Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher*, Robert Tibshirani†, and Gilbert Chu*‡

*Departments of Medicine and Biochemistry, Stanford University, 269 Campus Drive, Center for Clinical Sciences Research 1115, Stanford, CA 94305-5151; and †Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305

Communicated by Bradley Efron, Stanford University, Stanford, CA, February 6, 2001 (received for review December 1, 2000)

Microarrays can measure the expression of thousands of genes to identify changes in expression between different biological states. Methods are needed to determine the significance of these changes while accounting for the enormous number of genes. We describe a method, Significance Analysis of Microarrays (SAM), that assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance, the false discovery rate (FDR). When the transcriptional response of human cells to ionizing radiation was measured by microarrays, SAM identified 34 genes that changed at least 1.5-fold with an estimated FDR of 12%, compared with FDRs of 60 and 84% by using conventional methods of analysis. Of the 34 genes, 19 were involved in cell cycle regulation and 3 in apoptosis. Surprisingly, four nucleotide excision repair genes were induced, suggesting that this repair pathway for UV-damaged DNA might play a previously unrecognized role in repairing DNA damaged by ionizing radiation.

DNA microarrays contain oligonucleotide or cDNA probes for measuring the expression of thousands of genes in a single hybridization experiment. Although massive amounts of data are generated, methods are needed to determine whether changes in gene expression are experimentally significant. Cluster analysis of microarray data can find coherent (logical) patterns of gene expression (1) but provides little information about statistical significance. Methods based on conventional *t* tests provide the probability (*P*) that a difference in gene expression occurred by chance (2, 3). Although $P = 0.01$ is significant in the context of experiments designed to evaluate small numbers of genes, a microarray experiment for 10,000 genes would identify 100 genes by chance. This problem led us to develop a statistical method adapted specifically for microarrays, Significance Analysis of Microarrays (SAM).

SAM identifies genes with statistically significant changes in expression by assimilating a set of gene-specific *t* tests. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with scores greater than a threshold are deemed potentially significant. The percentage of such genes identified by chance is the false discovery rate (FDR). To estimate the FDR, nonsense genes are identified by analyzing permutations of the measurements. The threshold can be adjusted to identify smaller or larger sets of genes, and FDRs are calculated for each set. To demonstrate its utility, SAM was used to analyze a biologically important problem: the transcriptional response of lymphoblastoid cells to ionizing radiation (IR).

Materials and Methods

Preparation of RNA. Human lymphoblastoid cell lines GM14660 and GM08925 (Coriell Cell Repositories, Camden, NJ) were seeded at 2.5×10^5 cells/ml and exposed to IR 24 h later. RNA was isolated, labeled, and hybridized to the HuGeneFL GeneChip microarray according to manufacturer's protocols (Affymetrix, Santa Clara, CA).

Microarray Hybridization. Each gene in the microarray was represented by 20 oligonucleotide pairs, each pair consisting of an oligonucleotide perfectly matched to the cDNA sequence, and a second oligonucleotide containing a single base mismatch. Because gene expression was computed from differences in hybridization to the matched and mismatched probes, expression levels were sometimes reported by the GeneChip analysis suite software as negative numbers.

Northern Blot Hybridization. Total RNA (15 μg) was resolved by agarose gel electrophoresis, transferred to a nylon membrane, and hybridized to specific radiolabeled DNA probes, which were prepared by PCR amplification.

Results

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 h after exposure to a modest dose of 5 Gy of IR. RNA samples were labeled and divided into two identical aliquots for independent hybridizations, A and B. Thus, data for 6,800 genes on the microarray were generated from eight hybridizations (U1A, U1B, U2A, U2B, I1A, I1B, I2A, and I2B).

We scaled the data from different hybridizations as follows. A reference data set was generated by averaging the expression of each gene over all eight hybridizations. The data for each hybridization were compared with the reference data set in a cube root scatter plot. We chose the cube root scatter plot because it resolved the vast majority of genes that are expressed at low levels and permitted the inclusion of negative levels of expression that are sometimes generated by the GeneChip software. A linear least-squares fit to the cube root scatter plot was then used to calibrate each hybridization.

After scaling, a linear scatter plot was generated for average gene expression in the four A aliquots (U1A, I1A, U2A, and U2A) vs. the average in the four B aliquots (U1B, I1B, U2B, and U2B), a partitioning of the data that eliminates biological changes in gene expression (Fig. 1*A*). The linear scatter plot confirmed that the data were generally reproducible but failed to resolve genes expressed at low levels. Better resolution of these genes was achieved by the cube root scatter plot (Fig. 1*B*), which revealed three salient features: the large percentage of genes (24%) assigned negative levels of expression, the large percentage of genes with low levels of expression, and the low signal-to-noise ratio at low levels of expression.

To assess the biological effect of IR, a scatter plot was generated for average gene expression in the four irradiated states vs. the four unirradiated states (compare Fig. 1 *B* and *C*). A few of the potentially significant changes in gene expression are indicated by arrows in Fig. 1*C*, but the effect was not easily quantified, and a method was needed to identify changes with statistical confidence.

Abbreviations: SAM, significance analysis of microarrays; FDR, false discovery rate; IR, ionizing radiation; FWER, family-wise error rate.

