Cross-Site Comparison of Gene Expression Data Reveals High Similarity

Tzu-Ming Chu,[1] Shibing Deng,[1] Russ Wolfinger,[1] Richard S. Paules,[2] and Hisham K. Hamadeh[3]

[1]SAS Institute Inc., Cary, North Carolina, USA; [2]National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, USA; [3]Amgen Inc., Thousand Oaks, California, USA

Consistency and coherence of gene expression data across multiple sites depends on several factors such as platform (oligo, cDNA, etc.), environmental conditions at each laboratory, and data quality. The Hepatotoxicity Working Group of the International Life Sciences Institute Health and Environmental Sciences Institute consortium on the application of genomics to mechanism-based risk assessment is investigating these factors by comparing high-density gene expression data sets generated on two sets of RNA from methapyrilene (MP) experiments conducted at Abbott Laboratories and Boehringer-Ingelheim Pharmaceuticals, Inc. using a single platform (Affymetrix Rat Genome U34A GeneChip) at seven different sites. This article focuses on the evaluation of data quality and statistical models that facilitate the comparison of such data sets at the probe level. We present methods for exploring and quantitatively assessing differences in the data, with the principal goal being the generation of lists of site-insensitive genes responsive to low and high doses of MP. A combination of numerical and graphical techniques reveals important patterns and partitions of variability in the data, including the magnitude of the site effects. Although the site effects are significantly large in the analysis results, they appear to be primarily additive and therefore can be adjusted in the statistical calculations in a way that does not bias conclusions regarding treatment differences. *Key words:* **cross-site comparison, gene expression, hepatotoxicity, ILSI, toxicogenomics.** *Environ Health Perspect* **112:449–455 (2004). doi:10.1289/txg.6787 available via *http://dx.doi.org/* [Online 15 January 2004]**

Advancing the general knowledge base of mechanisms and markers of hepatotoxicity is of great interest to all parties involved in this consortium. The Hepatotoxicity Working Group of the International Life Sciences Institute (ILSI) Health and Environmental Sciences Institute (HESI) Committee on the Application of Genomics to Mechanism-Based Risk Assessment is investigating these factors by comparing high-density gene expression data sets generated on two sets of RNA from two independent *in vivo* experiments where rats were dosed with methapyrilene (MP) conducted at either Abbott Laboratories (site A; Abbott Park, IL) or Boehringer-Ingelheim Pharmaceuticals, Inc. (BIPI; site B; Ridgefield, CT).

Most microarray studies are designed with large "*p*" (number of genes) and small "*n*" (number of arrays) characteristics. Two issues of concern arise when investigators work with data having such characteristics. The first issue is of statistical inference power where the aim is to minimize both false-positive and false-negative rates. Increasing sample size (i.e., *n*) can afford better statistical inference power; however, this remedy is often cost prohibitive. The second issue arises in the attempt to address the first concern by increasing sample size by incorporating data sets generated at disparate sites and times. Thus, the second concern is about the consistency of such data sets generated across multiple sites and whether the same or similar conclusions can be drawn. An across-site microarray study can be useful for addressing this issue, which is another way to increase sample size. Conceptually, the complexities among data generated across different sites are higher than those of data generated within one site. The above two issues are related; however, the second one may be more general and of increasing concern as microarray data sets become increasingly available and the desire to compare and contrast across studies increases.

Male Sprague-Dawley rats [CRL: CD(SD)IGS VAF/Plus] (Charles River Laboratories, Kingston, NY) approximately 6–7 weeks of age were assigned to nine study groups (four rats/group) and dosed by gavage for 1, 3, or 7 days with water (vehicle), 10 mg/kg/day MP, or 100 mg/kg/day MP (Figure 1). Dose selection was based on published and unpublished studies; the high dose of MP was chosen to yield hepatotoxicity, and a nontoxic low dose was selected. In general, the BIPI study yielded more hepatotoxicity than the study conducted at Abbott Laboratories, as defined by clinical pathology parameters and microscopic examinations of hematoxylin and eosin–stained liver sections. No significant histopathological alterations were observed in livers of rats treated with 10 mg/kg/day MP at the 1- and 3-day time points compared with alterations in livers of the control groups. In comparison, MP treatment with 10 mg/kg/day for 7 days resulted in minimal portal mononuclear infiltrates, minimal hepatocellular periportal necrosis, and minimal microvesicular hepatocellular vacuolization. At the 100-mg/kg/day dose, all rats showed early minimal mononuclear portal infiltrates, minimal hepatocellular periportal necrosis, and mild to moderate periportal microvesicular vacuolization at 1 and 3 days of exposure. The severity of the lesions increased at day 7, and mild hyperplasia became evident. In addition, in the 100-mg/kg/day MP dose group at 7 days of exposure, moderate mononuclear portal infiltrates were noted, and the number of enlarged periportal hepatocytes with microvesicular vacuolization increased. The severity of hepatocellular periportal necrosis at the 7-day time point also increased, accompanied by increased numbers of hepatocellular mitotic figures. Bile duct hyperplasia was observed in animals in the 100-mg/kg/day dose group at 3 and 7 days. Minimal bile duct hyperplasia was seen at the 3-day time point and increased in severity to mild by 7 days. Levels of alanine aminotransferase, aspartate aminotransferase, and sorbitol dehydrogenase increased in high-dose animals in a time-dependent manner. Total bilirubin tended to be elevated in the high-dose group with continued dosing. All the above parameters were reflective of liver toxicity.

An initial amount of 5–20 μg total RNA derived from livers of rats used in those studies was used for the synthesis of double-stranded cDNA with a commercially available kit (Superscript Choice System;

This article is part of the mini-monograph "Application of Genomics to Mechanism-Based Risk Assessment."

Address correspondence to H.K. Hamadeh, Amgen Inc., One Amgen Center Dr., Mail Drop 5-1-A, Thousand Oaks, CA 91320 USA. Telephone: (805) 447-4818. Fax: (805) 499-2936. E-mail: hhamadeh@amgen.com

We thank our colleague H. Chen, at SAS Institute, Inc., for creating some of the figures in this article. We also thank the Hepatotoxicity Working Group of the ILSI Health and Environmental Sciences Institute's Committee on the Application of Genomics to Mechanism-Based Risk Assessment testing program, a scientific consortium organized to facilitate further development and advances in genomics and proteomic methodologies to increase the utility of gene expression data for mechanism-based risk assessment.

