

A systematic statistical linear modeling approach to oligonucleotide array experiments

Tzu-Ming Chu ^{a,c}, Bruce Weir ^{a,b}, Russ Wolfinger ^{b,c,*}

^a *Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA*

^b *Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA*

^c *SAS Institute Inc., Cary, NC 27513, USA*

Received 2 May 2001; received in revised form 14 June 2001; accepted 20 November 2001

Abstract

We outline and describe steps for a statistically rigorous approach to analyzing probe-level Affymetrix GeneChip data. The approach employs classical linear mixed models and operates on a gene-by-gene basis. Forgoing any attempts at gene presence or absence calls, the method simultaneously considers the data across all chips in an experiment. Primary output includes precise estimates of fold change (some as low as 1.1), their statistical significance, and measures of array and probe variability. The method can accommodate complex experiments involving many kinds of treatments and can test for their effects at the probe level. Furthermore, mismatch probe data can be incorporated in different ways or ignored altogether. Data from an ionizing radiation experiment on human cell lines illustrate the key concepts. © 2002 Published by Elsevier Science Inc.

Keywords: Microarray; Mixed model; Split plot; Probe

1. Introduction

The GeneChip from Affymetrix is currently a popular commercial oligonucleotide technology for studying gene expression. Although scientific progress with this remarkably miniaturized platform has been considerable, many subtle issues remain regarding the proper analysis and interpretation of the data it produces. Many investigators struggle with such issues as the

* Corresponding author. Address: Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA. Tel.: +1-919 531 6695.

E-mail address: russ.wolfinger@sas.com (R. Wolfinger).

reliability of a ‘present’ or ‘absent’ call, proper incorporation of mismatched probe information, and appropriate comparisons across many chips. In this paper, we outline a systematic approach for handling data generated from GeneChip experiments, with a particular emphasis on statistical model selection and gene significance testing.

The GeneChip contains probe sets representing unique genes, and each probe set consists of 20 probe pairs. Each probe pair consists of a perfect match (PM) oligonucleotide probe, which is designed exactly complementary to a preselected 25mer of the target gene, and a mismatch probe (MM), which is identical to PM except for one single nucleotide difference at position 13. According to Lockhart et al. [9], the purpose of the mismatch probe is to serve as an internal control of hybridization specificity. Affymetrix provides basic software to summarize the expression information of the probe set measurements by averaging the difference or log ratio of PM and MM after deleting those extreme measurements which exceed three standard deviations from the mean. Schadt et al. [12] address many of the important issues and provide useful extensions to Affymetrix’s methods.

While summary methods for one or two chips are certainly useful, a statistically optimal approach for experimental data involving many chips requires that we consider all PM and MM data simultaneously. This provides 40 times the data compared to traditional summary methods and gives more power for statistical inference. However, several questions arise regarding the statistical relationships of PM and MM within a probe pair, between probe pairs within a probe set, and between probe sets across arrays. Do they have a linear relationship? Are the amounts of cross-hybridization similar for PM and MM probes? To what extent does MM serve as an internal quality control? Li and Wong [6] investigate these and other questions in the context of a multiplicative model for the measurements, whereas Efron et al. [2] consider scaled logarithms. Lazardis et al. [5] suggest using PM information only. In this paper, we draw on the rich tradition of statistical linear models and propose methods for potentially complex experiments involving many chips. Our methods are also related to Kerr et al. [4] and Wolfinger et al. [16] that propose analysis of variance (ANOVA) applying on cDNA spotted microarrays. The former uses only fixed effects and has all genes being incorporated in one large model. The latter suggest a two-step mixed models to normalize data in array level and, then, to analyze the residuals from first model by forming ‘single-gene’ mixed model for finding significant differential genes and extracting those interesting effects. Our methods are closer to Wolfinger et al. [16]. We apply regression to normalized data and propose a template mixed model for each single gene for further analysis.

In the next section, we outline and discuss a step-by-step approach to handling oligonucleotide array data. We then use the ionizing radiation response data from Tusher et al. [14] as an example to illustrate these steps and provide a detailed comparison to results of their ‘significance analysis of microarrays’ (SAM) method.

2. Analysis steps

2.1. Identify the experimental design

Prior to any formal data analysis, an early and detailed understanding of the statistical experimental design is crucial for maximizing information gain from the data. This entails identification of all real and potential effects impacting both the location and dispersion of the data,

how these effects interrelate, and how they affect the experimental units. The effects most commonly of interest are those changed experimentally, and can involve treatment and genotype (cell line) effects. In addition to the experimental factors, the designs we consider also include broad effects on entire arrays and probe-specific effects for each gene. Experimental design has a long and successful history in the statistics literature, and we employ traditional designs such as the split plot used in agricultural field trials; refer to [13].

2.2. *Extract numerical data from the image*

This is obviously the first and one of the most critical steps to properly investigating array data. For sake of brevity and emphasis in this paper, we assume that this step has been completed in a satisfactory fashion and those reliable numerical intensities corresponding to each PM and MM probes are available.

2.3. *Formulate and fit a statistical model*

This is the key step in our approach and requires careful consideration. The goal is to derive a statistical model that adequately accounts for all aspects of the experimental design yet is simple enough to be interpretable. Doing this allows the researcher to make rigorous quantitative assessments about effects influencing the data that properly separate true signals from experimental and biological noise. As a reasonable starting point, we recommend the classical mixed linear model as a suitably flexible framework; refer to [8,10,15], for theoretical background, examples, and references.

An important initial decision in formulating a linear model involves determining precisely what data values will be modeled for each gene. This usually involves a transformation to make the statistical modeling assumptions reasonable and an adjustment for gross chip-wide effects. As a default method, we recommend a log base 2 (\log_2) transformation for individual PM and MM measurements. If PM–MM differences are desired, accommodations must be made for negative values before applying the log transform. Since our proposed statistical model will be additive, using a log transformation on the response can be interpreted as fitting a multiplicative model on the original scale [6], and resulting statistical estimates are interpretable as fold changes. To adjust for gross array-level effects, we also recommend centering the logged values so that they have mean 0. We find that this can be the simplest and reasonable way for normalization (Fig. 1). However, this adjustment involves an assumption that the averaged logged expression level is the same within chip. More complicated smoothing spline normalizations are also possible [1,7,12,17], although we are not sure they add much value if all of the experimental data for a gene are considered together.