‡To whom reprint requests should be addressed. E-mail: chu@cmgm.stanford.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

www.pnas.org/cgi/doi/10.1073/pnas.091062498

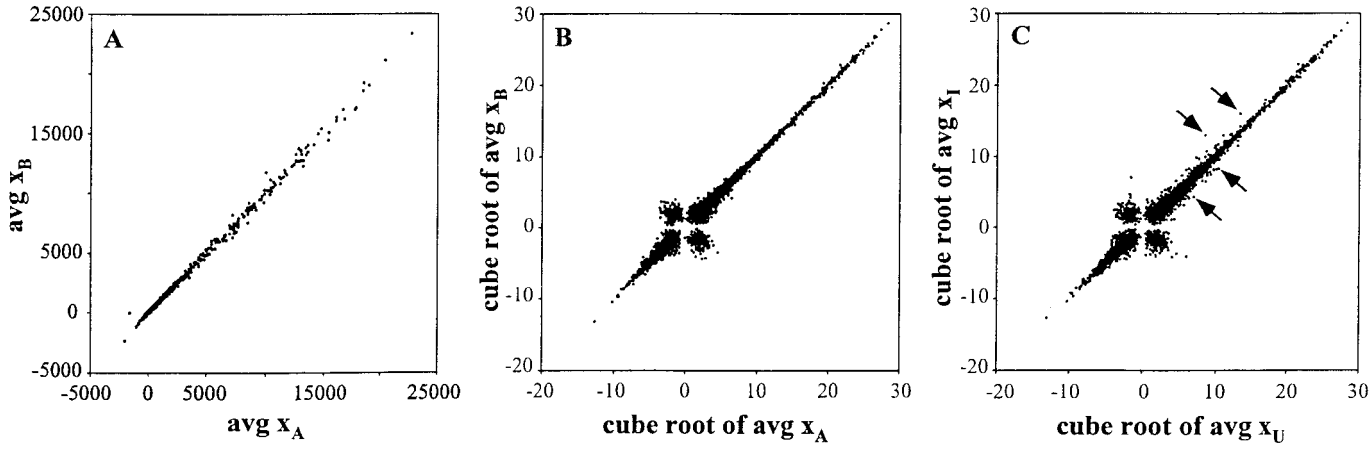

Fig. 1. Gene expression measured by microarrays. (*A*) Linear scatter plot of gene expression. Each gene (*i*) in the microarray is represented by a point with coordinates consisting of average gene expression measured from the four A hybridizations (avg x_A) and the average gene expression in the four B hybridizations (avg x_B). (*B*) Cube root scatter plot of gene expression. The average gene expression from the A and B hybridizations have been plotted on a cube root scale to resolve genes expressed at low levels. (C) Cube root scatter plot of average gene expression from the four hybridizations with uninduced cells (avg x_U) and induced cells 4 h after exposure to 5 Gy of IR (avg x_I). Some of the genes that responded to IR are indicated by arrows.

Our approach was based on analysis of random fluctuations in the data. In general, the signal-to-noise ratio decreased with decreasing gene expression (Fig. 1). However, even for a given level of expression, we found that fluctuations were gene specific. To account for gene-specific fluctuations, we defined a statistic based on the ratio of change in gene expression to standard deviation in the data for that gene. The "relative difference" $d(i)$ in gene expression is:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} \qquad [1]$$

where $\bar{x}_I(i)$ and $\bar{x}_U(i)$ are defined as the average levels of expression for gene (i) in states I and U, respectively. The "gene-specific scatter" $s(i)$ is the standard deviation of repeated expression measurements:

$$s(i) = \sqrt{a\left\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2\right\}} \qquad [2]$$

where Σ_m and Σ_n are summations of the expression measurements in states I and U, respectively, $a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$, and n_1 and n_2 are the numbers of measurements in states I and U (four in this experiment).

To compare values of $d(i)$ across all genes, the distribution of $d(i)$ should be independent of the level of gene expression. At low expression levels, variance in $d(i)$ can be high because of small values of $s(i)$. To ensure that the variance of $d(i)$ is independent of gene expression, we added a small positive constant s_0 to the denominator of Eq. **1**. The coefficient of variation of $d(i)$ was computed as a function of $s(i)$ in moving windows across the data. The value for s_0 was chosen to minimize the coefficient of variation. For the data in this paper, this computation yielded $s_0 = 3.3$.

Scatter plots of $d(i)$ vs. $s(i)$ are shown in Fig. 2. The scatter plot for relative difference between states I and U is shown in Fig. 2*A*. By contrast, the scatter plot for relative difference between cell lines 1 and 2 shows more marked changes in Fig. 2*B*. These relative differences exceeded random fluctuations in the data, as measured by the relative difference between hybridizations A and B in Fig. 2*C*.

Although the relative difference computed from hybridizations A and B provided a control for random fluctuations, additional controls were needed to assign statistical significance to the biological effect of IR. Instead of performing more experiments, which are expensive and labor intensive, we generated a large number of controls by computing relative differences from permutations of the hybridizations for the four irradiated and four unirradiated states. To minimize potentially confounding effects from differences between the two cell lines, we analyzed the data by using the 36 permutations that were balanced for cell lines 1 and 2. Permutations were defined as balanced when each group of four experiments contained two experiments from cell line 1 and two experiments from cell line 2. Fig. 2 *C* and *D* are examples of balanced permutations.

