BIOINFORMATICS

Vol. 19 no. 14 2003, pages 1808–1816
DOI: 10.1093/bioinformatics/btg242

Controlling false-negative errors in microarray differential expression analysis: a PRIM approach

Steve W. Cole[1,3,*], Zoran Galic[1,3] and Jerome A. Zack[1,2,3]

[1]Department of Medicine, [2]Department of Microbiology, Immunology, and Molecular Genetics, David Geffen School of Medicine at UCLA, and [3]UCLA AIDS Institute, Los Angeles, CA 90095-1678, USA

Received on November 18, 2002; revised on February 25, 2003; accepted on March 27, 2003

ABSTRACT

Motivation: Theoretical considerations suggest that current microarray screening algorithms may fail to detect many true differences in gene expression (Type II analytic errors). We assessed 'false negative' error rates in differential expression analyses by conventional linear statistical models (e.g. *t*-test), microarray-adapted variants (e.g. SAM, Cyber-T), and a novel strategy based on hold-out cross-validation. The latter approach employs the machine-learning algorithm Patient Rule Induction Method (PRIM) to infer minimum thresholds for reliable change in gene expression from Boolean conjunctions of fold-induction and raw fluorescence measurements.

Results: Monte Carlo analyses based on four empirical data sets show that conventional statistical models and their microarray-adapted variants overlook more than 50% of genes showing significant up-regulation. Conjoint PRIM prediction rules recover approximately twice as many differentially expressed transcripts while maintaining strong control over false-positive (Type I) errors. As a result, experimental replication rates increase and total analytic error rates decline. RT-PCR studies confirm that gene inductions detected by PRIM but overlooked by other methods represent true changes in mRNA levels. PRIM-based conjoint inference rules thus represent an improved strategy for high-sensitivity screening of DNA microarrays.

Availability: Freestanding JAVA application at http://microarray.crump.ucla.edu/focus

Contact: coles@ucla.edu

INTRODUCTION

Early approaches to analyzing microarray expression data focused on reducing large numbers of assayed transcripts into a small number of groups showing distinct expression profiles (Eisen *et al*., 1998; Tamayo *et al*., 1999; Alter *et al*., 2000). However, these analyses produced unreliable results at the level of individual genes because such 'unsupervised learning algorithms' provide no mechanism for controlling analytic errors (e.g. a *p*-value estimate of the 'false positive' errors). In response to this problem, researchers began to employ a more stringent hypothesis-testing approach aimed at controlling Type I analytic errors, or false declarations of change. Second generation analyses utilized conventional inferential statistics (Kerr *et al*., 2000) or modifications of the univariate general linear model (GLM) (Dudoit *et al*., 2000; Long *et al*., 2001; Tusher *et al*., 2001) to control Type I error. However, the increased stringency of these analyses and the poor sensitivity of the GLM in the presence of high noise and limited replicates (Miller, 1986) suggest that second generation screening strategies may overlook many true differences in gene expression (committing Type II 'false negative' errors). Theoretical power analyses indicate that GLM techniques will fail to detect more than 70% of genes showing 2-fold up-regulation in typical microarray data structures (e.g. gene-specific *t*-tests analyzing five paired test versus control samples with Type I error controlled at $p < 0.05$ and a coefficient of variation in replicate change scores of ~100%, as observed empirically below) (Winer, 1971). In microarray screening, this problem is aggravated by increasing stringency to control for thousands of parallel hypothesis tests.

A high rate of Type II error undermines several basic applications of microarray technology, including efforts to map gene expression networks and identify phenotypically influential genes. In network mapping, Type II errors constitute a failure to recognize existing links, which leads to underestimates of network connectivity and faulty conclusions about system stability, redundancy, path lengths, and block structure (Harary, 1969). In efforts to uncover influential genes, high Type II error rates increase the likelihood that researchers will overlook key results even when they are present in the data (e.g. in searching for a single viral receptor by comparing gene expression in infectable versus uninfectable cell types, or seeking one causally significant gene among

*To whom correspondence should be addressed at: 11-394 Factor Building, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095-1678, USA.

a large background of spurious correlates or consequences). For such applications, it is scientifically crucial that screening algorithms minimize Type II errors without incurring any substantial increase in false positives.

Here, we present a novel approach to microarray differential expression analysis that employs the machine learning algorithm Patient Rule Induction Method (PRIM) (Friedman and Fisher, 1999) to infer 'rules' about which profiles of change in observed data are likely to recur in a subsequent replication study. The basic form of these rules is derived from analysis of visual processing algorithms used by experienced microarray analysts to infer reliable change from graphical data displays (as detailed in the Web Supplement, *Visual pattern analysis of differential expression*). The key insights are that human visual analysts, (1) attend to multiple measures of change simultaneously (e.g. fold-change *and* raw change) and, (2) focus on minimum bounds for change rather than the expected (mean) values. The PRIM approach automates analysts' reliability induction processes using an explicitly defined statistical criterion for replication likelihood. Monte Carlo studies show that this approach maintains tight control over Type I errors while substantially reducing the Type II error rates incurred by other second-generation microarray screening analyses.