Once a response value is chosen, additive analysis-of-variance effects are specified to partition its variability. In the mixed model setting, one must decide whether these effects are ‘fixed’ or ‘random’. Fixed effects are those effects with a well-defined, finite number of levels and only these finite levels are of interest in the experiment. For fixed effects, we estimate each level and do testing among all levels or comparisons between levels to see if they are significant. Random effects are those effects considered to be drawn from an infinite population having some probability distribution, usually normal. For random effects, we estimate the parameters of this probability distribution (variance components in the normal case) and possibly also individual effect estimates

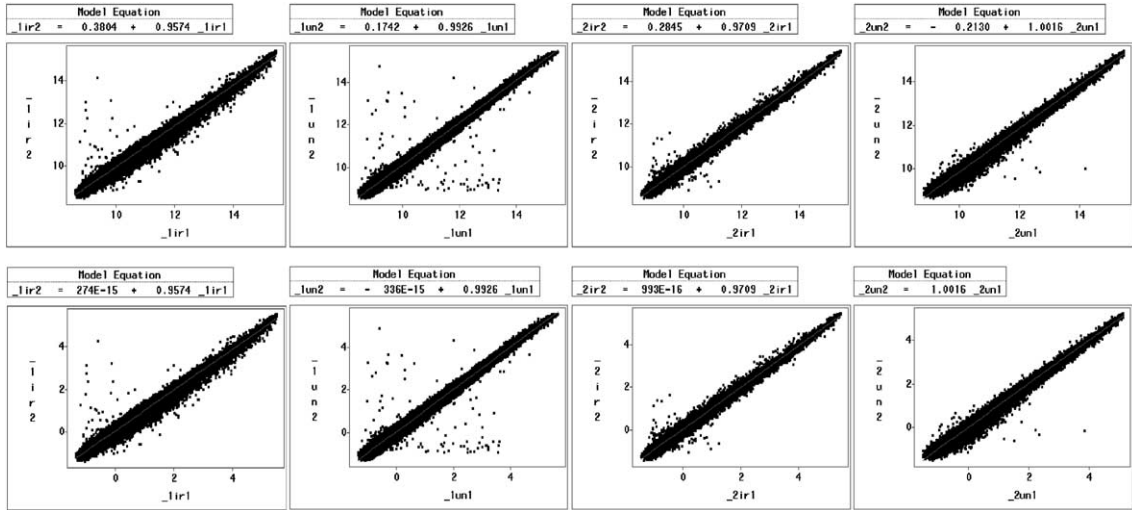


Fig. 1. Scatter plots of four experimental effects between two replicate arrays before (top 4) and after (bottom 4) standardization. Here, each X or Y variable identifies those eight arrays. ‘_1ir2’ indicates the array was applied by cell line I with irradiation treatment in replicate array II.

properly shrunken towards zero. Inclusion of random effects also allows inferences about the fixed effects to be made to broader populations.

For GeneChip experiments, we typically consider cell line, treatment, and probe effects to be fixed, and because of potentially complex experimental sources of variation such as cross-hybridization, it is typically sensible to include two-way interactions of these effects as well. Effects impacting arrays can be considered random, reasoning that they are the accumulation of small experimental sources of noise. Putting these all together, the following linear mixed model serves as an initial template for the data from a single gene:

$$Y_{ijkl} = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + \varepsilon_{ijkl}. \quad (2.1)$$

Here, Y_{ijkl} is the transformed and centered expression measurement of the i th cell line applying the j th treatment at the k th probe in the l th replicate. Y can be the centered $\log_2(\text{PM})$ measurements if we do not wish to incorporate any MM information, or Y can be the centered \log_2 differences of PM–MM pairs (suitably adjusted for negative values) if we believe that MM serves directly as an additive internal control on the original scale. A somewhat intermediate position explored by Efron et al. [2] is to let Y take the form $\log(\text{PM}) - 0.5 \log(\text{MM})$, and this can be used directly or generalized by including $\log(\text{MM})$ as a covariate in the right-hand side of the model.

The symbols L , T , LT , P , LP , TP and A in (2.1) represent cell line, treatment, cell line–treatment interaction, probe, cell line–probe interaction, treatment–probe interaction, and array effects, respectively. The $A_{l(ij)}$ s are assumed to be independent and identically distributed normal random variables with mean 0 and variance σ_a^2 . The ε_{ijkl} s are assumed to be independent identically distributed normal random variables with mean 0 and variance σ^2 , and are independent of the $A_{l(ij)}$ s. We will elaborate on these effects and on variations of the model in the context of our example in the next section. For fitting the model, standard maximum likelihood methods are usually best and can be accessed through software like Proc Mixed [11].

2.4. Check assumptions, remove outliers, reformulate and refit the model if necessary

Because we make probabilistic assumptions in the preceding model, it is wise to perform some diagnostic checking on results of the model to verify that it adequately represents the data. In (2.1), the randomness in each observation Y_{ijkl} is represented by two terms, $A_{l(ij)}$ and ε_{ijkl} . According to our normality assumptions in (2.1),

$$A_{l(ij)} + \varepsilon_{ijkl} \sim N(0, \sigma_a^2 + \sigma^2), \quad (2.2)$$

$$\text{Cov}(A_{l(ij)} + \varepsilon_{ijkl}, A_{l'(i'j')} + \varepsilon_{i'j'k'l'}) = \begin{cases} \sigma_a^2 + \sigma^2 & \text{if } (i, j, k, l) = (i', j', k', l'), \\ \sigma_a^2 & \text{if } (i, j, l) = (i', j', l') \text{ but } k \neq k', \\ 0 & \text{otherwise.} \end{cases} \quad (2.3)$$

A standard way of checking these assumptions is to examine the residuals from the fitted model. The nature and definition of residuals are more complicated in mixed models than in standard linear models because there are multiple sources of random error. For simplicity, and as a first step, we recommend inspecting residuals formed by subtracting the fitted fixed effects from the observed data and then standardizing these values by an estimate of their variance, $\sigma_a^2 + \sigma^2$. This allows the residuals from all genes to be plotted as groups or together; see Fig. 2 in the next section. Model departures are often apparent in such plots by a non-random scatter around the zero horizontal line. Standardized residuals with magnitude larger than three are potential outliers and can be eliminated from the analysis. Li and Wong [6] discuss systematic ways to eliminate outliers using a multiplicative model, and we have work underway in our similar log-additive context.