The authors declare they have no competing financial interests.

Received 6 October 2003; accepted 15 December 2003.

Invitrogen Life Technologies, Carlsbad, CA, or Roche Molecular Biochemicals, Mannheim, Germany) in the presence of a T7-$(dT)_{24}$ DNA oligonucleotide primer. After synthesis, the cDNA was purified by phenol/chloroform/isoamyl alcohol extraction and ethanol precipitation. The purified cDNA was then transcribed *in vitro* (ENZO Life Sciences, Farmingdale, NY or Ambion Diagnostics, Austin, TX) in the presence of biotinylated ribonucleotides to form biotin-labeled cRNA. The labeled cRNA was then purified on an affinity resin [RNeasy, Qiagen (Valencia, CA)], quantified, and fragmented. Ten to 20 µg labeled cRNA was hybridized for approximately 16 hr at 45°C to an expression probe array. The array was then washed and stained with streptavidin-P-phycoerythrin (SAPE; Molecular Probes, Inc., Eugene, OR). The signal was amplified using a biotinylated goat antistreptavidin antibody (Vector Laboratories, Burlingame, CA), and the array received a final staining with SAPE. The GeneChip Fluidics Workstation 400 (Affymetrix, Inc., Santa Clara, CA) was used to stain the arrays. The array was then scanned twice using a confocal laser scanner (GeneArray Scanner 2500, Hewlett Packard, or Agilent, Foster City, CA), which resulted in one scanned image.

Many factors can contribute to the heterogeneity of data sets, including but not limited to differences in platform (oligo, cDNA, etc.), environmental conditions at each laboratory, and data quality. Meta-analysis performed by statistically integrating results from different data sets (Choi et al. 2003; Ghosh et al. 2003) is one solution for across-site microarray studies. When the experimental design settings are homogenous across sites, pooling data sets for a comprehensive analysis provides direct comparisons across sites, with better statistical inference power. The major challenges to this end are how to normalize data sets and how to define and use variation across sites.

Five well-adapted normalization methods, including cyclic Loess (modified from Dudoit et al. 2002), contrast-based method (Astrand, unpublished data), quantile normalization (Irizarry et al. 2003), the scaling method in Affymetrix MAS 5.0 (Affymetrix, Inc. 2002), and the nonlinear method (Li and Wong 2001; Schadt et al. 2001) are reviewed by Bolstad et al. (2003). Quantile normalization seemed slightly better than the others in the three types of comparisons performed. Quantile normalization makes the distribution of each chip the same by aggressively removing specific sources of variation (such as those associated with chip-to-chip differences) and artificially generating ideal distributions. Another less aggressive way to deal with this is to normalize the location and scaling of the distribution. We prefer the second approach and have implemented that in this article. In our opinion, quantile normalization has a potential drawback in that it can reduce the consistency of the expression profile of the same probe set across arrays. This consistent expression profile is one of the essential characteristics of the GeneChip probe data (Li and Wong 2001). Lowering this consistency may increase the variation in downstream statistical modeling. An interquartile range normalization is applied in this article toward the same goal as quantile normalization (i.e., to obtain consistent but not identical data distribution among chips). This approach makes data comparable across sites and preserves a certain level of site effects when combining the data. The factor of site can be easily adapted in an analysis of variance (ANOVA) model. In using variation components naturally involved in the data, the mixed-model approach provides flexibility (Chu et al. 2002) and robustness (Chu et al., in press) for this task. Tan and coworkers (Tan et al. 2003) conducted a similar comparison study that demonstrated mildly positive concordance between Affymetrix and Amersham (Piscataway, NJ) short oligo arrays and Agilent cDNA arrays.