SYSTEMS AND METHODS

Microarray data

We assessed analytic error rates in several differential expression data sets including comparisons of HIV-infected versus uninfected thymocytes (three replicate observations of 7070 transcripts surveyed by Affymetrix HuGene-FL high density oligonucleotide arrays), normal versus cancerous breast tissue [five replicate observations of 5584 genes surveyed by spotted cDNA microarrays (Perou *et al*., 2000)], immature versus mature T lymphocytes (three replicate observations of 7070 transcripts, Affymetrix), and quiescent versus activated T lymphocytes (three replicate observations of 7070 transcripts, Affymetrix). Raw data are available from http://microarray.crump.ucla.edu/focus (Affymetrix studies) and http://genome-www.stanford.edu/sutech (cDNA arrays). Technical details of experiments and microarray data collection are contained in the Web Supplement (*Experimental methods*) and published papers (Perou *et al*., 2000).

GLM analyses

Following most second-generation microarray analyses (Dudoit *et al*., 2000; Kerr *et al*., 2000; Li and Wong, 2001; Long *et al*., 2001; Tusher *et al*., 2001), we fit a standard GLM statistical model to log-transformed expression values to identify genes with a mean change in expression exceeding a biologically significant threshold, Δ_{thresh} (e.g. >2-fold increase at $p < 0.05$). Denoting the assayed expression of gene g under condition c in the rth experimental replicate as x_{gcr} ($g = 1$–G genes, $c = 0$ for control or 1 for experimental conditions, and $r = 1$–R replicates), the model expresses each observation as the sum of independent effects representing the mean expression intensity across all genes (μ), gene-specific differences in basal expression (γ_g), generalized differences in expression intensity across experimental conditions (β_c), gene-specific effects of experimental conditions (δ_{gc}), and a residual term capturing all other sources of variation (ε_{gcr}):

$$\log x_{gcr} = \mu + \gamma_g + \beta_c + \delta_{gc} + \varepsilon_{gcr}. \quad (1)$$

ε is modeled as a random normal variate with a mean of 0 and a standard deviation of σ. Gene is g identified as differentially expressed if the difference $\Delta_g = \delta_{g1} - \delta_{g0}$ exceeds Δ_{thresh} (e.g. $\log 2$-fold) at a specified level of significance (e.g. $p < 0.05$). This model subsumes most standard statistical analyses including the paired- and independent sample t-tests and array-wide analysis of variance (ANOVA). Variants differ mainly in their methods for estimating σ and are further detailed in the Web Supplement (*Data analysis*). As in previous studies (Kerr *et al*., 2000), GLM models fit the data well, with R^2 goodness-of-fit statistics exceeding 0.85 for all data sets examined (e.g. Supplemental Table 1).

Monte Carlo analysis of Type I and II errors

Theoretical power analyses make assumptions about the magnitude of 'true' change (e.g. all $\Delta_g = 0$), but realized Type II error rates depend upon the empirical distribution of true changes (i.e. how many genes' expression are actually altered by the studied manipulation?). We estimated the true change distribution in empirical microarray data by fitting the GLM model of Equation (1). We then used the resulting parameter estimates to generate 200 Monte Carlo data sets corresponding to each empirical data set (Web Supplement, *Monte Carlo studies*) (Bratley *et al*., 1983). The 'true' magnitude of change in expression of gene g was specified by $\Delta_g = \delta_{g1} - \delta_{g0}$, and noisy 'observations' of x_{gcr} were generated via Equation (1), with ε_{gcr} drawn from a random normal distribution with the empirical value of σ (or gene-specific values of σ_g in heteroscedastic models; see Web Supplement, *Data analysis*). Monte Carlo data were then analyzed by alternative screening strategies, and declared results for each gene were compared with its true status to assess rates of Type I error (significant change indicated by analysis, but no true change at the level of Δ_g), Type II error (no significant change indicated by analysis, but a true change at the level of Δ_g), and total analytic error (Type I + Type II).

Consistent with theoretical power analyses, Monte Carlo studies showed that conventional GLM analyses overlook 60–70% of true differences in gene expression (Table 1, Columns 1 and 2, and Supplemental Tables 2–4). Higher

Table 1. Monte Carlo assessment of analytic error for alternative microarray screening tools

