Journal of Biopharmaceutical Statistics, 17: 309–327, 2007

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400601177327

Taylor & Francis
Taylor & Francis Group

NONINFERIORITY TESTS BASED ON CONCORDANCE CORRELATION COEFFICIENT FOR ASSESSMENT OF THE AGREEMENT FOR GENE EXPRESSION DATA FROM MICROARRAY EXPERIMENTS

Chen-Tuo Liao and Chia-Ying Lin
Division of Biometry, Institute of Agronomy, National Taiwan University, Taipei, Taiwan

Jen-pei Liu
Division of Biometry, Institute of Agronomy, National Taiwan University, Taipei, Taiwan and Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan

Microarray is one of the breakthrough technologies in the twenty-first century. Despite of its great potential, transition and realization of microarray technology into the clinically useful commercial products have not been as rapid as the technology could promise. One of the primary reasons is lack of agreement and poor reproducibility of the intensity measurements on gene expression obtained from microarray experiments. Current practices often use the testing the hypothesis of zero Pearson correlation coefficient to assess the agreement of gene expression levels between the technical replicates from microarray experiments. However, Pearson correlation coefficient is to evaluate linear association between two variables and fail to take into account changes in accuracy and precision. Hence, it is not appropriate for evaluation of agreement of gene expression levels between technical replicates. Therefore, we propose to use the concordance correlation coefficient to assess agreement of gene expression levels between technical replicates. We also apply the Generalized Pivotal Quantities to obtain the exact confidence interval for concordance coefficient. In addition, based on the concept of noninferiority test, a one-sided $(1-\alpha)$ ***lower confidence limit for concordance correlation coefficient is employed to test the hypothesis that the agreement of expression levels of the same genes between two technical replicates exceeds some minimal requirement of agreement. We conducted a simulation study, under various combinations of mean differences, variability, and sample size, to empirically compare the performance of different methods for assessment of agreement in terms of coverage probability, expected length, size, and power. Numerical data from published papers illustrate the application of the proposed methods.***

Key Words: Agreement; Concordance correlation coefficient; Generalized pivotal quantity; Noninferiority.

Received August 18, 2006; Accepted November 29, 2006

The views expressed in this article are personal opinions of the authors and may not necessarily represent the position of the National Taiwan University and National Health Research Institutes, Taiwan.

Address correspondence to Jen-pei Liu, Division of Biometry, Institute of Agronomy, National Taiwan University, Taipei, Taiwan; Fax: 886-2-3366-4791; E-mail: jpliu@ntu.edu.tw

1. INTRODUCTION

Microarray is one of the most important technologies in life science developed in the past decade. Based on the complementary property of DNA and reverse transcription reaction, microarray can simultaneously measure the expression levels of tens of thousands of genes. Hence, it not only revolutionizes the biological and medical research but also provides a mechanism to understand genetic nature and biological phenomena of organisms. In addition, it paves the way to breakthrough applications. For example, in vitro multivariate index assay-based multiplex diagnostic biochips have already been used in targeted clinical trials for evaluation of the efficacy and safety of individualized treatments for the patients with breast cancer. (FDA, 2006; MINDACT, 2006; Sprarano et al., 2006) However, despite of its immense promising potential, not until recently, the US Food and Drug Administration (FDA) approved the first diagnostic device based on microarray technology for diagnosis of gene subtypes for cytochrome P450 2D6 and 2C19. One of the primary reasons is lack of agreement and poor reproducibility of the intensity measurements on gene expression obtained from these multivariate index assays because of complicated techniques and lengthy procedures. Its usefulness in clinical practice has been questioned and criticized (Ioannidis, 2005).

In addition, the result of a multivariate index assay is a composite function of expression levels of the individual genes selected into the multiplex biochip. Therefore, in order to have consistent results for the multivariate index assay, the expression levels of each component genes should be also consistent under the same operating conditions within or between laboratories. For the technical replicates of the same sample or tissue, Tan et al. (2003) reported that consistent results using one microarray platform performed in one laboratory cannot be reproduced in other laboratories with the same or different platforms. Sometimes, consistent results can not be even obtained from the technical replicates of the same sample within the same laboratory using the same platform under the same operating condition. On the other hand, Michiels et al. (2005) reported that out of the seven published studies to predict prognosis of cancer patients, the performance of DNA microarray analysis of five studies is no better than flopping a coin. In addition, the other two studies barely beat horoscopes (Ioannidis, 2005).

As a result, agreement and reproducibility have recently drawn a lot of attention in microarray experiments. For example, Dobbin et al. (2005), Irizarry et al. (2005), Larkin et al. (2005), and Toxicogenomics Research Consortium (2005) examined the agreement on measurements of gene expressions between laboratories and across different platforms. Testing the hypothesis of zero Pearson correlation coefficient (PCC) is one of the most common statistical methods to assess comparability of gene expression levels between technical replicates across laboratories. However, to evaluate comparability on gene expressions between laboratories, it is to assess the agreement of the measurements of the technical replicates for the same genes of the same samples from different laboratories or platforms. Hence, objective for evaluation of comparability is to investigate the closeness or equivalence of gene expression levels between technical replicates of the same samples obtained under the same operating conditions within or between laboratories. Although Pearson correlation coefficient is an excellent statistic for

evaluation of linear association, it is location and scale invariant. Hence, it cannot detect changes in accuracy and precision (Bland and Altman, 1986) and cannot be used for assessment of agreement of gene expression levels between technical replicates which requires evaluation of equivalence in both accuracy and precision. On the other hand, at the 5% level for testing the null hypothesis of no linear correlation, with 10,000 genes in a microarray experiment, an estimated Pearson correlation coefficient as low as 0.02 can reach the statistical significance. Therefore, hypothesis of zero linear correlation is not appropriate for evaluation of agreement of gene expression levels between technical replicates of the same samples. In order to meet the minimal requirement of agreement, the hypothesis for assessment of agreement of gene expression levels between technical replicates should be formulated as the noninferiority hypothesis which not only the linear association exceeds a prespecified threshold but also the means and variability between technical replicates are equivalent within some pre-determined limits.

On the other hand, the concordance correlation coefficient, proposed by Lin (1989, 1992) and Lin et al. (2002) is a product of Pearson correlation coefficient and a factor consisting of location and scale shifts. Therefore, it can be employed to evaluate the agreement of gene expression levels between the technical replicates of the same samples. However, it is an asymptotic method and its performance on the coverage probability of the confidence interval in the small samples is not fully investigated. Therefore, we apply the concept of the Generalized Pivotal Quantities (GPQs) to obtain the exact confidence interval for concordance correlation coefficient (Weerahandi, 1993, 1995). In addition, based on the concept of noninferiority test, a one-side $(1-\alpha)$ lower confidence limit for concordance correlation coefficient is employed to test the hypothesis that agreement of gene expression levels between two technical replicates exceeds some minimal requirement of agreement.

