8 - \mathbf{M}_6) \tag{3.4}$$

Symbolically, the vector $\mathbf{Z}$ of expression scores (2.5) is obtained via

$$\begin{array}{ccccccc} \{\text{original data}\} & \rightarrow & \mathbf{M} & \rightarrow & \mathbf{D} & \rightarrow & \mathbf{Z}. \\ 6810 \times 20 \times 2 \times 8 & & 6810 \times 8 & & 6810 \times 4 & & 6810 \end{array} \tag{3.5}$$

Now let the "null difference matrix" $\mathbf{d}$ be the 6810×4 matrix obtained by differencing within the aliquot splits,

$$\mathbf{d} = (\mathbf{M}_2 - \mathbf{M}_1, \mathbf{M}_4 - \mathbf{M}_3, \mathbf{M}_6 - \mathbf{M}_5, \mathbf{M}_8 - \mathbf{M}_7), \tag{3.6}$$

so for example the first column of $\mathbf{d}$ records differences between the B and A splits of the unirradiated wildtype 1 experiments. We define "null scores" $\mathbf{z} = (z_1, z_2, \ldots, z_{6810})'$ by

$$\{\text{original data}\} \rightarrow \mathbf{M} \rightarrow \mathbf{d} \rightarrow \mathbf{z}, \tag{3.7}$$

with the understanding that except for the substitution of $\mathbf{d}$ for $\mathbf{D}$, the arrows in (3.7) indicate the same mappings as in (3.5).

We will use the empirical distribution of the null scores $\{z_i\}$ to estimate the null density $f_0(z)$ in (3.3). One could just as well take $\mathbf{M}_1 - \mathbf{M}_2$ as $\mathbf{M}_2 - \mathbf{M}_1$ in (3.6), etc., and in fact our numerical algorithm employs random sign permutations of the columns of $\mathbf{d}$ to improve the estimation of f_0. Remark C discusses strategies that might be used for f_0's estimation in other situations.

3.1 Bayesian inference

An application of Bayes' rule to the mixture model (3.3) gives the aposteriori probability $p_1(Z)$ that a gene with score Z was affected by the treatment,

$$p_1(Z) = 1 - p_0 f_0(Z)/f(Z). \tag{3.8}$$

The ratio $f_0(Z)/f(Z)$ can be estimated directly from the $\{Z_i\}$ and $\{z_i\}$ empirical distributions. The probabilities p_0 and $p_1 = 1 - p_0$, are unidentifiable without strong parametric assumptions, but this will turn out to be less problematic than it might seem. The constraint that $p_1(Z)$ be nonnegative for all Z does restrict p_0 and p_1,

$$p_1 \geq 1 - \min_Z \{f(Z)/f_0(Z)\}. \tag{3.9}$$

We can now describe the construction of the curve Prob{Event$|Z$} in Figure 1, skipping the technical details which appear in Remark B:

(a) The 6810 scores $\{Z_i\}$ were constructed according to (3.5), using probe reduction (2.6) and gene reduction (2.8).

(b) The null scores $\{z_i\}$ were constructed in the same way, but beginning with (3.7) rather than (3.5). (Actually 20 versions of the $\{z_i\}$ were generated, based on 20 independent row-wise sign permutations of **d**.)

(c) A logistic regression technique was used to estimate the ratio $f_0(z)/f(z)$ based on the relative densities of the $\{Z_i\}$ to the $\{z_i\}$.

(d) Relationship (3.9) gave an estimated lower bound for p_1, $p_1 \geq .189$.

(e) The solid curve Prob{Event$|Z$} is (3.8), with f_0/f estimated from the logistic regression, and p_0 equaling its estimated maximum value $.811 = 1 - .189$.

In comparing different data reductions, the main goal of this paper, it is convenient to always have the same marginal distribution for Z. To this end, the raw scores $\{Z_i\}$ from (2.8) were monotonically transformed to have a nearly perfect $N(0,1)$ distribution, say by transformation $m(Z)$, and then the null scores were transformed according to the same $m(z)$. (Notice that the crucial ratio $f_0(z)/f(z)$ remains the same under such transformations.)

We will always make the empirical distribution of the $\{Z_i\}$ almost perfectly $N(0,1)$, using a normal scores transformation, implying for example that $42 = 6810 \cdot (1 - \Phi(2.5))$ of the 6810 genes have $Z_i > 2.5$, with Φ the standard normal cumulative distribution function.

Figure 2 shows the estimates of f_0, f_1, and f contributing to Figure 1; $f(Z)$ is a standard $N(0,1)$ density, by construction, while $f_0(z)$ is a less dispersed density. This is what we hoped for of course: the Z's should be more dispersed than the z's since they reflect the disturbing effects of the radiation treatment. The large values of Prob$\{\text{Event}|Z\}$ in the tails of Figure 1 come from (3.8), and the small ratio of $f_0(z)$ to $f(Z)$. A good choice of data reductions makes $f_0(z)/f(z)$ small for $|z|$ large, and we will use this criteria to guide our choices of the probe and gene reductions in what follows.

Looking again at (3.8),

$$p_1(Z) \geq 1 - f_0(Z)/f(Z), \tag{3.10}$$

since this corresponds to $p_0 = 1$, the largest possible value. The dotted curve in Figure 1 is $1 - f_0(Z)/f(Z)$. This is not much less than the solid curve for large values of $|Z|$, giving for example Prob$\{\text{Event}|Z\} \geq .997$ for gene #2715.

Our Bayesian analysis is actually "empirical Bayes" in the sense that the crucial ratio $f_0(Z)/f(Z)$ in (3.8) is estimated from the data rather than from apriori assumptions. Newton et al. (2000) carry out a similar analysis, but using specific Bayesian modeling assumptions beyond (3.1) - (3.3).

3.2 Frequentist Inference

The simple model (3.1) - (3.3) can also be analyzed from a frequentist hypothesis-testing point of view. Notice that we can express the likelihood ratio $f_1(Z)/f_0(Z)$ as

$$\frac{f_1(Z)}{f_0(Z)} = \frac{1}{p_1}\left[\frac{f(Z)}{f_0(Z)} - p_0\right], \tag{3.11}$$

which is an increasing function of $f(Z)/f_0(Z)$. Therefore the optimal Neyman-Pearson test comparing the null hypothesis density f_0 with the alternative density f_1 rejects the null hypothesis for $f(Z)/f_1(Z) \geq c_0$, the constant c_0 being chosen to give the desired test size.

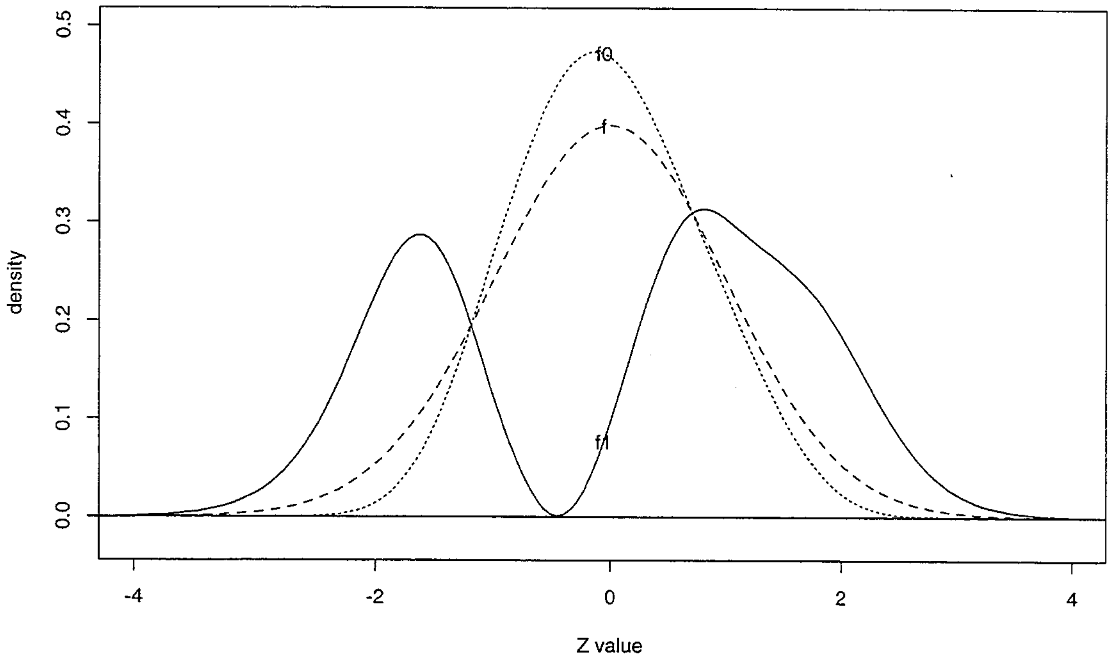

Figure 2: *Estimates of* $f(\cdot)$, $f_0(\cdot)$ *and* $f_1(\cdot)$ *for the situation of Figure 1, model (3.3);* $p_1 = .189$, *its minimum possible value.*

For a given choice of c_0 define the sets $A_0 = \{Z : f(Z)/f_0(Z) \le c_0\}$ and $A_1 = \{Z : f(Z)/f_0(Z) > c_0\}$, and the errors of the first and second kinds

$$\alpha_0 = \text{Prob}_{f_0}\{A_1\} \quad \text{and} \quad \beta_0 = \text{Prob}_{f_1}\{A_0\}. \tag{3.12}$$

Remark F verifies the following Lemma.
Lemma Defining

$$b_0 = \text{Prob}_f\{A_0\} = \int_{A_0} f(Z)dZ, \tag{3.13}$$

we have

$$(1-\alpha_0) - \beta_0 = \frac{(1-\alpha_0) - b_0}{p_1}. \tag{3.14}$$

Figure 3 illustrates relationship (3.14).