Prototype data set	Screening tool[a]			
	Gene-specific *t* 95% CI	ANOVA 95% CI	$PRIM_{Min}$	$PRIM_{Mean}$
HIV infection				
Type I error rate[b] (6745 genes <2-fold)	0.40%	0.80%	2.00%	2.20%
Type II error rate[b] (325 genes >2-fold)	75.0%	72.3%	55.8%	55.7%
Total error rate[b]	3.9%	4.1%	4.1%	4.0%
Yield[b]	110	144	279	289
κ[b]	0.359	0.365	0.452	0.440
Breast cancer				
Type I error rate (5054 genes <2-fold)	0.10%	0.20%	0.90%	1.40%
Type II error rate (722 genes >2-fold)	84.6%	65.9%	50.9%	44.3%
Total error rate	8.3%	6.7%	5.8%	5.6%
Yield	87	205	326	389
κ	0.247	0.474	0.594	0.630
T cell development				
Type I error rate (6335 genes <2-fold)	0.10%	0.40%	2.10%	2.30%
Type II error rate (735 genes >2-fold)	52.3%	49.9%	40.9%	40.4%
Total error rate	5.6%	5.6%	6.1%	6.3%
Yield	359	394	568	587
κ	0.615	0.625	0.633	0.629
Immunologic activation				
Type I error rate (5971 genes <2-fold)	0.30%	0.60%	4.20%	4.30%
Type II error rate (1099 genes >2-fold)	57.3%	56.8%	26.4%	25.8%
Total error rate	9.1%	9.4%	7.6%	7.6%
Yield	486	514	1058	1072
κ	0.549	0.544	0.706	0.707
Average				
Type I error rate	0.23%	0.50%	2.30%	2.55%
Type II error rate	67.3%	61.2%	43.5%	41.6%
Total error rate	6.7%	6.5%	5.9%	6.2%
Yield	261	314	558	585
κ	0.443	0.502	0.596	0.602

[a]Screening functions are detailed in the Web Supplement and all are tuned to detect an increase in expression of 2-fold or greater. Gene-specific t = 95% confidence interval (CI) based on gene-specific t-test ($p < 0.05$); ANOVA = 95% CI derived from fully randomized ANOVA with error pooled across all transcripts (model: log expression level = Gene + Condition + Gene * Condition + error); $PRIM_{Min}$ = PRIM prediction rule estimating 80% likelihood of >2-fold increase in replicate observation from minimum observed raw- and fold-changes; $PRIM_{Mean}$ = PRIM prediction rule estimating 60% likelihood of >2-fold increase in replicate observation from mean observed raw- and geometric mean fold-change.

[b]Table entries represent mean value over 200 Monte Carlo data sets corresponding to each prototype data set. Results are presented for the heteroscedastic Monte Carlo model (sampling variability differing across transcripts). Comparable results emerged from homeoscedastic models (constant sampling variability). Type I error rate = percent of increases declared >2-fold that actually show a true mean change <2-fold ('false positive'). Type II error rate = percent of true >2-fold increases not detected by analysis ('false negative'). Total error rate = frequency of Type I or Type II error/total number of transcripts analyzed. Yield = number of transcripts identified as differentially expressed. κ = Cohen's chance-corrected measure of accuracy (0 = no better than expected by chance; 1 = perfect detection performance; 0.5 = 50% better than expected by chance).

Type II error rates emerged for microarray-adapted variants of the GLM such as Significance Analysis of Microarrays (Tusher *et al*., 2001) and Cyber-T (Long *et al*., 2001) (data not shown). To verify that differences overlooked by GLM analyses represented true biological changes, we conducted reverse-transcriptase PCR (RT-PCR) to measure mRNA levels for several transcripts showing $\Delta_g > 2$-fold (detailed in the Web Supplement, *Biological verification of Type II errors*). As shown in Table 2, results indicated consistent changes in mRNA level despite failure of GLM analyses to indicate statistically significant differences.

ALGORITHM AND IMPLEMENTATION

Replication inference via PRIM

Analyses of intuitive reliability inference by experienced microarray users (Web Supplement, *Visual pattern analysis of differential expression*) suggest that it may be possible to identify replicable changes in gene expression from a multivariate data representation including both absolute and relative change—$ratio_{gr} = x_{gr1}/x_{gr0}$ and $difference_{gr} = x_{gr1} - x_{gr0}$. The resulting Boolean conjunction rules structurally resemble the 'prediction boxes' produced by the

Table 2. RT-PCR verification of transcripts recovered by PRIM but missed by all other analyses[a]

Accession	Gene name		Differential expression ratio[b]			
			Replicate 1	Replicate 2	Replicate 3	Geometric mean
X00695	Interleukin 2 (IL2)					
		Microarray[c]	51.3	10.2	5.0	13.9
		RT-PCR[d]	508.2	143.9	45.5	149.3
M20137	Interleukin 3 (IL3)					
		Microarray	250.2	4.7	10.6	23.1
		RT-PCR	>500-fold[e]	>500	>500	>500
U43672	Interleukin 18 receptor (IL18R1)					
		Microarray	18.4	628.5	10.3	49.2
		RT-PCR	23.0	12.9	16.6	17.0
L40379	Thyroid receptor interactor protein 10 (TRIP10)					
		Microarray	15.0	199.9	5.64	25.7
		RT-PCR	1.7	2.2	4.3	2.5
X74987	Ribonuclease L inhibitor (RNASELI)					
		Microarray	1.4	1.4	1.3	1.3
		RT-PCR	24.5	2.9	3.0	5.9
D16611	Corporoporphyrinogen oxidase (CPO)					
		Microarray	3.1	8.1	10.1	6.3
		RT-PCR	2.0	4.0	1.6	2.4

