Interlaboratory Comparability Study of Cancer Gene Expression Analysis Using Oligonucleotide Microarrays

Kevin K. Dobbin,1,2 David G. Beer,4 Matthew Meyerson,8 Timothy J. Yeatman,9 William L. Gerald,10 James W. Jacobson,1 Barbara Conley,1 Kenneth H. Buetow,3 Mervi Heiskanen,3 Richard M. Simon,2 John D. Minna,12 Luc Girard,12 David E. Misek,5 Jeremy M.G. Taylor,7 Samir Hanash,5 John D. Minna,12 Luc Girard,12 David E. Misek,5 Jeremy M.G. Taylor,7 Samir Hanash,5 Katsuhiko Naoki,8 D. Neil Hayes,8 Christine Ladd-Acosta,13 Steven A. Enkemann,9 Agnes Viale,11 and Thomas J. Giordano6

1Cancer Diagnosis Program, 2Biometric Research Branch, and 3Center for Bioinformatics, National Cancer Institute, Bethesda, Maryland; Departments of 4Surgery, 5Pediatrics, and 6Pathology, University of Michigan Medical School; 7Department of Biostatistics, University of Michigan, Ann Arbor, Michigan; 8Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Pathology, Harvard Medical School, Boston, Massachusetts; 9Department of Surgery, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, Tampa, Florida; Departments of 10Pathology and 11Molecular Biology, Memorial Sloan-Kettering Cancer Center, New York, New York; 12Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, Texas; and 13Whitehead Institute-Massachusetts Institute of Technology Center for Genome Research, Cambridge, Massachusetts

ABSTRACT

A key step in bringing gene expression data into clinical practice is the conduct of large studies to confirm preliminary models. The performance of such confirmatory studies and the transition to clinical practice requires that microarray data from different laboratories are comparable and reproducible. We designed a study to assess the comparability of data from four laboratories that will conduct a larger microarray profiling confirmation project in lung adenocarcinomas. To test the feasibility of combining data across laboratories, frozen tumor tissues, cell line pellets, and purified RNA samples were analyzed at each of the four laboratories. Samples of each type and several subsamples from each tumor and each cell line were blinded before being distributed. The laboratories followed a common protocol for all steps of tissue processing, RNA extraction, and microarray analysis using Affymetrix Human Genome U133A arrays. High within-laboratory and between-laboratory correlations were observed on the purified RNA samples, the cell lines, and the frozen tumor tissues. Intraclass correlation within laboratories was only slightly stronger than between laboratories, and the interclass correlation tended to be weakest for genes expressed at low levels and showing small variation. Finally, hierarchical cluster analysis revealed that the repeated samples clustered together regardless of the laboratory in which the experiments were done. The findings indicate that under properly controlled conditions it is feasible to perform complete tumor microarray analysis, from tissue processing to hybridization and scanning, at multiple independent laboratories for a single study.

INTRODUCTION

Gene expression microarrays have been successfully used to identify novel taxonomies for cancer (1–5) and gene expression signatures associated with clinical outcome (4–9). These successes indicate that, within the particular laboratories carrying out these studies, the data from different microarray assays were comparable enough to identify biological heterogeneity between the tumors. However, interlaboratory comparability does not guarantee that data from tumor specimens analyzed in different laboratories will be comparable. The tumor marker literature has numerous examples of assays that produce consistent results within a particular laboratory but are inconsistent when data from different laboratories are compared (10, 11).

A study is being planned to confirm previously reported associations of gene expression signatures with patient outcome in stage I lung adenocarcinomas (4–6). Four laboratories will participate in the confirmatory study where 600 tumors will be analyzed with the goal of combining data from all sites. The main goal of this preliminary laboratory comparability study was to determine whether differences between the laboratories in tissue processing, RNA extraction, generation of labeled target, hybridization, and scanning result in comparable gene expression measurements from the same samples. If the data from the four laboratories were not comparable, then an alternative experimental design would be required for the confirmation study. The preliminary study was designed to identify sources of variation in gene expression measurements from frozen tissues, cell line samples, and purified RNA samples analyzed with Affymetrix Human Genome U133A arrays.