To find significant changes in gene expression, genes were ranked by magnitude of their $d(i)$ values, so that $d(1)$ was the largest relative difference, $d(2)$ was the second largest relative difference, and $d(i)$ was the ith largest relative difference. For each of the 36 balanced permutations, relative differences $d_p(i)$ were also calculated, and the genes were again ranked such that $d_p(i)$ was the ith largest relative difference for permutation p. The expected relative difference, $d_E(i)$, was defined as the average over the 36 balanced permutations, $d_E(i) = \Sigma_p d_p(i)/36$.

To identify potentially significant changes in expression, we used a scatter plot of the observed relative difference $d(i)$ vs. the expected relative difference $d_E(i)$ (Fig. 3*A*). For the vast majority of genes, $d(i) \cong d_E(i)$, but some genes are represented by points displaced from the $d(i) = d_E(i)$ line by a distance greater than a threshold Δ. For example, the threshold $\Delta = 1.2$ illustrated by the broken lines in Fig. 3*A* yielded 46 genes that were "called significant." These 46 genes are shown in the context of the scatter plot for $d(i)$ vs. $s(i)$ (Fig. 3*B*) and in the scatter plot for the cube root of gene expression $\bar{x}_I(i)$ vs. $\bar{x}_U(i)$ (Fig. 3*C*). Genes identified by $d(i)$ do not necessarily have the largest changes in gene expression.

To determine the number of falsely significant genes generated by SAM, horizontal cutoffs were defined as the smallest $d(i)$ among the genes called significantly induced and the least negative $d(i)$ among the genes called significantly repressed. The number of falsely significant genes corresponding to each permutation was computed by counting the number of genes that exceeded the horizontal cutoffs for induced and repressed genes. The estimated number of falsely significant genes was the average of the number of genes called significant from all 36 permutations. For $\Delta = 1.2$, the permuted data sets generated an average of 8.4 falsely significant genes, compared with 46 genes called significant, yielding an estimated FDR of 18% (Table 1). As Δ decreased, the number of genes called significant by SAM increased but at the cost of an

STATISTICS

GENETICS

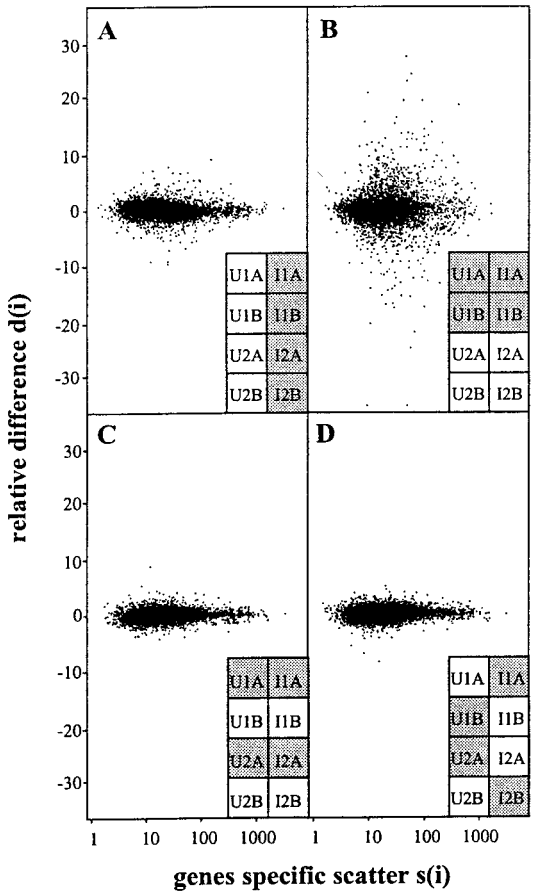

Fig. 2. Scatter plots of relative difference in gene expression $d(i)$ vs. gene-specific scatter $s(i)$. The data were partitioned to calculate $d(i)$, as indicated by the bar codes. The shaded and unshaded entries were used for the first and second terms in the numerator of $d(i)$ in Eq. **1**. (*A*) Relative difference between irradiated and unirradiated states. The statistic $d(i)$ was computed from expression measurements partitioned between irradiated and unirradiated cells. (*B*) Relative difference between cell lines 1 and 2. The statistic $d(i)$ was computed from expression measurements partitioned between cell lines 1 and 2. (*C*) Relative difference between hybridizations A and B. The statistic $d(i)$ was computed from the permutation in which the expression measurements were partitioned between the equivalent hybridizations A and B. (*D*) Relative difference for a permutation of the data that was balanced between cell lines 1 and 2.

increasing FDR. (Omitting s_0 from Eq. **1** produced higher FDRs of 45, 35, and 28% for $\Delta = 0.6$, 0.9, and 1.2.)

