ORIGINAL ARTICLE

Statistical Evaluation of Quality Performance on Genomic Composite Biomarker Classifiers

Jen-Pei Liu,[1,2] Li-Tien Lu[1]*

Background/Purpose: After completion of the Human Genome Project, genomic composite biomarker classifiers (GCBCs) became available. However, quality performance of GCBCs varies. We propose statistical methods for evaluation of the quality performance of GCBCs on selection of differentially expressed genes, agreement and reproducibility.

Methods: For detection of differentially expressed genes, an interval hypothesis was employed to take into account both biological and statistical significance. The concordance correlation coefficient (CCC) was used to evaluate the agreement of expression levels of technical replicates. The intraclass correlation coefficient (ICC) was suggested to assess the reproducibility between laboratories.

Results: A two one-sided test procedure was proposed to test the interval hypothesis. Statistical methods based on the generalized pivotal quantities for CCC and ICC were suggested to test the hypotheses for agreement and reproducibility. Simulation results demonstrated that all three methods could adequately control the type I error rate at the nominal level for assessment of differentially expressed genes, agreement and reproducibility.

Conclusion: Three appropriate statistical methods were developed for evaluation of quality performance on differentially expressed genes, agreement and reproducibility of GCBCs. [*J Formos Med Assoc* 2008;107 (12 Suppl):S28–S34]

Key Words: agreement, differentially expressed genes, genomic composite biomarker classifier, reproducibility of results

Pharmacogenomics and microarrays are two of the most important scientific breakthroughs of the last decade and have great potential in detection and treatment of diseases, and many other applications. On the other hand, because of genetic variations and genetic-by-environmental interaction, patients respond differently to the same treatment or therapeutic regimen. After completion of the Human Genome Project (HGP), disease targets at the molecular level have been identified. Thus, biochip products based on heritable DNA markers, mutations, and expression patterns for detection of diseases have become possible. One of the unique characteristics is that they are *in vitro* diagnostic devices composed of multiple parallel assays for multiple markers assayed simultaneously. In addition, one of the features of these *in vitro* diagnostic devices is their ability to classify patients through their molecular signature into the following four categories: (1) treatment is efficacious without toxicity; (2) treatment is efficacious with toxicity; (3) treatment is not efficacious without toxicity; and (4) treatment is not efficacious with toxicity. As a result,

[1]Division of Biometry, Graduate Institute of Agronomy, National Taiwan University, Taipei, and [2]Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan, Taiwan.

Received: July 30, 2008
Revised: September 15, 2008
Accepted: September 19, 2008

***Correspondence to:** Dr Jen-Pei Liu, Division of Biometry, Graduate Institute of Agronomy, National Taiwan University, 1, Section 4, Roosevelt Road, Taipei, Taiwan. E-mail: jpliu@ntu.edu.tw

these *in vitro* diagnostic devices based on multiple DNA markers, mutations, and expression patterns are referred to as genomic composite biomarker classifiers (GCBCs).

However, the quality performance of GCBCs varies.[1–3] For example, Ma et al[4] suggested the use of the ratio of the expression levels of two genes for prediction of clinical outcome in patients with early-stage breast cancer after receiving tamoxifen. However, their findings have not been reproduced by other investigators.[5] On the other hand, for evaluating the risk of recurrence in patients with early-stage breast cancer treated with hormonal therapy, the Oncotype DX breast cancer assay uses the expression levels of 21 genes measured by RT-PCR.[6,7] However, MammaPrint® assesses the risk of distant metastasis determined by the molecular signature provided by a 70-gene microarray. The diagnostic results provided by the GCBCs are not only important for the prognosis and prediction of clinical outcomes, but are also vital for selection of the optimal treatment modality. Therefore, the quality performance of the GCBCs becomes crucial. Consequently, the US Food and Drug Administration (FDA) recently issued the following guidance to ensure the quality of GCBCs: (1) *Gene Expression Profiling Test System for Breast Cancer Prognosis* (9 May 2007); (2) *Pharmacogenetic Tests and Genetic Tests for Heritable Markers* (19 June 2007); (3) *In Vitro Diagnostic Multivariate Index Assays* (26 July 2007); and (4) *Statistical Guidance on Reporting Results from Studies Evaluating Diagnosis Tests* (13 March 2007).

