The intraclass correlation coefficient applied for evaluation of data correction, labeling methods, and rectal biopsy sampling in DNA microarray experiments

Linette Pellis, Nicole L. W. Franssen-van Hal, Jan Burema, and Jaap Keijer

DNA microarray technology is becoming a commonly and widely used hybridization based technique in research and clinical applications (4). The technique originated from the need to address the functions of the large numbers of genes that have been identified in large-scale sequencing programs. The use of high-precision robots allowed researchers to immobilize tens of thousands of DNAs on extremely small areas of a solid base, generating high-density DNA microarrays. Finally, the advent of fluorescent technology made it possible to address the microarrays in a relatively simple, safe, and accurate manner. All of this has led to a great increase in the throughput of analysis of biological molecules and samples. Applications of DNA microarrays range from genetic mapping studies and mutational analysis to genome-wide monitoring of gene expression (5).

The technology was first used for assessment of cellular mRNA expression levels. The goal of this application is to determine gene expression of the “whole” genome in any cell or tissue sample of interest (14). For this, isolated (m)RNA is labeled and hybridized to the DNA microarray. After washing and scanning of the DNA microarray, a raw fluorescence data image is obtained which should represent overall gene expression level of the original sample. Usually, not every gene is represented on the DNA microarray, and hybridization results are influenced by multiple factors, namely, the labeling method, hybridization conditions, the sequence of the gene, and target features. Since in most cases not absolute expression levels but differences in expression between samples are determined, problems can be minimized by direct comparison of two samples, labeled with different fluorophores (e.g., Cy3 and Cy5) on one array (7). To compare multiple samples (each separately labeled with, e.g., Cy5), a common reference sample is labeled (labeled with Cy3) and can be used in each hybridization. In the ideal situation the reference represents each spotted gene on the DNA microarray, but in practice this is a pool representing each investigated sample.

For expression analysis to be efficient and reliable, reproducible laboratory protocols and validated procedures for data normalization are required (8). Labeling procedures are constantly optimized and new procedures are being developed, primarily to reduce the amount of input material that is required (10, 21, 23). Critical to the success of these protocols is that they are reproducible. Therefore, it is striking that no simple, standard protocol is available to assess the reproducibility. The last step before comparing the gene expression of different samples is data normalization, which is performed on the obtained raw fluorescent data. Data normalization makes use of the Cy3 reference sample images and allows direct comparison of the Cy5 sample values on different slides (1, 7). This is a widely used approach, but again, no simple standard method is available to assess whether data normalization is performed correctly. Different methods have been reported to address the way data normalization and labeling methods are performed (17, 22). These methods range from simple linear correlation coefficient ($R^2$) analysis to sophisticated statistical methods (12, 13, 22). When evaluating different labeling methods, one is interested in the systematic deviation (i.e., difference in the amount of labeled RNA samples), which preferably should be as small as possible. However, when the $R^2$ is used in DNA microarray data analysis, its feature to be insensitive to the systematic deviation makes it unsuitable for the purpose of assessing reproducibility (2). Another limitation of the $R^2$ is that it only facilitates the comparison of two
samples at the same time. An attractive alternative is the use of the intraclass correlation coefficient (ICC). This is a relatively simple statistical procedure used to determine the reproducibility of a measurement of a variable (3, 18, 20). This correlation is based on variance components analysis and measures the homogeneity within groups relative to the total variation. The ICC is large when there is little variation within the groups compared to variation among group means, where groups consist of replicate measurements. A small ICC occurs when within-group variation is large compared with between-group variability, indicating that some unknown variable has introduced nonrandom effects in the different groups. The maximum value of the ICC is 1, and the minimum value is theoretically 0 (3, 13, 18, 20). The ICC is routinely used in epidemiological studies to address the test-retest reliability, validity of questionnaires, interlaboratory concordance, and correlation of plasma/tissue levels to disease status. Having seen how the ICC is utilized for reliability, reproducibility, and validation analysis (6, 15, 16), we decided to investigate whether it can be used to assess technical variation in DNA microarray technology and, more specifically, to assess the reproducibility of sample RNA labeling methods and to optimize data normalization. We decided to extend its use to assessment of biological variation. To this end, the within- and between-person variation in gene expression in small biological samples was estimated. The analysis of gene expression in biological samples is used for clinical as well as epidemiological studies. Since human material is often scarce, it is necessary to determine how many biopsies should be taken to acquire sufficient accuracy in the assessment of the tissue that is analyzed.