In next section, Pearson linear correlation coefficient and concordance correlation coefficient will be reviewed. We argue the reasons why testing the hypothesis of zero Pearson linear correlation coefficient fails to address the issue of agreement. In section 3, the exact confidence interval for concordance correlation coefficient based on the concept of GPQs is derived. In addition, a noninferiority test for evaluation of agreement of gene expression levels between technical replicates of the same samples is formulated. A testing procedure based on the lower $(1-\alpha)$ confidence limit for concordance correlation coefficient is proposed for the noninferiority hypothesis In section 4, under various combinations of mean differences, variability, and sample size, a simulation study was conducted to empirically compare the performance of different methods for assessment of agreement in terms of coverage probability, expected length, size, and power. In section 5, numeric data from published data illustrate the proposed methods. Discussion and final remarks are provided in the final section.

2. PEARSON CORRELATION COEFFICIENT AND CONCORDANCE CORRELATION COEFFICIENT

In what follows, we use replicates as a generic term to indicate either the technical replicates of a sample within the same laboratory using the same platform or the technical replicates of the same sample from different laboratories or

from different platforms. Suppose that for the same sample, Y_{ij} be the expression measurement of gene j (Y_{1j}, Y_{2j}) for replicate $i, j = 1, \ldots, n$; $i = 1, 2$. In addition, the paired measurements of technical replicates of gene j, are independently identically distributed (i.i.d.) as a bivariate normal distribution with mean vector (μ_1, μ_2) and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Then Pearson correlation coefficient is defined as

$$\rho = \frac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{\mathrm{Var}(Y_1)\mathrm{Var}(Y_2)}} = \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}}, \tag{2.1}$$

The sample estimator of Pearson correlation coefficient is obtained by substituting the sample estimates into (2.1) as

$$\hat{\rho} = S_{12}/S_{11}S_{22}, \tag{2.2}$$

where

$$S_{11} = \sum_{i=1}^{n}(Y_{1j} - \overline{Y}_{1.})^2, \quad S_{22} = \sum_{i=1}^{n}(Y_{2j} - \overline{Y}_{2.})^2, \quad S_{12} = \sum_{i=1}^{n}(Y_{1j} - \overline{Y}_{1.})(Y_{2j} - \overline{Y}_{2.}),$$

and $\overline{Y}_{1.}$ and $\overline{Y}_{2.}$ are the mean for gene expression levels for replicates 1 and 2, respectively.

The asymptotic property of $\hat{\rho}$ is obtained through the inverse hyperbolic tangent transformation. In other words, $z_{\hat{\rho}} = (1/2)\ln[(1+\hat{\rho})/(1-\hat{\rho})]$ is asymptotically normal with mean $z_\rho = 0.5\ln[(1+\rho)/(1-\rho)]$ and variance $\sigma^2_{z_{\hat{\rho}}} = 1/(n-3)$, where ln is the nature logarithm based on e.

Pearson correlation coefficient is an excellent parameter to evaluate the linear relationship between the pair measurements of gene expression levels between two different replicates of the same sample. In other words, it is to measure whether Y_{1j} increases (or decreases) in a linear fashion as Y_{2j} increases (or decreases). In addition, Pearson correlation coefficient is invariant to location and scale changes. Therefore, the same Pearson correlation coefficient is obtained if a linear transformation is performed for Y_{1j}. It follows that a high Pearson correlation coefficient will be obtained if Y_{1j} and Y_{2j} are related in a linear manner even when they are far apart in both location and variability. Three cases given Table 1 illustrate the major deficiency of using Pearson correlation coefficient in assessment of agreement. For Case I, $Y_{1i} = Y_{2i}$, while $Y_{1i} = 2Y_{2i}$ and $Y_{1i} = 4Y_{2i}$, respectively for Cases II and III. Since Pearson correlation coefficient is invariant to the scale shift, despite of the fact that the squared Euclidean distance increases from 0 in Case I to 30 in Case II, and to 270 in Case III, it remains 1. Suppose that the threshold for the diagnosis of a certain disease based on a multivariate index assay is 3, (i.e., >3). As demonstrated in Table 1, although the Pearson correlation coefficient is 1 between $Y1$ and $Y2$ for all cases, the results of diagnosis using $Y2$ are completely different from those using $Y1$ for Cases II and III. Therefore, the

Table 1 Measurements of agreement

	Case I		Case II		Case III	
	Y1	Y2	Y1	Y2	Y1	Y2
	1	1	1	2	1	4
	2	2	2	4	2	8
	3	3	3	6	3	12
	4	4	4	8	4	16
ρ	1		1		1	
ED	0		30		270	
ρ_C	1		0.4		0.13	

ED: eclidean distance.

invariant property of Pearson correlation coefficient fails to evaluate equivalence in accuracy and precision between the technical replicates. Therefore, it can not be used to represent or assess the agreement of the expression levels obtained from two technical replicates of the same genes for the same sample where the agreement is to assess the closeness Y_{1j} to Y_{2j} for $j = 1, \ldots, n$.

The following hypothesis of no linear association based on Pearson correlation coefficient is currently used to evaluate comparability of gene expression levels between technical replicates in literature:

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_a : \rho \neq 0. \tag{2.3}$$

However, the hypothesis of zero Pearson correlation coefficient can only detect whether a linear association exists in gene expression levels between two technical replicates. As a minimal requirement, tt should test whether the degree of the linear association exceeds a minimal required magnitude, say 0.8 or 0.9. Therefore, Pearson correlation coefficient in conjunction with the hypothesis of no linear association is not an appropriate method for assessment of agreement of gene expression levels between technical replicates from the same samples. Because of these shortcomings, we suggest that evaluation of agreement of gene expression levels between technical replicates from the same sample be formulated in terms of noninferiority hypothesis such that a pre-specified minimal requirement of agreement must be satisfied. In addition, the minimal requirement of agreement should consist of minimal threshold for the linear association and equivalence limits in both means (accuracy) and variability (precision) on expression levels of the same genes between technical replicates.

Lin (1989, 1992) proposed the concordance correlation coefficient for assessment of agreement in assay validation. Since microarray is a parallel assay with multiple analytes, the concordance correlation coefficient can be used to assess the agreement of gene expression levels between two technical replicates. The concordance correlation coefficient is defined as

$$\rho_c = \frac{2\sigma_{12}}{\sigma_{11} + \sigma_{22} + (\mu_1 - \mu_2)^2} = \rho C_b \tag{2.4}$$

where $C_b = [(v + 1/v + u^2)/2]^{-1}$, $v = \sqrt{\sigma_{11}/\sigma_{22}}$ is the scale shift, and $u = (\mu_1 - \mu_2)/\sqrt[4]{\sigma_{11}\sigma_{12}}$ denotes the location shift relative to the scale.