Among the many issues with regard to the quality performance of GCBCs, we focused only on the following three fundamental subjects: selection of differentially expressed genes, agreement of measurements of expression levels between technical replicates, and reproducibility between laboratories.

As demonstrated above, the number of genes or biomarkers varies from one GCBC to another. Therefore, the first task for development of a GCBC is to identify the molecular markers that truly differentiate patients with different molecular signatures, which can predict clinical outcomes or can correlate with responses to treatments. Traditional statistical approaches to identify differentially expressed genes only consider statistical significance. However, a gene identified by statistical significance does not imply that it is of any biological importance or can classify patients with different clinical outcomes or treatment responses. Therefore, we suggest that statistical formulation for the hypothesis on identification of differentially expressed genes must take into consideration the biological meaning and statistical significance simultaneously.

The Pearson correlation coefficient (PCC) is the most commonly employed statistical tool for evaluation of agreement of measurements of expression levels between technical replicates of the same gene, from the same sample, under the same operating conditions. PCC is an excellent measure for detection of linear associations, and it remains unchanged if the measurements of expression levels are added or multiplied by a constant. In other words, it is location and scale invariant. However, agreement of the measurements of expression levels between technical replicates requires reflection of changes in both means and variability. Therefore, the PCC cannot be used to assess agreement of gene expression levels between technical replicates. We suggest using the concordance correlation coefficient (CCC)[8] to evaluate the agreement of measured expression levels between technical replicates of the same genes, from the same samples, under the same operating conditions. On the other hand, for assessment of reproducibility between laboratories, the between-laboratory and between-sample variability are the most important sources of variation.[9] Therefore, based on the intraclass correlation coefficient (ICC), we propose formulating a hypothesis for assessment of reproducibility between laboratories under a two-way random effects model without interaction. For these three fundamental quality performance issues for GCBCs, we first state the corresponding statistical hypothesis. The statistical methods are then presented with either the simulation results or numerical examples.

Materials and Methods

Currently, most available statistical methods for identification of differentially expressed genes, including the *t* test,[10] permutation *t* test,[11] or significance analysis of microarray (SAM),[12] are based on the traditional hypothesis of equality. However, the hypothesis of equality can only detect whether the difference in the average expression levels is zero between groups of patients with different molecular signatures. Thus, it fails to take into account the magnitude of biologically meaningful fold changes. In addition, the false-positive rate for identifying differentially expressed genes can be very high because of simultaneously testing tens of thousands of genes at the same time with a small number of samples.

In order to include the magnitude of biological meaningful fold change in the formulation of statistical hypothesis, we define a non-differential zone in which the expression level of a gene cannot distinguish between two groups of patients with different molecular signatures. The upper boundary of the non-differential zone is a minimal biologically meaningful upper threshold above which a gene is overexpressed (overexpressed zone). Similarly, the lower boundary of the non-differential zone is a maximal biologically meaningful upper threshold under which a gene is underexpressed (underexpressed zone). A gene is said to be differentially expressed between two groups of patients if the difference in the average expression levels between the two groups is either larger than a minimal biologically meaningful upper threshold or smaller than a maximal biologically meaningful lower threshold. Based on this concept, Liu and Chow[13] formulated a hypothesis for identifying the differentially expressed genes as the interval hypothesis, by simultaneously taking both the minimal biologically meaningful fold changes and statistical significance into consideration. The null hypothesis of the interval hypothesis is the non-differential zone. On the other hand, the alternative hypothesis of the interval hypothesis consists of both overexpressed and underexpressed zones. Based on the interval hypothesis, a two one-sided tests (TOST) procedure has also been suggested by Liu and Chow.[13] However, the structures of the correlations of the expression levels of the different genes are not incorporated in their proposed TOST procedure, which requires the normality assumption. Therefore, Liu et al[14] proposed applying the multivariate permutation method to improve the TOST procedure (permuted TOST procedure).