The gist of the Lemma is that the Neyman-Pearson (α, β) figure for testing f_1 vs f_0 has a simple relationship to the corresponding figure for testing f vs f_0. It is the latter's ratio $f(Z)/f_0(Z)$ that we can easily estimate from the data, as in step (c) of the Bayesian algorithm, leading to an "empirical Neyman-Pearson" figure for f_1 versus f_0 via (3.14). Notice that p_1 in (3.14) cannot be too small since $(1-\alpha_0) - \beta_0$ must be nonnegative. It turns out that the lower bound (3.9) on p_1 also expresses this constraint, see Remark F.

With 6810 genes under investigation, multiple testing difficulties make traditional p-values hard to interpret. Dudoit et al. (2000) discuss some of the problems. In a similar spirit Goss et al. (2000) developed a frequentist methodology "SAM" (Significance Analysis of Microarrays) specifically aimed at the multiple testing situation. SAM begins with scores $\{Z_i\}$ and null scores $\{z_i\}$ as in steps (a) and (b) of the Bayes algorithm following (3.9), and then proceeds as follows:

(c) The ordered values of the scores, $Z_{(1)} \le Z_{(2)} \ldots \le Z_{(6810)}$, are computed, and similarly for each sign-permutation version of the null scores, say

$$z_{(1)}(b) \le z_{(2)}(b) \ldots \le z_{6810}(b) \qquad b = 1, 2, \ldots, B, \tag{3.15}$$

where $\mathbf{z}(b)$ is the bth of B permutations ($B = 50$ in the following example.).

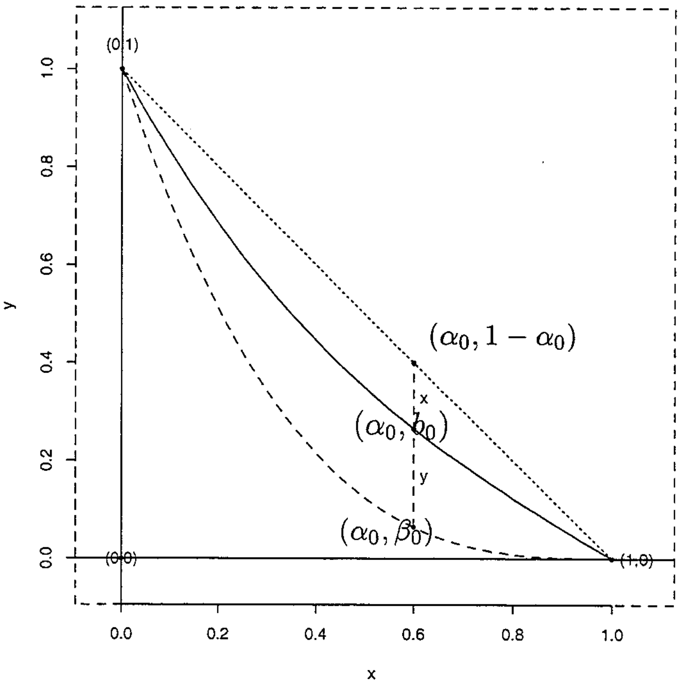

Figure 3: *Lemma (3.13) - (3.14) gives the Neyman-Pearson figure for testing* f_1 *versus* f_0 *(dashed curve) as a scaled-up version of the corresponding figure for testing* f *vs* f_0 *(solid curve.) The ratio of lengths* $(y+x)/x$ *equals* $1/p_1$.

Table 2: *Frequentist analysis of situation discussed in Figure 1, SAM methodology, as illustrated in Figure 4. False Detection Rate (FDR) becomes very small for threshold values exceeding 0.3.*

threshold	Mean # false pos	# true positive	False Detection Rate
0.1	804.78	3856	0.209
0.2	71.04	1038	0.0684
0.3	0.26	421	0.00062
0.4	0.12	204	0.00059
0.5	0.00	102	0.00000

(d) "Expected null scores"

$$\bar{z}_{(i)} = \sum_{b=1}^{B} z_{(i)}(b)/B \tag{3.16}$$

are computed, and the points $(\bar{z}_{(i)}, Z_{(i)})$ plotted for $i = 1, 2, \ldots 6810$. Figure 4 shows the SAM plot when the Z's and z's are computed as in Figure 1.

(e) Two "exceedance numbers" are calculated for each of several possible threshold values t: the *true positive* number

$$\text{True } (t) = \#\{i : |Z_{(i)} - \bar{z}_{(i)}| > t\}, \quad \bar{z}_{(i)} = \sum_{b=1}^{B} z_{(i)}(b)/B \tag{3.17}$$

and the average *false positive* number

$$\text{False } (t) = \frac{1}{B} \sum_{b} \#\{i : |z_{(i)}(b) - \bar{z}_{(i,b)}| > t\} \tag{3.18}$$

where $\bar{z}_{(i,b)} = \Sigma_{c \neq b} z_i(c)/(B-1)$, the average *not* including $\mathbf{z}(b)$'s contribution. (Goss et al. 2000) actually use a more conservative method of counting exceedences, but the two methods agree here.)

(f) Finally the *False Detection Rate*, "FDR",

$$\text{FDR } (t) = \text{Fake } (t)/\text{True } (t) \tag{3.19}$$

is computed for increasing values of t. (See Remark H for more details). Table 2 gives FDR's for Figure 1's situation.

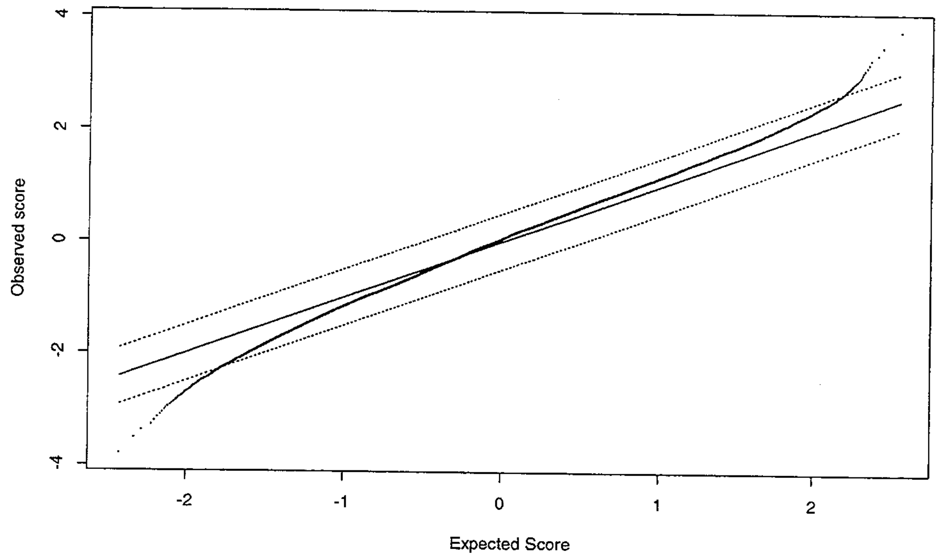

Figure 4: *Observed scores $Z_{(i)}$ versus the expected null scores $\bar{z}_{(i)}$, (3.16), for the probe and gene reductions (2.6), (2.8) used in Figure 1. Dashed lines are ±.5 units above and below the main diagonal. Table 2 shows that 102 of the $Z_{(i)}$ values fall outside the dashed lines, while none of the corresponding null values (3.18) do.*

The implication of Table 2 is that genes corresponding to $Z_{(i)}$ values beyond some large threshold value are likely to be genuinely affected. Taking $t = 0.5$ as in Figure 4, these are genes with

$$Z_{(i)} < -1.78 \quad \text{of} \quad Z_{(i)} > 2.16. \tag{3.20}$$

Looking back at Figure 1, 14 of the 18 Northern Blot genes are correctly identified by this rule, three of the four errors being close to the boundaries (3.20). The posterior probabilities Prob{Event$|Z$} corresponding to (3.20) are 0.62 and 0.75 respectively.

In our situation there is a broad agreement between the empirical Bayes methodology, SAM, and Neyman-Pearson arguments: all three suggest that scores $Z_{(i)}$ that are extreme compared to the distribution of the null scores indicate genuine gene activity. Both Figure 1 and Table 2 reflect the fact that $f_0(Z)/f(Z)$, (3.3) goes to zero as $|Z|$ gets big.

We have thresholded the empirical Bayes probabilities (at say .90) in order to obtain a list of "significant" genes. However interpretation of this probability must also be face the multiple comparison problem. To investigate this, we repeated the calculation of Figure 1 for 16 datasets, having differences $(D^*_{i2}, D^*_{i2}, D^*_{i3}, D^*_{i4})$. Each $\mathbf{D}^*_i = \pm\mathbf{d}_i$ as in (3.6) and hence is "null data". Table 3 shows the number of significant genes at each probability level, for both the actual and null data.