[a]Each transcript failed to be detected by ANOVA, paired and two-sample *t*-tests, Significance Analysis of Microarrays, and Cyber-T (screening functions detailed in Web Supplement). All screening tools were tuned to detect >2-fold up-regulation in mean expression levels for mature versus immature T lymphocytes ($p < 0.05$, experiment-wide error rate = 0.25 for Cyber-T, false discovery rates up to 25% for Significance Analysis of Microarrays). Each transcript was identified by $PRIM_{Min}$ and $PRIM_{Mean}$ prediction rules tuned to identify >60% probability of >2-fold upregulation.
[b]Fold-increase: mature/immature T lymphocytes.
[c]Microarray fold-determination based on ratio of Affymetrix Microarray Suite average difference values.
[d]RT-PCR fold-determination based on ratio of mRNA quantities after normalization to GAPDH mRNA levels.
[e]IL3 transcripts were undetectable in all immature T lymphocyte samples. All mature T lymphocyte values were greater than 5000 copies IL3 per 10 000 GAPDH copies.

machine-learning algorithm PRIM (Friedman and Fisher, 1999). Friedman and Fisher developed PRIM to identify Boolean product prediction regions associated with exceptionally high values of a criterion variable (y). Predictions take the form,

$$\hat{y}_n = \begin{cases} 1 & \text{if } \boldsymbol{x}_n \in \{\cap_{p=1\text{-}P}(x_{np} \in [t_p^-, t_p^+])\}, \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

with n indexing the N observations, $\hat{y}_n$ denoting the rule's predicted value of y_n, and x_{np} representing one of P measured predictors ($\boldsymbol{x}_n$ representing the vector of all x values for observation n). t_p^- and t_p^+ denote lower and upper bounds on a continuous subspace of the predictor dimension p, and the Boolean conjunction of all P subspaces constitutes a hypercube segregating observations according to realized values of $\boldsymbol{x}_n$ (is $t_p^- < x_{np} < t_p^+$ for all P predictors?). PRIM iteratively optimizes t_p^- and t_p^+ to identify regions of $\boldsymbol{x}$ space associated with high values of y. Fitting begins with a 'prediction box' $\boldsymbol{b}^0$ containing all observations within the current t bounds for each of the P predictors. At step 1, PRIM removes some fraction α (e.g. 5%) of the observations from the lower end of predictor dimension 1 (i.e. shift t_1^- from its current value of the 0th percentile up to the αth percentile of x_1). This creates a new prediction box $\boldsymbol{b}_1^1$, and PRIM computes the mean y value for observations whose $\boldsymbol{x}_n$ vector falls within that box. This process is repeated for the lower bound of all other x dimensions, and for the upper bound of all x dimensions (i.e. depressing t_p^+ from its current value at the 100th percentile down to the $100 - \alpha$th percentile value of x_p), until there are $2P$ candidate boxes $\boldsymbol{b}^1$. Whichever candidate box produces the highest mean y value is adopted as the actual $\boldsymbol{b}^1$, and the entire process is repeated until (a) the number of observations within $\boldsymbol{b}$ becomes small (<10 in our implementation), or (b) the mean y for in-box observations cannot be further increased. This iterated 'peeling' of the initial box is then followed by a series of 'pasting' operations in which t_p^- and t_p^+ are expanded outward by α% of the in-box observations until the mean y drops below some pre-specified threshold c_{thresh}. The current box $\boldsymbol{b}$ is taken as the final PRIM rule predicting y from observed $\boldsymbol{x}$.

Treating reliability as a prediction problem (Snedecor and Cochran, 1989), we asked PRIM to generate rules forecasting a high probability of future change in gene g's expression (e.g. an 80% probability of >2-fold change) given existing observations of its absolute and relative change—the $ratio_{gr}$ and $difference_{gr}$ employed by visual analysts. Following classic cross-validation, we hold out a randomly chosen replicate as a criterion and ask PRIM to predict the

incidence of change in that replicate from data on the remaining replicates (e.g. hold out replicate z; set $y_g = 1$ if $ratio_{gz} >$ 2-fold and $y_g = 0$ otherwise; let $\boldsymbol{x}_g =$ the vector of all measured values of $difference_{gr}$ and $ratio_{gr}$ such that $r \neq z$; ask PRIM to identify the largest subspace of $\boldsymbol{x}$ such that Pr{$ratio_{gz} >$ 2-fold|$difference_{g1}$, $difference_{g2}, \ldots,$ $difference_{gz-1}$, $difference_{gz+1}, \ldots,$ $difference_{gR}$, $ratio_{g1}$, $ratio_{g2}, \ldots,$ $ratio_{gz-1}$, $ratio_{gz+1}, \ldots,$ $ratio_{gR}$} $>$ 80%). This approach is repeated for 10 random hold-out replicates (z values sampled with replacement), and a final rule is defined by the mean *difference* and *ratio* of the final box $\boldsymbol{b}$ from each analysis. Visual pattern recognition algorithms were found to be particularly sensitive to minimum change across replicates (Web Supplement, *Visual pattern analysis of differential expression*), so we also tested PRIM rules predicting change in the hold-out sample from the minimum *ratio* and minimum *difference* observed in other samples ($PRIM_{Min}$: Pr{$ratio_{gz} > 2|\min_j\{ratio_{gj}\}_{j=1 \text{ to } z-1 \text{ and } z+1 \text{ to } R}$, $\min_k \{difference_{gk}\}_{k=1 \text{ to } z-1 \text{ and } z+1 \text{ to } R}$). An example of the resulting PRIM rule might be, '*there is an 80% probability that expression of gene g will increase >2-fold in a future replicate IF(gene g showed a minimum change >2.3-fold AND a minimum change >756 fluorescence intensity units in already observed replicates)*.' A variant, $PRIM_{Mean}$, predicted change in the criterion sample from the Boolean product of the mean *difference* and the minimum *ratio*. (Taking the minimum *ratio* across replicates operates as a Boolean product ensuring that all replicate values of $ratio_g$ equal or exceed that value.) The Web Supplement includes further details (*PRIM models of perceptual mapping functions*) and a free-standing JAVA implementation is available at http://microarray.crump.ucla.edu/focus.