The CCC can be represented as a ratio. The numerator is the covariance of the expression levels between two technical replicates. The denominator is the sum of variances of the expression levels of two technical replicates, plus the square of the difference of the average expression levels between two technical replicates. The range of CCC is from −1 to 1. A CCC of 1 indicates a perfect positive agreement and a CCC of −1 implies a perfect negative agreement. On the other hand, a CCC of 0 reveals no agreement. Liao et al[15] proposed a noninferiority hypothesis for evaluation of agreement of expression levels between two technical replicates. The null hypothesis of the noninferiority hypothesis is that CCC is less than a minimal threshold for agreement, say 0.9, while the alternative hypothesis is that CCC is greater than the minimal threshold of CCC for agreement. The null hypothesis is rejected and agreement of expression levels between technical replicates can be claimed at the 5% significance level if the lower 95% confidence limit for CCC is greater than the minimal threshold. The confidence limit for CCC can be constructed using either the asymptotic approach[8] or by the method of generalized pivotal quantities (GPQ).[16]

For evaluation of reproducibility between laboratories, a two-way random-effects model is employed to estimate the between-laboratory, between-sample and the error variances. The ICC is the ratio of the between-laboratory variance to the total variance, which is the sum of the between-laboratory, between-sample, and error variances. The range of ICC is from 0 to 1. An ICC of 1 implies a perfect reproducibility. Again, a noninferiority hypothesis is used to assess the between-laboratory reproducibility. The null hypothesis of

the noninferiority hypothesis is that the ICC is no greater than the minimal threshold of ICC for reproducibility, say 0.5. On the other hand, the alternative hypothesis is that the ICC is greater than the minimal threshold. The null hypothesis of the noninferiority hypothesis is rejected, and the between-laboratory reproducibility can be concluded at the 5% significance level if the lower 95% confidence limit for ICC is greater than the minimal threshold. The confidence limit for CCC can be constructed using either the modified large sample (MLS) approach[17] or by the GPQ method which can handle the imbalances. Technical details of the GPQ for ICC can be obtained from the authors upon request.

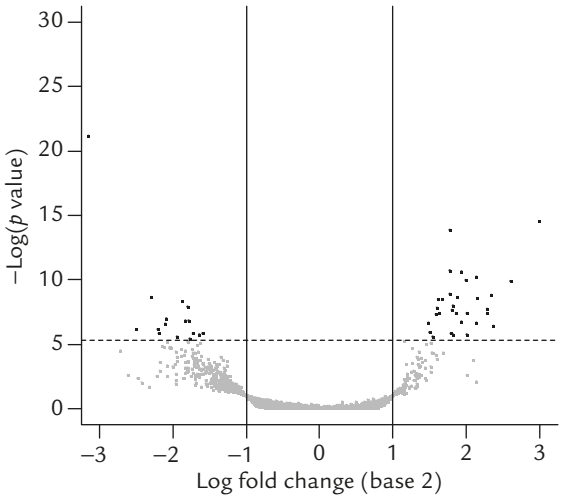

Figure 1. Results of the TOST procedure.

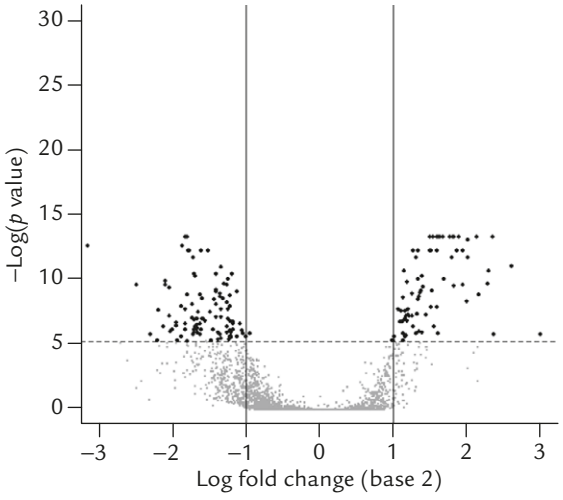

Figure 2. Results of the permuted TOST procedure.