Table 3: *Empirical Bayes model: number of significant genes at each probability level, for both the actual data, and average for null data*

Probability threshold	Mean # false	#Actual
0.900	0.38	79
0.950	0.13	49
0.975	0.06	34
0.990	0.00	25

For example with a probability threshold of .95, the average number of false positives is only .13.

With appropriate thresholds, will the SAM method and empirical Bayes model yield the same significant genes? In most cases, both measures are monotone functions of $Z > 0$ as Z increases, and similarly for $Z < 0$. However neither are symmetric in $|Z|$, and they shouldn't be, there is no apriori reason that the data should contain both overexpressed and underexpressed genes.

Hence we need to know more about the form of the threshold rules. For a large class of exponential family models, the log-likelihood ratio is linear in z:

$$\log\left[f_1(z)/f_0(z)\right] = \beta_0 + \beta_1 z \tag{3.21}$$

This holds for example if $f_0(z)$ is the density of $N(0, \sigma^2)$ and $f_1(z)$ the density of $N(\mu, \sigma^2)$. For a set of 1000 observations $Z_i \sim N(\mu, 1), z_i \sim N(0, 1)$, Figure 5 shows a plot of $E[Z_{(i)} - z_{(i)}]$ versus $E[Z_{(i)}]$ for $Z_{(i)} > 0$ and for different values

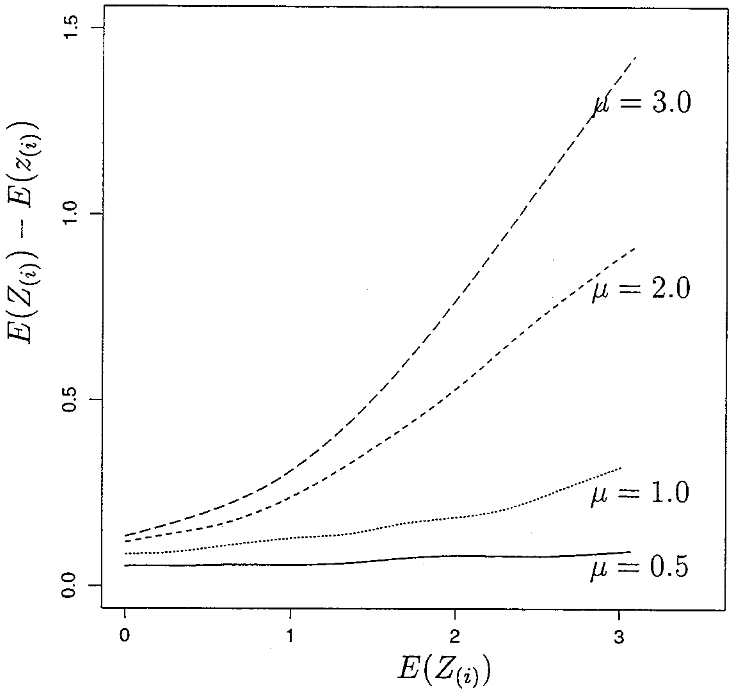

Figure 5: *Gaussian model: f_0 equals the density of $N(0, \sigma^2)$, $f_1(z)$ equals the density of $N(\mu, \sigma^2)$. Shown are the expected $E[Z_{(i)}) - z_{(i)}]$ versus $E[Z_{(i)}]$, for $Z_{(i)} > 0$ and for different values of μ*

of μ. [These quantities were computed as averages over 20 simulations]. We see that this difference is also approximately linear, at least for $Z > 1$ with the slope a function of μ. Hence the two methods will yield approximately the same genes in this situation.

SAM, and other frequentist hypothesis testing methods, are best suited to finding a small number of affected genes amidst a great majority of unaffecteds. They are at a disadvantage in situations like the radiation experiment where perhaps thousands of the genes have been affected by the treatment. This better suits the empirical Bayes methodology, which will be emphasized in our applications here.

4 Efficient Estimation of Gene Effects

One can imagine a great variety of probe and gene reductions (2.4)-(2.5) that might be superior to the choices (2.6)-(2.8) used in Figures 1 and 4. This Section examines a range of possibilities, comparing them in terms of the inferences they produce for the radiation experiment.

4.1 Gene Reduction

We begin by examining variations of gene reduction (2.8) keeping the probe reduction (2.6) fixed. Figure 5 compares different choices of "a_0"; a_0 equal to the 90th percentile of the 6810 S_i values, the 50th percentile, the 5th percentile, $a_0 = 0$, and $a_0 \to \infty$. The choice $a_0 = 0$ makes Z_i in (2.8) proportional to the one-sample t-statistic for the four differences $(D_{i1}, D_{i2}, D_{i3}, D_{i4})$, while $a_0 \to \infty$ makes Z_i equivalent to the numerator $\bar{D}_i$. The vertical axis in Figure 5 is the logit $\log\{p/(1-p)\}$ of $p = \text{Prob}\{\text{Event}|Z\}$, instead of the probability itself as in Figure 1. (The actual probability level is indicated at the right. Using logits emphasizes differences for large $|Z|$ values, the ones that identify legitimate gene effects.) The event probabilities in Figure 5, and elsewhere in what follows, are based on formula (3.8) with $p_0 = 1$, and so are actually the lower bound (3.10).

Figure 5 shows that the best choice of a_0 is the one we used before, a_0 the 90th percentile. This manifests itself as higher values of $\text{Prob}\{\text{Event}|Z\}$ at both ends of the Z scale. The density $f_0(z)$ in Figure 2 is more concentrated around zero than it is say for the disastrous choice $a_0 = 0$, lowering $f_0(z)/f(z)$ in the tails and raising $p_1(z)$, (3.10). Figure 3, the Neyman-Pearson curve, will also be more favorable, giving smaller values of β_0 at any fixed value of α_0.

The difference between $a_0 = .90$ and $a_0 = .50$ looks small in Figure 6, and it is reasonable to ask if it is statistically significant. We tested this by means of a bootstrap analysis. Bootstrap data sets were formed by the "row resampling" method described in Section 5. Then Figure 6 was recomputed using the bootstrap data, and the differences between the various curves recalculated. All of this was done 25 times, providing bootstrap standard errors for the various differences.

The differences shown in Figure 6 proved genuine. For example, measured at $Z = 3$ the "$a_0 = .90$ minus $a_0 = .50$" difference had point estimate and bootstrap standard error 0.68 ± 0.13; the corresponding statistics at $Z = 3$

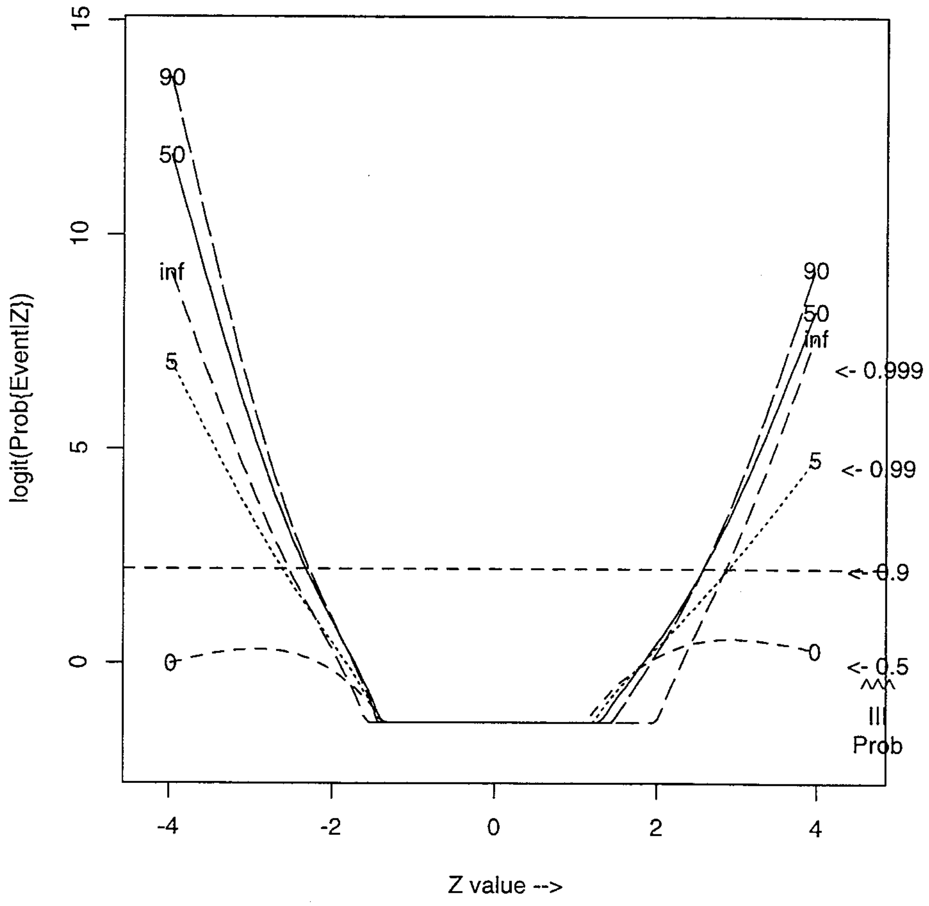

Figure 6: *Choice of* a_0 *in the gene mapping* $Z_i = \bar{D}_i/(a_0 + S_i)$, *(2.8); vertical axis is logit of* $\mathrm{Prob}\{\mathrm{Event}|Z\}$, *estimated as in (3.8) with* $p_0 = 1$; *"90" indicates* a_0 *equaling 90th percentile of the 6810* S_i *values, etc.; "inf" is limit as* $a_0 \to \infty$. *We see that 90 is the best choice in terms of maximizing* $\mathrm{Prob}\{\mathrm{Event}|Z\}$ *for large* $|Z|$; $a_0 = 0$ *is worst. All choices used probe reduction (2.6). The vertical axis is truncated at lower bound* $\mathrm{Prob}\{\mathrm{Event}|Z\} = .20$.

were 0.224 ± 0.105.