2.5. Perform basic statistical inference and filter out insignificant genes

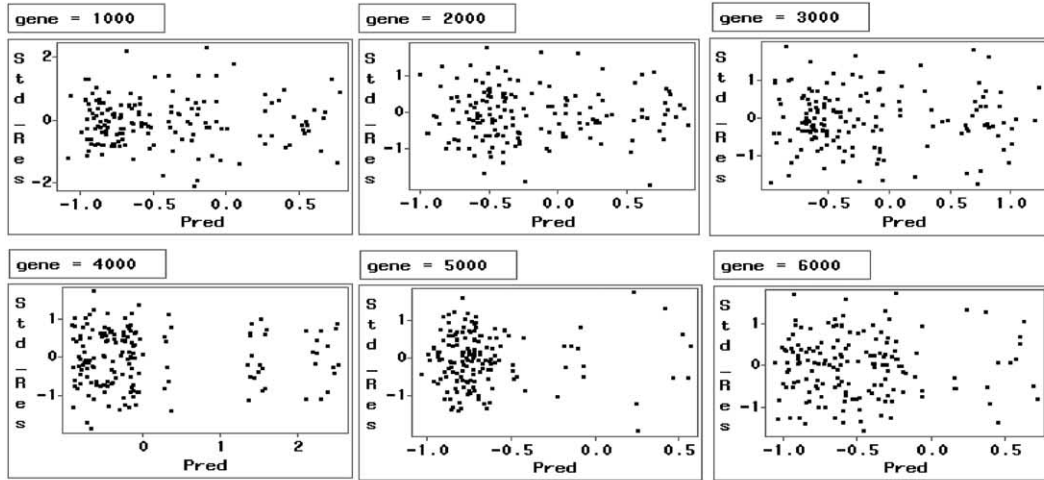
After suitable model determination and validation, an investigator can study the results of the fitted model for each gene. General statistical tests for the individual fixed effects in the model are available, as are custom estimates of the fold change and significance for specific treatment comparisons. A useful graphical display is the ‘volcano plot’, which plots negative log (base 10) p -values on the y -axis versus estimated \log_2 fold change on the x -axis; see Fig. 4 in the next section and [3,16] for examples. The plot takes the shape of a ‘V’ because larger fold changes tend to be more significant, although usually the most significant genes do not exhibit the greatest fold change.

A simple procedure for statistically filtering genes is to draw a horizontal cutoff line on a volcano plot to represent a desired false positive rate for the test under consideration. Genes corresponding to points below this line are not considered in subsequent analyses. This method can produce dramatically different results from filtering genes on the basis of fold change alone.

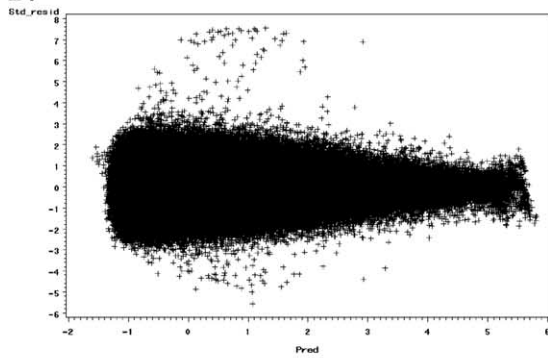
2.6. Perform additional analyses of statistically filtered data

Again, for sake of brevity, we do not elaborate or discuss this step, which usually involves analyses such as clustering and principal components, except to say that appropriate statistical filtering and inference must be done prior to this step to help ensure its validity.

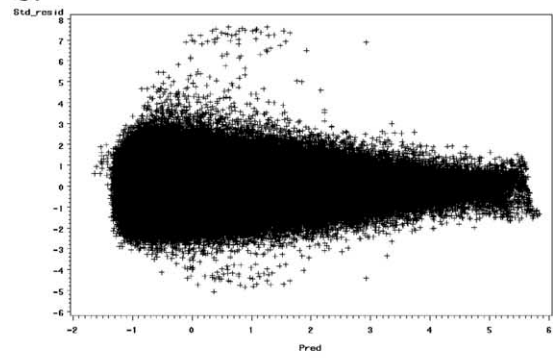
A.



B.



C.



D.

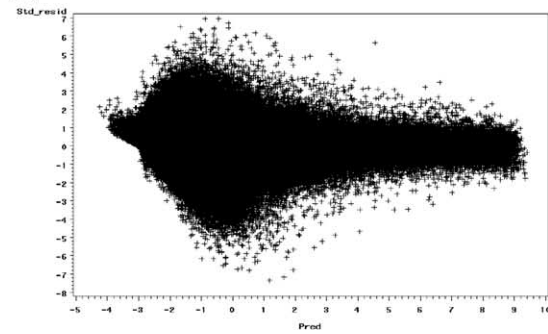


Fig. 2. (A) Standardized residual plots for gene 1000, 2000, 3000, 4000, 5000, and 6000 from Model I. (B–D) ‘Submarine plots’ of standardized residuals of all genes from Models I, II, and III.

3. Ionizing radiation example

We now illustrate the preceding steps using the ionizing radiation data from Tusher et al. [14]. The data arise from eight Affymetrix GeneChips with 7129 probe sets (genes) in each and were designed to study transcriptional responses of human cells to ionizing radiation.

3.1. Identify the experimental design

There are two experimental effects, treatment and cell line, with two levels each, and two replicate arrays for each effect combination. At the array level, this is a 2×2 experiment with 2 replicates. Combined with 20 probes for a probe set, there are four factors as stated below.

1. Two levels of radiation treatment (irradiated, unirradiated).
2. Two levels of cell line (line I and line II).
3. Twenty PM–MM probe pairs in a probe set (P1–P20).
4. Two replicate arrays (array I and array II).

This results in a total of 160 observations for each gene, and Table 1 shows an example design layout. This is a split plot design (refer to [8,13]), and the whole plot units are the arrays. Radiation treatment and cell line are the whole-plot effects, and probe is the sub-plot effect.

3.2. Extract numerical data from the image

As indicated in the previous section, we are skipping this step and assuming that reliable numerical data are available.

3.3. Formulate and fit a statistical model

The linear mixed model is a ‘perfect match’ for data arising from a split-plot design. As is common practice, we consider both the whole- and sub-plot effects as fixed and the whole-plot

Table 1
Design layout (within a gene)—There are four experimental effects (two lines and two treatments) applied on eight whole plot units (probe sets) below and each whole plot unit has 20 sub-plot units (probes)

	Treatment I (irradiated)				Treatment II (unirradiated)			
<i>Replicate A I</i>								
Line I	P1	P2	...	P20	P1	P2	...	P20
Line II	P1	P2	...	P20	P1	P2	...	P20
<i>Replicate A II</i>								
Line I	P1	P2	...	P20	P1	P2	...	P20
Line II	P1	P2	...	P20	P1	P2	...	P20

experimental units (arrays) as random. Working from the basic model template described previously in (2.1), we consider the following three models:

Model I:

$$\log_2(\text{PM}_{ijkl}) = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + \beta \log_2(\text{MM}_{ijkl}) + A_{l(ij)} + \varepsilon_{ijkl}. \quad (3.1)$$

Model II:

$$\log_2(\text{PM}_{ijkl}) = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + \varepsilon_{ijkl}. \quad (3.2)$$

Model III:

$$\log_2(D_{ijkl}) = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + \varepsilon_{ijkl}. \quad (3.3)$$

In Model I, we use array-centered $\log_2(\text{PM})$ as the response variable and array-centered $\log_2(\text{MM})$ as a covariate, with β as its coefficient. In Model II, we only use the PM probe data and no mismatch probe data. In Model III, we use the centered \log_2 of the difference (D) of the PM–MM pair as the response variable, and before doing the log transformation on D , we truncate those measurements less than 10 to 10 to avoid the negative value problem. Over one third (38.5%) of the data are truncated in this fashion.