In this paper, we show that the ICC is a relatively simple statistical measure that can be used to estimate methodological and biological variation as exemplified by addressing the validity of our data normalization procedure, by comparing the reproducibility of different labeling methods and by analyzing variation of gene expression in human rectal biopsies.

MATERIAL AND METHODS

Sample preparation. Caco-2 and HT-29 cells (ATCC, Manassas, VA) were inoculated in 75-cm² culture flasks at 40,000 cells/cm². Cells were grown at 37°C in air with 5% CO₂ and 100% relative humidity in Dulbecco’s modified Eagle’s medium (Sigma, Zwijndrecht, The Netherlands) supplemented with NaHCO₃ (3.7 g/l, Sigma), nonessential amino acids (1×: ICN, Zoetermeer, The Netherlands), FCS (5%; Invitrogen, Breda, The Netherlands), penicillin (5,000 U, Sigma), and streptomycin (21-mer), and/or 150.0 ng random hexamers in a reverse transcription reaction was performed for 2 h at 37°C in a final volume of 25 μl. The reaction mixture contained the RNA template with the annealed primer, 1× first-strand buffer (Invitrogen), 10 mmol/l dithiothreitol, 0.5 mmol/l dATP, 0.5 mmol/l dGTP, 0.5 mmol/l dTTP, 0.04 mmol/l Cy3-labeled dCTP (or Cy5-labeled CTP, Amersham), 15 U of RNase OUT (Invitrogen), and 150 U of SuperScript II reverse transcriptase (Invitrogen). The labeled cDNA obtained was purified by an ethanol precipitation performed at room temperature. The pellet was dried and dissolved in 10 mmol/l Tris-HCl (pH 8.0) and 0.5 mmol/l EDTA). After a 3-min boiling step, the cDNA was immediately put on ice, and 2.5 μl of 1 mol/l NaOH was added. The cDNA was then incubated for 10 min at 37°C to break down the remaining RNA. To neutralize the pH, 2.0 μl of 1 mol/l HCl and 2.5 μl of 1 mol/l Tris-HCl (pH 8.0) were added. Finally, an additional ethanol precipitation at room temperature was performed, and the resulting cDNA pellet was dissolved in 10 μl TE, pH 8.0.

Direct labeling. The RNA labeling protocol is based on the protocol from Schena et al. (19). Either total RNA or mRNA was labeled by incorporation of either Cy3-dCTP or Cy5-dCTP during reverse transcription. A combination of HT-29 and Caco-2 mRNA (1.0 μg) or total RNA (40.0 μg) was used for the labeling. In short, 1.0 μg of sample poly(A)⁺ mRNA (10 μg total RNA mixed with 1.0 μg control luciferase mRNA (Promega, Leiden, The Netherlands), 2.0 μg oligo(dT)₂₁ primer (Isogen, Maarssen, The Netherlands), and/or 150.0 ng random hexamers (Invitrogen) in a final volume of 13.5 μl, heated for 3 min at 65°C (RNA denaturation) and 10 min at 25°C (primer annealing), then immediately put on ice. Then, a reverse transcription reaction was performed for 2 h at 37°C in a final volume of 25 μl. The reaction mixture contained the RNA template with the annealed primer, 1× first-strand buffer (Invitrogen), 10 mmol/l dithiothreitol, 0.5 mmol/l dATP, 0.5 mmol/l dGTP, 0.5 mmol/l dTTP, 0.04 mmol/l Cy3-labeled dCTP (or Cy5-labeled CTP, Amersham), 15 U of RNase OUT (Invitrogen), and 150 U of SuperScript II reverse transcriptase (Invitrogen). The labeled cDNA obtained was purified by an ethanol precipitation performed at room temperature. The pellet was dried and dissolved in 100 μl Millipore-filtered water. All cDNAs were purified using the PCR purification protocol.
Microarrays construction. An in-house-produced subtracted cDNA library, enriched for genes which are expressed in differentiated and undifferentiated Caco-2 cells (A. Peijnenburg et al., unpublished results), were printed on silylated slides (CEL Associates, Houston, TX) using a PixSys 7500 arrayer (Cartesian Technologies, Durham, NC). Arrays were spotted by passive dispensing using quill pins (ChipMaker 3; TeleChem, Sunnyvale, CA), resulting in a spot diameter of 0.12 mm at a volume of about 0.5 nl. After printing, microarrays were allowed to dry at room temperature for at least 3 days. Free aldehyde groups were blocked with NaBH₄ according to the method of Schena et al. (19). The microarrays, used for the validation of the normalization method and assessment of different labeling methods, contained 1,152 spotted genes in duplicate. For the assessment of biological variation, the microarrays that were used contained 2,304 single-spotted genes.