From (2.4), the concordance correlation coefficient consists of two components. The first component is Pearson correlation coefficient, ρ, which measures the linear association of gene expression levels between two technical replicates. The second component is a function of two measures for accuracy and precision, respectively. The value of u^2 is the relative squared difference in means scaled by the square root of the product of variances of the two technical replicates, while v is the ratio of the standard deviations of gene expression levels between technical replicates. As a result, the concordance correlation coefficient not only measures the linear association but also addresses the accuracy and precision of gene expression levels between two technical replicates. Because the concordance correlation coefficient is to take into account both the location and scale shift, as demonstrated in Table 1, it decreases from 1 in Case I, to 0.4 in Case II, and to 0.13 in Case III. In fact, the concordance correlation coefficient is equal to Pearson correlation coefficient only if $\mu_1 = \mu_2$, and $\sigma_{11} = \sigma_{22}$ as demonstrated in Case I of Table 1. Because of these properties, the concordance correlation coefficient can be used to evaluate the agreement of gene expression levels between two technical replicates of the same sample.

The sample estimator of ρ_c is given as

$$\hat{\rho}_c = \frac{2S_{12}}{S_{11} + S_{22} + (\overline{Y}_1 - \overline{Y}_2)^2}. \tag{2.5}$$

Lin (1989) showed that under the bivariate normal distribution, the asymptotic distribution of the inverse hyperbolic tangent transformation of $\hat{\rho}_c$

$$\widehat{Z}_{\hat{\rho}_c} = \frac{1}{2}\ln\frac{1 + \hat{\rho}_c}{1 - \hat{\rho}_c} \tag{2.6}$$

follows a normal distribution with mean $Z_{\rho_c} = (1/2)\ln[(1 + \rho_c)/(1 - \rho_c)]$, and variance

$$\sigma^2_{\widehat{Z}_{\hat{\rho}_c}} = \frac{1}{n-2}\left[\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{2\rho_c^3(1-\rho_c)u^2}{\rho(1-\rho_c^2)^2} + \frac{\rho_c^4 u^4}{2\rho^2(1-\rho_c^2)^2}\right]. \tag{2.7}$$

An asymptotic $(1 - \alpha)\%$ confidence interval for ρ_c is obtained from the inverse-transformation of the lower and upper limit of the $(1 - \alpha)\%$ confidence interval for Z_{ρ_c} based on $\widehat{Z}_{\hat{\rho}_c}$ and the estimator of $\sigma^2_{\widehat{Z}_{\hat{\rho}_c}}$.

3. NONINFERIORITY TEST AND EXACT CONFIDENCE INTERVAL

Although the concordance correlation coefficient can be used as a parameter for assessment of agreement of gene expression levels between two technical replicates, the hypothesis to detect whether the concordance correlation coefficient is zero or not is not an appropriate hypothesis for assessment of agreement. The gene expression levels between two technical replicates are said to be in agreement if the concordance correlation coefficient exceeds some prespecified minimal requirement

for agreement. As a result, evaluation of agreement of gene expression levels between two technical replicates should be formulated in the following one-sided noninferiority hypothesis:

$$H_0 : \rho_C \leq \rho_{CL} \quad \text{vs.} \quad H_a : \rho_C > \rho_{CL}, \tag{3.1}$$

where $\rho_{CL} > 0$ is some prespecified minimal requirement of agreement.

ρ_{CL} can be determined by the minimal threshold of the linear association and the equivalence limits of means and standard deviations of gene expression levels between two technical replicates. In addition, the magnitudes of linear association and equivalence limits should also meet the requirements of quality control and quality assurance for agreement between technical replicates for assay validation of microarray experiments. For example, the minimal requirement for the linear association as measured by Pearson correlation coefficient is 0.95. In addition, the ratio of the standard deviation of expression levels for the first technical replicate to that of the second is at least 0.8 and the relative squared difference in means scaled by the square root of the product of variances of the two technical replicates is within 0.25, it follows that the minimal requirement for agreement represented by the concordance correlation coefficient is

$$\rho_{CL} = \frac{0.95}{[(0.8 + 1/0.8 + 0.25^2)/2]} = 0.8994 \approx 0.90.$$

Hence, the corresponding noninferiority hypothesis for assessment of agreement of gene expression levels between two technical replicates becomes

$$H_0 : \rho_C \leq 0.90 \quad \text{vs.} \quad H_a : \rho_C > 0.90. \tag{3.2}$$

One approach to testing the noninferiority hypothesis of agreement is to use the confidence limit approach because not only the confidence limit can provide interval estimation for the concordance correlation coefficient but also it can be used as a test statistics for hypothesis in (3.1). In other words, if the lower $(1 - \alpha)\%$ confidence limit for ρ_C is greater than ρ_{CL}, then H_0 is rejected at the α significance level and one can conclude that gene expression levels of the two technical replicates meet the minimal requirement of quality control for agreement.

The asymptotic normality of $\hat{\rho}_c$ reviewed above allows us to construct an asymptotic confidence interval for the concordance correlation coefficient. However, its coverage probability and expected length in the finite samples have not been fully investigated. Hence, we apply the QPQs proposed by Weerahandi (1993, 1995) to construct an exact confidence interval for concordance correlation coefficient. A GPQ for the concordance correlation coefficient is given as

$$R_{\mu\rho_C} = \frac{2R_{12}}{R_{11} + R_{22} + R_{\mu^2}}, \tag{3.3}$$

where R_{11}, R_{12}, R_{22}, R_{μ^2} and derivation of $R_{\mu\rho_C}$ are provided in Appendix.

It follows that an equal-tailed $100(1 - \alpha)\%$ confidence interval for ρ_C can be obtained as the $100(\alpha/2)\%$ and $100(1 - \alpha/2)\%$th percentiles of the sampling

distribution of $R_{\mu\rho_C}$ from the Monte-Carlo algorithm with large number of replications, say 10,000.

4. SIMULATION STUDY

A simulation study was conducted to investigate and compare performance of the exact confidence intervals based on GPQ approach and the asymptotic confidence intervals in terms of coverage probability and expected length. In addition, with respect to the noninferiority hypothesis with a minimal requirement of quality control for agreement, we also compare the empirical size and power for the testing procedures using the lower exact and asymptotic confidence limits. Fortran 90 and IMSL STAT/LIBRARY Fortran subroutines were used in the simulation study.