4.2 Probe Reduction

Next we considered variations on probe reduction (2.6), now keeping the gene reduction fixed as in (2.8) with percentile $a_0 = .90$.

Figure 7 compares probe reductions of the form

$$M_{ik} = \underset{j}{\text{mean}}\{s(pm_{ijk}) - c \cdot s(mm_{ijk})\}, \tag{4.1}$$

with s either the log function or the identity function. For example curve 2 in the left panel used $M_{ik} = \underset{j}{\text{mean}}\{pm_{ijk} - mm_{ijk}\}$ while the dotted curve in the right panel used $M_{ik} = \underset{j}{\text{mean}}\{\log(pm_{ijk})\}$. Our preferred choice (2.6-2.8) is curve 1, "$c = .5$ & logs". The "Affy" curve in the left panel was based on the algorithm provided by Affymetrix, which is similar to the "$c = 1$ no logs" choice, but with a provision for removing apparent outliers among the 20 $pm_{ijk} - mm_{ijk}$ differences before averaging. As a simple comparison statistic the number

$$N90 = \#\{\text{genes}: \ \text{Prob}\{\text{Event}|Z\} \geq .90\} \tag{4.2}$$

for each probe reduction is shown below the panels.

Figure 6 indicates a substantial advantage to taking logs, and a mild advantage to using $c = .5$ rather than $c = 1$ or $c = 0$. The comparison between $c = .5$ and $c = 1$ is close on the log scale, but other comparisons, one of which we will see in the next subsection, reinforce the superiority of $c = .5$.

We also tried using trimmed means instead of ordinary means in (4.1). When applied on the log scale this form of robustification made almost no difference to our results. However trimming might be more useful in other experimental situations, and Remark G discusses a data-based method for choosing the best trimming proportion.

4.3 Separate Analysis of the Two Wildtypes

The radiation experiment comprised two wildtype samples, wildtype1 corresponding to the plates (U1A, U1B, I1A, I1B) and wildtype2 corresponding

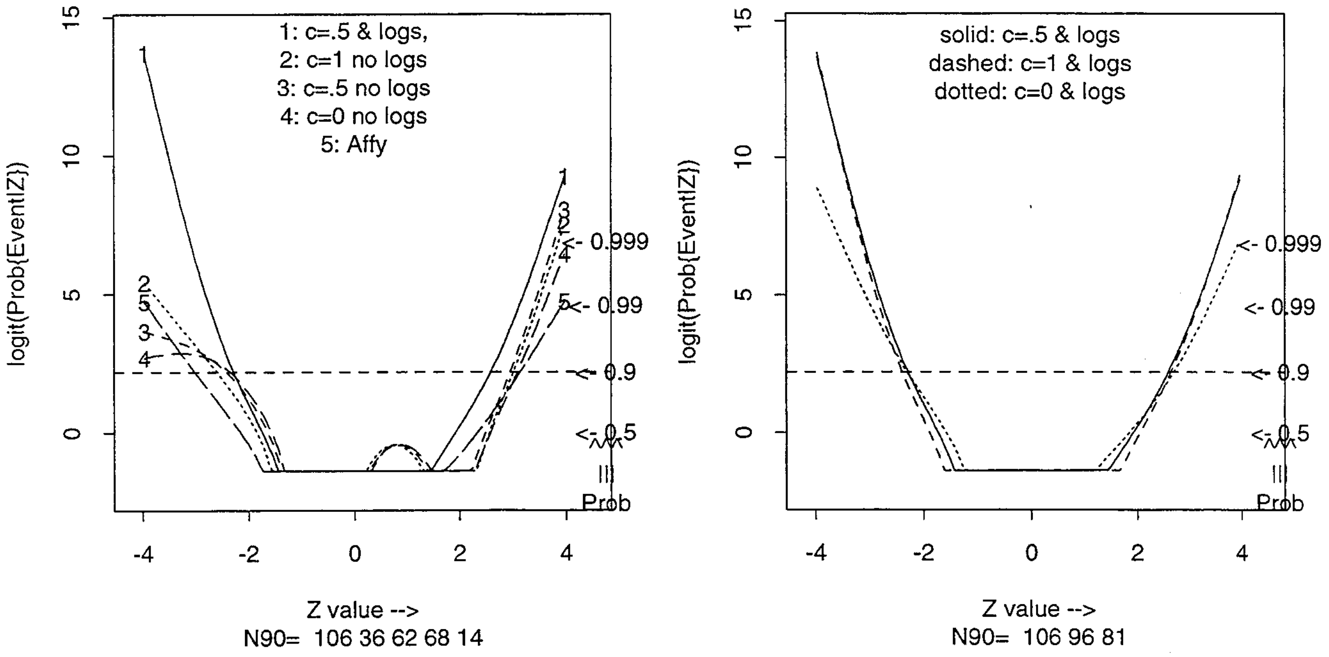

Figure 7: *Comparison of various probe reductions (gene reduction fixed as in (2.8), a_0 = 90th percentile). The solid curve in both panels is the choice (2.6) used previously; constant "c" is multiple of mm level subtracted from pm level, eg "c = 1 no logs" uses* $M_{ik} = \underset{j}{\text{mean}}\{pm_{ijk} - mm_{ijk}\}$. *"Affy" based on the probe reduction software provided with the Affymetric Genechip.* **N90** *indicates number of genes having* $\text{Prob}\{\text{Event}|Z\} \geq .90$.

to (U2A, U2B, I2A, I2B). The left panel of Figure 8 shows the analysis based on $c = .5$ & logs applied to the separated wildtype data. For example curve 3, "wildtype1 only", worked with the first four columns of the matrix M from (2.6), computing the Z sums as in (2.8), $a_0 = .90$. (The matrices $\mathbf{D}$ and $\mathbf{d}$ in (3.5) and (3.7) now have just two columns each.)

The results are rather shocking. Basing the analysis on only the wildtype2 data gives *better* results than using all of the data. This is made more believable when we see that wildtype1 by itself is almost useless.

The right panel of Figure 8 repeats the comparison made on the right of Figure 7, but using only the wildtype2 data. Now the choice $c = .5$ is clearly superior to $c = 1$. Our other comparisons, as in Figures 6 and 7, were rerun using only the wildtype2 data, giving similar results.

The trouble with the wildtype1 data seemed to be concentrated in its "A" aliquot. Table 3 displays the correlation matrix of the four differences (3.4) obtained from (2.6). $\mathbf{D}_1$, the difference I1A - U1A for the A aliquot of wildtype1, is anomalous, being negatively correlated with $\mathbf{D}_2$, $\mathbf{D}_3$, and $\mathbf{D}_4$.

	$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$	$\mathbf{D}_4$
$\mathbf{D}_1$ = I1A - U1A	1	-0.22	-0.06	-0.11
$\mathbf{D}_2$ = I1B - U1B		1	0.50	0.52
$\mathbf{D}_3$ = I2A - U2A			1	0.64
$\mathbf{D}_4$ = I2B - U2B				1

Table 3 Correlation matrix of the four differences (3.4); $\mathbf{D}_1$, the difference I1A - U1A for the A aliquot of wildtype1, is not correlated with the other three.

5 Accuracy and Regression to the Mean

The Bayesian inference mapping represented by the solid curve in Figure 1 assigns posterior probability

$$\text{Prob}\{\text{Event}_i | Z_i\} \equiv p_1(Z_i) \tag{5.1}$$

to the event "gene i affected by the radiation". Because this is an *empirical* Bayes analysis, $p_1(Z_i)$ is only an estimate. This Section discusses its accuracy in the usual sense of bias and variability.