Microarray hybridization. Prior to hybridization, microarrays were prehybridized in hybridization buffer at 42°C for several hours. After prehybridization, slides were rinsed twice in Millipore-filtered water, once in isopropanol and dried by centrifugation (2 min, 470 g). The hybridization was performed in a Geneframe (1 × 1 cm², 25-µl hybridization volume; Westburg, Leusden, The Netherlands). A 1:1 (vol/vol) mixture of Cy3- and Cy5-labeled cDNAs was hybridized to each array. Arrays were hybridized overnight at 42°C in a humid hybridization chamber. After hybridization, slides were washed at room temperature, first in 1× SSC/0.1% SDS (5 min) and subsequently in 0.1× SSC/0.1% SDS (5 min) and 0.1× SSC (1 min), and then dried by centrifugation (2 min, 470 g).

Microarray scanning. Microarrays were scanned using a confocal laser scanner (ScanArray 3000; General Scanning, Watertown, MA) containing a HeNe 543-nm laser for Cy3 measurement and a RhNe 633-nm laser for Cy5 measurement. Scans were made with a pixel resolution of 10 µm, a laser power of 90%, and a photomultiplier tube voltage of 80%. The software package ArrayVision (version 7.0; Imaging Research, Ontario, Canada) was used for image analysis of the TIFF files, as generated by the scanner. Density values of each spot, multiplied by the area and the background (surrounding entire template), were collected and stored for further data processing in Microsoft Excel.

Experimental setup and data normalization. All arrays were hybridized with a Cy5-labeled sample cDNA and a Cy3-labeled reference cDNA. For validation of the normalization method and the assessment of different labeling methods, the reference cDNA was a mixture of directly labeled mRNA of HT-29 and Caco-2 cells. A mixture of indirectly labeled total RNA from rectal biopsies and IIT-29 and Caco-2 cells was used as reference cDNA for assessment of biological variation. The reference cDNA was pooled after fluorescent labeling and subsequently subdivided and hybridized on all arrays. The reference hybridization signals of the spots should under ideal circumstances be identical on each slide. In practice this signal will differ because of 1) fluctuations in the amount of DNA spotted and 2) variations in the hybridization conditions within a slide, between slides, and between different experiments (random variation). For these reasons, corrected sample hybridization signals, Cy5spot(x,slide(X)), were calculated for each spot x on slide X according to Eq. 1 in Fig. 1 of the first correction, where N is the total number of hybridized slides, Cy3spot(x,slide(X)) and Cy3spot(x,slide(X)) are the measured Cy3 and Cy5 hybridization signals of spot x on slide X, respectively, and Cy3median is the median of the hybridization values of spot x on all slides hybridized in the experiment. To correct for differences in labeling efficiency between samples and for inaccuracies in the amount of sample mRNA used in the labeling reaction (systematic labeling deviation), sample hybridization values were also corrected for the median Cy5 signal according to equation in Fig. 1 of the second correction, where n is the total number of spots on the array, Cy5median corr,spot(x,slide(X)) is the median of the Cy5 signals of all spots on slide X after the first correction, and

\[ \text{ICC}_{i} = \frac{\text{MS}_b - \text{MS}_w}{\text{MS}_b + (N - 1)\text{MS}_w}, \]

where MS_b is the between-spots mean square (spot 1, ..., n), MS_w is the within-spots mean square (spot 1, ..., n in slide 1, ..., N), and N is the number of replicative spots. \( \text{ICC}_i \) varies from 0 to 1, and an value of ICC_i ≥ 0.95 is usually interpreted as good agreement. The ICC score is a measure of linear association between two variables. One feature of \( \text{ICC}_i \) is that it disregards systematic error; this implies that the ICC_i observes the average of the cloud of points as if it was on the y = x line.