We are aware of no published studies that evaluate the interlaboratory comparability for microarray data on human tumor specimens. Piper et al. (12) assessed interlaboratory comparability of microarray measurements of yeast cell cultures grown in various conditions. There have additionally been some small, unpublished studies on limited numbers of extracted RNA samples. Because tissue handling and RNA extraction are considered major sources of variability in the
To assess the results of the interlaboratory comparability, we needed some estimate of the intralaboratory comparability to use as a baseline. Previous studies of intralaboratory comparability did not use the standardized protocols developed for this experiment or involve all aspects of the tumor assay process, including tissue processing (selecting and physically cutting the tissue to be assayed from the frozen tumor block, etc.), extraction of RNA from the tissue specimen, preparation of labeled cRNA target (reverse-transcription, labeling, fragmentation, etc.), and array hybridization, washing, and scanning. To address all of these issues, we designed an experiment to determine the intralaboratory and interlaboratory comparability of DNA microarrays at four laboratories using a standardized protocol.

MATERIALS AND METHODS

Study Design

Figure 1 shows the study design. Twelve tissues, predominantly tumors, were divided into six replicate blocks each, and each laboratory received one block from each tissue and one additional block from six of the tissues. The second tissue blocks were randomized to the four sites so that all tissues were analyzed six times. Five lung adenocarcinoma cell line pellets were divided into eight subsamples, and two subsamples of each were distributed to each laboratory. Five purified lung adenocarcinoma RNA samples were each divided into four aliquots, and one aliquot was distributed to each laboratory. Altogether, each of the four sites received 18 frozen tissue blocks, 10 cell line samples, and 5 RNA samples for a total of 33 microarrays per site. The 33 samples for each site were divided into two sets to be analyzed on different weeks; the first set contained one block from each of the 12 tissues, one sample from each of the 5 cell lines, and 3 of the RNA samples; the second set contained one block from each of six different tissues (within-laboratory replicates), replicate samples from each of the five cell lines, and the two remaining RNA samples. In addition, eight arrays were run with Stratagene universal human reference RNA. Of the total 140 microarrays, 4 failed and were discarded. Three of the four failed arrays were identified by the laboratories on the blinded samples. These three arrays were associated with the same frozen tumor tissue sample, which was a melanoma metastatic to lymph node. The high melanin content in this tumor was thought to be responsible for the high array failure rate. The fourth failed array was identified by National Cancer Institute based on extremely poor Affymetrix MAS 5.0 quality metrics. This array was associated with a cell line, and the reason for the failure is not known. The study was blinded as far as possible by assigning each distributed sample a numerical identifier, and the codes to break the identifiers were not provided until after all laboratories had uploaded the Affymetrix CEL files to the National Cancer Institute Director’s Challenge Web site (http://gedp.nci.nih.gov/).

The question being addressed in this study is whether a single tumor section, measured at different laboratories, would yield similar microarray measurements. The study was designed to address the sources of technical variation resulting from carrying out the analysis. It is not feasible to measure the same section at different laboratories, so instead we distributed different sections from the same tumor to different laboratories. We selected large homogeneous tissues for the study, which should show small expression variation across sections and hence show reproducible results for the within-laboratory replicates. Selecting homogeneous samples minimized the contribution of tissue heterogeneity to the variability in the analyses and ensured that we would have a noise-free baseline against which to assess the interlaboratory reproducibility.