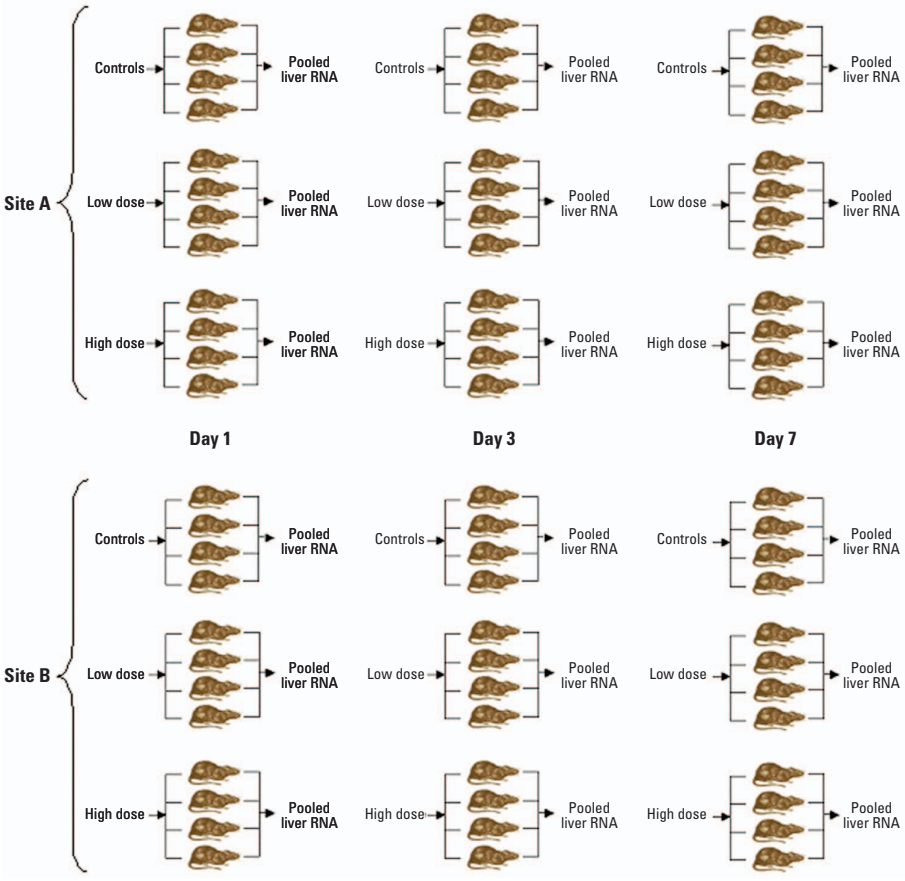

Figure 1. Schematic diagram of experimental design of studies conducted at Abbott Laboratories (site A) and Boehringer-Ingelheim Pharmaceuticals, Inc. (site B). Pooled samples were derived from livers corresponding to rats in the study. Pooling was conducted at the RNA level, where equal amounts of RNA were pooled from four replicate rats at each dose–time point. Sites 2, 3, 6, and 8 received samples from each dose–time combination derived from studies run at both sites A and B, whereas sites 4, 5, and 7 received samples corresponding to all dose–time combinations derived from site B only.

Data Consistency and Normalization

RNA samples were analyzed independently by seven different Affymetrix platform sites using RGU34A expression probe arrays (Affymetrix) containing 8,799 probe sets interrogating primarily annotated genes, for a total of 99 chips. Each probe set consisted of 16 probes; thus, each chip resulted in 140,784 probes. The rat sequences used for the design of the RGU34A expression probe array were derived from the UniGene Database build #34 (created from Genbank 107/dbEST 11/18/98) and supplemented with additional annotated gene sequences from Genbank 110 (http://www.ncbi.nih.gov/GenBank). UniGene clusters are represented by a sample sequence that is the most complete and

most 3′ sequence in the cluster. The oligonucleotide probes are 25mers, and 16 probe pairs per sequence are used. The detection sensitivity is 1:100,000, measured by the detection in a comparative analysis between a complex RNA containing spiked control transcripts and a complex RNA with no spikes [Anonymous. GeneChip Rat Genome U34 Set data sheet (Affymetrix 2002)]; detection is quantitative over more than three orders of magnitude (Lockhart et al. 1996). Each site is identified as a user in the HESI consortium in Table 1. Sites 2, 3, 6, and 8 analyzed RNA samples from both *in vivo* studies performed at sites A and B, whereas the remaining sites analyzed only the RNA samples from site B. Each RNA set was analyzed using nine Affymetrix chips, one for each of the three dose levels and three time points, with the exception of site 3, where all nine chips were used in three technical replicates of the RNA samples from three doses from the third time point from the *in vivo* study performed at site A only. Data for perfect match probe intensities from .CEL files were used for analysis. In this article we focus on the point that site effects exist but can be statistically accounted for and adjusted. It would be interesting to investigate whether using different normalization methods or outcome variables (such as perfect match and mismatch) results in different site-effect magnitude, but this was not a major goal of this article. Data used in this report may be further analyzed by interested scientists and can be accessed via the Internet (http://dir.niehs.nih.gov/microarray/ilsi-datasets/home.htm or the European Bioinformatics Institute ArrayExpress database at http://www.ebi.ac.uk/arrayexpress/).

A $\log_2$ transformation was applied to all data before any analysis process. As a first attempt to inspect data consistency across sites, box plots of all chips were generated in Figure 2A for comparison of the distribution of each chip. As shown in Figure 2A, the within-site chip-to-chip variation is higher in sites 3 and 5. In addition, the range of data varies significantly from site to site. To further inspect the correlations among chips, a subgroup of 10 chips corresponding to control animals (the first level of dose factor, i.e., dosed with vehicle alone or the 0 mg/kg dose) at day 1 (the first level of time factor) was chosen to compute the interchip correlation coefficients. Figure 3A shows the scatterplots of the $\log_2$ perfect-match probe intensities among the 10 chips. The red ellipse curve within each plot indicates the 95% density curve based on bivariate normal distribution (i.e., 95% of data is inside the ellipse). The pairwise correlation coefficients ranged from 0.92 to 0.98 except for those from site 3 (B01_3) or site 5 (B01_5), which ranged from 0.82 to 0.90.

Another method used to inspect the consistency of these chips is examination of Oligo B2 (Affymetrix 2002). The Oligo B2 contains the Poly-A Controls (*dap*, *lys*, *phe*, *thr*, and *trp*) and the Hybridization Controls (*bioB*, *bioC*, *bioD*, and *cre*) as part of the GeneChip Eukaryotic Hybridization Control Kit. Details on Oligo B2 can be found in the reference provided. Briefly, Oligo B2 serves as spike-in controls. The Poly-A Controls can be spiked into a complex RNA sample and carried through the sample preparation process. The Hybridization Controls are prepared in staggered concentrations (1.5, 5, 25, and 100 pM for *bioB*, *bioC*, *bioD*, and *cre*, respectively) independent of RNA sample preparation and are spiked into the hybridization cocktail. Although in the Affymetrix reference it states that the variation in B2 hybridization intensities across the array is normal and does not indicate a variation in hybridization efficiency, we have often observed that these controls are expressed consistently across chips (unpublished data). Figure 3B presents the scatterplots of Oligo B2 controls among the

Table 1. User identifiers of seven sites in HESI consortium.