Monte Carlo analyses show that conjoint PRIM rules recover ~2 times as many differentially expressed genes as do conventional GLM analyses while maintaining Type I error rates below the nominal $p < 0.05$ (Table 1 and Supplemental Tables 2–4). In Monte Carlo analyses based on the breast cancer data, for example, $PRIM_{Mean}$ identified a mean of 389 up-regulated genes versus 87 recovered by t-tests and 205 by array-wide ANOVA. Similar increases in yield were observed in Monte Carlo analyses corresponding to each data set examined (Table 1, all differences $p < 0.001$). As a result of reduced Type II error, PRIM also produced lower total error rates (Type I + Type II) and increased predictive accuracy ($\kappa = 0.60$ versus 0.44–0.50 for GLM analyses; Table 1). PRIM-based indications of change also showed significantly higher replication rates than did GLM analyses in split-half reliability studies (Supplemental Tables 4 and 5). For example, among 34 transcripts identified by t-test as showing >2-fold up-regulation in the first three replicates of the breast cancer data, only 13 reappeared in analysis of the remaining replicates (38% replication). In contrast, 66 of 108 differences identified by $PRIM_{Mean}$ were replicated (61%). PRIM rules showed especially strong advantages with limited replicates (<8 per condition) and consistently outperformed GLM statistical models in total error rates for magnitudes of change ranging between >1-fold and >100-fold, for signal-to-noise ratios ranging between 1/10th and 5 times those empirically observed, and across experimental settings involving differing numbers of up-regulated genes (Table 1 and Supplemental Tables 2–4). PRIM rules even outperformed the ANOVA models used to generate the Monte Carlo data, underscoring the poor sensitivity of GLM analyses in typical microarray settings. RT-PCR studies verified that gene expression differences recovered by PRIM but overlooked by other methods represent true biological differences in mRNA expression (Table 2).

DISCUSSION AND CONCLUSION

The present studies show that PRIM-based conjoint prediction rules can substantially outperform conventional GLM statistical analyses in the high-noise limited-replicate setting of microarray differential expression screening. The GLM's difficulties stem from the fact that indications of *change* are quite noisy even when individual measurements of expression *level* are reliable (Fig. 1A and B). GLM methods are known to show poor sensitivity in detecting qualitatively consistent changes if they show great quantitative variability (e.g. 10-fold increase in one replicate and 100-fold in another, Fig. 1B) (Miller, 1986). As shown in Figure 1C, PRIM Boolean product rules are much more tolerant of quantitative inconsistency across replicates as long as the qualitative characteristics of change remain constant. In addition, PRIM rules mimic visual pattern recognition processes in using absolute change in signal intensity to modify confidence in predictions made on the basis of qualitatively consistent fold changes (Fig. 2A). Univariate GLM models do not take such information into account, and are thus prone to overpredicting the replicability of small raw changes and underpredicting the replicability of large ones (Fig. 2A). Another advantage stems from PRIM's ability to 'learn from the data' where to set reliability thresholds while GLM analyses rely upon statistical theories assuming a stable relationship between average change and its standard deviation (Fig. 2C). As a result, PRIM conjoint prediction rules routinely outperform GLM statistical analyses and microarray-adapted variants (e.g. SAM and Cyber-T) in analytic yield (transcripts recovered), total error rates (Type I + Type II), and experimental replication rates. RT-PCR studies confirm that differences detected by PRIM but overlooked by all other methods represent true biological changes, and that other methods' failure to detect such differences represent *bona fide* Type II errors.

These data join previous studies in documenting high sampling variability in microarray-based measures of *change* despite reliable measurement of individual observations (Dudoit *et al*., 2000; Lee *et al*., 2000; Long *et al*., 2001;