Description of Tumor Tissue Samples, Cell Line Samples, and RNA Samples

Frozen tissues were collected at the Department of Pathology of the University of Michigan Health System from surplus tissue from surgical specimens. Selected tissues were large and grossly and histologically homogeneous to reduce variability due to tumor heterogeneity between replicates.
As the goal of our study was to examine variability within replicates from a tissue sample and not between different tissue samples, it was decided that the tissues could be of varying histogenesis and/or differentiations. Thus, 12 tissues, predominantly tumors, of diverse histogenesis and differentiation were selected, although 2 lung carcinomas were specifically included. Six frozen OCT blocks were prepared from each of the 12 tissue samples. Tissues were coded A-L and de-identified from the patient after the diagnosis was recorded. RNA quality was roughly assessed for each of the 12 tissues by gel electrophoresis. The OCT blocks were distributed to the four laboratories on dry ice, and each laboratory prepared a frozen section and selected the area to be used for RNA extraction. The tissues selected included two primary lung squamous cell carcinomas; adrenal cortical adenoma; primary gastric adenocarcinoma; normal liver; recurrent renal cell carcinoma, chromophobe cell type; primary malignant gastrointestinal stromal tumor; uterine leiomyoma; primary ovarian papillary serous adenocarcinoma; metastatic renal cell carcinoma, clear cell type; primary large cell lymphoma; and metastatic melanoma to lymph node.

RNA samples extracted from lung adenocarcinoma tumors were obtained from Ardais Corp. (Lexington, MA). The tumors were from a variety of patients with various stages of disease.

Cell line samples were obtained from the University of Texas Southwestern Medical Center at Dallas. The lung adenocarcinoma tumor cell lines used were NCI-H2009, NCI-H1437, NCI-H2087, NCI-H2347, and HCC78. All but HCC78 are deposited in the American Type Culture Collection (Manassas, VA; http://www.atcc.org). Cells were grown to

**Specimen Preparation and Laboratory Procedures**

Tumors were sectioned using a cryostat, and several 5 to 8 µm cryostat sections were quickly obtained and then stained with hematoxylin to select the regions for use in RNA isolation. Typically, these specimens were large enough to obtain all of the material for RNA from a single region (e.g., 5-10 mm³). Care was taken to avoid warming the specimen no more than to −18°C for the shortest time possible. The selected regions were at least 60% tumor cells (tumor cellularity), and tumors having mixed histology (i.e., adenocarcinomas) were not used. The tumor portions chosen for RNA isolation were obtained by cutting out that region of the tumor using a razor blade or scalpel cooled using dry ice. The material was then placed in labeled tubes and maintained at −80°C until RNA isolation.

**RNA Isolation.** RNA was isolated using 1 mL volume of Trizol (Invitrogen, Inc., Carlsbad, CA) reagent per sample using the protocol provided by the manufacturer. After RNA precipitation and 70% ethanol wash, the pellet was resuspended in RNase-free water and further purified using the RNeasy columns (Qiagen, Inc., Valencia, CA) as described by the manufacturer. RNA was eluted from the columns with RNase-free water. RNA concentration was determined by spectrometry, one 5 µg aliquot of the sample was used for reverse transcription, and another 1 µg portion was used for assessing RNA quality using the Agilent bioanalyzer.

**cRNA Synthesis and Hybridization: First-Strand Synthesis.** Briefly, 5 µg of total RNA were resuspended in Ambion (Austin, TX) DEPC-treated water and quantified by A260. The quality was determined using both the A260 ratio and Agilent bioanalyzer. RNA was converted into double-stranded cDNA by reverse transcription using a cDNA synthesis kit (Invitrogen). The oligo(dT)24 primer [Affymetrix, T7-oligo(dT) Promoter Primer kit] containing a T7 RNA polymerase promoter located 5’ to the poly(T) was used. The temperature of incubation was 42°C for 1 hour in a PCR machine. Following incubation, the mixture was quick spun in a centrifuge and placed on ice, and cold premixed second-strand reagents were added.

**cRNA Synthesis and Hybridization: Second-Strand Synthesis.** The Invitrogen cDNA synthesis kit was used for the second-strand synthesis. The incubations were done at 16°C for 2 hours in a PCR machine, and the reaction was stopped using 10 µL of 0.5 mol/L EDTA (Sigma Chemical Co., St. Louis MO). The reactions were immediately treated using a Sample Cleanup module (Affymetrix) and resuspended in 22 µL DEPC water. Labeled cRNA was generated from the cDNA sample by an in vitro transcription reaction that was supplemented with biotin-11-CTP and biotin-16-UTP (Enzo, Farmingdale, NY, via Affymetrix Enzo BioArray High-Yield Transcript Labeling kits) for 6 hours at 37°C in a PCR machine with no shaking. All of the labeled cRNA were purified using Affymetrix GeneChip Sample Cleanup module. The cRNA was quantified (using A260), and 15 µg were fragmented in a total volume of 40 µL fragmentation buffer at 94°C for 35 minutes using a PCR machine. The fragmentation buffer [200 mmol/L Tris-acetate (pH 8.1), 500 mmol/L KOAc, 150 mmol/L MgOAc] was provided in the Affymetrix GeneChip Sample Cleanup module.