User ID	Site	RNA sample applied
2	Novartis AG	A, B
3	Roche Molecular Biochemicals	A, B
4	Wyeth Research	B
5	AstraZeneca Pharmaceuticals, Inc.	B
6	Schering-Plough Research Institute	A, B
7	Boehringer-Ingelheim Pharmaceuticals, Inc.	B
8	Pfizer Inc	A, B

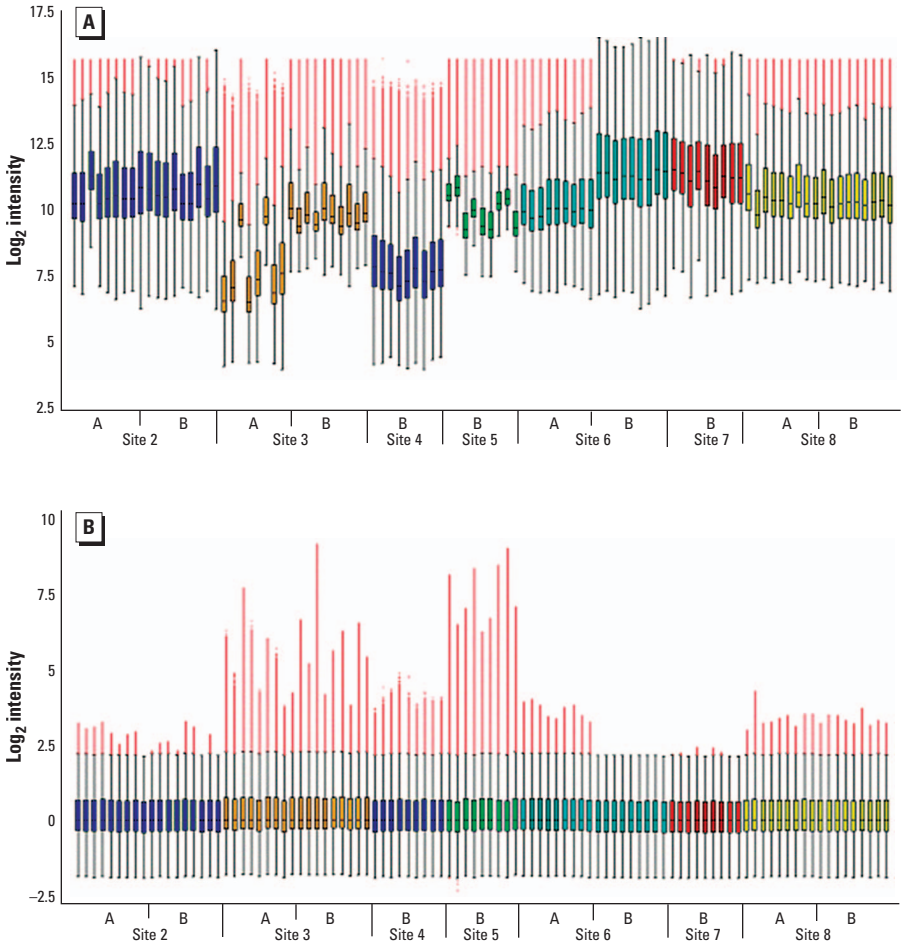

Figure 2. (*A*) Box plots of $\log_2$ perfect-match probe intensities. Each site indicated by color is listed from left (site 2) to right (site 8). Within site, the order of chips is sorted by RNA sample, dose, and time point, sequentially. Marks on x-axis indicate the associated site and RNA sample. (*B*) Box plots of $\log_2$ perfect-match probe intensities after interquartile normalization.

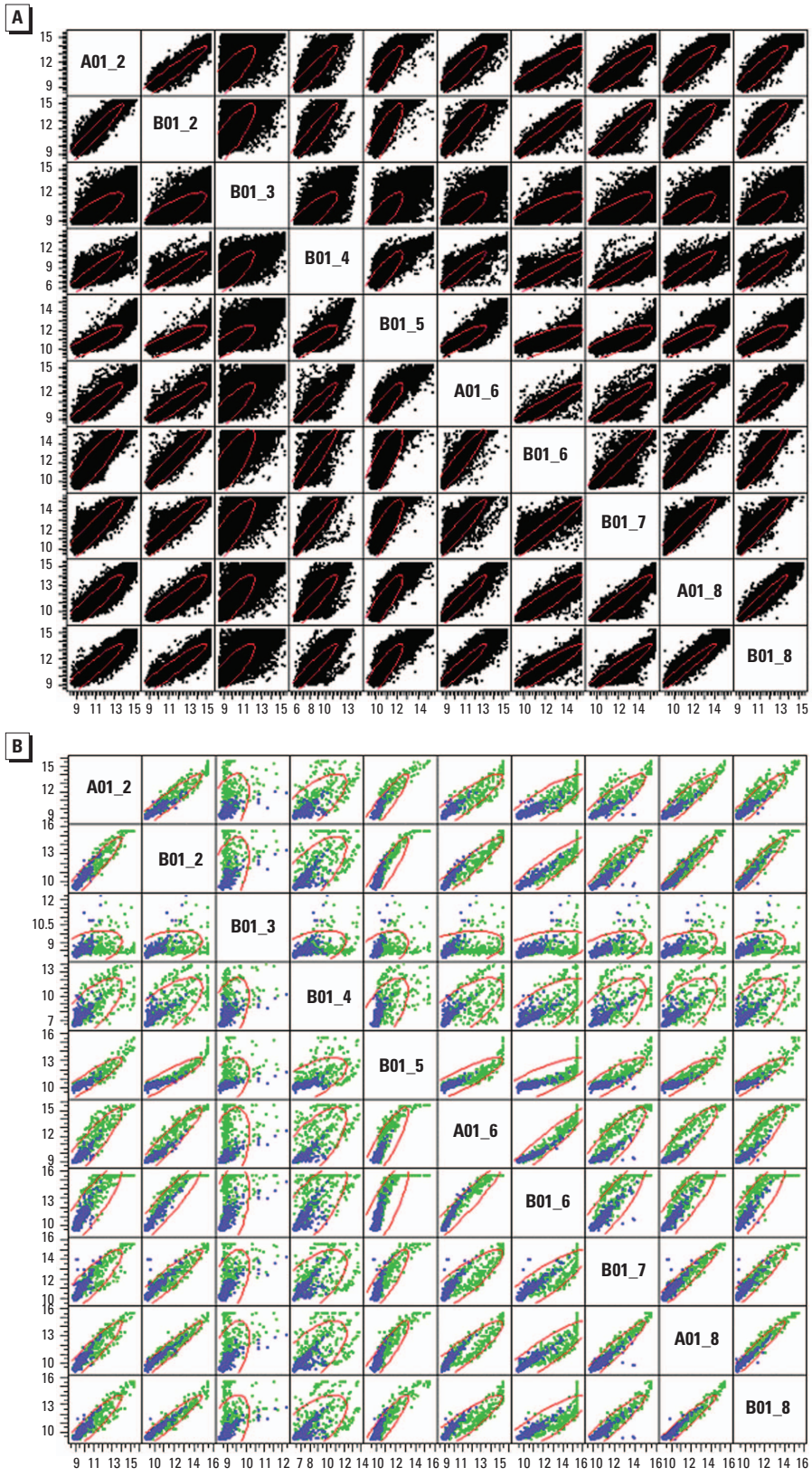

Figure 3. (*A*) Scatterplots of $\log_2$ perfect-match probe intensity among the 10 chips with controlling dose level on day 1. The first letter and the last digital of the marks in the diagonal indicate the RNA sample used and the site identifier, respectively. (*B*) Scatterplots of spiked-in probes among the 10 chips. Poly-A and Hybridization Controls are indicated by blue and green, respectively.