Following Lin (1989), five cases of μ_1, μ_2, σ_{11}, σ_{22}, and σ_{12} given in Table 2 were considered in the simulation studies. These five cases represent different degrees of location and scale shifts. In addition, the expression levels of different genes may be correlated. Therefore, correlations of 0 and 0.2, and sample size of 10, 20, 50, and 100 were chosen to investigate the impact of correlation and sample size on coverage probability. For each combination, 10,000 random samples of bivariate normal vectors were generated. The empirical coverage probability is calculated as the proportion of the 10,000 95% confidence intervals that include the pre-specified value of ρ_C given in Table 2. The empirical expected length is estimated as the average of the differences between the upper and lower limits of the 10,000 95% confidence intervals. For 10,000 random samples, a 95% confidence interval is said to provide coverage probability of 95% if its empirical coverage probability is greater than 0.9464. In addition, the empirical size and power for testing the noninferiority hypothesis with $\rho_{CL} = 0.90$ was also investigated in the simulation study with $\rho_C = 0.85$, and from 0.90 to 0.99 by an interval of 0.01. We also consider

Table 2 Specifications of the parameters in simulation study

ρ_C	$\mu_1 - \mu_1$	Σ
0.95	0	$\begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$
0.905	$\sqrt{0.1}$	$\begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$
0.887	$\sqrt{0.1}$	$\begin{pmatrix} 1.1^2 & 0.95 \times 1.1 \times 0.9 \\ 0.95 \times 1.1 \times 0.9 & 0.9^2 \end{pmatrix}$
0.747	$\sqrt{0.1}$	$\begin{pmatrix} 0.9^2 & 0.8 \times 0.9 \times 1.1 \\ 0.8 \times 0.9 \times 1.1 & 1.1^2 \end{pmatrix}$
0.360	$\sqrt{0.25}$	$\begin{pmatrix} \left(\frac{4}{3}\right)^2 & 0.5 \times \frac{4}{3} \times \frac{2}{3} \\ 0.5 \times \frac{4}{3} \times \frac{2}{3} & \left(\frac{2}{3}\right)^2 \end{pmatrix}$

the difference in means of 0, 0.3, and 0.5 for empirical size and power. Again, 10,000 random samples of bivariate normal vectors were generated for each combination of correlations, sample size, differences in means, and values of ρ_C. The empirical size and power were estimated as the proportion of the 10,000 random samples that reject the null hypothesis (3.2) at the 5% significance level.

Table 3 presents the coverage probability and expected length of the 95% confidence interval by the asymptotic and exact methods for uncorrelated cases. From Table 3, for uncorrelated data, when the sample size is 10 or 20, all empirical coverage probabilities of the 95% asymptotic confidence interval are smaller than 0.9464. For sample size greater than 20, there is still 50% of the empirical probabilities of the 95% asymptotic confidence interval smaller than 0.9464. On the other hand, all empirical coverage probabilities of the 95% exact confidence interval based on GPQs are greater than 0.9464 for all sample sizes in all ranges of the concordance correlation coefficient investigated in the simulation study. The expected lengths of the exact confidence interval are in general larger than those of the asymptotic confidence interval. However, as sample size increases beyond 20, the difference in the expected length between the asymptotic and exact confidence interval becomes negligible. On the other hand, the expected length decreases as the sample size increases. In addition, the expected lengths of both methods increase as the location and scale shifts become large. In summary, the exact confidence interval derived from GPQ can provide sufficient coverage probability for the concordance correlation coefficient. However, the simulation study showed that the 95% asymptotic confidence interval for the concordance correlation coefficient fails

Table 3 Coverage probability and expected length for uncorrelated data

		Coverage probability		Expected length	
N	ρ_C	Exact	Asymptotic	Exact	Asymptotic
10	0.950	0.9761	0.9315	0.2481	0.2014
	0.905	0.9742	0.9236	0.3072	0.2982
	0.887	0.9805	0.9248	0.3272	0.3072
	0.747	0.9710	0.9227	0.6507	0.5829
	0.360	0.9594	0.9208	0.8275	0.7137
20	0.950	0.9486	0.9435	0.1242	0.1115
	0.905	0.9583	0.9366	0.1758	0.1766
	0.887	0.9646	0.9394	0.1858	0.1833
	0.747	0.9617	0.9412	0.4045	0.3872
	0.360	0.9544	0.9347	0.5458	0.5191
50	0.950	0.9473	0.9467	0.0630	0.0603
	0.905	0.9500	0.9405	0.0986	0.0993
	0.887	0.9562	0.9472	0.1054	0.1052
	0.747	0.9518	0.9432	0.2368	0.2332
	0.360	0.9525	0.9463	0.3394	0.3343
100	0.950	0.9498	0.9506	0.0412	0.0402
	0.905	0.9490	0.9457	0.0669	0.0672
	0.887	0.9526	0.9483	0.0718	0.0718
	0.747	0.9494	0.9462	0.1625	0.1613
	0.360	0.9520	0.9497	0.2385	0.2367

Table 4 Coverage probability and expected length for correlated data correlation = 0.2

		Coverage probability		Expected length	
N	ρ_C	Exact	Asymptotic	Exact	Asymptotic
10	0.950	0.9642	0.9060	0.2964	0.2434
	0.905	0.9754	0.8867	0.3576	0.3464
	0.887	0.9773	0.8795	0.3659	0.3429
	0.747	0.9592	0.8779	0.7150	0.6381
	0.360	0.8952	0.8096	0.7692	0.6490
20	0.950	0.9053	0.9007	0.1516	0.1368
	0.905	0.9358	0.8769	0.2075	0.2084
	0.887	0.9322	0.8670	0.2135	0.2104
	0.747	0.9244	0.8739	0.4542	0.4347
	0.360	0.8295	0.7769	0.5006	0.4737
50	0.950	0.8412	0.8608	0.0779	0.0746
	0.905	0.8732	0.8325	0.1182	0.1189
	0.887	0.8502	0.8068	0.1216	0.1214
	0.747	0.8437	0.8142	0.2709	0.2668
	0.360	0.6233	0.6008	0.3093	0.3043
100	0.950	0.7589	0.7815	0.0510	0.0499
	0.905	0.7711	0.7381	0.0811	0.0814
	0.887	0.7287	0.6903	0.0834	0.0833
	0.747	0.7254	0.7086	0.1875	0.1862
	0.360	0.3662	0.3490	0.2175	0.2159

to provide sufficient coverage probability even when the sample size is as large as 100.