**Preparation of the Hybridization Cocktail.** Fragmented cRNA (15 µg) was used to prepare 300 µL hybridization cocktail (100 mmol/L MES, 1 mol/L NaCl, 20 mmol/L EDTA, 0.01% Tween 20) containing 0.1 mg/mL (3 mL/300 mL) of herring sperm DNA (Promega, Madison, WI, 10 µg/mL) and 500 µg/mL acetylated bovine serum albumin (3 mL/300 mL, Invitrogen, 50 mg/mL). EDTA was obtained from Sigma Chemical, Tween 20 from Pierce Chemical (Rockford, IL), and DEPC-treated water from Ambion. Control cRNA used for comparison of hybridization efficiency between arrays and to standardize the quantitation of measured transcript levels is included as component of Eukaryotic Hybridization Control kit (Affymetrix, 20×) and uses 15 mL/300 mL hybridization cocktail. Before hybridization, the cocktails were heated to 94°C for 5 minutes, equilibrated at 45°C for 5 minutes, and clarified by centrifugation (16,000 × g) at room temperature for 5 minutes. Aliquots of each sample (10 µg fragmented cRNA in 200 µL hybridization cocktail) were prehybridized to U133A arrays at 45°C for 60 minutes and then hybridized for 16 to 18 hours in a rotisserie oven at 60 × g. The arrays were then washed using nonstringent wash buffer (6 × saline-sodium-phosphate-EDTA) at 25°C followed by stringent wash buffer [100 mmol/L MES (pH 6.7), 0.1 mol/L NaCl, 0.01% Tween 20] at 50°C. After staining with streptavidinphycoerytin (Molecular Probes, Eugene, OR), the arrays were washed again with 6 × saline-sodium-phosphate-EDTA and incubated with biotinylated anti-streptavidin IgG followed by
a second staining with streptavidin-phycocerythrin and a third washing with 6× saline-sodium phosphate-EDTA. The arrays were scanned using the GeneArray scanner (Affymetrix). Data analysis was done using AffyArray software. Features on the oligonucleotide arrays were carefully reviewed to confirm expression levels and exclude hybridization or washing artifacts.