3.4. Check assumptions, remove outliers, reformulate and refit the model if necessary

As a quick check on data standardization requirements, Fig. 1 plots pair of replicated \log_2 intensities for the four combinations of cell line and irradiation treatment. The fitting regression equations are shown above each plot. The R^2 values defined in (3.4) from left to right are 0.9794, 0.9801, 0.9879, 0.9819 and are not changed by standardization. The linearity of the plots indicates that a simple mean standardization is suitable. Those points far from diagonal line indicate that there may be undesirable outliers in data, but we retain them for subsequent comparison purposes.

Using Model I as an illustrative example, Fig. 2(A) displays standardized residual plots of genes 1000, 2000, 3000, 4000, 5000, and 6000. These plots exhibit random scatter around the zero horizontal line and indicate no significant departures from model assumptions. Fig. 2(B)–(D) plot all of the standardized residuals together for the three models in what we have nicknamed a ‘submarine plot’. These ‘submarine plots’ are not regular residual plots, as the residuals are pooled from different models. However, they can be useful for observing genome-wide features of gene variability. The ‘bubbles’ in these plots represent potential outliers with large positive or negative standardized residuals. A rough rule of thumb is to eliminate observations with standardized residuals having magnitude larger than 3, but doing this results in little to no changes in the most highly significant genes considered later. For subsequent comparison purposes, therefore, we filter no outliers. The absence of points in the lower left portion of Fig. 2(D) is an artifact of the truncation at 10 rule we used to analyze the data in Model III.

Also, it is interesting that while residual plots of single gene show no evidence of pattern, the ‘submarine plot’ suggests heteroscedacity. We try to test for normal distribution by Kolmogorov–Smirnov statistic with Bonferroni’s correction on standardized residuals for each gene excluding those residuals larger than 3 or less than -3 . Almost all genes (98.51%) pass the normal distribution test in Model I. Background correction may be a reason for observing ‘submarine’ which

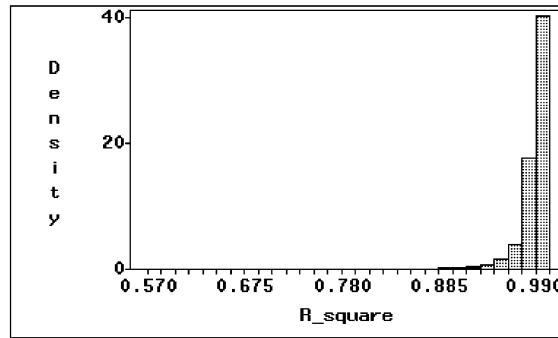


Fig. 3. Histogram of R^2 values from Model I for all genes excluding 160 genes that have 1/5 data missing.

has bigger head (larger dispersion for small measurement) and small tail (smaller dispersion for large measurement). According to Affymetrix's algorithm, the probe intensities are calculated by averaging out the pixel intensities and subtracting the background correction term, which averages out the lowest 2% pixel intensities. Those small measurements are expected to be more sensitive to background corrections than are large measurements.

For a quick assessment of goodness-of-fit, we calculate R^2 values of each gene and draw a histogram for Model I in Fig. 3. The definition of R^2 in the mixed model is somewhat ambiguous, so here we apply an ordinary model concept of R^2 , and define it as

$$R^2 = 1 - \frac{\sum R_{ijkl}^2}{\sum (Y_{ijkl} - \bar{Y})^2}, \quad (3.4)$$

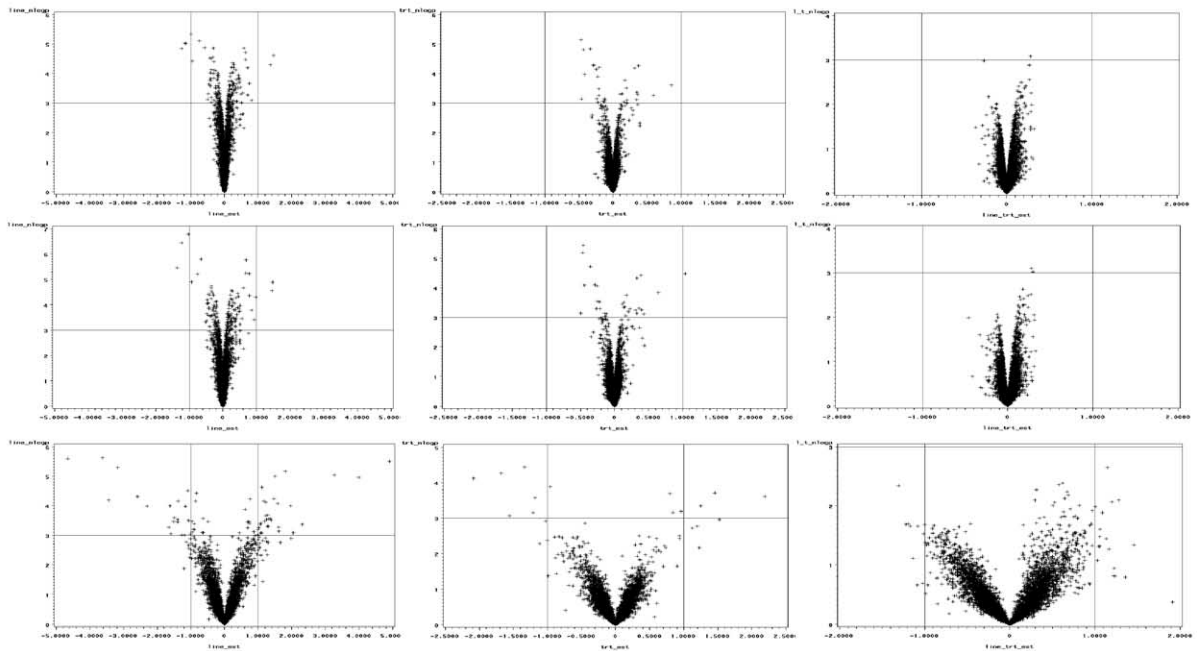
where \bar{Y} is the average of all Y_{ijkl} s. and R_{ijkl} is the residual term of Y_{ijkl} from the model. This histogram excludes those 160 genes having 1/5 of probe data missing. The first percentile is 0.86, indicating that Model I fits the data from almost all genes very well.