Table 4 provides the coverage probability and expected length of both the exact and asymptotic confidence intervals when the expression levels among different genes are correlated with a common correlation coefficient of 0.2. From Table 4, the all coverage probabilities of the 95% asymptotic confidence intervals are below 0.9060. On the other hand, when sample size is 10 and ρ_C is greater than 0.747, the coverage probability of the 95% exact confidence interval exceeds 0.9464. However, for all other combinations, the coverage probability of the 95% exact confidence interval is below 0.9053. In addition, the coverage probability decreases either as the sample size increases or as the location and scale shifts increase. The behavior of expected lengths under the correlated expression levels is similar to that under independent data. In general, the correlation of expression levels among different genes will severely decreases the coverage probability for both the exact and asymptotic confidence interval except for the exact confidence interval when sample size is 10 and the concordance correlation coefficient is higher than 0.747.

Table 5 presents the result of the empirical size using both the 95% exact and asymptotic lower confidence limits for testing the inferiority hypothesis in (3.2) with $\rho_{CL} = 0.90$ at the 5% significance level under uncorrelated and correlated data. Form Table 5, the empirical sizes of both asymptotic and exact methods are smaller than 0.05. However, the empirical sizes of the exact method are smaller than those of the asymptotic procedure. This indicates that the exact method is more conservative than the asymptotic procedure in testing the noninferiority hypothesis

Table 5 Empirical size

		Uncorrelated		Correlated (Correlation = 0.2)	
n	$\mu_1 - \mu_2$	Exact	Asymptotic	Exact	Asymptotic
10	0	0.0221	0.0338	0.0195	0.0361
	0.3	0.0239	0.0390	0.0183	0.0340
	0.5	0.0246	0.0412	0.0239	0.0391
20	0	0.0261	0.0415	0.0223	0.0348
	0.3	0.0293	0.0397	0.0243	0.0363
	0.5	0.0302	0.0412	0.0268	0.0395
50	0	0.0355	0.0455	0.0240	0.0329
	0.3	0.0356	0.0439	0.0285	0.0347
	0.5	0.0338	0.0419	0.0305	0.0373
100	0	0.0363	0.0456	0.0210	0.0258
	0.3	0.0396	0.0448	0.0239	0.0275
	0.5	0.0395	0.0447	0.0314	0.0370

for evaluation of agreement of gene expression levels between technical replicates of the same samples. Although the empirical sizes obtained when the expression levels among different genes are correlated are smaller than those under the independent expression levels, the impact of correlated expression levels among different genes on the empirical size, as demonstrated in Table 5, is rather minimal. Figures 1

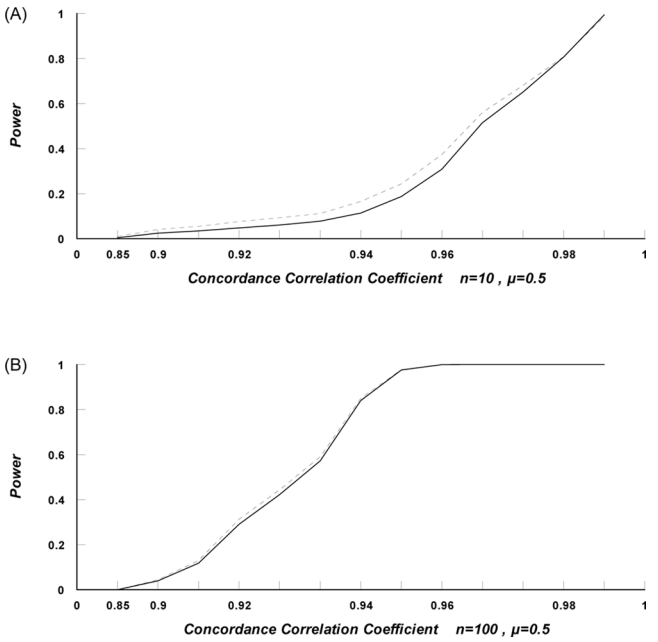

Figure 1 Power curve for testing noninferiority hypothesis with the lower limit of 0.9 for uncorrelated expression levels; n is the number of genes, and μ is the difference in mean. (solid curve: exact method; dash curve: asymptotic method).

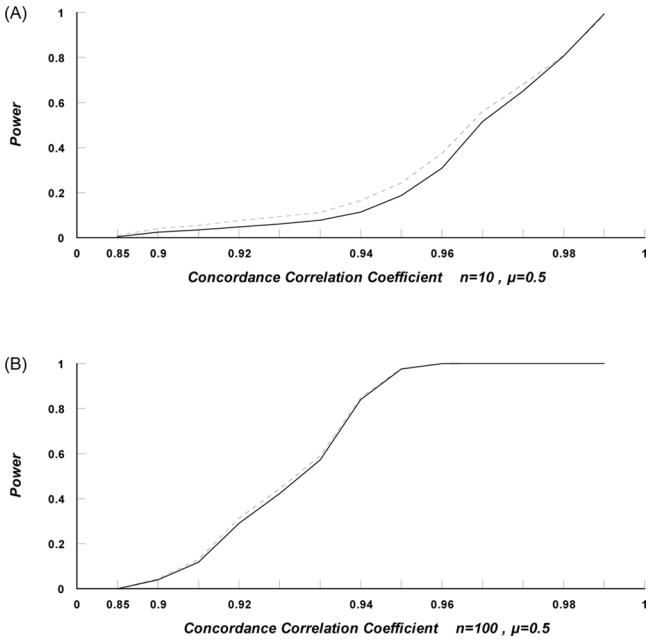

Figure 2 Power curve for testing noninferiority hypothesis with the lower limit of 0.9 for correlated expression levels (correlation $= 0.2$); n is the number of genes, and μ is the difference in mean (solid curve: exact method; dash curve: asymptotic method).

and 2 provide the empirical power curves when difference in means is 0.5 and sample size is either 10 or 100 for both uncorrelated and correlated expression levels respectively. When sample size is 10, the empirical power of the asymptotic method is higher than that of the exact method. However, the maximum difference in the power curve does not exceed 6%. When sample size increases, both methods provide almost identical power. Similar to the empirical size, the empirical power under the correlated expression levels in general is smaller than that under the independent data. However, a comparison between Figures 1 and 2 indicates that the impact of correlation on expression levels among different genes on power is negligible.