To gain insight into characteristics of the ICC for the analysis of DNA microarrays, a data set was generated by labeling a sample two times independently with Cy5. This was mixed afterward with Cy3-labeled RNA from the same reference pool, then hybridized to separate but identical slides. From this microarray data set three scatter plots were made (Fig. 2). The first scatter plot was obtained from the raw microarray data (Cy5 values, Fig. 2A). Correcting this data for the random error (Fig. 1, Eq. 1) resulted in the second scatter plot (Fig. 2B). The third scatter plot (Fig. 2C) was obtained with the data of the second scatter plot after adjusting this for the systematic error (Fig. 1, Eq. 2). Subsequently, the \( R^2 \) and ICC were calculated for each scatter plot. The uncorrected raw microarray data set rendered the smallest ICC and \( R^2 \) of all three scatter plots (ICC = 0.910 and \( R^2 = 0.977 \)). After correcting the data for the random error (Fig. 1, Eq. 1), both ICC and \( R^2 \) improved (ICC = 0.936 and \( R^2 = 0.989 \)). In contrast, when correcting for the systematic error (Fig. 1, Eq. 2), only the ICC improved (0.936 vs. 0.989). This shows that
the ICC is sensitive for both the random variation and the systematic deviation, whereas the $R^2$ is only sensitive for the random variation. Thus the ICC can be used to assess systematic deviation, whereas the $R^2$ cannot. Since the number of replicates is not limited for calculation of the ICC, more than two samples can be compared at once. The $R^2$ can only be calculated for two samples at the same time and thus only allows comparison for each scatter plot separately. Because the ICC is sensitive for both random variation and systematic deviation, it is a useful tool to obtain insight in both technical and biological variation.

Assessment of technical variation: data normalization. Different normalization methods have been described for DNA microarray data sets (1, 2, 12, 17). Working with Cy5/Cy3 ratios corrected by the mean or median Cy5 signal on each array is often used for DNA microarray data normalization (8, 17). To assess the best method, correcting based on the mean or median signal, we performed a hybridization experiment, applied the normalization steps (Fig. 1, Eqs. 1 and 2) with either the mean or median, and calculated the ICC for each method. Five arrays were hybridized with identical samples labeled with Cy5 independently, and a Cy3-labeled reference cDNA was used to allow comparison between the arrays. When correcting the data for the random error (Fig. 1, Eq. 1) with the mean or median Cy3 signal of each spot, no difference in ICC was observed. If subsequently correction for the systematic error (Fig. 1, Eq. 2) is performed with the mean or median value of the overall Cy5 signal, then a $\Delta ICC_{\text{mean}}$ of 0.030 (SD = 0.017) and $\Delta ICC_{\text{median}}$ of 0.037 (SD = 0.012) was obtained. Correction for the systematic error with median value gave a slight improvement compared with the mean value. Since use of the median value for the correction has the advantage that outliers will not disturb the corrections, median values were used for both corrections in the data normalization in all further experiments.

Assessment of technical variation: labeling methods. Various different labeling methods have been described for microarray hybridization experiments (9, 19, 21, 23). These methods can be divided in direct or indirect labeling methods. In the direct labeling method the fluorescent dye is incorporated during the cDNA synthesis, whereas in the indirect labeling method coupling of fluorescent dye to cDNA occurs afterwards. Since the fluorescent dyes are very bulky, indirect labeling may provide an advantage, since less bulky nucleotide

---

Fig. 1. Procedure for data normalization, divided into a first (random error) and second (systematical error) correction. First correction (Eq. 1) is accomplished by using the median value of the Cy3 signal (virtual slide) for each spot separately, whereas the second correction (Eq. 2) makes use of the median of the overall Cy5 signal.