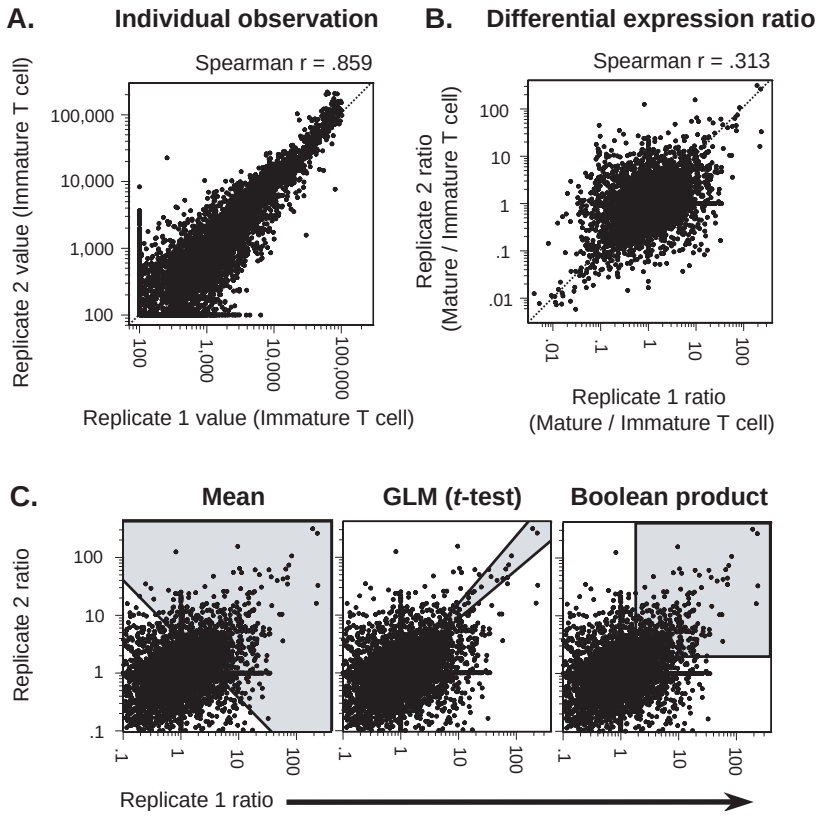

Fig. 1. Stability microarray measurements and its impact on screening. (**A**) Reliability plot for two replicate observations from the control condition of the T cell development study. Each data point represents one of 7070 transcripts assayed by Affymetrix high density oligonucleotide arrays, with values floored at 100 to reduce spurious variability (dashed diagonal = perfect reliability, realized reliability assessed by rank correlation). Results are representative of all data sets examined (median replicate CV ranged from 6 to 8%). (**B**) Reliability of $\log_{10}$-transformed differential expression ratios from the T cell development study (X = ratio from replicate 1, Y = ratio from replicate 2). Transcripts on the dashed diagonal show quantitatively similar change across replicates. Points on the '+' perpendiculars show substantial change in one replicate and negligible change in another. Note that it is not uncommon for transcripts to show 3- to 10-fold up-regulation in one experimental replicate and 3- to 10-fold down-regulation in another (points in the upper left and lower right quadrants). Comparable results were observed in all data sets examined, with median CV of replicate change measurements ranging from 77 to 168%. (**C**) Capture regions for alternative microarray screening tools are compared to empirically observed change ratios. Screens based on an *average* 2-fold change across replicates capture large numbers of unreliable results ('+' pattern and upper left or lower right quadrants). A 95% GLM confidence interval for mean change >2-fold (paired-*t* test on log-transformed expression values) stringently excludes genes showing qualitatively consistent change of variable magnitude (e.g. 10-fold increase in one replicate and 30-fold increase in another). PRIM Boolean product rules capture genes showing qualitatively consistent change of variable magnitude (upper right quadrant) while rejecting those that show qualitatively inconsistent change ('+' pattern and points in the upper left and lower right quadrants). PRIM bounds estimate 60% probability of >2-fold change in a future replicate.

Newton *et al*., 2001; Tusher *et al*., 2001). In each of the data sets examined here, more than 50% of the genes surveyed showed coefficients of variation in gene induction that exceed 100% (standard deviation of replicate change scores greater than their mean). Analyses beyond the scope of this presentation show that much of this variability stems from systematic differences across individuals in basal gene expression and response to experimental manipulations. Despite the biological noise in gene-induction measurements, the present studies reveal a low rate of Type I error for both gene-specific and array-wide GLM analyses (Table 1 and Supplemental Tables 2–4). Standard calculations suggest that parallel 95% confidence intervals for a 7000 gene microarray should produce an average of 350 spurious results at a $p < 0.05$. However, such calculations assume that no genes show any true change in expression. In empirical data, the magnitude of true change varies and standard calculations may overestimate actual Type I error rates. Given the change distributions of typical microarray data, the present Monte Carlo studies show that the incidence of Type I error is actually closer to 1%