Statistical Analysis

All data are publicly available for download at http://gedp.nci.nih.gov/ (experiment IDs 615-618). Affymetrix MAS 5.0 gene summaries were obtained for each array under the default parameter settings. Arrays were normalized as described by Wright et al. (13); genes with >50% Affymetrix “present” calls across arrays were identified, and for each array, the expression levels were multiplied by a constant to make the median of the identified genes close; signal values below 25 were truncated to 25, and the base two logarithm of the normalized intensities served as the signal. No filtering of genes was necessary because the truncation eliminated missing data and large negative log-intensity values. Comparability of two microarray measurements on the same sample was assessed by the Pearson correlation coefficient and the root mean square array measurements on the same sample was assessed by the large negative log-intensity values. Comparability of two microarray measurements on the same sample was assessed by the Pearson correlation coefficient and the root mean square deviation (14), defined as \[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \], where \( G \) is the number of genes on the array and \( x_i \) and \( y_i \) are the normalized log intensity readings for gene \( i \) on two different arrays. Pearson correlation was used because it is a common and intuitive measure in microarray studies. This was supplemented by examining the root mean square deviation because it captures aspects of comparability missed by correlation (i.e., systematically higher or lower expression values across genes). Comparability of repeated measurements on the same gene for pairs of replicated samples within a single laboratory, and for pairs of replicated samples between two laboratories, was assessed by intraclass correlation (ICC; refs. 15, 16). The ICC is calculated by fitting an ANOVA model for each gene: \[ \log_2(G_{i,j}) = S_{i} + L_{i} + E_{i,j}, \] where \( Y \) is the normalized intensity, \( S \) is the sample effect, \( L \) is the laboratory effect, and \( E \) is random error. ANOVA models were fit to the frozen tumor specimens and cell lines separately. ANOVA is generally robust, and the assumptions of normal and homogeneous error variances seemed adequate for the vast majority of genes based on \( \chi^2 \) goodness of fit tests and Fligner test for homogeneity of variances. Fitting the model results in estimates of the variance contributed by the different components, \( \sigma^2_g \), \( \sigma^2_s \), and \( \sigma^2_l \), for variation attributable to sample, laboratory, and measurement error, respectively. These were estimated from mean squares with adjustment for the partial replication and missing data (see Supplementary Material for details). The estimate of the ICC between laboratories is then \[ \hat{r}_{l}^2 = \frac{\sigma^2_g}{\sigma^2_l + \sigma^2_s + \sigma^2_g}, \] and within each laboratory is \[ \hat{r}_{s}^2 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_g}. \] In an ideal experiment, variation attributable to laboratory, \( \sigma^2_l \), and measurement error, \( \sigma^2_g \), would both be near 0, resulting in within-laboratory and between-laboratory ICC values close to 1. Values of the ICC close to 1 indicate good comparability, and values close to 0 indicate poor comparability. Within-laboratory ICC serves as the baseline against which between-laboratory ICC will be assessed. If the variation attributable to laboratory, \( \sigma^2_l \), is large, then the between-laboratory ICC will be small compared with the within-laboratory ICC. If the laboratories are comparable, then the between-laboratory ICC and the within-laboratory ICC will be similar. Separate within-laboratory and between-laboratory ICCs are calculated for each gene.

As a further test of the comparability of the data across sites, we ran hierarchical agglomerative cluster analysis using average linkage separately on the tumor tissue samples, the cell line pellets, and the purified RNA samples. The distance metric \( (1 - \text{correlation}) \) was used, although Euclidean distance yielded similar results.

RESULTS

Comparability of Gene Expression Profiles for Repeated Sample Measurements

Figure 2 shows a box plot of the correlations between the pairs of repeated measurements taken on the same samples. A similar figure for root mean square deviations is in the Supplementary Material. The median correlations of the within-laboratory gene expression profiles were 0.95, 0.96, and 0.97 for the tumor tissue, cell lines, and Stratagene reference RNA based on 21, 19, and 8 pairs of replicated samples, respectively. (No within-laboratory replicates on the tumor RNA were done.) Within-laboratory reproducibility was similar for each of the four laboratories, so the within-laboratory data are combined (see Supplementary Material for data on laboratory-to-laboratory variability). A similar pattern held for the root mean square deviations. Overall, the within-laboratory comparability is very good. The median correlations tend to be high, and the scatter of the correlations is fairly narrow.

A somewhat oversimplified interpretation is that within-laboratory variation associated with preparation of labeled cRNA target (reverse-transcription, labeling, fragmentation, etc.), array hybridization, washing, and scanning results in a 0.97 median correlation and a tight scatter of the correlations. Adding in the variability due to extraction of RNA from the tissue specimen reduces the median correlation slightly to 0.96 and produces

![Fig. 2 Correlations between repeated measures on identical samples.](image-url)
Hierarchical Clustering Reflects the Biological Variation within the Different Types of Samples

To determine the ability of between-laboratory studies to accurately discriminate among cancer samples, we did hierarchical clustering on the three types of samples: the samples for which purified tumor RNA samples were provided to each laboratory (Fig. 3A), those for which cell line pellets were provided (Fig. 3B), and those for which tumor tissue samples were provided (Fig. 3C). In each of these cases, the hierarchical cluster analysis accurately grouped the related samples together despite the many variables that could separate these samples. It is also interesting to note in Fig. 3C that the primary tumor specimens cluster close together, with the metastatic (f) and normal liver (c) samples distant from the primary tumor group; in interpreting Fig. 3C, recall that the distance between two clusters is represented by the vertical height of the connection between them, not the horizontal distance between them, so that, for instance, the metastatic and normal liver clusters on the right side are far distant from each other.