Our method for setting thresholds provides asymmetric cutoffs for induced and repressed genes. The alternative is the standard *t* test, which imposes a symmetric horizontal cutoff, with $d(i) > c$ for induced genes and $d(i) < -c$ for repressed genes. However, the asymmetric cutoff is preferred because it allows for the possibility that $d(i)$ for induced and repressed genes may behave differently in some biological experiments.

SAM proved to be superior to conventional methods for analyzing microarrays (Table 1 and Fig. 4*A*). First, SAM was compared with the approach of identifying genes as significantly changed if an *R*-fold change was observed. In this "fold change" method, $r(i) = \bar{x}_I(i)/\bar{x}_U(i)$, and gene (i) was called significantly changed if $r(i) > R$ or $r(i) < 1/R$. To permit computation of $r(i)$ from negative values for gene expression, $\bar{x}_I(i)$ and $\bar{x}_U(i)$ were converted to 10 when their values were negative or less than 10. The results of this procedure yielded unacceptably high FDRs of 73–84%.

Another approach attempts to account for uncertainty in the data by identifying genes as significantly changed if an *R*-fold change is observed consistently between paired samples (4). To apply this "pairwise fold change" method to our four data sets before IR and four data sets after IR, changes in gene expression were declared significant if 12 of 16 pairings satisfied the criteria $r(i) > R$ or $r(i) < 1/R$. Despite the demand for consistent changes between paired samples, this method yielded FDRs of 60–71%.

To understand why fold-change methods fail, note that the vast majority of genes are expressed at low levels where the signal-to-noise ratio is very low (Fig. 3*C*). Thus, 2-fold changes in gene expression occur at random for a large number of genes. Conversely, for higher levels of expression, smaller changes in gene expression may be real, but these changes are rejected by fold-change methods. The pairwise fold-change method provides modest improvement but remains inferior to SAM.

Of the 46 genes most highly ranked by SAM ($\Delta = 1.2$), 36 increased or decreased at least 1.5-fold (R = 1.5). The number of falsely significant genes that met these two criteria was 4.5, corresponding to a FDR of 12% (Table 1). Fas was identified three times as alternately spliced forms, leaving 34 independent genes (Table 2). As an indication of biological validity, 10 of the 34 genes have been reported in the literature as part of the transcriptional response to IR. TNF-α was reported to be induced by other investigators (5) but was repressed here. Quantitative reverse transcription–PCR confirmed this result.

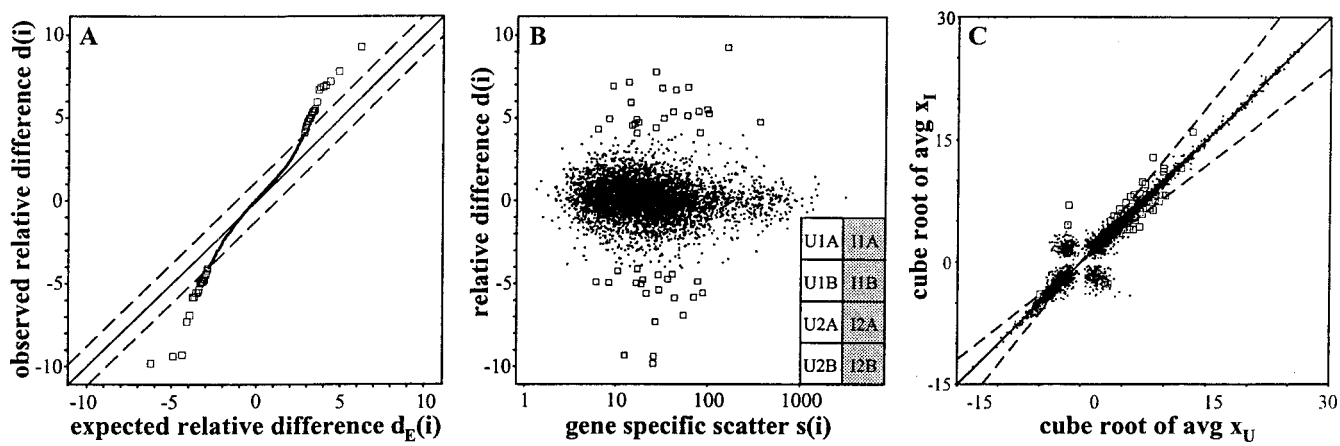

Fig. 3. Identification of genes with significant changes in expression. (*A*) Scatter plot of the observed relative difference $d(i)$ versus the expected relative difference $d_E(i)$. The solid line indicates the line for $d(i) = d_E(i)$, where the observed relative difference is identical to the expected relative difference. The dotted lines are drawn at a distance $\Delta = 1.2$ from the solid line. (*B*) Scatter plot of $d(i)$ vs. $s(i)$. (*C*) Cube root scatter plot of average gene expression in induced and uninduced cells. The cutoffs for 2-fold induction and repression are indicated by the dashed lines. In *A*–*C*, the 46 potentially significant genes for $\Delta = 1.2$ are indicated by the squares.