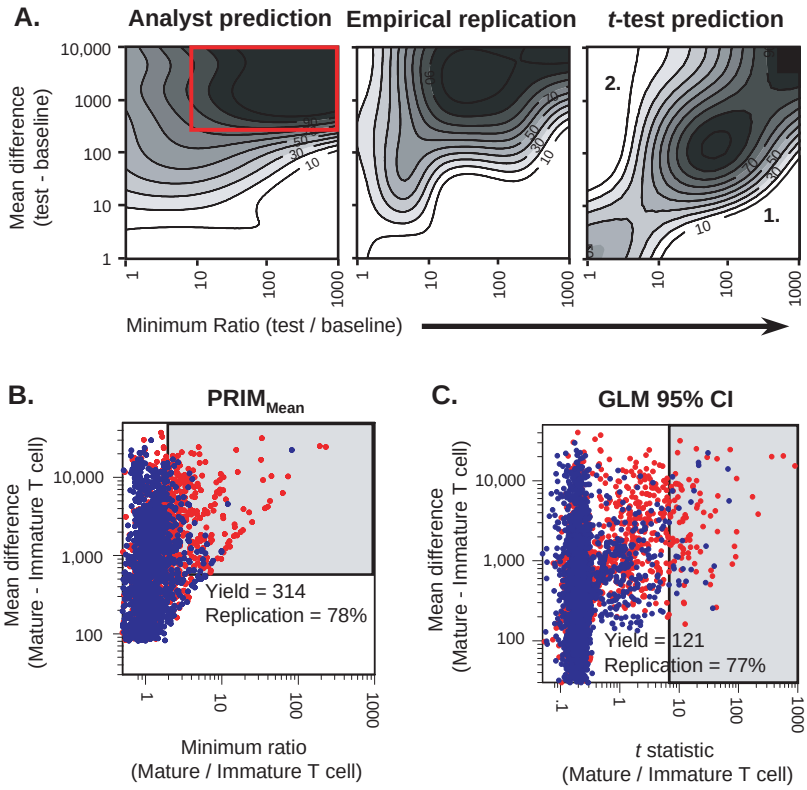

Fig. 2. Conjoint inference rules based on raw- and fold-change. (**A**) Experienced microarray data analysts estimated the likelihood that a subsequent replicate experiment would show >3-fold increase in expression based on profile-plot displays of two replicate observations on each of 231 genes assayed in the T cell development study. Raw- and fold-change were parametrically varied, and analysts' estimated probability of >3-fold increase in a subsequent replicate were mapped according to the average difference and minimum change ratio in observed replicates (contours = 5% increments in forecast probability). Analysts' forecasts (left panel) show a similar topography to actual probabilities of >3-fold change as measured in a held-out test sample (center panel). In contrast, *t*-tests (right panel) substantially underestimate the probability of future >3-fold induction for large mean differences with high replicate variability (region 2, Type II errors) while over-predicting replication of small mean differences (region 1, Type I errors). The rectangular region in the left panel compares a PRIM-derived prediction rule (60% replication) with analysts' estimates. (**B**) The structure of a PRIM conjoint inference rule is displayed for the full T cell development data set. Each point displays the minimum ratio change (*X*-axis) and mean raw change (*Y*-axis) for one of 7070 transcripts measured in two experimental replicates. Points are colored red if they showed >2-fold increase in a third hold-out criterion replicate and blue otherwise. PRIM rules identify regions in which the ratio of red to blue points exceeds a specified criterion (e.g. 60%). Such prediction bounds capture greater numbers of differentially expressed transcripts ('Yield') than do GLM statistical methods such as the paired *t*-test (**C**). Similar results emerged for independent-sample *t*-tests and array-wide ANOVA (Table 1, and Supplemental Tables 2–4).

at the nominal $p < 0.05$. It should be noted that we follow the statistical literature in defining Type I error rates as the number of falsely declared differences among genes showing no true difference. Microarray analysts have sometimes defined 'false positive' error rates as what statisticians term the 'false discovery rate'—the number of genes showing no true difference as a fraction of total declared differences. Using the later definition, both GLM analyses and PRIM rules produce false discovery rates in the range of 5–15% [consistent with previous reports (Lee *et al*., 2002)].

Despite low false positive error rates, an average 60% of GLM-declared differences fail to replicate in parallel analyses of data from the same experiment (Supplemental Tables 5 and 6). The explanation for this paradox lies in the high incidence of Type II error. Because each GLM analysis fails to identify 60–70% of true differences, the specific result set that is recovered can show substantial sampling variability. Replication failures thus stem from poor detection of true differences, rather than from high rates of false positives. Low Type I error rates should console microarray researchers who have sought increasingly complex methods for reducing false discovery (Tusher *et al*., 2001). However, high 'false negative' error rates suggest that alternative screening tools should be considered when scientific objectives require the most comprehensive recovery of differentially expressed genes (e.g. network mapping and gene hunting).

Conventional statistical analyses build a structural model of the data as a basis for replication inference (e.g. Equation 1). In contrast, the machine learning approach pursued here seeks an empirical model of replication likelihood based on prognostic features identified by human visual data analysts. This approach is not model-free, but descends from a distinct lineage of set-theoretic models describing cognitive processes involved in pattern recognition and similarity judgments (Tversky, 1977; Gati and Tversky, 1982). Cognitive heuristic models often outperform linear statistical models in the analysis of noisy data with limited replicates—the circumstances for which human pattern recognition is optimized (Kahneman *et al*., 1987). In the conjoint data representation utilized here, raw differences serve as a 'technical' screen to discriminate meaningful changes from assay noise, and relative change discriminates substantively significant results from weak perturbations of highly expressed transcripts. Arbitrary versions of such 'difference and ratio' rules have been employed in some previous studies (Hakak *et al*., 2001; Dadgostar *et al*., 2002) and a similar idea underlies 'pre-filtering' for cluster analysis. The PRIM approach establishes a clear statistical foundation for such screening rules by identifying capture bounds that maintain a desired level of reliability for a substantively specified magnitude of change (e.g. 80% likelihood of >2-fold increase). The resulting rules are easily interpreted by non-statisticians (Friedman and Fisher, 1999) and adapt automatically to differing microarray platforms (e.g. spotted cDNA arrays versus Affymetrix GeneChips), alternative experimental designs (e.g. controlled experiments in a single cell line versus observational studies of heterogeneous patient samples), and variations in assay noise (e.g. differences in reagent batches, hybridization conditions, etc.). PRIM rules also admit the inclusion of any measure of data reliability that may be of interest to researchers, such as the standard deviation of change scores, t statistics or F ratios, intra-condition coefficients of variation, and minimum or maximum expression values. Although the present studies have focused on two-group comparisons, the PRIM approach can also be extended to more complex experimental designs through conjoint analysis of contrast values, regression slopes, or dispersion indices (Miller, 1986; McCullagh and Nelder, 1991).