The dramatic similarity revealed by the hierarchical clustering analysis should not, however, be overinterpreted. The cell line and purified RNA are from lung adenocarcinoma samples, but the tumor tissue samples represent a variety of different tumor and tissue types. Consequently, on average, one would expect smaller correlations between the tumor tissue samples than between the cell line or tumor RNA samples or among a collection of all lung adenocarcinoma tumor tissue. For example, the median within-laboratory correlation between the different lung adenocarcinoma RNA samples, the different lung adenocarcinoma cell lines, and the different frozen tumor tissue samples was 0.88, 0.88, and 0.79, respectively. This implies that when performing cluster analysis using the distance metric (1 − correlation) the distances between the tumor tissue samples would tend to be larger than between a more homogeneous group of lung adenocarcinoma samples; hence, it would be “easier” for a cluster analysis algorithm to distinguish between the tumor tissues used in the study, and the results of cluster analysis on the tumor tissues would be overly optimistic for this reason.

To get a more realistic picture, note that in Fig. 2 the correlations between repeated measures on the same samples decreases by a median of 0.01 when moving from either the cell line or purified RNA to the frozen tissue specimens. This suggests estimating the correlations between biologically different frozen lung adenocarcinoma tissue samples by adding 0.01 to the correlations between the RNA samples and to the correlations between the cell lines. The resulting estimated first quartile, median, and third quartile correlations between different frozen lung adenocarcinoma tissue samples, lung adenocarcinoma cell lines, and the different frozen tumor tissue samples was 0.86, 0.87, and 0.88, respectively. Comparing this with the observed correlations between the repeated observations on the tumor tissues, which had first quartile, median, and third quartile of 0.92, 0.93, and 0.95, respectively, suggests that a more homogeneous set of lung adenocarcinoma tissue samples would still cluster almost entirely by tumor across sites.

Comparability of Repeated Measurements on the Same Gene across Samples

For each gene, we used ANOVA to separate the sources of variation into three types: biological variation due to differences in gene expression among the different biological specimens (the 12 different tissue specimens); laboratory variation resulting from laboratory-to-laboratory variability; and error variation resulting from measurement error, which will be present even when the same sample is measured multiple times in the same laboratory.

Results for the frozen tumor tissues are shown in Fig. 4. In general, the variability attributable to the laboratories was the smallest source of variation followed by variation due to measurement error and finally by biological variation as the largest. This suggests that the different laboratories are contributing some extra variation into the measurements but that variation tends to be relatively small compared with the other sources of variability already present in this type of data obtained from primary tumors. Similar results held for the cell line samples and are given in the Supplementary Material.
Table 1 presents summary statistics on the ICC. Intuitively, the ICC is similar to Pearson correlation. In fact, the within-laboratory ICC for a gene is the same as the Pearson correlation. The mean ICCs between laboratories are very similar to within-laboratories, whereas the medians are slightly farther apart. The separation is attributable to greater variability and skewness in the within-laboratory data due to the smaller sample sizes within each laboratory compared with the combined data across laboratories. One can see that there is some evidence that the center of these distributions may be different; but comparing the quartiles, there is clearly a large level of overlap between within-laboratory and between-laboratory ICC.