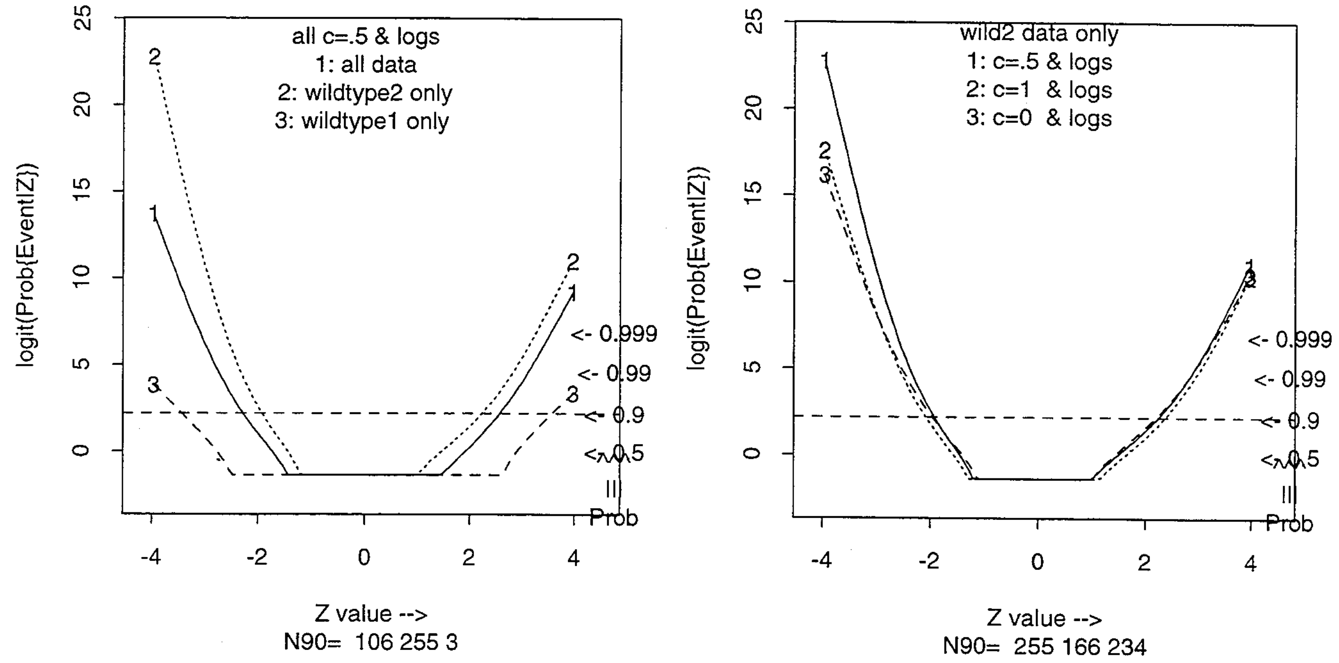

Figure 8: *Left Panel: Separate analyses of data from the two wildtype samples. Using just the wildtype2 data (curve 2) gives better results than using all data (curve 1); wildtype1 data (curve 3) is almost useless. Right panel: Comparison as in right panel of Figure 7 , using only wildtype2 data; now choice* $c = .5$ *is markedly superior to* $c = 1$.

There are two sources of error in $p_1(Z_i)$. First of all the solid curve $p_1(Z)$ in Figure 1 is obtained using estimates of p_0, $f(\cdot)$ and $f_0(\cdot)$ in (3.8). Secondly, Z_i itself is an estimate. The raw score from (2.8), say $Y_i = \bar{D}_i/(a_0 + S_i)$, was transformed by the normal scores mapping to the Z_i's actually used in Figure 1,

$$Z_i = \Phi^{-1}\left(\frac{\text{rank}_i - .5}{n}\right), \tag{5.2}$$

where rank_i is the rank of Y_i among the $n = 6810$ Y scores, and Φ is the standard normal c.d.f. Repeating the experiment would probably produce different Y_i's and Z_i's.

We begin with a simple normal Bayesian model for the accuracy of Z_i. Suppose that each gene has a "true score" μ_i, with the μ's having prior distribution.

$$\mu_i \sim N(0, A), \tag{5.3}$$

and that conditional on μ_i, Z_i has normal distribution

$$Z_i|\mu_i \sim N(\mu_i, V), \tag{5.4}$$

the variances A and V being given. Let

$$\zeta_i = \mu_i/\sqrt{A} \tag{5.5}$$

so that $\Phi(\zeta_i)$ is the percentile rank of μ_i among all the true scores, as in (5.2).

If ζ_i were known we could use it in place of Z_i to get a more accurate estimate $p_1(\zeta_i)$ for the probability of Event_i. Familiar Bayesian calculations yield the distribution ζ_i given Z_i,

$$\textit{Lemma} \quad \zeta_i|Z_i \sim N\left(C\frac{Z_i}{\sqrt{A+V}}, 1 - C^2\right), \; C = \sqrt{\frac{A}{A+V}}. \tag{5.6}$$

$A+V$ is the marginal variance of Z_i, so in our setup $A+V = 1$ by construction (5.2), giving a simpler form of the lemma,

$$\zeta_i|Z_i \sim N(\sqrt{1-V}\; Z_i, V). \tag{5.7}$$

If V is known we can use the formula to improve the estimate of $\text{Prob}\{\text{Event}_i|Z_i\}$, by averaging $p_1(\zeta_i)$ over distribution (5.7).

Formula (5.7) describes both the bias and variance of Z_i. The fact that $E\{\zeta_i|Z_i\} = \sqrt{1-V}\ Z_i$ is a manifestation of regression to the mean. With thousands of genes being studied simultaneously, choosing as "interesting" those with the largest scores Z_i (or $-Z_i$) tends to overstate their true effects.

In order to use (5.7) we need to know V. To this end, a small bootstrap analysis was carried out, using what we will call "row resampling", see Remark D. Let $\mathbf{x}_i$ be the 20×8 matrix for gene i having jkth entry

$$x_{ijk} = \log(pm_{ijk}) - .5 \cdot \log(mm_{ijk}),\ i = 1, 2, \ldots, 20, k = 1, 2, \ldots, 8. \quad (5.8)$$

The average of the 20 rows of $\mathbf{x}_i$ is $\mathbf{M}_i = (M_{i1}, M_{i2}, \ldots, M_{i8})$, the ith row of $\mathbf{M}$ in (2.6). The bootstrap matrix $\mathbf{M}^*$ was formed by taking $\mathbf{M}_i^*$ to be the average of 20 rows of $\mathbf{x}_i$ randomly selected with replacement from the actual 20 rows. $\mathbf{M}^*$ then gave bootstrap scores $\{Z_i^*\}$ via (2.8). Finally the algorithm (a)-(e) following (3.9) gave bootstrap inference curve $p_1(Z)^*$, using the bootstrap maximum value p_0^* at step (e). $B = 25$ bootstrap replications were generated in this way.

We can now estimate the variance "V" of the Z_i scores. Let i_ℓ be the index of the gene corresponding to the ℓth smallest (most negative) Z score in the original analysis of Figure 1, so for example Z_{i_1} is the smallest of the 6810 scores. The estimated variance for Z_{i_ℓ} is the usual empirical variance estimate based on the 25 $Z_{i_\ell}^*$ values.

Figure 9's left panel shows the variance estimates V_{i_ℓ} for $\ell = 1, 2, \ldots, 150$, the 150 genes with the most negative Z values, similar results holding for the 150 largest Z values. The right panel shows V_{i_ℓ} for $\ell = 25, 60, 95, \ldots, 6745, 6780$, spanning the range of Z. The variances are large near the center of the Z scale, averaging about .24 near $Z = 0$, but, fortunately, much smaller near the extremes.This means that the Z values are more accurate for $|Z|$ large, which are the cases that give big estimated event probabilities.

The average variance V_i in the right panel of Figure 9 was 0.209. We can set $A = 1 - 0.209 = 0.791$ in (5.3), satisfying the constraint that the composite marginal distribution of the Z_i's have variance 1. Following the Lemma, (5.6), we take

$$\zeta_i|Z_i \sim N\left(C_i \frac{Z_i}{\sqrt{A+V_i}}, 1 - C_i^2\right),\ C_i = \sqrt{\frac{A}{A+V_i}}, \quad (5.9)$$

See Remark E.