5. NUMERICAL EXAMPLE

Dobbin et al. (2005) investigated the comparability of expression levels of cancer genes using oligonucleotide microarray among four different laboratories. Two technical replicates for each sample of five cell line pellets were analyzed with Affymetrix Human Genome U133A arrays at each of the four laboratories (Lab 615, Lab 616, Lab 617, and Lab 618). Because the expression levels of the housekeeping genes are one of the importance measures for assessment of quality for the data derived from microarray experiments, therefore we select the normalized intensities on the log2 scale from 100 housekeeping genes of cell line H1437 obtained from each of the four laboratories to illustrate the proposed methods for evaluation of agreement of the two technical replicates within

Table 6 Summary of statistics of log 2 intensity by technical replicates and laboratory

	Replicate 1		Replicate 2	
Laboratory	Mean	Standard deviation	Mean	Standard deviation
615	12.2269	1.4108	12.1802	1.4662
616	12.1935	1.5599	12.2846	1.5654
617	12.2846	1.3988	12.2984	1.5007
618	12.1558	1.3340	12.1618	1.4286

laboratory. All data in Dobbin et al. (2005) are publicly available for download at http://gedp.nci.nih.gov (experiment IDs 615–618).

Table 6 provides the descriptive statistics by technical replicate and laboratory. As shown in Table 6, the location and scale shifts between technical replicates are quite negligible. Table 7 presents the results for evaluation of agreement of the expression levels between two replicates within each laboratory by the asymptotic and exact procedures. Based on the log2 intensity, the results of Pearson correlation coefficient, concordance correlation coefficient and the proposed GPQ-based method along with the 95% one-side lower confidence limits of agreement are provided in Table 7. Pearson correlation coefficients of four laboratories are all above 0.98 which indicate a very high positive linear relationship between the gene expression levels between the two technical replicates for all four laboratories. The concordance correlation coefficient ranges form 0.9775 of lab 617 to 0.9918 of lab 616. Because very little location and scale shifts exist between technical replicates, the estimated concordance correlation coefficients are smaller than but very close to the Pearson correlation coefficients. The 95% lower confidence limits on the log2 scale by both exact and asymptotic methods are identical to the third decimal point for all four laboratories. In addition, all 95% lower confidence limits by both methods are above 0.90 for four laboratories. Therefore, if ρ_{CL} is set at 0.90, then one can claim that at the 5% significance level, the expression levels of technical replicates for the 100 housekeeping genes of cell line H1437 meet the minimal requirement of quality control for agreement for all four laboratories. In other words, with respect to the 100 housekeeping genes, an excellent agreement exists between the two technical replicates for all four laboratories.

Table 7 Concordance correlation coefficient based on log 2 intensity by method and laboratory

	Concordance correlation coefficient			
	Pearson		95% lower confidence limit	
Laboratory	Correlation	Estimate	Asymptotic	Exact
615	0.9874	0.9862	0.9809	0.9804
616	0.9935	0.9918	0.9886	0.9885
617	0.9804	0.9775	0.9694	0.9687
618	0.9890	0.9867	0.9820	0.9817

Table 8 Concordance correlation coefficient based on log2 intensity with a location shift of 5 by method and laboratory

	Concordance correlation coefficient			
	Pearson		95% lower confidence limit	
Laboratory	Correlation	Estimate	Asymptotic	Exact
615	0.9874	0.1368	0.1087	0.1132
616	0.9935	0.1660	0.1332	0.1391
617	0.9804	0.1418	0.1127	0.1127
618	0.9890	0.1299	0.1031	0.1081

For the purpose of illustration only, a value of 5 is added to the log2 expression levels of replicate 1 for all four laboratories. A value of 5 on the log2 scale implies a 32-fold differential on the original scale in expression levels between the two technical replicates. The results for evaluation of agreement by adding a constant of 5 to the log 2 expression levels of replicate 1 is given in Table 8. From Table 8, Pearson correlation coefficients remain unchanged despite a 32-fold differential in the expression levels between the two technical replicates. However, the concordance correlation coefficient decreases to a range between 0.1299 and 0.1660. Again, the 95% lower confidence limits by both asymptotic and exact methods are very close with a range from 0.1031 to 0.1332 for the asymptotic method and from 0.1081 to 0.1391 for the exact procedure respectively. If a minimal required limit of 0.9 for agreement is used, based on concordance correlation coefficient, agreement in the expression levels between two technical replicates for 100 housekeeping genes of cell line H1437 can not be concluded at the 5% significance level for all four laboratories.

6. DISCUSSION

Despite of its great potential, the breakthrough technology of microarray has not been rapidly transferred into urgently needed medical diagnostic biochip products. One of the primary reasons is its lack of agreement of assay results because of its complex nature and complicated experimental processes and procedures. The first issue of the microarray assays is the quality of the arrays, namely, the correct percent and amount of DNA sequence of probes printed in a given array. The other issues are the standardization of processes which include tissue processing, extraction of RNA from the tissue specimen, preparation of labeled cRNA target (reverse-transcription, labeling, fragmentation, and so on), array hybridization, washing, scanning, and normalization (Shi et al., 2004). Only recently, researchers started to investigate the agreement and reproducibility on measurements of gene expressions between technical replicates from microarray experiments between laboratories and across different platforms. (Dobbin et al., 2005; Irizarry et al., 2005; Larkin et al., 2005; Toxicogenomics Research Consortium, 2005).

Pearson correlation coefficient is one of the primary parameter to examine agreement of microarray expression data between technical replicates. However,

we have argued that Pearson correlation coefficient is a parameter to investigate linear association and is not a parameter for assessment of agreement. On the other hand, concordance correlation coefficient not only can investigation the linear association but also can detect the location and scale shifts. We therefore propose to use the concordance correlation coefficient to assess agreement in the framework of a noninferiority hypothesis. However, selection of the lower limit for the noninferiority hypothesis should address the minimal required threshold for the linear association as well as the acceptable equivalence limits for mean and variability differences for gene expression levels between technical replicates. It should be determined jointly by biologists, clinicians, biostatisticians, and other research personal for quality control and quality assurance in development of diagnostic biochip products based on microarray technology.

However, there is no reason to assume that the expression levels of different genes in a given array are statistically independent. Our simulation results show that the correlation of expression levels among genes severely reduces the coverage probability of both asymptotic and exact confidence intervals for the concordance correlation coefficient as well as the coverage probability of asymptotic confidence interval for Pearson correlation coefficient. On the other hand, the impact of correlated expression levels of different genes is rather minimal on the size and power of both asymptotic and exact procedures for the noninferiority hypothesis in assessment of agreement. Therefore, the impact of dependence on assessment of agreement using concordance correlation coefficient requires further research, especially for the situations with ten of thousands of genes. In addition, agreement of the observed expression levels between technical replicates does not imply that they agree with the true unknown expression levels because both can be either close to or far from the true expression levels of genes. Since microarray is a high-dimensional parallel assay with ten of thousands of analytes, application of the traditional assay validation methods to diagnostic biochip products based on microarray technology remains a great challenge.

APPENDIX

First, we provide a brief review about the GPQs.