3.5. Perform basic statistical inference and filter out non-significant genes

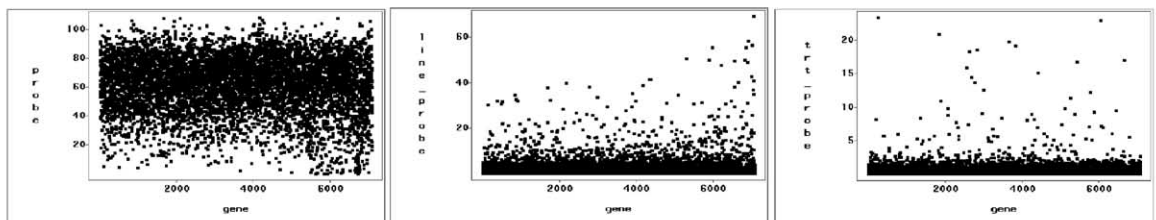
We first consider significance tests for all of the fixed effects in Fig. 4. The 'volcano' plots of Fig. 4(A) illustrate the relationship of the estimates of three whole-plot effects (cell line, treatment, cell line–treatment interaction) and their significance compared among all three models for all genes. The horizontal lines indicate p -values equal to 0.001 and the vertical lines indicate two-fold changes of effect estimates. In Fig. 4(A), we can see that there are only a few points outside the vertical lines and many points above the horizontal lines, especially, in the cell line and treatment volcano plots. The small magnitudes of the fold change estimates are remarkable and appear to represent excellent sensitivity in this experiment; see also the fold-change estimates in Tables 2 and 3. In addition, estimates from Model III are much more variable, indicating that direct subtraction of MM adds a lot of noise to the PM data.

The three plots in Fig. 4(B) show the significance of probe, cell line–probe interaction, and treatment–probe interaction effects for Model I, each of which results from F -tests with (19, 94) degrees of freedom. Results from Models II and III are similar. In Fig. 4(B), the probe main effects are highly significant, and even after a Bonferroni adjustment, 7088 (99.42%) are significant

A.



B.



C.

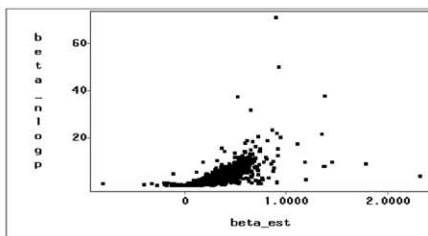


Fig. 4. (A) ‘Volcano’ plots of cell line, treatment, and cell line–treatment interaction effect among Models I (top row), II (middle row), and III (bottom row) for all genes. The x -axis is the \log_2 estimate of the difference of levels in the effect. The y -axis is the negative $\log p$ -value. The horizontal blue lines indicate the testing p -value equal to 0.001 and the vertical red lines indicate two-fold change of effect estimates. (B) Significance plots of the probe and its interaction effects applying Model I for all genes. The same plots applying Models II and III are similar to these three plots. The x -axis is the gene number. The y -axis is the negative $\log p$ -value. (C) Significance plot of the covariate effect applying Model I. The x -axis is the estimate of the parameter, β of $\log_2(\text{MM})$ covariate.

at the 5% level. This indicates huge variability in probe effectiveness. There is also evidence of some large interactions with the probe effect and a few examples are depicted in Fig. 5. In Fig. 5(A), PM curves (solid) show significant different in probes 9 and 14 comparing top (cell line I) and bottom (cell line II) rows; whereas, MM curves (dotted) show significant different in probe 9. In Fig. 5(B), PM curves (solid) show significant different in probes 11, 12, 13, and 14 comparing top (irradiation treatment) and bottom (unradiation treatment) rows; whereas, no significant different in MM curves (dotted). After a Bonferroni adjustment, 666 (9.34%) and 42 (0.59%) of the genes have significant cell line–probe and treatment–probe interactions, respectively. Causes for these interactions such as cross-hybridization may warrant further investigation.

Table 2
Most significantly induced genes

Gene no.	Accession no.	Rank				N log			Fold	Gene description
		Tusher	I ^a	II ^b	III ^c	I ^a	II ^b	III ^c	I ^a	
2863	U18300 ^d	4	1	2	6	4.26	4.43	3.16	1.31	p48, xeroderma pigmentosum group E gene
6684	X83490 ^d	2	2	3	4	4.19	4.33	3.34	1.26	Fas (alternate splice deleting exins 3 & 4)
2357	M92424 ^d	18	3	5	9	3.78	3.76	2.71	1.14	mdm2
2715	U09579 ^d	1	4	1	3	3.61	4.48	3.60	2.06	p21
5800	M58509	NA	5	15	14	3.51	3.18	2.42	1.15	Adrenodoxin reductase gene
3850	U82987	17	6	17	1	3.36	3.10	3.71	1.35	bcl-2 binding component 3 (bbc3)
1610	L42176	26	7	8	24	3.32	3.37	2.04	1.11	DRAL mRNA
1453	L29008	NA	8	6	100	3.31	3.51	1.30	1.10	L-iditol-2 dehydrogenase
170	D00762	NA	9	7	88	3.30	3.44	1.32	1.09	Proteasome subunit HC8
5915	L08096	29	10	12	2	3.29	3.25	3.69	1.32	CD27 ligand mRNA
1154	J05614 ^d	13	11	4	5	3.26	3.83	3.19	1.56	PCNA, proliferating cell nuclear antigen
4598	X77794 ^d	9	12	10	18	3.13	3.31	2.19	1.28	Cyclin G1
6683	X83492 ^d	14	13	14	37	3.11	3.22	1.64	1.15	Fas (alternate splice deleting exins 4 & 7)
1395	L20971	123	14	18	39	3.05	3.06	1.64	1.12	Phosphodiesterase mRNA
3148	U39400	8	15	20	12	3.02	2.92	2.46	1.15	NOF1
5431	U72649	NA	16	11	15	2.99	3.29	2.27	1.17	BTG2
1883	M28209	NA	17	23	28	2.96	2.72	1.92	1.09	GTP-binding protein (RAB1)
6089	M60974 ^d	10	18	26	8	2.95	2.66	2.77	1.29	gadd45
5469	X63717 ^d	6	27	32	13	2.60	2.61	2.43	1.22	APO-1 cell surface antigen
276	D21089	7	44	73	19	2.33	2.04	2.16	1.36	XPC, xeroderma pigmentosum group C gene
782	D90224	11	80	114	58	1.94	1.78	1.45	1.14	OX40 ligand, TNF ligand superfamily
6539	X85116	15	89	117	20	1.84	1.77	2.16	1.11	EPB72, integral membrane protein
3283	U48296	5	99	128	337	1.80	1.71	0.93	1.14	Protein tyrosine phosphatase PTP(CAAX1)
2946	U25138	12	2286	3136	2636	0.22	0.13	0.04	1.00	Maxi K potassium channel beta subunit
3320	U50136	16	2725	2548	2622	0.12	0.17	0.16	0.99	Leukotriene C4 synthase (LTC4S)
3270	U47621	3	3054	3225	3859	0.05	0.01	0.03	1.00	No 55 nucleolar autoantigen

^a Results from Model I.