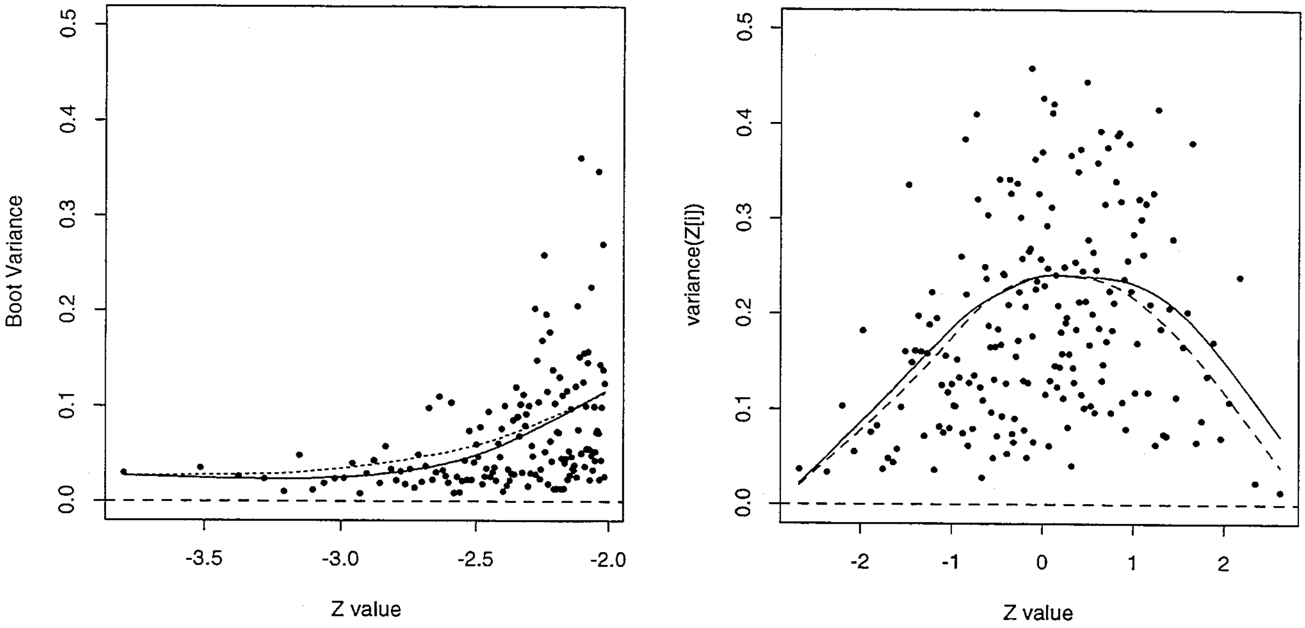

Figure 9: *Bootstrap Variance estimates V_i for the Z scores applying to Figure 1. Left panel: for genes i corresponding to the 150 smallest values Z_i. Right panel: for ℓth smallest, $\ell = 25, 60, 90, 130, \ldots, 6780$. Solid curves are smoothing splines, degrees of freedom 4, 5 respectively. Dotted curves are same for $J = 10$ bootstrap, explained in Remark D.*

The bootstrap analysis that gave the V_i's also produced 25 bootstrap versions $p_1(Z)^*$ of the curve $p_1(Z)$, (5.1). The variability of these curves was combined with the aposteriori variability of $\zeta_i|Z_i$ in (5.9) to give an 80% aposteriori range of values for $P\{\text{Event}_i|Z_i\}$, as detailed in Remark E. Figure 10 shows 80% intervals for the 18 Northern Blot genes featured in Figure 1. We see that the point estimates of $\text{Prob}\{\text{Event}_i|Z_i\}$ were highly accurate for $|Z| \geq 2.5$, but quite variable for smaller Z, especially for $|Z|$ near 2.

Northern Blot analyses produced a quantitative score "G_i" for each of the 18 genes, G standing for Gold Standard. G scores exceeding 1.30 were taken to indicate a positive effect of radiation on gene activity, the "+" symbols in Figures 1 and 10; likewise "−" for $G_i < 0.70$ and "o" for $0.70 \leq G_i \leq 1.30$. Figure 11 compares the Z_i scores from Figure 1 with $\log G_i$ for the 18 test genes. We see a nice monotone relationship, correlation 0.87.

The agreement seen in Figure 11 is impressive, especially considering the magnitude of the sampling errors displayed in Figure 10. Our gold standard, the Northern Blot score, is not pure gold, itself being subject to experimental error. There is only one flagrant disagreement in Figure 11, the "−" gene at $Z_i = -0.31$. The vector of differences (3.4) was $\mathbf{D}_i = (-1.59, 0.55, 0, 88, -0.83)$ for this gene, so that both wildtypes yielded aliquots of opposing signs. In contrast the "o" point at $Z_i = 2.51$, lying just below the "+" cutoff value $G = 1.30$, was consistently positive, $\mathbf{D}_i = (4.54, 2.81, 1.64, 2.65)$, strengthening our belief that this gene was positively affected by the radiation.

6 Remarks, Proofs, and Details

A. *Debrightening and Desumming* Some microarray plates are "brighter" than others in that they produce systematically larger expression levels. Following probe reduction (2.4) we debrightened the data by separately standardizing the columns of $\mathbf{M}$. That is, each column of $\mathbf{M}$ was linearly transformed to have mean 0 and empirical standard deviation 1.

"Desumming" corrects for another type of data inhomogeneity. Corresponding to $\mathbf{D}$ (3.4), let

$$\mathbf{S} = (\mathbf{M}_3 + \mathbf{M}_1, \mathbf{M}_4 + \mathbf{M}_2, \mathbf{M}_7 + \mathbf{M}_5, \mathbf{M}_8 + \mathbf{M}_6) \qquad (6.1)$$

A gene with larger $\mathbf{S}$ values tends to have larger values of $\mathbf{D}$, which undercuts the exchangeability across genes implicit in our empirical Bayes analyses.

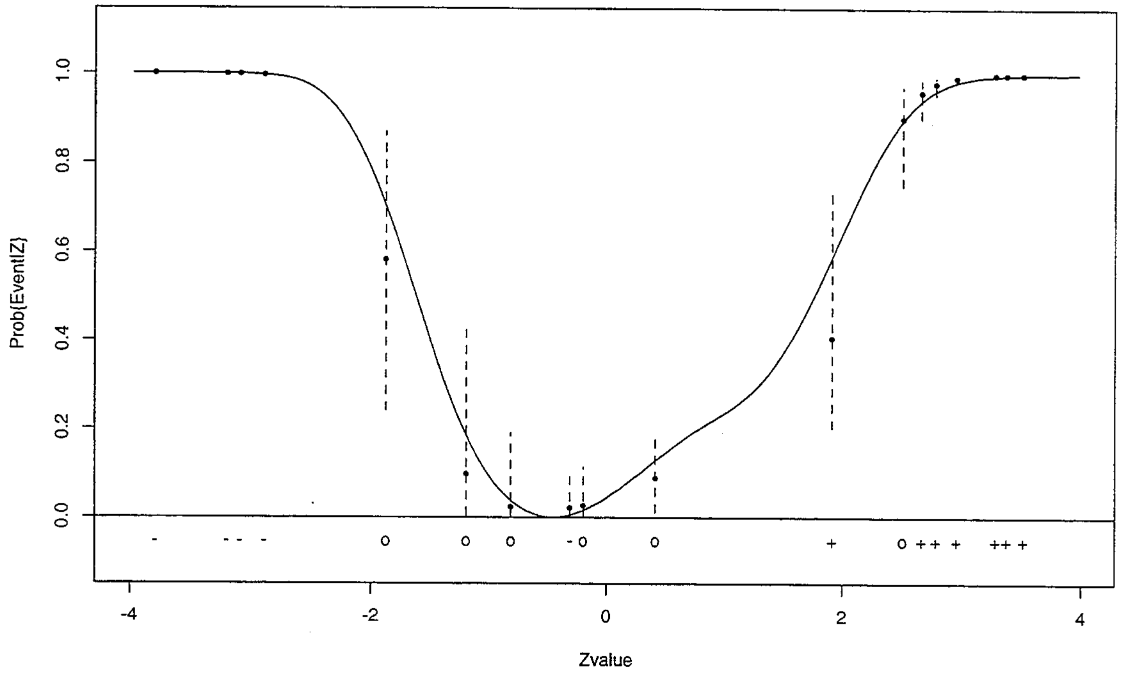

Figure 10: *80% aposteriori intervals for* Prob{Event$_i$|Z_i}*; for the 18 Northern Blot genes featured in Figure 1. solid curve is point estimate of* Prob{Event$_i$|Z_i} *from Figure 1. Solid dots are aposteriori medians. Based on 25 row resampling bootstraps as described in text. The intervals are very short for* $|Z| \geq 2.5$ *and widest for* $|Z|$ *near 2.*

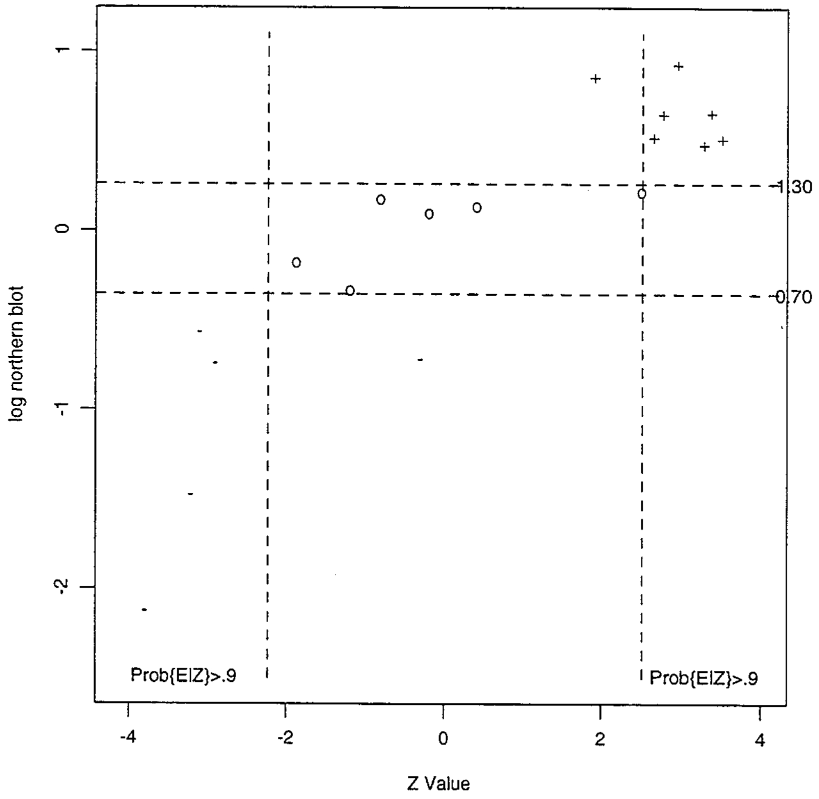

Figure 11: *Comparison of the Z scores from the analysis in Figure 1 with the logarithm of the Northern Blot results. Correlation 0.87. Z values outside the two vertical lines have* $\text{Prob}\{\text{Event}_i|Z_i\} \geq .90$.