In general, the ICCs in Table 1 are much smaller than the correlations in Fig. 2. This may seem counterintuitive, but the reason for the difference is that the experimental error in both cases will be similar, although the much greater dynamic range when considering expression across all genes (Fig. 2) compared with the range of expression for a single gene (Table 1) will produce much higher correlation in the former case than in the latter. However, a within-laboratory median correlation of 0.66 intuitively may seem small even when considered on a gene-specific basis. To investigate this further, Table 2 shows a breakdown of the ICCs by the overall expression level of the genes across the tumor tissue samples. From Table 2, one can see that for genes with low levels of expression (e.g., in the lowest 25% of genes), the ICC tends to be very small, with mean 0.31 between laboratories, indicating poor reproducibility of the measurements (although

---

**Fig. 3** Hierarchical cluster analysis dendrograms. A, purified RNA samples: u1, RNA sample u analyzed at laboratory 1; v2, RNA sample v analyzed at laboratory 2, etc. B, cell line samples: M1, cell line sample M analyzed at laboratory 1; etc. Outlier is the failed cell line array. C, tumor tissue samples; each letter is a different tumor; 12 tumors labeled a, b, ..., l.
The poor reproducibility is a result of the small biological variation in the expression of these genes across the samples. In fact, if biological variation is 0, then the ICC is 0 by definition, because the variation in measurements across samples is random experimental noise, and experimental errors are uncorrelated. Hence, the low ICC values for genes with low expression should not be interpreted as indicating increased noise in the measurements themselves but as indicating instead that the biological variation is so small for some genes that comparability of measurements across samples cannot be adequately assessed. Although there is no firm cutoff as to unacceptably low ICC, from Table 1 it is clear that for at least a quarter of genes the ICC is so low (lower quartile below 0.26) that it would require impractically large sample sizes to identify any small differential expression that may exist among these low expressed genes. For genes with high levels of expression (e.g., genes in the topmost 25%), the ICC tends to be much bigger, with mean 0.74. The same relation holds between the variance of the genes and the ICC; genes with higher overall expression variance showing higher ICC as well. The group with the highest 25% expression had higher biological SD (median 0.65 versus 0.60) than overall, so what is driving the high ICC values is real biological activity. Genes with the highest biological variation estimates split the samples into a group with very low expression and a group with very high expression with no overlap between the two and little variation within each group. Indeed, it is these high expression and high ICC genes, which also have high variance, that drive the correlation clustering because this metric is affected by variation.

In summation, between-laboratory reproducibility of gene expression measurements seemed very good compared with within-laboratory reproducibility. The pattern of reproducibility across genes was similar in both the tumor tissue samples and the cell line samples. Both between-laboratory and within-laboratory comparability seemed best for highly expressed genes. Genes with low expression across the samples were the ones that displayed systematic differences in expression on both between-laboratory and within-laboratory replicates.

**Table 1** ICCs for frozen tumor blocks: summary statistics for ICCs for the 22,283 probe sets on the microarrays

<table>
<thead>
<tr>
<th>ICC</th>
<th>Mean</th>
<th>Median</th>
<th>Lower quartile</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between labs</td>
<td>0.52</td>
<td>0.59</td>
<td>0.26</td>
<td>0.79</td>
</tr>
<tr>
<td>Within labs</td>
<td>0.53</td>
<td>0.66</td>
<td>0.24</td>
<td>0.88</td>
</tr>
</tbody>
</table>

**DISCUSSION**

We have presented the results of a study of the within-laboratory and between-laboratory reproducibility of data from microarray experiments. The study was designed with sufficient iterations at different replication levels to be able to distinguish different sources of variability in the data. The results of this study have implications for at least two important questions concerning this technology: (a) Can data from microarray experiments done at different laboratories be combined for analysis in a single study? This could greatly simplify the logistics of organizing and rapidly completing large confirmatory studies. (b) Is there potential for developing clinical tests based on microarray technology that might meet reproducibility and robustness criteria required by agencies such as the Food and Drug Administration? This study was designed to address the first question, and we have shown the adequacy of the reproducibility for this purpose. This study has not resolved the second question. However, our results do indicate a certain level of robustness and reproducibility in the microarray analyses, which is one prerequisite for the translation of this technology to the clinic.

Within-laboratory comparability of measurements was highest for the Stratagene RNA sample, lower for the cell line samples, and lowest for the tumor tissue sections. The additional within-laboratory variability associated with RNA extraction and tissue handling seemed to be relatively minor. Comparability between laboratories generally seemed to be lower than within laboratories for all three types of samples, but the loss in comparability seemed to be fairly minor and chiefly associated with sample labeling, hybridization, and scanning as opposed to tissue handling or RNA extraction.