**FIRST CORRECTION (Eq. 1)**

<table>
<thead>
<tr>
<th>reference Cy3</th>
<th>sample Cy5</th>
</tr>
</thead>
<tbody>
<tr>
<td>slide 1</td>
<td>slide 1</td>
</tr>
<tr>
<td>slide 2</td>
<td>slide 2</td>
</tr>
<tr>
<td>slide 3</td>
<td>slide 3</td>
</tr>
<tr>
<td>slide 4</td>
<td>slide 4</td>
</tr>
<tr>
<td>slide N</td>
<td>slide N</td>
</tr>
<tr>
<td>virtual slide</td>
<td>virtual slide</td>
</tr>
</tbody>
</table>

**SECOND CORRECTION (Eq. 2)**

<table>
<thead>
<tr>
<th>sample Cy5&lt;sub&gt;corr&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>slide 1</td>
</tr>
<tr>
<td>slide 2</td>
</tr>
<tr>
<td>slide 3</td>
</tr>
<tr>
<td>slide 4</td>
</tr>
<tr>
<td>slide N</td>
</tr>
</tbody>
</table>

**Eq. 1**

$$Cy_{5<sub>corr</sub>}^{\text{spot}(x),\text{slide}(x)} = \frac{Cy_{5}^{\text{med}(\text{median Cy5 signal})}}{Cy_{5}^{\text{med}(\text{median Cy5 signal})}} \times Cy_{5}^{\text{med}(\text{median Cy5 signal})}$$

**Eq. 2**

$$Cy_{5<sub>corr</sub>}^{\text{med}(\text{median Cy5 signal})} = Cy_{5}^{\text{med}(\text{median Cy5 signal})} \times Cy_{5}^{\text{med}(\text{median Cy5 signal})}$$
Each microarray. The raw microarray data was scatter plots compare the spot intensities of 1,054 genes that are represented on modifications can be used, which are likely to improve reverse transcriptase function and will result in a higher insertion of the modified nucleotides and longer cDNA strands. We compared the quality and reproducibility of oligo(dT)-primed direct labeling and indirect labeling using either mRNA or total RNA as input material. In addition, for labeling of mRNA, also a combination of oligo(dT) and random hexamer primers was tested in the indirect as well as in the direct labeling method. This approach aims at a higher labeling efficiency and cannot be used for total RNA since it would also result in labeling of ribosomal RNA. For an overview of the methods tested, see Table 1. In the indirect labeling method, aminoallyl nucleotides were incorporated during the cDNA reaction followed by reactive dye coupling. We also tested biotin nucleotide incorporation followed by streptavidin-dye coupling. However, in our hands the biotin-streptavidin-based technique resulted in very poor signal-to-noise ratios, due to high background signals. Therefore, this method was not taken along in further evaluation (data not shown).

Each labeling method was performed five times, independently, using Cy5 fluorescent dye. All methods were hybridized to separate but identical microarrays. The labeling methods were compared by using the Cy3-labeled reference cDNA, which was simultaneously hybridized on each array. To assess both random and systematic labeling deviation, only the first normalization step was performed (Fig. 1, Eq. 1). To assess only the random labeling variation, both normalization steps were performed (Fig. 1, Eqs. 1 and 2). As a measure for reproducibility, the ICC was calculated for each method. As a measure for quality, the average signal spot intensity, background, and signal-to-noise ratio were calculated for each method (Table 1).

In all three direct labeling methods a smaller ICC was obtained compared with the ICC of indirect labeling methods when both corrections were applied, whereas a higher ICC was obtained from the directly labeling methods when only the first correction was applied. The random labeling variation seems to play a greater role in the direct labeling methods, whereas the systematic deviation seems to play a bigger role in the indirect labeling methods (Table 1).

mRNA labeled with a combination of oligo(dT) and random hexamer primers had a better reproducibility and signal-to-noise ratio compared with mRNA labeled with only oligo(dT) primers, for both direct and indirect labeling method. The ICC for total RNA tended to be smaller than the ICC for mRNA. Based on these results, we decided to use the indirect labeling method in all subsequent experiments. The use of mRNA in combination with oligo(dT) and random hexamer primers gave the best results and is the preferred method if sufficient amounts of RNA are available. However, total RNA was used in subsequent experiments, since only limited amounts of RNA could be obtained from the rectal biopsies that were used and further mRNA isolation would result in too little input material.