Table 1. Comparison of methods for identifying changes in gene expression

Parameter	Number falsely significant	Number called significant	FDR
SAM			
$\Delta = 0.4$	134.9	288	47%
$\Delta = 0.5$	78.1	192	41%
$\Delta = 0.6$	56.1	162	35%
$\Delta = 0.9$	19.1	80	24%
$\Delta = 1.2$	8.4	46	18%
$\Delta = 1.2$; R = 1.5	4.5	34	12%
Fold change			
R = 2.0	283.1	348	81%
R = 2.5	137.8	169	82%
R = 3.0	76.8	99	78%
R = 3.5	46.7	64	73%
R = 4.0	29.3	35	84%
Pairwise fold change			
R = 1.2	245.6	355	69%
R = 1.3	155.4	220	71%
R = 1.5	76.2	118	65%
R = 1.7	44.8	70	64%
R = 2.0	22.8	38	60%

To increase the stringency for calling significant changes in gene expression, parameters for each method (Δ and R) were increased, as described in the text. The false discovery rate (FDR) was defined as the percentage of falsely significant genes compared to the genes called significant.

To test the validity of SAM directly, we performed Northern blots for genes that were randomly selected from the 46 and 57 genes most highly ranked by SAM ($\Delta = 1.2$) and the fold-change method (at least 3.6-fold change), respectively. Northern blots showed little correlation with the genes identified by the fold change method (Fig. 4*B*), but strong correlation with the genes identified by SAM (Fig. 4*C*). Indeed, Northern blots contradicted only 1 (maxiK) of 11 genes identified by SAM, consistent with our estimated FDR.

Nineteen of the 34 genes most highly ranked by SAM appear to be involved in the cell cycle. Three are known to be induced in a p53-dependent manner: p21, cyclin G1, and mdm2 (6–8). Six cell cycle genes were repressed: E2-EPF, p55cdc, cyclin B, ckshs2, cdc25, and wee1 (9, 10). Five genes encoding the mitotic machinery were also repressed: PLK-1, MKLP-1, MCAK, C-TAK1, CENP-E (11–13). Three genes involved in cell proliferation were induced or repressed: PTP(CAAX1), LPAP, and c-myc (14–18). Some responses appeared paradoxical. For example, cdc25 phosphatase and wee1 kinase have antagonistic effects on the phosphorylation state of cdc2, but both genes were repressed. Repression of these genes together with the mitotic genes may represent a damage response that dismantles the cell cycle machinery until the cell has repaired the damaged DNA.

Four of the 34 genes play roles in DNA repair, but none are involved in the repair of IR-induced double-strand breaks. Instead, the genes (p48, XPC, gadd45, PCNA) have roles in nucleotide excision repair, a pathway conventionally associated with UV-induced damage (19–22). We confirmed the induction of these genes by Northern blot (23–25). Fornace *et al.* reported defective removal of base damage induced by IR in xeroderma pigmentosum cells (26). Leadon *et al.* reported that a novel DNA repair pathway involving long excision repair patches of at least 150 nucleotides is activated by IR but not UV (27). Our results suggest that this novel pathway might include p48, XPC, gadd45, and PCNA.

Four of the 34 genes play roles in apoptosis (Fas, bbc3, TNF-α, OX40 ligand). The remaining genes may have previously unsuspected roles in the DNA damage response or may be among the estimated set of four falsely detected genes.

The 34 genes most highly ranked by SAM are only a subset of all of the genes that change 1.5-fold with IR. Indeed, we calculated the difference between the number of genes called significant and the number of falsely significant genes for decreasing $\Delta = 0.3$, 0.2, and 0.1, and found the differences to be 92, 170, and 184, respectively. Thus, SAM suggests that approximately 180 of the 6,800 genes on the microarray were induced or repressed by 5 Gy IR.

Discussion

SAM is a method for identifying genes on a microarray with statistically significant changes in expression, developed in the