10 chips in Figure 3A. A small percentage of the Hybridization Controls on the chips were saturated. The linearity (correlation) between chips was higher when there were no saturated intensities. The Poly-A Controls have a better consistency than Hybridization Controls. Because the Poly-A Controls are carried in RNA samples through the preparation process, they are better candidates to indicate the consistency of data. The correlation coefficients among the Poly-A Controls were calculated and listed in Table 2. These correlation coefficients revealed that sites 3, 4, and 5 have less consistency with other chips. However, the inspection here was based on assuming all experiments followed the same protocol.

For a global view of the correlations among all chips, a matrix with each entry as one minus the correlation coefficient of the corresponding pair of chips was calculated and used as the distance matrix for multidimensional scaling (MDS) analysis with the MDS procedure (SAS Institute 1999). The results are presented in Figure 4. This two-dimensional representation gives the relative location of each chip based on the distance matrix of a multidimensional space (a dimension of 99 in this case). The plotted points are their relative location on a two-dimensional map. The closer two points are to each other, the more similar they are. The chips from sites 3 and 5 are spotted apart from the others except for the three chips from site 3 on the margin.

For a quick summary of the consistency, the 99 chips were separated into 18 categories based on the unique site of *in vivo* studies, dose, and time combination, and the average within-category correlation coefficients were calculated. The results are shown in Figure 5. The average correlation coefficients in all categories were higher than 0.9. Categories that did not involve sites 3, 4, and 5 (A01, A02, A11, A12, A21, and A22 in Figure 5) because the site A samples were not analyzed at those sites, had correlation coefficients higher than 0.95.

Three major characteristics were revealed on inspection of the aforementioned probe data, namely, different within-site chip-to-chip variation, different site-to-site variation, and high within-treatment-group correlation across sites. The first characteristic can be handled with the mixed-model approach and will be discussed later in this article. The site-to-site variation can be reduced by normalization. As seen on examination of the box plots in Figure 2A, there are two factors that need normalization: the range of data and the within-chip variance (size of the box). An interquartile normalization with the median as the location parameter and

interquartile range as the scaling parameter was applied. Figure 2B presents the box plots after normalization. The observation of a high within-treatment-group correlation across sites provides an incentive to pool data across sites for more powerful statistical inferences.

Mixed-Model Analysis

The mixed-model approach provides flexible model specification for the ANOVA type of analysis with the ability to accommodate different correlation structures in the data. Chu et al. (2002) have more details for applying the mixed model on GeneChip probe data. The mixed model for the MP data applied here is as follows:

$$\begin{aligned} Y_{ijklp} &= R_i + D_j + T_k + S_l + RD_{ij} + RT_{ik} \\ &\quad + RS_{il} + DT_{jk} + DS_{jl} + TS_{kl} + P_p \\ &\quad + RP_{ip} + DP_{jp} + TP_{kp} + SP_{lp} \\ &\quad + A_{ijk(l)} + \varepsilon_{ijklp}, \\ A_{ijk(l)} &\sim N(0, \__l^2), \\ \varepsilon_{ijklp} &\sim N(0, \sigma^2). \end{aligned} \quad [1]$$

The indices *i*, *j*, *k*, *l*, and *p* indicate site of *in vivo* study, dose, time point, site, and probe number, respectively. The index of the gene was omitted in the model, as the model will be run on a per-gene basis. The dependent variable *Y* is the normalized perfect match probe intensity. The symbols *R*, *D*, *T*, *S*, *P* represent RNA samples, dose, time, site, and probe main effects, respectively. The symbols with two letters are the interactions of the two effects associated with the letters. The $A_{ijk(l)}$ is the *l*th within-site array random effect and is assumed to be normally distributed with mean 0 and variance σ_l^2. Specifying array random effect induces a correlation across all observations (probes) on the same chip (probe set). The ε_{ijklp} is a stochastic error and is assumed to be normally distributed with mean 0 and variance σ^2. The two random terms are assumed to be independent.

The interactions involving more than two effects can be included in Model 1. However, after fitting those higher interactions in the model for several genes, we observed that those interactions were not significant; therefore, they are not included in the model. The error term can be partitioned to associate with each site in a similar fashion to the array random effect. Because all the within-treatment-group correlations were higher than 0.9, partitioning error term becomes a minor issue. In addition, assuming that the errors are identically distributed can enhance the strength of pooling data to more accurately estimate the associated variance.

A desired outcome of this exercise was to find genes responding to different doses at different time points by performing statistical testing on dose, time, and dose-by-time interaction effects. Whether the significant genes selected are consistent across sites was also of particular interest to parties involved in this across-site microarray study. This can be achieved by testing dose-by-site and time-by-site interactions.

For comparison purposes, a subset of 18 chips from site 8 is extracted and fitted by a similar fashion as in Model 1 but without all site-involved effects. This single-site model is listed as follows:

$$\begin{aligned} Y_{ijkp} &= R_i + D_j + T_k + RD_{ij} + RT_{ik} + DT_{jk} \\ &\quad + P_p + RP_{ip} + DP_{jp} + TP_{kp} + A_{ijk} \\ &\quad + \varepsilon_{ijklp}, \\ A_{ijk} &\sim N(0, \sigma_a^2), \\ \varepsilon_{ijklp} &\sim N(0, \sigma^2). \end{aligned} \quad [2]$$

The significances of testing some effects in this case were compared with the results from Model 1. For fitting Models 1 and 2, standard maximum likelihood approaches are usually best and can be accessed through software like the MIXED procedure (SAS Institute 1999).