(Newton et al. (2000) adjust their data for a similar problem.) After debrightening, the individual columns of $\mathbf{D}$ were desummed as follows: a linear regression $|D_{ik}| = a_0 + a_1|S_{ik}| +$ error was fit individually to each column, and then each D_{ik} was transformed to

$$D_{ik}/(\widehat{a}_0 + \widehat{a}_1|S_{ik}|). \tag{6.2}$$

Similar transformations were made on the columns of $\mathbf{d}$, 3.6. It was the transformed $\mathbf{D}$ and $\mathbf{d}$ matrices that were used to compute the scores $\mathbf{Z}$ and $\mathbf{z}$ via (2.8).

B. *Logistic Regression Estimate of $f_0(z)/f(z)$.* The ratio $f_0(z)/f(z)$ in (3.8) was estimated by logistic regression. Given $B = 20$ replications of $\mathbf{z}$, all $n \cdot (1 + B) = 6810 \cdot 21$ scores Z_i and z_i were plotted on a line, with Z_i's considered as "successes" and $\mathbf{z}$'s as "failures". The probability $\pi(z)$ of a success at point z is given in terms of the densities (3.2),

$$\pi(z) = f(z)/(f(z) + Bf_0(z)), \tag{6.3}$$

so that (3.8) becomes

$$p_1(Z) = 1 - p_0 \frac{1 - \pi(Z)}{B\pi(Z)}. \tag{6.4}$$

With $n = 6810$ genes, the normal scores transformation resulted in $\max\{Z_i\} = -\min\{Z_i\} = 3.80$, while the null scores $\{z_i\}$ were confined to a smaller range, as in Figure 2. Our algorithm divided the range $[-4, 4]$ into 139 equal intervals, counted the number of Z_i's and z_i's in each interval, and estimated $\pi(z)$ by logistic regression, for use in (6.4). The regression function was a natural spline with 5 degrees of freedom, called by the Splus command $ns(x, df = 5)$, x being the 139 center points of the intervals. The choice of $B = 20$ $\mathbf{z}$ replications was based on an analysis like that in Figure 5, which showed considerable improvement for B increasing from 1 to 10, but little gain past 20.

C. *Estimating the Null Distribution* The null density $f_0(z)$ is supposed to describe the distribution of expression scores for genes unaffected by the treatment of interest. Basing f_0 on $\mathbf{d}$ in (3.6) seems natural for the $2 \times 2 \times 2$ radiation experiment, but other choices are possible and may be necessary for other experimental designs. In the separate analysis of the wildtype2 data in Section 4.3 for example, **d** was taken to be only the last two columns of (3.6).

As a less symmetric example, the trouble with $\mathbf{D}_1$ seen in Table 3 suggests basing the Z scores on just the last three columns of **D** in (3.4). We carried out such an analysis by also removing the second column from **d** in (3.6), the I1B-I1A column, (after a more extensive correlation study suggested that the I1A plate was the wildtype1 culprit). The resulting Prob$\{\text{Event}|Z\}$ curve was moderately superior to the wildtype2 curve in Figure 7, having $N90 = 353$.

As a simple but informative model suppose that

$$M_{ik} = \mu_i + \theta_i t_{ik} + \epsilon_{ik}, \tag{6.5}$$

where θ_i is the Treatment effect on gene i, t_{ik} indicates whether or not plate k employed the Treatment ($t_{ik} = 0$ for $k = 1, 2, 4, 6$ and 1 for $k = 3, 4, 7, 8$ in our experiment), and ϵ_{ik} is independent noise. Then Z_i in (2.8) is

$$Z_i = (\theta_i + \phi_i)/(a_o + S_i), \quad S_i = \left[\sum_{\ell}(\phi_{i\ell} - \phi_i)^2/(L-1)\right]^{\frac{1}{2}}, \tag{6.6}$$

where each $\phi_{i\ell}$ is the difference of two independent ϵ_{ik}'s, and ϕ_i is the average of $\phi_{i\ell}$ over the L components **D**, $L = 4$ in (3.4). The z_i's have the same expression except with $\theta_i = 0$ in (6.6), and we can see that $f_0(z)$ is a legitimate null hypothesis comparator for $f(Z)$.

The additive model (6.5) gives every column of **d** the same distribution, that of ϕ, but we might not trust the Treatment differences to really have the same distribution as the Control, I2B-I2A compared to U2B-U2A for example. Empirically this turned out not to be a problem for the radiation experiment, but if it had we might have used just the first and third columns of **d** in (3.6). Then we could calculate two versions of Z_i for each gene, from the first two and the last two columns **D** in (3.4), and use all $2 \cdot 6810$ Z's to estimate $f(Z)$.

The smallest possible comparative experiment would have just a single Treatment and single Control plate, so **M** in (3.5) would be $n \times 2$. In that case we could compare the density $f(Z)$ from $\mathbf{Z} = \mathbf{M}_2 - \mathbf{M}_1$ with $f_0(z)$, the empirical density obtained by randomly changing the signs of Z_i, say $z_i = C_i Z_i$ where the C_i were independently ±1 with probabilities 1/2. In terms of model (6.5), $f(Z)$ would be the empirical density of $\{\theta_i + \phi_i\}$, and $f_0(z)$ the empirical density of $\{C_i\theta_i + \phi_i\}$.

This is a weak experiment at best, but could be useful if, unlike the radiation experiment, the Treatment effects θ_i were mostly of one sign. It would be more effective to run a third plate, either Treatment or Control. If it were Treatment for example, then the difference of the two Treatment plates would provide a legitimate estimate of $f_0(z)$.

D. *Bootstrapping Microarrays* The bootstrap analysis of Section 5, resampling rows, can be motivated in terms of a much simpler situation: Suppose we observe uncorrelated real-valued random variables x_{ij},

$$x_{ij} \sim (\mu_i, \sigma_i^2) \quad \text{for} \quad i = 1, 2, \dots, n \quad \text{and} \quad j = 1, 2, \dots, J, \tag{6.7}$$

the notation indicating that x_{ij} has mean μ_i and variance σ_i^2, and that we wish to estimate the variance of $\bar{x} = \Sigma_i \Sigma_j x_{ij}/nJ$. The bootstrap data set $\{x^*_{ij}\}$ obtained by resampling J times within each of the n subsets $\mathbf{x}_i = \{x_{ij},\ j = 1, 2, \dots, J\}$, the analogy of row resampling, gives $\bar{x}^*$ bootstrap variance

$$\text{var}_*\{\bar{x}^*\} = \frac{1}{n^2 J} \sum_i \widehat{\sigma}_i^2 \quad \left[\widehat{\sigma}_i^2 = \sum_j (x_{ij} - x_i)^2/J\right], \tag{6.8}$$

which is a reasonable estimate of the actual variance $\text{var}\{\bar{x}\} = \Sigma_i \sigma_i^2/n^2 J$.

Result (6.8) depends on the x_{ij} being uncorrelated within the $\mathbf{x}_i$ subsets. Defining x_{ijk} as in (5.8), the analogous requirement for the row resampling of Section 5 is that rows of the 20×8 matrices $\mathbf{x}_i = \{x_{ijk},\ j = 1, 2, \dots, 20,\ k = 1, 2, \dots, 8\}$ be uncorrelated. In fact there was some evidence of row correlations, but not of a serious magnitude. As a check, the analysis in Section 5 was rerun taking adjacent *pairs* of rows as the resampling units, giving the "$J = 10$" curves in Figure 8. High adjacent correlations should then produce bigger bootstrap variances (effectively reducing the size of J in (6.8)), but this did not happen in Figure 8.

Other bootstrap schemes are available. If we had a full components of variance model for the microarray data, as in Li & Wong (2000), we could resample the model using its estimated variance components. Another, simpler, possibility is "resampling genes", in our case resampling as a unit the entire $20 \times 2 \times 8$ data vector for each gene. However this will not give variance estimates for the expression scores Z_i, since gene i's data always stays the same in the resamples. Moreover, in context (6.7), $\bar{x}^*$ will have marginal variance

$$\frac{\Sigma(\mu_i - \bar{\mu})^2}{n^2} + \frac{1}{n^2 J}\Sigma\sigma_i^2, \tag{6.9}$$

and will tend to give inflated bootstrap estimates if $\Sigma(\mu_i - \bar{\mu})^2$ is large.