Assessment of biological variation. The ICC was also applied to optimize human rectal biopsy sampling by investigating the variation in gene expression in rectal biopsies between and within persons. Since this question is targeted at assessment of random variation (difference in gene expression) and

 modifications can be used, which are likely to improve reverse transcriptase function and will result in a higher insertion of the modified nucleotides and longer cDNA strands. We compared the quality and reproducibility of oligo(dT)-primed direct labeling and indirect labeling using either mRNA or total RNA as input material. In addition, for labeling of mRNA, also a combination of oligo(dT) and random hexamer primers was tested in the indirect as well as in the direct labeling method. This approach aims at a higher labeling efficiency and cannot be used for total RNA since it would also result in labeling of ribosomal RNA. For an overview of the methods tested, see Table 1. In the indirect labeling method, aminoallyl nucleotides were incorporated during the cDNA reaction followed by reactive dye coupling. We also tested biotin nucleotide incorporation followed by streptavidin-dye coupling. However, in our hands the biotin-streptavidin-based technique resulted in very poor signal-to-noise ratios, due to high background signals. Therefore, this method was not taken along in further evaluation (data not shown).

Each labeling method was performed five times, independently, using Cy5 fluorescent dye. All methods were hybridized to separate but identical microarrays. The labeling methods were compared by using the Cy3-labeled reference cDNA, which was simultaneously hybridized on each array. To assess both random and systematic labeling deviation, only the first normalization step was performed (Fig. 1, Eq. 1). To assess only the random labeling variation, both normalization steps were performed (Fig. 1, Eqs. 1 and 2). As a measure for reproducibility, the ICC was calculated for each method. As a measure for quality, the average signal spot intensity, background, and signal-to-noise ratio were calculated for each method (Table 1).

In all three direct labeling methods a smaller ICC was obtained compared with the ICC of indirect labeling methods when both corrections were applied, whereas a higher ICC was obtained from the directly labeling methods when only the first correction was applied. The random labeling variation seems to play a greater role in the direct labeling methods, whereas the systematic deviation seems to play a bigger role in the indirect labeling methods (Table 1).

mRNA labeled with a combination of oligo(dT) and random hexamer primers had a better reproducibility and signal-to-noise ratio compared with mRNA labeled with only oligo(dT) primers, for both direct and indirect labeling method. The ICC for total RNA tended to be smaller than the ICC for mRNA. Based on these results, we decided to use the indirect labeling method in all subsequent experiments. The use of mRNA in combination with oligo(dT) and random hexamer primers gave the best results and is the preferred method if sufficient amounts of RNA are available. However, total RNA was used in subsequent experiments, since only limited amounts of RNA could be obtained from the rectal biopsies that were used and further mRNA isolation would result in too little input material.

Assessment of biological variation. The ICC was also applied to optimize human rectal biopsy sampling by investigating the variation in gene expression in rectal biopsies between and within persons. Since this question is targeted at assessment of random variation (difference in gene expression) and

 Table 1. Assessment of different RNA labeling methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>ICC^corr1</th>
<th>ICC^corr2</th>
<th>Signal</th>
<th>Background</th>
<th>Signal/Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.O.M</td>
<td>0.973</td>
<td>0.983</td>
<td>38.76</td>
<td>6.20</td>
<td>6.25</td>
</tr>
<tr>
<td>D.OR.M</td>
<td>0.985</td>
<td>0.991</td>
<td>73.97</td>
<td>7.12</td>
<td>10.39</td>
</tr>
<tr>
<td>D.O.T</td>
<td>0.971</td>
<td>0.980</td>
<td>80.04</td>
<td>6.33</td>
<td>12.65</td>
</tr>
<tr>
<td>L.O.M</td>
<td>0.951</td>
<td>0.988</td>
<td>12.00</td>
<td>1.14</td>
<td>10.53</td>
</tr>
<tr>
<td>L.OR.M</td>
<td>0.978</td>
<td>0.993</td>
<td>51.04</td>
<td>1.32</td>
<td>38.67</td>
</tr>
<tr>
<td>L.O.T</td>
<td>0.956</td>
<td>0.981</td>
<td>26.41</td>
<td>1.14</td>
<td>23.17</td>
</tr>
</tbody>
</table>