Suppose that X is a random variable whose distribution depends on a vector of unknown parameters $\boldsymbol{\zeta} = (\theta, \boldsymbol{\eta})$, where θ is a parameter of interest and $\boldsymbol{\eta}$ is a vector of nuisance parameter. Let $\mathbf{X}$ be a random sample from X and $\mathbf{x}$ be the observed value of $\mathbf{X}$. And let $R = r(\mathbf{X}; \mathbf{x}, \boldsymbol{\zeta})$ be a function of $\mathbf{X}$ and possible $\mathbf{x}, \boldsymbol{\zeta}$ as well. The random quantity R is said to be a GPQ for θ if it has the following two properties:

Property A: R has a probability distribution that is free of unknown parameters.

Property B: r_{obs} defined as $r_{obs} = r(\mathbf{x}; \mathbf{x}, \boldsymbol{\zeta})$ (this will be referred to as the observed pivotal) does not depend on nuisance parameters, $\boldsymbol{\eta}$.

As a result, a two-sided equal-tailed $100(1-\alpha)\%$ generalized confidence interval for θ is given by $(R_{\alpha/2}; R_{1-\alpha/2})$, where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of the distribution of R. The percentiles of R can be estimated using Monte-Carlo algorithms.

Mathew and Webb (2005) provide a set of mutually independent GPQs for the variances and covariance of a bivariate normal distribution. We may directly apply their results to our study. Define the following functions of parameters:

$$(\mu, \Sigma, \sigma_{11.2}) = \left(\mu_1 - \mu_2, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}, \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right). \tag{A.1}$$

Their estimators, denoted by $(\hat{\mu}, \widehat{\Sigma}, S_{11.2})$, are expressed as

$$(\hat{\mu}, \widehat{\Sigma}, S_{11.2}) = \left(\hat{\mu}_1 - \hat{\mu}_2, \begin{pmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{pmatrix}, S_{11} - \frac{S_{12}^2}{S_{22}}\right). \tag{A.2}$$

It can be verified that estimators $S_{22}, S_{11.2}, S_{12}$, and $\hat{\mu}$ are associated with pivotal quantities $U_{22}, U_{11.2}, Z_{S12}$, and Z_{μ} with the following known distributions:

$$U_{22} = \frac{S_{22}}{\sigma_{22}} \sim \chi^2_{n-1}, \quad U_{11.2} = \frac{S_{11.2}}{\sigma_{11.2}} \sim \chi^2_{n-2},$$

$$Z_{S_{12}} = \left(S_{12} - \frac{\sigma_{12}}{\sigma_{22}} S_{22}\right) / \sqrt{\sigma_{11.2} S_{22}} \sim N(0, 1), \quad \text{and}$$

$$Z_{\mu} = \frac{\hat{\mu} - \mu}{\sqrt{\frac{1}{n}(\sigma_{11} + \sigma_{22} - 2\sigma_{12})}} \sim N(0, 1)$$

where χ^2_{n-1} and χ^2_{n-2} denote the central chi-square distribution with $n-1$ and $n-2$ degrees of freedom, respectively, and $N(0, 1)$ is the standard normal distribution. Because $\hat{\mu}$ and $\widehat{\Sigma}$ are independent, it follows that $\hat{\mu}$ and $(S_{22}, S_{11.2}, S_{12})$ are also independent. Let $s_{22}, s_{11.2}, s_{12}, s_{11}$, and d denote the observed value of $S_{22}, S_{11.2}, S_{12}, S_{11}$, and, $\hat{\mu}$, respectively. Then the GPQs for $\sigma_{22}, \sigma_{12}, \sigma_{11}$, and μ are given by, respectively

$$R_{\sigma_{22}} = \frac{\sigma_{22}}{S_{22}} s_{22} = \frac{s_{22}}{U_{22}}, \tag{A.3}$$

$$\begin{aligned} R_{\sigma_{12}} &= \frac{\sigma_{22}}{S_{22}} s_{12} - \left[\sqrt{s_{11.2} s_{22}} \frac{(S_{12} - (\sigma_{12}/\sigma_{22}) S_{22})}{\sqrt{\sigma_{11.2} S_{22}}} \sqrt{\frac{\sigma_{11.2}}{S_{11.2}}} \frac{\sigma_{22}}{S_{22}}\right] \\ &= \frac{s_{12}}{U_{22}} - \left[\sqrt{s_{11.2} s_{22}} \frac{Z_{s_{12}}}{\sqrt{U_{11.2}}} \frac{1}{U_{22}}\right] \end{aligned} \tag{A.4}$$

$$R_{\sigma_{11}} = \frac{\sigma_{11.2}}{S_{11.2}} s_{11.2} + \frac{R^2_{S_{12}}}{R_{S_{22}}} = \frac{s_{11.2}}{U_{11.2}} + \frac{R^2_{\sigma_{12}}}{R_{\sigma_{22}}}. \tag{A.5}$$

From the results (A.3) to (A.5), we generate other set of GPQs $R_{\mu\sigma_{22}}, R_{\mu\sigma_{12}}, R_{\mu\sigma_{11}}$ with other pivotal quantities $U_{\mu 22}, U_{\mu 11.2}, Z_{\mu s_{12}}$, where $Z_{\mu S_{12}}$ is the standard normal variable, $U_{\mu 22}$ and $U_{\mu 11.2}$ are independent chi-square random variables with the

degrees of freedom $n-1$ and $n-2$, respectively. The a GPQ for μ is given as

$$\begin{aligned} R_\mu &= d - \frac{\hat{\mu}-\mu}{\sqrt{\frac{1}{n}(\sigma_{11}+\sigma_{22}-2\sigma_{12})}} \\ &\quad \times \sqrt{\frac{1}{n}\left[\left(\frac{s_{11.2}}{U_{\mu 11.2}}+\frac{R^2_{\mu\sigma_{12}}}{R^2_{\mu\sigma_{22}}}\right)+\left(\frac{s_{22}}{U_{\mu 22}}\right)-2\left(\frac{s_{12}}{U_{\mu 22}}-\sqrt{s_{11.2}s_{22}}\frac{Z_{\mu s_{12}}}{\sqrt{US_{\mu 11.2}}}\frac{1}{U_{\mu 22}}\right)\right]} \\ &= d - Z_\mu\sqrt{\frac{1}{n}(R_{\mu\sigma_{11}}+R_{\mu\sigma_{22}}-2R_{\mu\sigma_{12}})}. \end{aligned}$$

It can be shown that the observed values of $(R_{\sigma_{22}}, R_{\sigma_{12}}, R_{\sigma_{11}}, R_\mu)$ have distributions that are free of parameters $(\sigma_{22}, \sigma_{12}, \sigma_{11}, \mu)$, respectively. When $(S_{22}, S_{12}, S_{11}, \hat{\mu})$ are substituted by their observed values $(s_{22}, s_{12}, s_{11}, d)$, then $(R_{\sigma_{22}}, R_{\sigma_{12}}, R_{\sigma_{11}}, R_\mu)$ turns out to be $(\sigma_{22}, \sigma_{12}, \sigma_{11}, \mu)$. Hence, they too fulfill the requirements of Property A and Property B of GPQs described above.