Results

The statistical testing results of the 10 fixed effects that did not involve probe in Model 1 are presented in Figure 6A. Two plots, a histogram and a box, are drawn on negative $\log_{10}$ *p*-values for each effect among all genes. Table 3 presents the number of significant genes selected with controlling false-positive rate by Bonferroni's approach and controlling three different false discovery rates (FDR). The cutoff of the negative $\log_{10}$ *p*-value for 0.05 family-wide false-positive rate with Bonferroni's adjustment in this case is 6.245, which is indicated as the red horizontal line on each plot in Figure 6A. As expected, the site effect is highly significant for most of the genes (95.7%). The percentage of significant genes for both dose and time-by-site interactions were 0.11 and 0.19, respectively. This implies that only 27 of 8,799 genes showed a differential response to dose or time across sites. Therefore, the genes selected as differentially responsive to dose or time from the pooled data sets is consistent across sites

Table 2. Correlation coefficients of Poly-A Controls of the 10 chips selected.

	A01_2	B01_2	B01_3	B01_4	B01_5	A01_6	B01_6	B01_7	A01_8	B01_8
A01_2	1.00	0.90	0.71	0.66	0.78	0.84	0.83	0.80	0.89	0.90
B01_2		1.00	0.78	0.71	0.80	0.94	0.95	0.86	0.95	0.96
B01_3			1.00	0.59	0.69	0.82	0.77	0.73	0.78	0.78
B01_4				1.00	0.62	0.73	0.74	0.80	0.72	0.68
B01_5					1.00	0.79	0.77	0.72	0.80	0.78
A01_6						1.00	0.96	0.87	0.94	0.92
B01_6							1.00	0.89	0.94	0.94
B01_7								1.00	0.87	0.86
A01_8									1.00	0.95
B01_8										1.00

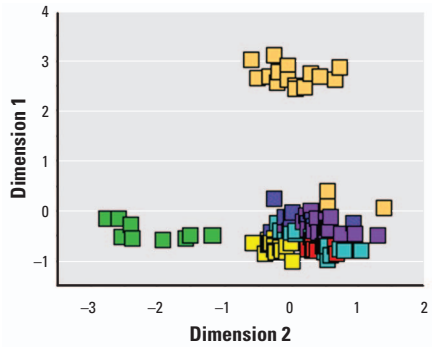

Figure 4. Two-dimensional representation of multidimensional scaling analysis on the 99 chips. Color indices are the same as in Figure 2.

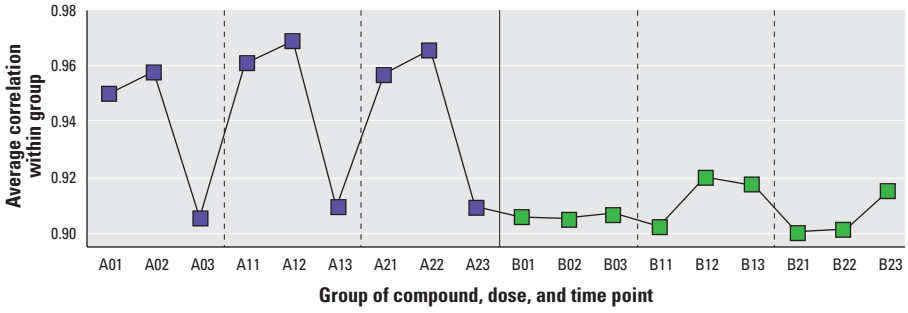

Figure 5. Average correlation within each combination of RNA sample, dose, and time point. The first, second, and third digits in the mark of x-axis indicate RNA sample, dose, and time-point-levels.

except for a very few. However, there were seven genes showing a very highly significant dose effect with a negative logarithm *p*-value larger than 10 that also showed a significant dose-by-site interaction. An explanation for this is that the extremely significant main effect often causes significant interactions with other effects. The results of testing the fixed effects of Model 2 with data from site 8 only are presented in Figure 6B. All the negative logarithm *p*-values are < 6 in this case. This implies that there were no genes showing any significant effect with Bonferroni's criterion when data from site 8 only were used.

Compared with Bonferroni's approach, which conservatively guarantees that the probability of only one (or more) false positives is less than 0.05 across all of the tests, FDR allows a certain proportion (the cutoff rate) of significant genes to be false discoveries. More significant genes were selected with FDR approaches. However, the proportion of genes with significant dose or time-by-site effects was still considerably low—1.3 and 2.2% in the case of setting 0.005 (1 in 200) as cutoff.

The results in Table 3 reveal that site-to-site effects, although significant in a large number of genes, appear to be only additive, as Model 1 fit the data well, with a median r^2 equal to 0.97, and only a few genes showed significant site-by-treatment interaction. Therefore, this can be adjusted for in the statistical calculations in a way that does not bias conclusions regarding treatment differences. In other words, each site ends up relaying a similar story regarding significantly differentially expressed genes.

Figure 7 presents the histograms and box plots of the standard deviations of the random components that are seven within-site array random factors, $A_{ijk(l)}$, and the stochastic errors, ε_{ijklp}, in Model 1. These plots provide global comparisons among the site-specific array variations. The data from sites 6 and 7 show less array variation, whereas the data from sites 3 and 5 show larger array variation. The median standard deviations of sites 2–8 and stochastic errors are 0.029, 0.089, 0.037, 0.087, 0.018, 0.018, 0.024, and 0.096, respectively. Judging from comparison of these medians, the array variations from sites 3 and 5 are about 23-fold ($0.089^2/0.018^2$) larger than the variations from sites 6 and 7.