The intraclass correlation coefficient (ICC), average signal spot intensity, background, and signal-to-noise ratio of different labeling methods using mRNA or total RNA as starting material are given. D, directly labeling; I, indirectly labeling; O, oligo dT primer; R, random hexamer primer; M, mRNA; T, total RNA; ICC^corr1, ICC after first correction; and ICC^corr2, ICC after both corrections (systematic deviation).
not the systematic deviation, both data corrections (Fig. 1, Eqs. 1 and 2) were performed on the microarray data set. Multiple biopsies \((n = 4–6)\) were taken from five different healthy persons for the within-person variation analysis. From 22 persons with intestinal complaints, but without visible symptoms, one rectal biopsy was obtained to assess the variation in gene expression of biopsies from different subjects. Total RNA was labeled using the indirect labeling method and hybridized to separate but identical microarrays in combination with identical reference cDNA on all slides. To determine the within-person variation, all biopsies of one individual were used separately in SPSS to obtain two ICCs, namely the single and average measured value. Using these, we calculated the ICCs ranging from one to six biopsies, for each of the five individuals separately (Fig. 3A). An average ICC of 0.870, of the five individuals, was found if one biopsy per person was taken. It should be noted that a variation in gene expression in one biopsy can only be obtained by calculation in relation to gene expression data of multiple biopsies. If two biopsies are used for analysis of variation, then the ICC becomes 0.930. Additional biopsies lead to a slightly increased ICC, but the biggest increase in ICC is found in the step from one to two biopsies.

Often in microarray data analysis, an \(R^2\) threshold of 0.9 is used (10, 21, 23), and in epidemiological studies for within-person analysis, an ICC threshold of 0.9 is used (16). For the ICC this means that the variance due to between-subject variations is at least 90% of the total variance.

The ICC of two biopsies is above 0.9. Therefore, it can be concluded that at least two rectal biopsies per person should be used when working with DNA microarrays. To establish whether the ICC can be improved by pooling biopsies of a person, the following analysis was done. The ICC was obtained from two individuals, with two biopsies per individual; these were used separately or the values were averaged per individual \((N = 4\) or \(N = 2\), ICC\(_1\) vs. ICC\(_2\)). This was also performed for three individuals, with two biopsies per individual; these were used separately or the values were averaged per individual \((N = 6\) or \(N = 3\), ICC\(_3\) vs. ICC\(_3\)). Use of the mean of two biopsies gave an improved ICC compared with the single biopsies \((\Delta\text{ICC} = 0.026, SD = 0.009)\). The reliability of the microarray data of biopsies from different persons also becomes greater when the number of biopsies increases (Fig. 3B). When combining all the subjects, biopsies from four different persons give an ICC above 0.9. This is well within the resolution of the analysis, since biopsies from 21 different persons are necessary to obtain an ICC that is identical to the technical variation \([0.981, \text{Table 1 (indirect labeling of total RNA)}]\). By dividing the subjects into groups, by age and gender, more homogeneous groups are created. In three cases...
this resulted in improved ICCs, except in the group of male subjects that are older than 60 yr; there, for unknown reasons, biopsies from six different persons are needed to give an ICC above 0.9 (0.908). By pooling all data, the diversity in this group (men over 60 yr old) is concealed by the uniformity of the other groups.

DISCUSSION

Assessment of intraclass correlation coefficient characteristics. The purpose of this study was to investigate whether the ICC can be used as a statistical method to optimize DNA microarray experiments. We attempted to evaluate whether the ICC can be used as a simple tool for the assessment of technical and biological variation in microarray experiments. We demonstrated that the ICC is sensitive for both systematic deviation and random variation. Therefore, we used the ICC to test our normalization method, to assess different labeling methods, and to optimize human rectal biopsy sampling.

Assessment of technical variation: data normalization. Data normalization by cobyhybridization of a separately labeled reference sample is widely used and accepted in DNA microarray analysis (8, 17); in most cases mean values are used. We compared data normalization performed with mean or median values. Since the ICC is sensitive for both random and systematic deviation, it could be established that the median value, rather than the mean value, of overall Cy5 values should be used for correction for the systematic error (Fig. 1, Eq. 2), in data normalization. The normalization step is based on the assumption that there are only minor differences in overall gene expression between samples and that there is symmetry in expression levels of up- and downregulated genes. The assumption that there are only minor differences in overall gene expression levels between samples is not valid in each experimental setup. First of all, the hybridized arrays should contain a large number of genes. Furthermore, the arrays should not contain a preselected group of genes that is expected to hybridize stronger to one sample than to the other. Finally, the total number of different RNA messengers expressed in the samples should be equal. When in one sample a lower variation in the type of RNA messengers is expressed compared with the other, the hybridization signals will be relatively high for the sample where a lower number of different mRNA transcripts are expressed, assuming that equal amounts of RNA are used in the reverse transcription reaction. For these exceptions, another method for the second data correction step should be used, for example, normalization based on a set of “housekeeping” genes or normalization based on a set of spiked controls, each with their own disadvantages.