The present results tend to support empirical observations that microarray analyses overlook many differentially expressed genes (Richmond *et al*., 1990; Newton *et al*., 2001; Taniguchi *et al*., 2001; Lee *et al*., 2002; Yuen *et al*., 2002). Our results suggest that such problems stem mainly from sub-optimal analytic strategies rather than insufficiencies in microarray measurement technology per se. As a result, it may be possible to re-screen existing data sets using more sensitive analytic algorithms if maximal recovery of differentially expressed genes is a scientific objective. These results also suggest that Type II errors will require more attention in gene network mapping to avoid underestimating network connectivity. The PRIM conjoint screening approach described here provides an initial step, and further research is needed to improve network 'link-definition' algorithms.

ACKNOWLEDGEMENTS

This research was supported by the University of California University wide AIDS Research Program (K99-LA-030), the Norman Cousins Center at UCLA, the James L. Pendelton Charitable Trust, and the National Institute of Allergy and Infectious Diseases (AI 33259, AI 49135, AI 52737). We thank C. Denny, T. Symensma, J. Fahey, M. Kemeny, S. Horvath and the UCLA Microarray Users' Group for their insightful discussions, and we gratefully acknowledge the technical assistance of A. Kacena, R. Dennis, M. Hazel, W. Wachsman, the San Diego Veterans' Administration Microarray Core, the UCLA Gene Expression Core, and the Interactive Media Lab at the Crump Institute for Biological Imaging.

REFERENCES

Alter,O., Brown,P.O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Bratley,P., Fox,B.L. *et al*. (1983) *A Guide to Simulation*. Springer Verlag, New York.

Dadgostar,H., Zarnegar,B. *et al*. (2002) Cooperation of multiple signaling pathways in CD40-regulated gene expression in B lymphocytes. *Proc. Natl Acad. Sci. USA*, **99**, 1497–1502.

Dudoit,S., Yang,Y.H. *et al*. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Berkeley, University of California, Berkeley, Department of Statistics.

Eisen,M.B., Spellman,P.T. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Friedman,J.H. and Fisher,N.I. (1999) Bump hunting in high-dimensional data. *Stat. Comput.*, **9**, 123–143.

Gati,I. and Tversky,A. (1982) Representations of qualitative and quantitative dimensions. *J. Exp. Psychol.: Human Perception and Performance*, **8**, 325-340.

Hakak,Y., Walker,J.R. *et al*. (2001) Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proc. Natl Acad. Sci. USA*, **98**, 4745–4751.

Harary,F. (1969) *Graph Theory*. Addison-Wesley, Reading, MA.

Kahneman,D., Slovic,P. *et al*. (1987) *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Kerr,M.K., Martin,M. *et al*. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Lee,M.T., Kuo,F.C. *et al*. (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Lee,T.I., Rinaldi,N.J. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Long,A.D., Mangalam,H.J. *et al*. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *J. Biol. Chem.*, **276**, 19937–19944.

McCullagh,P. and Nelder,J.A. (1991) *Generalized Linear Models*. Chapman & Hall, London.

Miller,R.G. (1986) *Beyond ANOVA: Basics of Applied Statistics*. Wiley, New York.

Newton,M.A., Kendziorski,C.M. *et al*. (2001) On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Perou,C.M., Sorlie,T. *et al*. (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.

Richmond,C.S., Glasner,J.D. *et al*. (1990) Genome-wide expression profiling of *Escherichia coli* K-12. *Nucleic Acids Res.*, **27**, 3821–3835.

Snedecor,G.W. and Cochran,W.G. (1989) *Statistical Methods*. Iowa State University, Ames Iowa.

Tamayo,P., Slonim,D. *et al*. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Taniguchi,M., Miura,K. *et al*. (2001) Quantitative assessment of DNA microarrays—comparison with Northern blot analyses. *Genomics*, **71**(34–39).

Tusher,V.G., Tibshirani,R. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Tversky,A. (1977) Features of similarity. *Psychol. Rev.*, **84**, 327–352.

Winer,B.J. (1971) *Statistical Principles in Experimental Design*. McGraw-Hill, New York.

Yuen,T., Wurmbach,E. *et al*. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.*, **30**, e48.