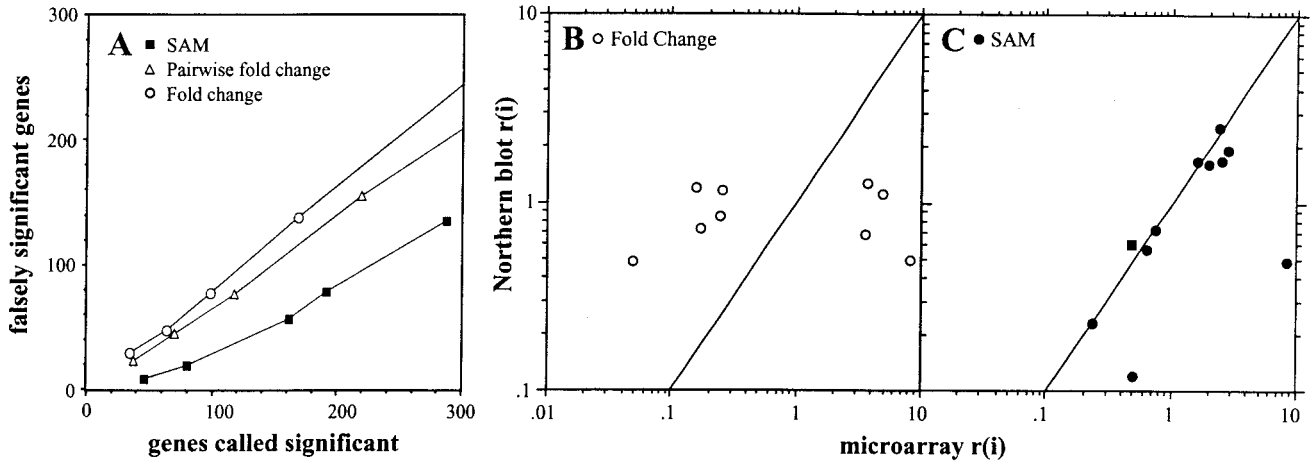

Fig. 4. Comparison of SAM to conventional methods for analyzing microarrays. (*A*) Falsely significant genes plotted against number of genes called significant. Of the 57 genes most highly ranked by the fold change method, 5 were included among the 46 genes most highly ranked by SAM. Of the 38 genes most highly ranked by the pairwise fold change method, 11 were included among the 46 genes most highly ranked by SAM. These results were consistent with the FDR of SAM compared to the FDRs of the fold change and pairwise fold change methods. (*B*) Northern blot validation for genes identified by the fold change method. Values of $r(i)$ are plotted for genes chosen at random from the 57 genes most highly ranked by the fold change method. (*C*) Validation for genes identified by SAM. Results are plotted for genes chosen at random from the 46 genes most highly ranked by SAM. Genes analyzed by Northern blot are represented by circles. TNF-α was validated by using a PreDeveloped TaqMan assay (PE Biosystems) and is represented by a square. The straight lines in *B* and *C* indicate the position of exact agreement between Northern blot and microarray results.

Table 2. Genes with changes in expression called significant by SAM

Rank	Accession	$d(i)$	$r(i)$	$s(i)$	$\bar{x}_U(i)$	$\bar{x}_l(i)$	Gene*
Induced Genes							
1	U09579†	9.2	3.4	158	633	2119	p21/cip1; cyclin-dependent kinase inhibitor
2	X83490†	7.7	2.5	26	155	381	Fas alternate splice deleting exons 3 and 4
3	U47621	7.1	2.9	13	61	178	No55; nucleolar autoantigen
4	U18300†	6.9	1.9	59	448	869	p48; gene mutated in xeroderma pigmentosum group E
5	U48296	6.7	1.6	30	354	583	PTP(CAAX1); protein tyrosine phosphatase
6	X63717†	6.6	2.2	43	254	561	Fas; member of TNF receptor superfamily
7	D21089	5.5	2.4	96	392	930	XPC; gene mutated in xeroderma pigmentosum group C
8	U39400	5.4	1.7	41	349	581	NOF1; neighbor of FAU
9	X77794†	5.3	1.6	99	964	1499	cyclin G1; G2/M phase arrest cyclin
10	M60974†	5.1	2.5	58	203	516	gadd45; growth arrest and DNA damage inducible protein
11	D90224	4.9	2.8	32	96	270	OX40 ligand; TNF ligand superfamily member
12	U25138	4.9	9.0‡	16	-4	90	Maxi K; potassium channel beta subunit
13	J05614 †	4.8	1.7	352	2358	4043	PCNA; proliferating cell nuclear antigen
14	X83492†	4.8	1.8	17	117	213	Fas alternate splice deleting exons 4 and 7
15	X85116	4.5	2.0	15	83	163	EPB72; erythrocyte membrane protein
16	U50136	4.4	1.6	26	231	359	LTC4S; leukotriene C4 synthase
17	U82987	4.2	33.6‡	161	-4	336	bbc3; bcl-2 binding component 3
18	M92424†	4.0	2.8	16	45	125	mdm2; p53 binding protein
Repressed Genes							
1	U01038	-9.8	0.50	25	551	275	PLK-1; polo-like kinase 1
2	M91670	-9.3	0.61	25	693	425	E2-EPF; ubiquitin carrier protein
3	U68233	-9.3	0.48	12	275	133	HRR-1; member of nuclear receptor subfamily 1, group H
4	U05340	-6.9	0.39	54	642	253	p55cdc; homolog of yeast cdc20 cell cycle protein
5	M25753†	-5.8	0.23	43	345	78	cyclin B; cdc2-interacting cyclin
6	X97267	-5.8	0.49	69	818	400	LPAP; lymphocyte phosphatase-associated phosphoprotein
7	S78187†	-5.5	0.63	20	357	224	cdc25; cdc2 phosphatase
8	X54942	-5.5	0.52	87	1026	534	ckshs2; cdc28 protein kinase 2
9	U63743	-5.0	0.59	19	265	158	MCAK; mitotic centromere-associated kinesin
10	D86973	-4.9	0.64	16	264	168	GCN1; general control of amino-acid synthesis 1
11	X62048†	-4.9	0.44	8	99	43	wee1; cdc2 kinase
12	M80359	-4.9	0.40‡	6	25	-19	C-TAK1; microtubule affinity-regulating kinase 3
13	U28386	-4.8	0.58	77	928	541	hSRP1α; receptor for nuclear localization sequences
14	X02910	-4.8	0.37	20	170	63	TNF-α; tumor necrosis factor α
15	D31764	-4.7	0.26	36	247	64	hEphB1b; Eph-like receptor tyrosine kinase
16	X67155	-4.4	0.30	28	199	60	MKLP-1; mitotic kinesin-like protein 1
17	HG3523	-4.4	0.51	40	391	200	c-Myc, alternate splice 3
18	Z15005	-4.2	0.28‡	10	36	-20	CENP-E; centromere protein E, putative kinetochore motor

*Gene functions: Black = cell cycle; Dark gray = apoptosis; Light gray = DNA repair.
†Genes previously reported to respond transcriptionally to IR.
‡To compute $r(i) = \bar{x}_l(i)/\bar{x}_U(i)$, negative levels of expression were reset to a value of 10.

context of an actual biological experiment. SAM was successful in analyzing this experiment as well as several other experiments with oligonucleotide and cDNA microarrays (data not shown).