^b Results from Model II.

^c Results from Model III. Fold I indicates the treatment fold change according to Model I. NA indicates the gene has less than 1.5 fold change and is filtered out prior to the SAM analysis.

^d Genes previously reported to respond transcriptionally to ionizing radiation. Tusher rank values from [14].

Table 3
Most significantly repressed genes

Gene no.	Accession no.	Rank				N log			Fold	Gene description
		Tusher	I ^a	II ^b	III ^c	I ^a	II ^b	III ^c	I ^a	
1860	M25753 ^d	5	1	2	3	5.14	5.19	4.12	1.39	Cyclin B
2645	U05340	4	2	3	2	4.84	4.72	4.27	1.27	p55cdc; present in dividing cells
2576	U01038	1	3	1	1	4.80	5.44	4.43	1.37	PLK, polokinase homolog
2789	U14518	19	4	4	7	4.28	4.11	3.06	1.22	Centromere protein-A (CENP-A)
5136	Z36714	29	5	16	114	4.22	2.99	1.43	1.15	Cyclin F
4153	X14850	58	6	6	17	4.15	4.06	2.33	1.19	Histone H2A.X
3682	U73379	20	7	5	5	3.98	4.09	3.57	1.35	Cyclin-selective ubiquitin carrier protein
6702	X97267	6	8	7	16	3.90	3.54	2.36	1.19	Lymphosphatase assoc phosphoprotein
6815	HG1980	NA	9	8	37	3.49	3.53	1.97	1.16	'Tubulin' Beta 2
4214	X51688	62	10	10	39	3.26	3.30	1.96	1.15	Cyclin A
3535	U63743	9	11	14	42	3.15	3.04	1.93	1.10	MCAK, mitotic centromere-associated kinesin
2353	M91670	2	12	9	12	3.14	3.30	2.46	1.20	Ubiquitin carrier protein (E2-EPF)
4273	X54942	8	13	11	4	3.14	3.14	3.88	1.41	ckshs2, cks1 protein homolog
1612	L42324	67	14	13	28	3.13	3.12	2.08	1.09	G protein-linked receptor gene (GPCR) gene
203	D13633	38	15	15	35	3.07	3.04	1.99	1.11	KIAA0008 gene
5409	U37426	116	16	19	10	3.05	2.95	2.48	1.10	Kinesin-like spindle protein HKSP (HKSP)
6377	X62534	NA	17	17	24	3.03	2.96	2.19	1.16	HMG-2
2306	M86699	33	18	18	20	2.93	2.95	2.27	1.12	Kinase (TTK) mRNA
4453	X67155	16	19	23	8	2.92	2.91	2.91	1.14	MKLP-1, mitotic kinesin-like protein-1
2988	U28386	13	34	33	50	2.51	2.50	1.88	1.25	hSRP1alpha, NLS receptor
5063	Z15005	18	36	40	27	2.43	2.30	2.14	1.10	CENP-E putative kinetochore motor
4370	X62048 ^d	11	46	48	43	2.22	2.19	1.93	1.04	wee1 kinase
4039	X02910	14	49	60	23	2.19	2.03	2.22	1.09	Tumor necrosis factor (TNF-alpha)
2511	S78187 ^d	7	56	74	52	2.09	1.93	1.85	1.11	cdc25 phosphatase
5847	HG3523	17	71	139	67	1.96	1.54	1.69	1.08	c-Myc, alternate splice form 3
2245	M80359	12	106	330	79	1.70	1.16	1.61	1.05	C-TAK1, cdc25c associated protein kinase
361	D31764	15	136	297	119	1.59	1.19	1.39	1.05	hEphB1b, Eph-like receptor tyrosine kinase
674	D86973	10	2581	3082	935	0.16	0.05	0.57	1.00	GCN1, translational regulator of GCN4
3615	U68233	3	805	1285	1208	0.86	0.69	0.47	0.98	HRR-1 farnesol receptor

^a Results from Model I.

^b Results from Model II.

^c Results from Model III. Fold I indicates the reciprocal of fold change for comparison purpose.

^d Genes previously reported to respond transcriptionally to ionizing radiation.

Fig. 4(C) shows the significance of using $\log_2(\text{MM})$ as a covariate in Model I. The estimates of β range from -1 to 2 , and can be viewed as gene-specific generalizations of the constant 0.5 value considered by Efron et al. [2]. However, only 422 (5.92%) of these coefficients are significantly different from zero at the 5% level with a Bonferroni correction.

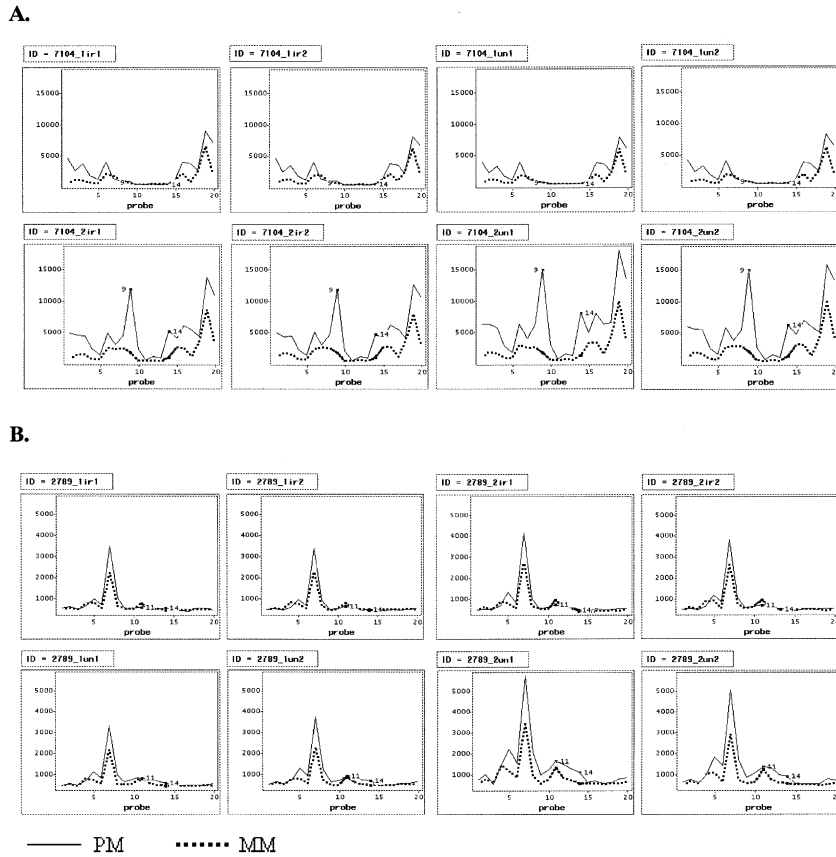


Fig. 5. Plots of original expression measurements of PM (solid) and MM (dotted). (A) Significant cell line–probe interaction in gene 7104 (X3068). (B) Significant treatment–probe interaction in gene 2789 (U14518).