Figure 8 presents the comparison of significance of dose and time effects from Models 1 and 2. The red lines on plots are regression-fitted lines with slopes 0.14 and 0.16 on Figure 8A (comparing significance of dose effect) and Figure 8B (comparing significance of time effect), respectively. Pooling data across sites increases statistical inference power significantly. Judging from the inverse of slopes, the negative log *p*-values increase 7.14- and 6.25-fold for dose and time effects, respectively, when pooling data across seven sites, with the number of chips applied increasing from 18 to 99.

Discussion

Combining data across sites typically provides more powerful statistical inference. However, consistency of data sets is an essential issue for analyzing pooled data sets across sites. A robust normalization method

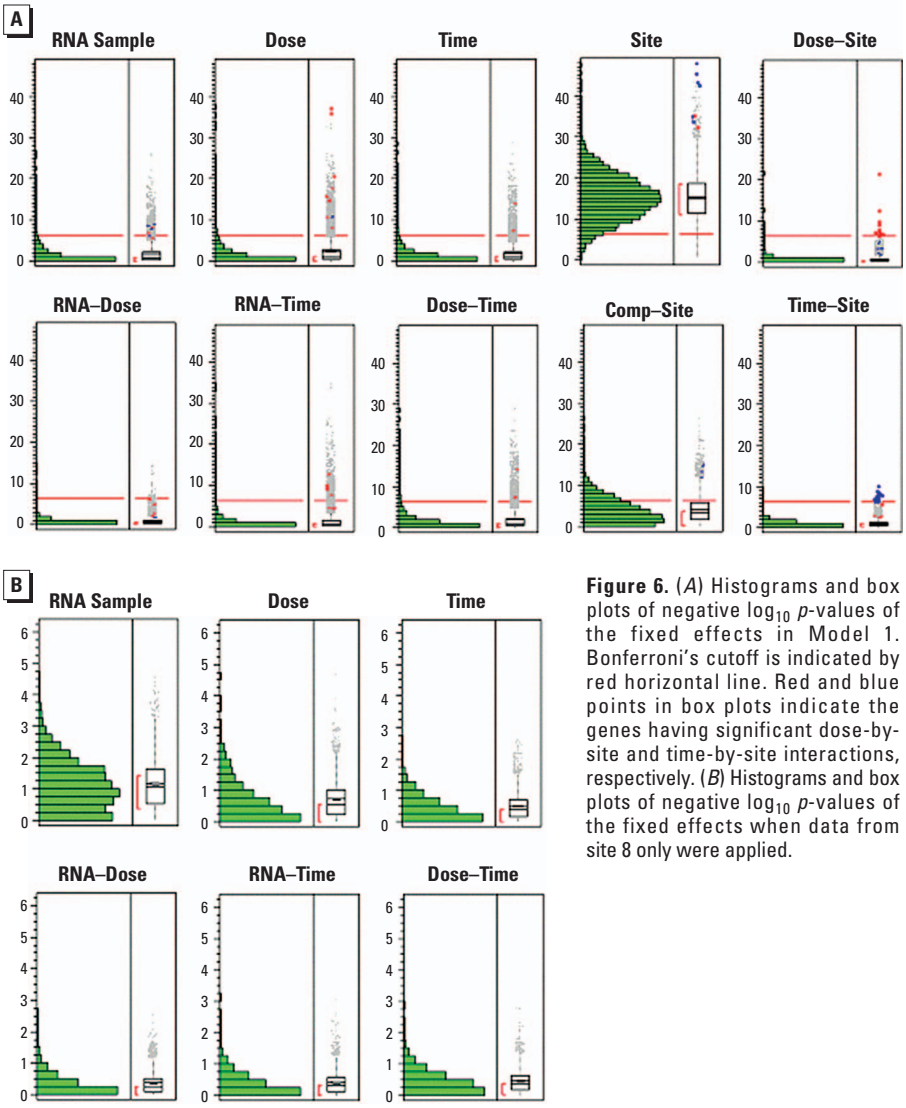

Figure 6. (*A*) Histograms and box plots of negative $\log_{10}$ *p*-values of the fixed effects in Model 1. Bonferroni's cutoff is indicated by red horizontal line. Red and blue points in box plots indicate the genes having significant dose-by-site and time-by-site interactions, respectively. (*B*) Histograms and box plots of negative $\log_{10}$ *p*-values of the fixed effects when data from site 8 only were applied.

Table 3. Number of significant genes selected by the 10 effects without probe involved.

	Number of significant genes[a]			
Effect	Bonferroni	FDR (0.005)[b]	FDR (0.01)[b]	FDR (0.05)[b]
R	466 (5.3)	1,620 (18.4)	1,851 (21.0)	2,605 (29.6)
D	787 (8.9)	1,900 (21.6)	2,159 (24.5)	3,067 (34.9)
T	173 (2.0)	735 (8.4)	890 (10.1)	1,539 (17.5)
S	8,422 (95.7)	8,746 (99.4)	8,762 (99.6)	8,792 (99.9)
RD	33 (0.38)	274 (3.1)	352 (4.0)	714 (8.1)
RT	361 (4.1)	972 (11.0)	1,144 (13.0)	1,738 (19.8)
RS	1,921 (21.8)	5,031 (57.2)	5,443 (61.9)	6,509 (73.3)
DT	507 (5.8)	1,504 (17.1)	1,735 (19.7)	2,690 (30.6)
DS	10 (0.11)	112 (1.3)	149 (1.7)	421 (4.8)
TS	17 (0.19)	190 (2.2)	308 (3.5)	970 (11.0)

Abbreviations: D, dose; DS, dose–site; DT, dose–time; R, RNA sample; RD, RNA sample - dose; RS, RNA sample - site; RT, RNA sample - time; S, site; T,time; TS, time–site.
[a]Results of false discovery rate (FDR), with 0.005, 0.01, and 0.05 as the cutoffs. [b]The percentage of significant genes of each effect is listed inside parentheses after the counts of significant genes.

is desirable to make data sets across sites more comparable. The interquartile range normalization is suitable for data with high correlation but inconsistent data range across chips. This normalization was used on the data applied here to achieve consistent ranges across the majority of the data and to preserve the phenomena of significant site variation.