The random variable $\hat{\mu}^2$ converges in distribution to a normal distribution with mean $(\mu_1-\mu_2)^2$ and variance $4(\mu_1-\mu_2)^2\ \frac{1}{n}(\sigma_{11}+\sigma_{22}-2\sigma_{12})$, i.e.,

$$Z_{\mu^2} = \frac{\hat{\mu}^2-(\mu_1-\mu_2)^2}{\sqrt{4(\mu_1-\mu_2)^2\frac{1}{n}(\sigma_{11}+\sigma_{22}-2\sigma_{12})}} \sim N(0,1), \quad \text{as } n\to\infty.$$

Let d^2 denote the observed value of $\hat{\mu}^2$; and $R_{\mu^2\sigma_{22}}, R_{\mu^2\sigma_{12}}, R_{\mu^2\sigma_{11}}$ can be generated with other GPQs $U_{\mu^2 22}, U_{\mu^2 11.2}, Z_{\mu^2 s_{12}}$, where $Z_{\mu^2 s_{12}}$ is the standard normal random variable, and $U_{\mu^2 22}$ and $U_{\mu^2 11.2}$ are independent central chi-square random variables with the $n-1$ and $n-2$ degrees of freedom, respectively. Following Quiroz (2004), a GPQ for $\hat{\mu}^2$ is then given by

$$\begin{aligned} R_{\mu^2} &= d^2 - \frac{\hat{\mu}^2-\mu^2}{2|\mu_1-\mu_2|\sqrt{\frac{1}{n}(\sigma_{11}+\sigma_{22}-2\sigma_{12})}} \\ &\quad \times 2|R_\mu|\sqrt{\frac{1}{n}\left[\left(\frac{s_{11.2}}{U_{\mu^2 11.2}}+\frac{R^2_{\mu^2\sigma_{12}}}{R_{\mu^2\sigma_{22}}}\right)+\left(\frac{s_{22}}{U_{\mu^2 22}}\right)-2\left(\frac{s_{12}}{U_{\mu^2 22}}-\sqrt{s_{11.2}s_{22}}\frac{Z_{\mu^2 s_{12}}}{\sqrt{U_{\mu^2 11.2}}}\frac{1}{U_{\mu^2 22}}\right)\right]} \\ &= d^2 - 2Z_{\mu^2}|R_\mu|\sqrt{\frac{1}{n}(R_{\mu^2\sigma_{11}}+R_{\mu^2\sigma_{22}}-2R_{\mu^2\sigma_{12}})}. \end{aligned} \tag{A.6}$$

The observed value of R_{μ^2} has distributions free of parameters μ^2 too. When $\hat{\mu}^2$ is substituted by it observed values d^2, then R_{μ^2} turns out to be μ^2. Hence, it fulfills the requirements of Property A and Property B described above for GPQ. A GPQ for the concordance correlation coefficient is then given as:

$$R_{\mu\rho_C} = \frac{2R_{12}}{R_{11}+R_{22}+R_{\mu^2}}. \tag{A.7}$$

Because $R_{\mu\rho_C}$ is a function that only depends on $(R_{\sigma_{22}}, R_{\sigma_{12}}, R_{\sigma_{11}}, R_{\mu^2})$, it too satisfies the conditions for being a GPQ.

ACKNOWLEDGMENT

We want to thanks the anonymous reviewers for their thorough and thoughtful review and comments which greatly improve the presentation of the manuscript. This research is partially supported by the Taiwan National Science Council Grant: NSC95 2118-M-002-007-MY2 to Jen-pei Liu.

REFERENCES

Bland, J. M., Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurements. *Lancet* 1:307–310.

Dobbin, K. K., Beer, D. G., Myerson, M, Yeatman, T. J., Gerald, W. L., Jacobson, J. W. (2005). Interlaboratory comparability study pf cancer gene expression analysis using oligonucleotide. *Clinical Cancer Research* 11:565–572.

Members of the Toxicogenomics Research Consortium (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods* 2:351–356.

Ioannidis, J. P. A. (2005). Microarrays and molecular research: noise discovery. *Lancet* 365:454–455.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C. et al. (2005). Multi-laboratory comparison of microarray platforms. *Nature Methods* 2:345–349.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R., Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. *Nature Methods* 2:337–343.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–268.

Lin, L. I. (1992). Assay validation using the concordance correlation coefficient. *Biometrics* 48:599–604.

Lin, L. I., Hedayat, A. S., Sinha, B., Yang, M. (2002). Statistical methods in assessing agreement: models, issues, and tools. *Journal of the American Statistical Association* 97:257–270.

Mathew, T., Webb, D. W. (2005). Generalized p values and confidence intervals for variance components: applications to army test and evaluation. *Technometrics* 47:312–322.

Michiels, S., Koscielny, S., Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365:488–492.

MINDACT Design and MINDACT trial overview. http://www.breast/internationalgroup.org/transbig.html. Accessed on June 5 2006.

Quiroz, J. (2004). Assessment of equivalence using a concordance correlation coefficient in a repeated measurements design. *Journal of Biopharmaceutical Statistics* 15:913–928.

Shi, L., Tong, W., Goodsaid, F., Frueh, F. W., Fang, H., Han, T., Fuscoe, J. C., Casciano, D. A. (2004). QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Review of Molecular Diagnosis* 4:761–777.

Sprarano, J., Hayes, D, Dees, E. (2006). Phase III randomized study of adjuvant combination chemotherapy and hormonal therapy versus adjuvant hormonal therapy alone in women with previously resected axillary node-negative breast cancer with various levels of risk for recurrence (TAILORX Trial). http://www.cancer.gov/clinicaltrials/ECOG-PACCT-1. Accessed on June 5 2006.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempick, R. A., Raaka, B. M., Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research* 31:5676–5684.

The U.S. Food and Drug Administration (2006). Draft Guidance on In Vitro Diagnostic Multivariate Index Assays. http://www.fda.gov./cdrh/ovid/guidance/1610. pdf. Accessed on September 30, 2006.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association* 88:899–905.

Weerahandi, S. (1995). *Exact Statistical Methods for Data Analysis*. New York: Spring-Verlag.