Results

Example 1: identification of differentially expressed genes

The data set of Luo et al[18] was used to illustrate application of the TOST and permuted TOST methods. It consisted of normalized gene expression ratios obtained from a collection of 25 prostate tissue samples comprising 16 prostate cancers and nine benign prostatic hyperplasia (BPH) specimens. The data were obtained from http://research.nhgri.nih.gov/microarray/prostate/supplement/images/6500GeneListw=CRs&QSs.xls. A common reference design was used for this series of experiments. A total of 5854 genes with quality scores greater than zero for at least three prostate cancer specimens and three BPH specimens were used in the analysis. Log-transformation (base 2) was the scale used for analysis. For the purpose of illustration, the non-differential zone for a gene between the prostatic cancer and BPH was [−1, 1] on the log scale. In other words, a gene was differentially expressed between the patients with prostatic cancer and those with BPH if its true fold change was greater than 2 or smaller than 1/2.

The results of the TOST based on the interval hypothesis with the minimal biologically meaningful threshold of 1 (log2 base) are presented in Figure 1. At the 5% nominal significance level, there were a total of 47 genes (0.8%) with p values calculated from the TOST procedure < 0.05. As can be seen in Figure 1, the minimal observed fold change with $p < 0.05$ obtained from the TOST procedure was > 2.83 (i.e. 1.5 log2). However, as mentioned above, the TOST is based on the normal assumption and does not consider the correlation structures of expression levels among genes. Therefore, as demonstrated by this example, it can be quite conservative in the identification of differentially expressed genes. Figure 2 provides the results of the permuted TOST procedure. A total of 181 genes were identified as differentially expressed by the permuted TOST procedure at the 5% significance level. The interval hypothesis

Table. Concordance correlation coefficient based on log2 intensity by method and laboratory

Laboratory	Estimate	95% lower confidence limit	
		Asymptotic	Exact
615	0.9862	0.9809	0.9804
616	0.9918	0.9886	0.9885
617	0.9775	0.9694	0.9687
618	0.9867	0.9820	0.9817

Reproduced with permission from Liao et al.[15]

directly takes into consideration the minimal biologically meaningful threshold; therefore, these 181 differentially expressed genes identified by the permuted TOST not only possessed biologically meaningful fold changes with a magnitude of at least 2, but also took into account the variability of observed fold changes and reached statistical significance.

Example 2: assessment of agreement

Liao et al[15] used the dataset of Dobbin et al[9] to demonstrate the application of CCC to the assessment of agreement between two technical replicates of samples of five cell line pellets at each of four laboratories (Lab 615, Lab 616, Lab 617, Lab 618). The array platform used was the Affymetrix Human Genome U133A arrays. Due to the importance of the expression levels of the housekeeping genes for quality control of the data derived from microarray experiments, the normalized intensities on the log2 scale from 100 housekeeping genes of cell line H1437 obtained from each of the four laboratories were used in the analysis.

The results for evaluation of agreement of the expression levels between two replicates within each laboratory by the asymptotic and the GPQ methods are presented in the Table. The CCC ranged from 0.9775 in Lab 617 to 0.9918 in Lab 616. The 95% lower confidence limits on the log2 scale by both the GPQ and asymptotic methods were almost identical to the third decimal point for all four laboratories. In addition, all 95% lower confidence limits by both methods were >0.90 for the four laboratories. Therefore, if the minimal threshold of CCC for agreement was 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 met the minimal requirement of quality control for agreement for all four laboratories. In other words, for the 100 housekeeping genes, an excellent agreement existed between the two technical replicates at all four laboratories.

Reproducibility of simulation results

The results of our simulation studies indicated that the performance of both MLS and GPQ approaches was very similar with respect to probability coverage, type I error rate and power. For example, the probability coverage of the 95% confidence intervals constructed by both MLS and GPQ methods was at least 95%. This phenomenon also implied that both methods were conservative in their assessment of reproducibility between laboratories. In other words, the type I error rate for falsely claiming to meet the minimal threshold of reproducibility was in fact below the nominal 5% level. Figure 3 presents the power curves of both MLS and GPQ methods when the minimal threshold of reproducibility was 0.5, and the numbers of laboratories and samples were equal to 10. Figure 3 reveals that although the difference in power was relatively small, the GPQ approach was uniformly more powerful than the MLS method.