An alternative means for normalizing data across sites is to use a universal reference sample. The universal reference is typically a type of mRNA pool from all the mice in the experiment and is distributed to each site of the consortium to serve as baseline for normalization. Data can be normalized within each site to the universal reference array, using linear or nonlinear methods. The advantage of a universal reference is that it serves as a bridge to bring the data from all chips across sites to be comparable; however, extra costs are involved in preparation, distribution, and maintenance of the pooled reference as well as the expense of running more arrays. In addition, data from the reference are subject to nonconstant within-site sources of variability and do not provide a gold standard for comparison. This concern will be more serious for sites with high within-site array-to-array variation

Rather than using the data from the whole array for normalization, another alternative is to use a portion of the data considered to be invariant across arrays to generate a scoring function for normalization (Li and Wong 2001; Schadt et al. 2001). Those probes in Affymetrix provided as Oligo B2 can be considered invariant with known concentrations. Again, the quality of those controls is key to the success of this approach, and we are currently investigating ways of implementing this approach.

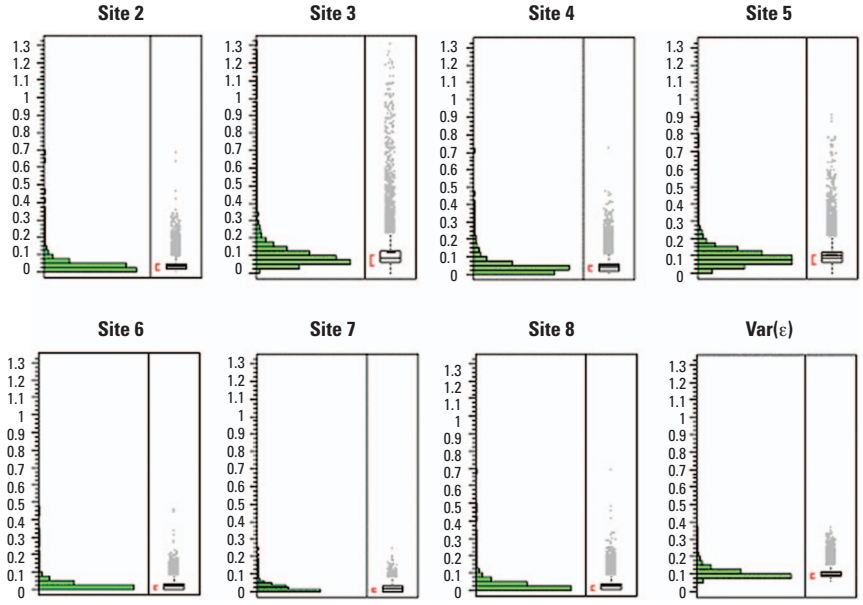

Figure 7. Histograms and box plots of the standard deviations of the random components, seven within-site array random factors and the stochastic errors, in Model 1.

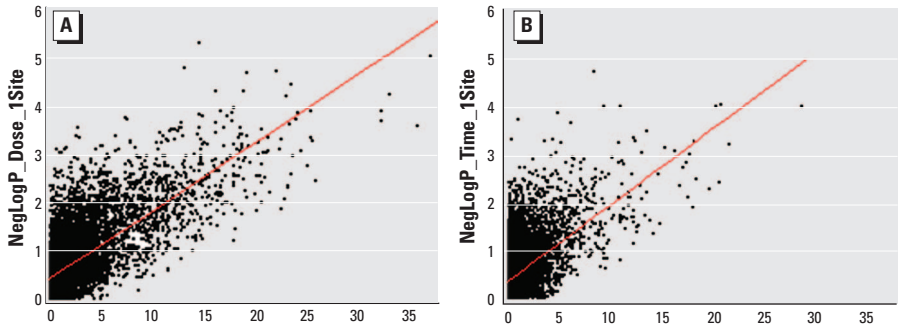

Figure 8. Scatterplots of negative $\log_{10}$ *p*-values of testing dose (*A*) and time (*B*) effects from Model 1 with data from seven sites and Model 2 with data from site 8 only. The red lines are regression-fitted lines with slopes 0.14 and 0.16 on *A* and *B* plots, respectively.

The mixed model provides a flexible method to adjust site effects and to use different array variations between sites. Significant site effects were revealed by this analysis as expected; however, only a few genes showed significant interaction effects between sites and treatments, dose, or time. In other words, each site tends to tell the same story regarding the list of significantly differentially expressed genes. This is a primarily positive result from this study and lends hope to the prospect of gaining power by combining study results. Similar studies are needed to extend this type of analysis to investigation of cross-platform data sets.

References

Affymetrix, Inc. 2002. GeneChip Expression Analysis: Data Analysis Fundamentals. Santa Clara, CA:Affymetric, Inc. Available: http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf [accessed 7 October 2003].

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high-density oligonucleotide array data based on bias and variance. Bioinformatics 19:185–193.

Choi JK, Yu U, Kim S, Yoo OJ. 2003. Combining multiple microarray studies and modeling interstudy variation. Bioinformatics 19 (suppl 1):I84–I90.

Chu T, Weir B, Wolfinger RD. 2002. A systematic statistical linear modeling approach to oligonucleotide array experiments. Math Biosci 176:35–51.

Chu T, Weir BS, Wolfinger RD. In press. Comparison of Li-Wong and loglinear mixed models for the statistical analysis of oligonucleotide arrays. Bioinformatics.

Dudoit S, Yang YH, Callow MJ, Speed TP. 2002. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat Sin 12:111–139.

Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM. 2003. Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. Funct Integr Genomics 3:180–188.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264.

Li C, Wong WH. 2001. Model-base analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 98:31–36.

Lockhart D, Dong H, Byrne M, Follettie M, Gallo M, Chee M, et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680.

SAS Institute Inc. 1999. SAS/STAT Software, Version 8. Cary, NC:SAS Institute, Inc.

Schadt E, Li C, Eliss B, Wong WH. 2001. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem (suppl 37):120–125.

Tan PK, Downey TJ, Spitznagel EL, Xu P, Fu D, Dimitrov DS, et al. 2003. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 31:5676–5684.