**Table 2** Tumor tissue ICC by expression level group: summary statistics for the ICCs for the 22,283 probe sets on the microarray broken down by median probe set intensity quartile over the arrays

<table>
<thead>
<tr>
<th>Expression level group</th>
<th>Mean</th>
<th>Median</th>
<th>Lower quartile</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest 25% genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between laboratories</td>
<td>0.31</td>
<td>0.19</td>
<td>0.05</td>
<td>0.54</td>
</tr>
<tr>
<td>Within laboratories</td>
<td>0.25</td>
<td>0.10</td>
<td>-0.03</td>
<td>0.65</td>
</tr>
<tr>
<td>25-50% genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between laboratories</td>
<td>0.43</td>
<td>0.42</td>
<td>0.19</td>
<td>0.68</td>
</tr>
<tr>
<td>Within laboratories</td>
<td>0.44</td>
<td>0.54</td>
<td>0.16</td>
<td>0.80</td>
</tr>
<tr>
<td>50-75% genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between laboratories</td>
<td>0.62</td>
<td>0.66</td>
<td>0.47</td>
<td>0.80</td>
</tr>
<tr>
<td>Within laboratories</td>
<td>0.62</td>
<td>0.74</td>
<td>0.46</td>
<td>0.89</td>
</tr>
<tr>
<td>Highest 25% genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between laboratories</td>
<td>0.74</td>
<td>0.78</td>
<td>0.66</td>
<td>0.86</td>
</tr>
<tr>
<td>Within laboratories</td>
<td>0.74</td>
<td>0.84</td>
<td>0.66</td>
<td>0.93</td>
</tr>
</tbody>
</table>

**Fig. 4** Tumor tissue data only: estimated SD of error contributed by laboratory, measurement error, and biological variation for each of the 22,283 probe sets on the array.
Microarray data between laboratories seems comparable under the protocols used. For future studies, it seems that standardizing equipment, protocols, and reagents associated with preparation of labeled cRNA target (reverse-transcription, labeling, fragmentation, etc.) and array hybridization, washing, and scanning across sites may be important for ensuring comparability.

Much effort was made to standardize the laboratory protocols across the laboratories and to ensure that the Affymetrix scanners, reagents, etc., were as consistent as possible at the sites. The laboratories involved in this study all have extensive experience with Affymetrix gene chips. Three are medium-sized core facilities and one is a large, high-throughput core facility. Thus, it does not seem that these results can necessarily be generalized to less controlled situations or used to justify combining publicly available data from previous studies for analysis. In addition, it is important to bear in mind that some preliminary studies have suggested that comparability across platforms or even across probe set summaries within the same platform may be poor (17), although others have had some success combining data across platforms (18).

It is perhaps surprising how small a role biological variation within the tumors seems to have played, because there was high correlation between all six tumor sections across the tumors (with the exception of the failed arrays of the problematic melanoma). Because the tumors were selected based on apparent homogeneity, these tumors may be particularly homogeneous in gene expression. Nevertheless, the hierarchical clustering analysis was striking because the tumor samples consistently clustered together despite the use of independent tumor sections, the selection of tumor regions at different institutions, and the independent RNA extraction, target labeling, and array hybridization and scanning results. This argues for the reproducibility and potential utility of microarray analysis.

If gene expression signatures are to be clinically useful in the future, they must be converted into standardized assays that are reproducible. How this transition step will occur is not yet clear. It may involve taking the results of microarray studies and developing assays based on different technologies (such as quantitative reverse transcription-PCR). However, the results of this study indicate that it may also be possible to create a standardized assay based on microarray technology that is reproducible enough for clinical use.

ACKNOWLEDGMENTS

We thank Fredrick Pollock and Affymetrix for technical support for this study, Dr. Chantale T. Guy and Ardais Corp. for the generous gift of purified RNA from lung tumors, and Drs. Tracy G. Lugo and Sheila E. Taube for the careful reading of the article.

REFERENCES