In the statistics of multiple testing (28–30), the family-wise error rate (FWER) is the probability of at least one false positive over the collection of tests. The Bonferroni method, the most basic method for bounding the FWER, assumes independence of the different tests. An acceptable FWER could be achieved for our microarray data only if the corresponding threshold was set so high that no genes were identified. The step-down correction method of Westfall and Young (29), adapted for microarrays by Dudoit *et al.* (http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html), allows for dependent tests but still remains too stringent, yielding no genes from our data.

Westfall and Young (29) define "weak control" to be control of the FWER when all of the null hypotheses are true (i.e., when there are no changes in gene expression). "Strong control" is control of the FWER when any subset of the null hypotheses is true. Under certain conditions, weak control implies strong control. In fact, the step-down correction method exerts both weak and strong control.

The method of Benjamini and Hochberg (31) assumes independent tests and guarantees an upper bound for the FDR (with both weak and strong control) by a step-up or step-down procedure applied to the individual P values. For our data, the P value for each gene is calculated from permutations of the eight experiments. Because of the limited number of permutations, the FDR is too "granular", and we identified either zero or 300 significant genes, depending on how the P value was defined. A similar granular result was obtained for the adaptation to dependent tests by Benjamini *et al.* [*The Control of the False Discovery Rate in Multiple Testing Under Dependency* (Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv). http://www.math.tau.ac.il/~ybenja/].

SAM does not have strong or weak control of the FWER. Instead, SAM provides an estimate of the FDR for each value of the tuning parameter Δ. The estimated FDR is computed from permutations of the data and hence assumes that all null hypotheses are true, allowing for the possibility of dependent tests. It seems plausible that this estimated FDR approximates the strongly controlled FDR when any subset of null hypotheses is true. However, we have not proven this in general. It is possible for SAM to give an estimate of the FDR that is greater than 1. However, this has not occurred in our experience. Indeed, SAM provides a reasonably accurate estimate for the true FDR. To confirm this, we constructed artificial data sets in which a subset of genes was induced over a background of noise. SAM successfully identified the induced genes and estimated the FDR with reasonable accuracy.

Although this paper analyzes a simple two-state experiment, SAM can be generalized to other types of experiments by defining $d(i)$ in a different way. Suppose the data includes gene expression $x_j(i)$ and a response parameter y_j, in which $i = 1, 2, \ldots, m$ genes, $j = 1, 2, \ldots, n$ states. The generalized statistical parameter still takes the form $d(i) = r(i)/[s(i) + s_0]$, except that the definitions of $r(i)$ and $s(i)$ change.

To identify genes with changes in expression in an experiment with three or more states, the parameter $d(i)$ is defined in terms of the Fisher's linear discriminant. One goal might be to identify genes whose expression in one type of tumor is different from its expression in other types of tumors. Suppose that a set of n samples consists of K nonoverlapping subsets, such that the response parameter $y_j \in \{1, \ldots, K\}$. Define $C(k) = \{j : y_j = k\}$. Let $n_k =$ number of observations in $C(k)$. The average gene expression in each subset is $\bar{x}_k(i) = \Sigma_{j \in C(k)} x_j(i)/n_k$ and the average gene expression for all n samples is $\bar{x}(i) = \Sigma_j x_j(i)/n$. Then define:

$$r(i) = \{[\Sigma_k n_k/\Pi_k n_k]\Sigma_k n_k[\bar{x}_k(i) - \bar{x}(i)]^2\}^{1/2} \quad [3]$$

$$s(i) = \{[\Sigma_k(1/n_k)/\Sigma_k(n_k - 1)]\Sigma_k\Sigma_{j \in C(k)}[x_j(i) - \bar{x}_k(i)]^2\}^{1/2} \quad [4]$$

SAM can be adapted for still other types of experimental data. For example, to identify genes whose expression correlates with survival time, $d(i)$ is defined in terms of Cox's proportional hazards function, in which some of the patients remain alive or are lost to follow-up at the time of the study. To identify genes whose expression correlates with a quantitative parameter, such as tumor stage, $d(i)$ can be defined in terms of the Pearson correlation coefficient. Another example includes the definition of $d(i)$ for paired data, such as gene expression in tumors before and after chemotherapy. In each case, the FDR is estimated by random permutation of the data for gene expression among the different experimental arms, i.e., permutations among the n arms of y_j. Thus, SAM is a robust and straightforward method that can be adapted to a broad range of experimental situations. SAM and the adaptations discussed above are available for use at http://www-stat-class.stanford.edu/SAM/SAMServlet.