E. *Aposteriori Intervals for Prob$\{Event_i | Z_i\}$* The construction of Figure 10 began by approximating the normal distribution (5.9) with a discrete distribution supported on 100 equally weighted, but unequally spaced, points ζ_{ih}, $h = 1, 2, \ldots, 100$. The bth of the 25 bootstrap curves $p_1(Z)^*$, say $p_1^{(b)}(\cdot)$, then gives

$$p_{ih}^{(b)} = p_1^{(b)}(\zeta_{ih}), \quad h = 1, 2, \ldots, 100, \tag{6.10}$$

a total of $2500 = 25 \cdot 100$ points in all. The heavy dots in Figure 10 are the medians of the point sets, while the dashed intervals are the ranges of the central 2000 points, excluding 250 at each extreme. Calling these "aposteriori intervals" relies on two assumptions, that the bootstrap curves $p_1(Z)^*$ can be thought of as samples from the aposteriori distribution starting from an uninformative prior on the true curve, as in Rubin (1979) and Efron (1982); and that this aposteriori distribution acts independently of (5.9).

Equation (5.9) is motivated in the following way. We begin with a generalized version of assumptions (5.3), (5.4),

$$\mu_i \sim N(0, A) \quad \text{and} \quad Z_i | \mu_i \sim N(\mu_i, V_i), \tag{6.11}$$

which as before gives the aposteriori distribution of $\zeta_i = \mu_i / \sqrt{A}$,

$$\zeta_i | Z_i \sim N\left(C_i \frac{Z_i}{\sqrt{A + V_i}}, \ 1 - C_i^2\right), \quad C_i = \sqrt{\frac{A}{A + V_i}} \ . \tag{6.12}$$

However Z_i is not directly observable. Instead we see the ith gene's relative ranking, say

$$\widetilde{Z}_i = \Phi'\left(\frac{\text{rank}_i - .5}{n}\right) \tag{6.13}$$

as in(5.2). Marginally $Z_j \sim N(0, A + V_j)$ so the expectation of rank$_i$ given Z_i is $\Sigma_j \Phi\left(Z_i / \sqrt{A + V_j}\right)$.

With n as large as 6810 it is reasonable to approximate rank$_i$ with its expectation, giving

$$\widetilde{Z}_i \doteq \Phi^{-1}\left(\frac{1}{n}\sum_j \Phi\left(\frac{Z_i}{\sqrt{A+V_j}}\right)\right) . \tag{6.14}$$

When (6.13) was numerically solved for Z_i given $\widetilde{Z}_i$, using the estimates of A and $\{V_i\}$ from the bootstrap analysis of Section 5, it was found that Z_i was always within 2% of $\widetilde{Z}_i$, leading to approximation (5.9).

F. *Neyman-Pearson Calculations* The Lemma in Section 3.2 follows directly from (3.11-3.13),

$$\begin{aligned} \beta_0 &= \int_{A_o} f_1(z)dz = \frac{1}{p_1}\int_{A_o}\left(\frac{f(z)}{f_0(z)} - p_0\right) f_0(z)dz \\ &= \frac{1}{p_1}[b_0 - (1-\alpha_0)p_0] = b_0 - \frac{p_0}{p_1}[(1-\alpha_0) - b_0], \end{aligned} \tag{6.15}$$

which is equivalent to (3.14).

The fact that the lower bound (3.9) on p_1 is equivalent to the condition that β_0 in (3.14) must be non-negative is an exercise in Neyman-Pearson theory. In order for the curve (α_0, β_0) to be the boundary of a Neyman-Pearson figure, $d\beta_0/d\alpha_0$ must be everywhere nonpositive. Using (3.14),

$$\frac{d\beta_0}{d\alpha_0} = \frac{1}{p_1}\left[p_0 + \frac{db_0}{d\alpha_0}\right] \leq 0 . \tag{6.16}$$

Suppose that the likelihood ratio $L(z) = f(z)/f_0(z)$ has densities $g(\ell)$ and $g_0(\ell)$ under f and f_0 respectively, and that L is supported on the interval $[L_{\text{min}}, L_{\text{max}}]$. Then

$$\begin{aligned} g(\ell) &= E\{f(z)|L(z)=\ell\} = E\{Lf_0(z)|L=\ell\} \\ &= \ell E\{f_0(z)|L=\ell\} = \ell g_0(\ell), \end{aligned} \tag{6.17}$$

the expectations being taken with respect to Lebesque measure in our case.

The likelihood ratio test at level α_0 rejects H_0 for say $L \geq L_0$, giving

$$\alpha_0 = \int_{L_0}^{L_{\max}} g_0(\ell)d\ell \quad \text{and} \quad b_0 = \int_{L_{\min}}^{L_0} g(\ell)d\ell = \int_{L_{\min}}^{L_0} \ell g_0(\ell)d\ell, \tag{6.18}$$

so

$$\frac{db_0}{d\alpha_0} = \frac{db_0/dL_0}{d\alpha_0/dL_0} = -L_0 \ . \tag{6.19}$$

Then condition (6.11) implies $p_0 - L_0 \leq 0$, or that

$$p_0 \leq L_{\min} \quad \text{and} \quad p_1 \geq 1 - L_{\min} \ , \tag{6.20}$$

which is (3.9).

G. *Ideal Linear Combinations* Probe reduction (2.6), generalized in (4.1), uses probe-wise means of the numbers

$$x_{ijk} = \log(pm_{ijk}) - .5 \cdot \log(mm_{ijk}) \tag{6.21}$$

to form the expression values M_{ik}. In place of means, we also tried using "L estimators", that is weighted linear combinations of the ordered values $x_{i(j)k}$, where $x_{i(20)k}$ is the largest of the 20 x_{ijk} numbers, etc:

$$M_{ik} = \sum_j w_j x_{i(j)k} \quad \Big[w_j \geq 0, \ \sum_j w_j = 1\Big]. \tag{6.22}$$

The search for an ideal weight vector $\mathbf{w}$ was carried out using a form of cross-validation: (1) Using only the wildtype2 data (plates 5, 6, 7, 8) specifications (2.6), 2.8) were used to produce scores Z_i for all 6810 genes; (2) The 6810×20 data matrices $\{x_{ij4}\}$ and $\{x_{ij2}\}$, experiments I1B and U1B, had their rows ordered and then the latter was subtracted from the former to give the 6810×20 difference matrix "$\boldsymbol{\delta}$"; (3) the rows $\boldsymbol{\delta}_i$ corresponding to the 1% (68) largest values of Z_i were averaged to give vector "$\boldsymbol{\delta}_{\text{high}}$", and likewise "$\boldsymbol{\delta}_{\text{low}}$" from the 68 most negative Z_λ's, yielding the difference vector

$$\mathbf{a} = \boldsymbol{\delta}_{\text{high}} - \boldsymbol{\delta}_{\text{low}}; \tag{6.23}$$

(4) The middle 30% (2043) of the Z_i's gave a 2043×20 matrix of corresponding $\boldsymbol{\delta}_i$ vectors, having 20×20 sample covariance matrix "$\mathbf{B}$"; (5) Finally, the figure of merit

$$Q(\mathbf{w}) = (\mathbf{a}'\mathbf{w})^2/\mathbf{w}'\mathbf{B}\mathbf{w} \tag{6.24}$$

was maximized over the choice of $\mathbf{w}$, giving the "ideal linear combination"

$$\mathbf{w}^{+} = \mathbf{B}^{-1}\mathbf{a}. \tag{6.25}$$

The idea here is that we expect $(\mathbf{a}'\mathbf{w})^2$ to be big if $\mathbf{w}$ is well-chosen, assuming that the extreme genes identified at Step (2) are genuinely affected. Dividing by $\mathbf{w}'\mathbf{B}\mathbf{w}$ in (6.24) accounts for the variance of the linear combination. The algorithm was carried out twice, the second time with the vectors $\boldsymbol{\delta}_i$ reversed for indices i with $Z_i < 0$. In both cases, but more clearly in the latter, $\mathbf{w}^{+}$ was seen to put less weight on the extreme probes j. This suggested using a trimmed mean in (6.22), with perhaps w_1, w_2, w_3 and w_{18}, w_{19}, w_{20} equaling zero. However the resulting analysis yielded results no better than curve 1 in Figure 7.

H. *False detection rates for SAM.* The false detection rate (FDR) is the proportion of changes in gene expression that are actually not significant among those changes declared to be significant. Define p to be the proportion of the total genes that are declared significant, b to be the proportion that are truly significant, and α to be the type I error rate. Then $\text{FDR} = \alpha(1-b)/p$. Although b is unknown, we can conclude that $\text{FDR} \leq \alpha/p$. Thus, there is a maximum possible value for the FDR, which is shown in the rightmost column of Table 2.

References

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., J., P., Marti, G., Moore, T., Hudsom, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R. Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000), 'Identification of molecularly and clinically distinct substypes of diffuse large b cell lymphoma by gene expression profiling', *Nature* **403**, 503–511.

Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2000), Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Unpublished.

Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc. Nat. Acad. Sci* **95**, 14863–14868.

Goss, V., Tibshirani, R. & Chu, C. (2000), Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. submitted.

Lee, M., Kuo, F., Whitmore, G. & Sklar, J. (2000), Importance of replication in microarray gene expression studies: statistical methods and evidence from a single cdna array experiment. To appear, Proc. Nat. Acad. Sci.

Li, C. & Wong, W. H. (2000), Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Unpublished.

Newton, M., Kendziorski, C., Richmond, C., Blatter, F. & Tsui, K. (2000), On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. To appear, J. Comp. Biology.