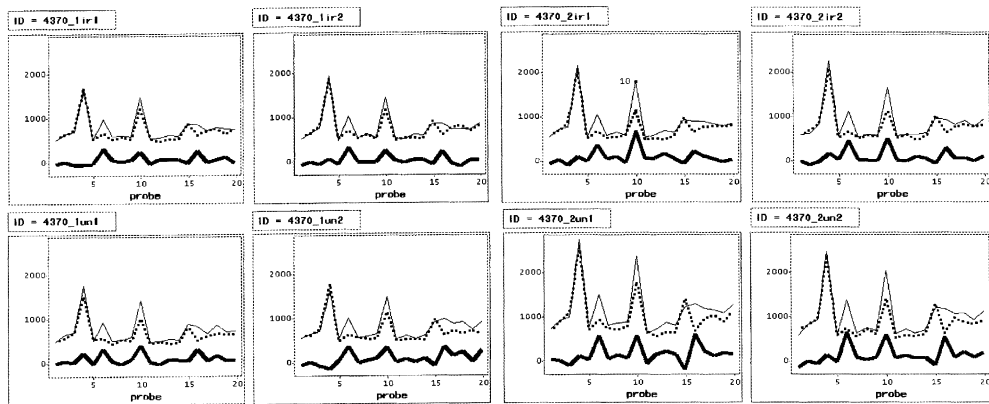
Tables 2 and 3 list the top induced and repressed genes according to our three models along with a comparison to the top genes from the SAM method of Tusher et al. [14]. The methods generally agree, although there are a few key differences that we discuss below. Tusher et al. [14] highlight twelve genes that were previously reported in the literature to respond transcriptionally to ionizing radiation. Nine of these are included in the top eighteen of Model I’s induced or repressed genes. The other three (X62048, S78187, X63717) are also marginally significant with negative log p -values larger than 2 but not as significant as the result of Tusher’s SAM.

One likely reason for these discrepancies is that the SAM results make use of Affymetrix’s summary measures across the probe level data whereas the mixed model results here delete no outliers. Fig. 6(A) uses gene X62048 as an example. Affymetrix applies a ‘three standard deviation rule’ for outlier deletion; that is, if the difference of probe pair exceeds three standard deviations within a probe set, it will be excluded for averaging. In Fig. 6(A), probe pair 10 in array ‘2ir1’ is an outlier according to the ‘three standard deviation rule’. Although it is questionable to say this probe pair is an outlier, excluding that point when calculating the average difference will lower the single expression measurement of gene X62048 in array ‘2ir1’, and this may increase the significance of repression when comparing the irradiated group to the untreated group. We also

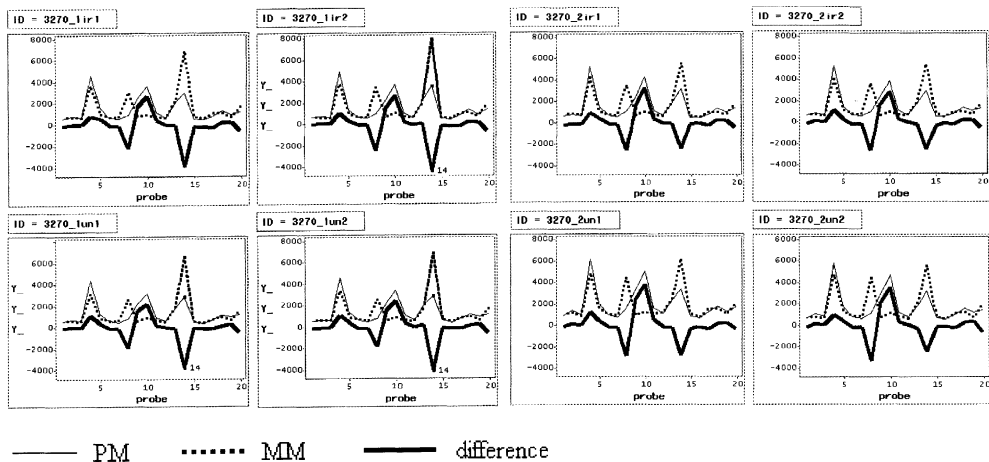
inspected several genes which were highly significant according to SAM but not significant for Models I, II, and III. Outliers appear to be at the root of these differences as well, and another example is in Fig. 6(B). The difference of probe pair 14 in arrays ‘1ir2’, ‘1un1’, and ‘1un2’ are outliers identified by ‘three standard deviation rule’ in this case. Resolutions of issues like these almost surely need to be done at the probe level, and indeed, independent analyses applied at the probe level produce almost identical results to ours (Virginia Tusher, - personal communication).

Comparing results from Models I, II and III, we see Models I and II are similar whereas some of the results from Model III are very different. The overall behavior of the treatment effects from these three models is displayed in Fig. 7(A). The correlation coefficients of the negative

A.



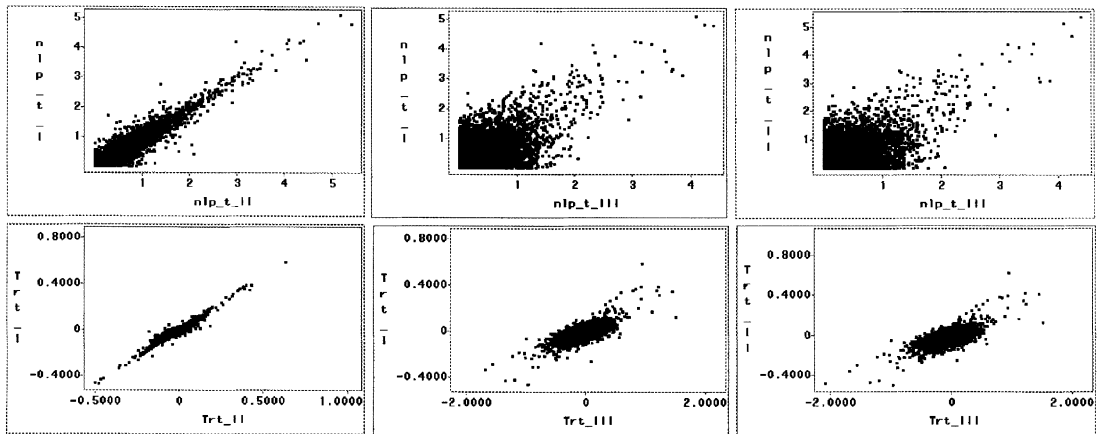
B.



— PM MM — difference

Fig. 6. Plots of original expression measurements of PM (solid), MM (dotted), and difference (bold) of PM and MM pair. (A) Probe profiles of gene 4370 (X62048). (B) Probe profiles of gene 3270 (U47621).