Assessment of technical variation: labeling methods. Total or mRNA that was indirectly labeled gave a better reproducibility and quality compared with direct labeling, although a larger systematic error was found. The reproducibility and quality was increased when mRNA was reversibly transcribed with the combination of oligo(dT) and random hexamer, for the direct as well as indirect labeling. Theoretically the reproducibility (ICC) has a maximum value of 1. In most papers a reproducibility for labeling methods of 0.9 and higher is considered to be sufficient (10, 21, 23). In view of the above, a good reproducibility was obtained for all labeling methods, particularly for the methods which use the indirect labeling method. The low background of the indirect labeling method can be ascribed to the removal of excess dye that is accomplished by a column, in contrast to precipitation in the direct labeling method. Adjusting for the systematic error led to enhanced ICCs, especially for the indirect labeling methods. This can be due to efficiency differences of the chemical coupling of the dye. An increase of the coupling time might overcome this problem.

Assessment of biological variation. As expected, the between-person variation was found to be greater than the within-person variation in human rectal biopsies. No difference was found between male and female rectal biopsies. The within-person variation in gene expression can be attributed to a difference in composition of the biopsies. The biopsies taken can vary in the blood and muscle content, but also in the constitution of the epithelium layer consisting of enterocytes, goblet, endocrine, and gut-associated lymphatic tissue cells. The above-mentioned difference can also have an effect on the between-person variation. Variation in the biopsies can also be due to the fact that the rectum might be affected by disease, despite the absence of visible symptoms in colon or rectum.

To decrease variation, two biopsies per subject should be taken for analysis. This was determined for individuals separately (Fig. 3A), but also, when multiple biopsies from different persons were combined into multiple groups and analyzed, the mean of two biopsies gave an improved ICC compared with using the single biopsies (ΔICC = 0.026, SD = 0.009). From the results presented in Fig. 3B, where a single biopsy per person was used for gene expression analysis, it can be determined that four persons per group are required to obtain a sufficiently homogeneous sample, based on an ICC cutoff of 0.9. However, by separating persons in groups, we found that in some cases (males of over 60 yr) four biopsies seemed insufficient. This is in agreement with Hwang et al. (11), who found in DNA microarray analysis of bone marrow samples from different lymphoid leukemia subtypes that seven persons per group are required to separate distinct disease states or other physiological differences with statistically significant reliability.

We also have used this data set to determine the minimum sample size in microarray experiments. Using the same threshold as Hwang et al. (0.95), we found that a minimum of seven subjects per subgroup should be used for this data set.

In view of all the data, we suggest that if the group that is sampled is diverse or not well characterized, as is generally the case in intervention trials or cohort studies, then the overall cutoff should be increased to 0.95. This implies that, if one biopsy per person is analyzed for gene expression, then at least eight different persons per group are needed. This number can be reduced to a minimum of six persons if two or more biopsies per person are sampled and averaged.

One should take also into account that six repeats are considered to be sufficient in our laboratory conditions, but this value must be determined by each experimenter, as it may depend on several local factors (array quality, probe labeling, hybridization conditions, scanning of the slides, etc.).

We showed that the ICC can be used for assessment of technical and biological variation in microarray experiments. After evaluation of the technical variation, we recommend that the indirect labeling should be used and whenever possible mRNA should be taken as input material with the combination
of oligo(dT) and random hexamer primers. Assessing the biological variation of human rectal biopsies revealed that two biopsies per person and at least six persons, in total per group, should be analyzed when studying gene expression for example in human dietary intervention trials.

ACKNOWLEDGMENTS

We thank Prof. Dr Chris Mulder for human rectal biopsy sampling and Prof. Dr Frans Kok and Maureen van den Donk for critical reading the manuscript. For construction and use of the cDNA microarray, we thank Dr. Ad Peijnenburg and Hakan Baykus.

GRANTS

This work was supported by MLDS Grant WS 99-72, the Dutch Digestive Diseases Foundation, and by Grant VCZ 980-10-020 from ZonMW, the Netherlands Organization for Health Research and Development.

REFERENCES