We thank Peter Jackson, Ron Davis, James Ferrell, Dean Felsher, Lisa DeFazio, Joe Budman, Jean Tang, Tom Tan, and Kerri Rieger for helpful discussions. This work was supported by the Burroughs Wellcome Clinical Scientist Award and by National Institutes of Health (NIH) Grant CA77302 to G.C., by NIH Small Business Technology Transfer grant CA75675 to G.C. and Affymetrix, and by the Stanford Genome Training Grant to V.T.

1. Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
2. Roberts, C., Nelson, B., Marton, M., Stoughton, R., Meyer, M., Bennett, H., He, Y., Dai, H., Walker, W., Hughes, T., Tyers, M., Boone, C. & Friend, S. (2000) *Science* **287,** 873–880.
3. Galitski, T., Saldanha, A., Styles, C., Lander, E. & Fink, G. (1999) *Science* **285,** 251–254.
4. Ly, D., Lockhart, D., Lerner, R. & Schultz, P. (2000) *Science* **287,** 2486–2492.
5. Weill, D., Gay, F., Tovey, M. & Chouaib, S. (1996) *J. Interferon Cytokine Res.* **16,** 395–402.
6. Harper, J. W., Adami, G. R., Wei, N., Keyomarsi, K. & Elledge, S. J. (1993) *Cell* **75,** 805–816.
7. Okamoto, K. & Beach, D. (1994) *EMBO J.* **13,** 4816–4822.
8. Prives, C. (1998) *Cell* **95,** 5–8.
9. Furnari, B., Rhind, N. & Russell, P. (1997) *Science* **277,** 1495–1497.
10. Liu, Z., Diaz, L., Haas, A. & Giudice, G. (1992) *J. Biol. Chem.* **267,** 15829–15835.
11. Lee, K., Yuan, Y., Kuriyama, R. & Erikson, R. (1995) *Mol. Cell. Biol.* **15,** 7143–7151.
12. Maney, T., Hunter, A., Wagenbach, M. & Wordeman, L. (1998) *J. Cell. Biol.* **142,** 787–801.
13. Wood, K., Sakowicz, R., Goldstein, L. & Cleveland, D. (1997) *Cell* **91,** 357–366.
14. Ding, I., Bruyns, E., Li, P., Magada, D., Paskind, M., Rodman, L., Seshadri, T., Alexander, D., Giese, T. & Schraven, B. (1999) *Eur. J. Immunol.* **29,** 3956–3961.
15. Cates, C., Michael, R., Stayrook, K., Harvey, K., Burke, Y., Randall, S., Crowell, P. & Crowell, D. (1996) *Cancer Lett.* **110,** 49–55.
16. Godfrey, W., Fagnoni, R., Harara, M., Buck, D. & Engleman, E. (1994) *J. Exp. Med.* **180,** 757–762.
17. Lord, J., McIntosh, B., Greenberg, P. & Nelson, B. (2000) *J. Immunol.* **164,** 2533–2541.
18. Prevot, D., Voeltzel, T., Birot, A., Morel, A., Rostan, M., Magaud, J. & Corbo, L. (2000) *J. Biol. Chem.* **275,** 147–153.
19. Aboussekhra, A., Biggerstaff, M., Shivji, M., Vilpo, J., Moncollin, V., Podust, V., Protic, M., Hubscher, U., Egly, J. & Wood, R. (1995) *Cell* **80,** 859–868.
20. Smith, M., Ford, J., Hollander, M., Bortnick, R., Amundson, S., Seo, Y., Deng, C., Hanawalt, P. & Fornace, A. J. (2000) *Mol. Cell. Biol.* **20,** 3705–3714.
21. Sugasawa, K., Ng, J., Masutani, C., Iwai, S., van der Spek, P., Eker, A., Hanaoka, F., Bootsma, D. & Hoeijmakers, J. (1998) *Mol. Cell* **2,** 223–232.
22. Tang, J., Hwang, B., Ford, J., Hanawalt, P. & Chu, G. (2000) *Mol. Cell* **5,** 737–744.
23. Kastan, M., Zhan, Q., El-Deiry, F., Jacks, T., Walsh, W., Plunkett, B., Vogelstein, B. & Fornace, A. (1992) *Cell* **71,** 587–597.
24. Hwang, B. J., Ford, J., Hanawalt, P. C. & Chu, G. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 424–428.
25. Xu, J. & Morris, G. (1999) *Mol. Cell. Biol.* **19,** 12–20.
26. Fornace, A., Dobson, P. & Kinsella, T. (1986) *Radiat. Res.* **106,** 73–77.
27. Leadon, S., Dunn, A. & Ross, C. (1996) *Radiat. Res.* **146,** 123–130.
28. Hochberg, Y. & Tamhane, A. (1987) *Multiple Comparisons Procedures* (Wiley, New York).
29. Westfall, P. & Young, S. (1993) *Resampling-Based Multiple Testing* (Wiley, New York).
30. Hsu, J. (1996) *Multiple Comparisons: Theory and Methods* (Chapman & Hall, London).
31. Benjamini, Y. & Hochberg, Y. (1995) *J. R. Stat. Soc. B* **57,** 289–300.