A.



B.

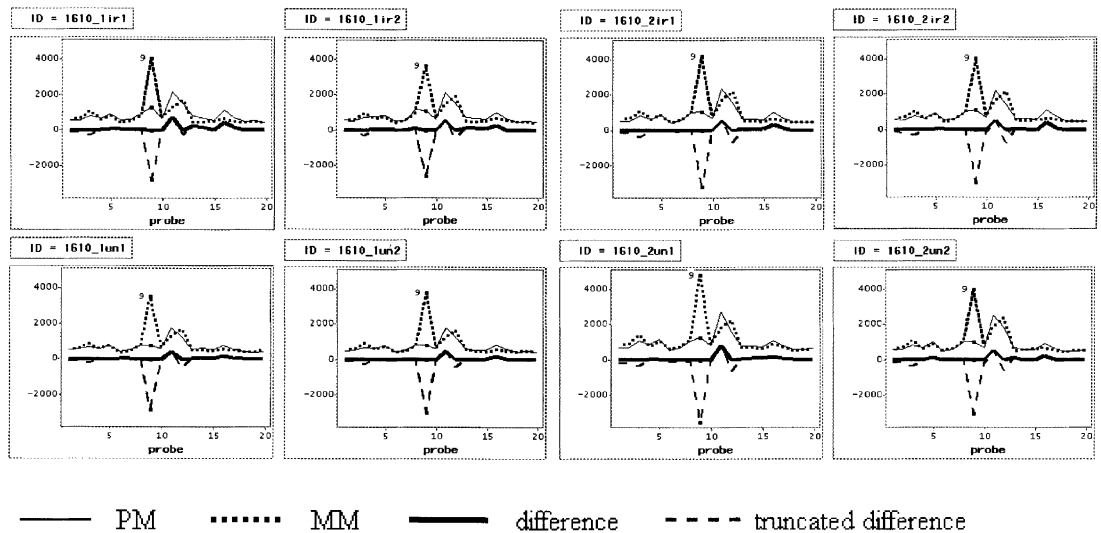


Fig. 7. (A) Scatter plots for comparing the negative log p -values (top row) and estimates (bottom row) of treatment effect among Models I, II, and III. (B) Probe profile of gene 1610 (L42176). The solid, dotted, bold and dashed lines represent the expression profiles of MM, PM, truncated difference, and difference, respectively. The difference of probe pairs 2, 9, and 12 are converted to 10.

log p -values of the model pairs (I, II), (I, III), and (II, III) are 0.95, 0.53, and 0.45 respectively. Here many of the discrepancies appear to be related to the truncation at 10 used in Model III to avoid negative difference values when taking logarithms, and Fig. 7(B) gives an example. Results like these shed doubt on the practical usefulness of trimming the data in this fashion. Also, including MM as a covariate does not change the results very much for this example, suggesting a lack of need for MM.

Fold-change estimates of the Model I treatment effects for genes are included in the next-to-last columns of Tables 2 and 3. Most are surprisingly small (between 1.1 and 1.3) and illustrate the potential of statistical methods to detect subtle changes.

3.6. Perform additional analyses on statistically filtered data

This step is omitted, except to say that clustering could now be performed on the treatment estimates from the previous step. Estimated treatment means appropriately adjusted for other effects in the model are very useful quantities for subsequent analyses like clustering and principal components, and can provide more accurate results than the raw data themselves.

4. Discussion

The differences noted above between SAM and our linear models are typically resolvable and of an order of magnitude smaller than differences between these methods and ones that make decisions based purely on fold change. It is absolutely critical to accommodate important sources of experimental variability when assessing GeneChip data, and failure to do so will surely result in higher rates of both false positives and false negatives. The systematic approach detailed here provides a good outline on how to analyze these data in a statistically sensible fashion, and the linear mixed modeling framework can flexibly accommodate most any kind of experimental design. Our hope is that statistical approaches like these will help researchers make the most of their microarray efforts.

Acknowledgements

We thank Virginia Tusher, Gilbert Chu, and Rob Tibshirani for kindly providing the ionizing radiation data and many helpful related comments. We also thank Rob for kindly spending much time in extracting the data at the probe level. Moreover, we thank reviewers who provided us with very useful comments to help us illustrating more complete about our approach. This work was supported by NIH grant GM45344.

References

- [1] S. Dudoit, Y.H. Yang, M.J. Callow, T.P. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical Paper, University of California, Berkeley, CA, 2000.
- [2] B. Efron, R. Tibshirani, V. Goss, G. Chu, Microarrays and their use in a comparative experiment, Technical Report, Stanford University, 2000.
- [3] W. Jin, R. Riley, R.D. Wolfinger, K.P. White, G. Passador-Gurgel, G. Gibson, Contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*, *Nat. Genet.* 29 (2001) 389.
- [4] M.K. Kerr, M. Martin, G.A. Churchill, Analysis of variance for gene expression microarray data, *J. Computat. Biol.* 7 (2000) 819.

- [5] E.N. Lazaridis, D. Sinibaldi, G. Bloom, S. Mane, R. Jove, A Simple method to improve probe set estimates from oligonucleotide arrays, University of South Florida, 2001.
- [6] C. Li, W.H. Wong, Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proc. Nat. Acad. Sci. USA* 98 (2001) 31.
- [7] C. Li, W.H. Wong, Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biol.* 2 (2001) research0032.1.
- [8] R.C. Littell, G.A. Milliken, W.W. Stroup, R.D. Wolfinger, SAS System for Mixed Models, SAS Institute, Cary, NC, 1996.
- [9] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, E.L. Brown, Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat. Biotechnol.* 14 (1996) 1675.
- [10] C.E. McCulloch, S.R. Searle, Generalized, Linear and Mixed Models, Wiley, New York, NY, 2001.
- [11] SAS Institute Inc., SAS/STAT Software Version 8, SAS Institute, Cary, NC, 1999.
- [12] E.E. Schadt, C. Li, C. Su, W.H. Wong, Analyzing high-density oligonucleotide gene expression array data, *J. Cell Biochem.* 80 (2000) 192.
- [13] R.G.D. Steel, J.H. Torrie, D.A. Dickey, Principles and Procedures of Statistics: a Biometrical Approach, 3rd Ed., McGraw-Hill, New York, 1997.
- [14] V. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Nat. Acad. Sci.* 98 (2001) 5116.
- [15] G. Verbeke, G. Molenberghs, Linear Mixed Models for Longitudinal Data, Springer, New York, NY, 2000.
- [16] R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, R.S. Paules, Assessing gene significance from cDNA microarray expression data via mixed model, *J. Computat. Biol.* 8 (2001) 625.
- [17] Y.H. Yang, S. Dudoit, P. Luu, T.P. Speed, Normalization for cDNA microarray data. Technical paper, University of California, Berkeley, CA, 2001.