Discussion

Quality performance of GCBCs is vital to the success of translational medicine. A well-performed GCBC can not only accurately predict the clinical outcomes of patients, but also optimize the treatment for each patient based on his or her own clinical or pharmacogenomic characteristics, to achieve the goal of personalized medicine. Three of the most important elements in quality performance of GCBCs are selection of truly differentially expressed genes, agreement of expression

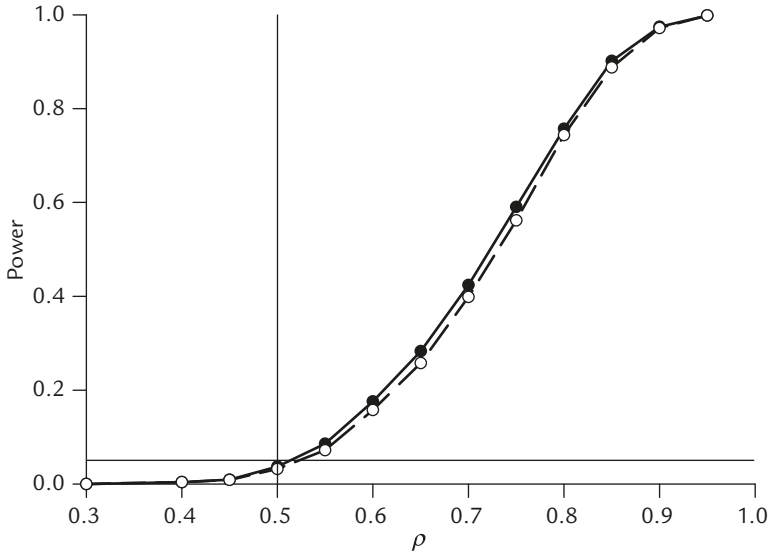

Figure 3. Power curve for testing the hypothesis that between-laboratory reproducibility (ρ) is >0.5. Solid curve = GPQ method; dashed curve = MLS method.

levels between technical replicates of the same samples obtained under the same operating conditions, and reproducibility between laboratories. We proposed criteria for evaluation of quality performance with respect to the three elements. Based on the suggested criteria, we formulated statistical hypotheses and proposed statistical methods with respect to each element. Simulation results and applications to real data demonstrated that the proposed hypotheses and statistical methods are an adequate approach for evaluation of quality of GCBCs, with respect to selection of truly differentially expressed genes, agreement and reproducibility.

Sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic (ROC) curve are other quality measures for evaluation of accuracy of GCBCs.[19,20] However, most of the current methods for statistical inference using the area under the ROC curve are derived based on a single biomarker. GCBCs, on the other hand, are *in vitro* diagnostic devices composed of multiple biomarkers or genomic markers. In addition, even for the prognosis or prediction of the same clinical outcome, the number of genomic markers is different from one GCBC to another. Therefore, evaluation of quality performance for a single GCBC, choice of thresholds for prognosis, and medical decision or comparison of accuracy between GCBCs based on area under the ROC curve turn out to be more complicated and require further research.

MammaPrint® is a qualitative *in vitro* diagnostic test that uses the expression profile of 70 genes from fresh frozen breast cancer tissue samples, based on microarray technology, to evaluate the risk of distant metastasis in patients with node-negative breast cancer.[21–24] The area under the ROC curves for time to distant metastases at 5 years and overall survival at 10 years are 0.681 and 0.648. In addition, the positive predictive value is only 0.22 for metastatic disease at 5 years. In other words, 78% of patients with a positive result by MammaPrint® will not have metastatic disease at 5 years. Although it is approved by the US FDA, the accuracy of MammaPrint® is, at best, mediocre. One of the possible reasons for the middling performance of MammaPrint® is that the duration for prognosis or prediction is 5–10 years. Many medical advances, such as introduction of a new effective treatment, can take place during such a long period. Therefore, research on the study design and analyses for evaluation of GCBCs for long-term prognosis or prediction require urgent attention.

Disclaimer

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

References

1. Shi L, Tong W, Goodsaid F, et al. QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Rev Mol Diagn* 2004;4:761–77.
2. Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. *J Natl Cancer Inst* 2005;97:866–7.
3. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol* 2005; 23:7332–41.
4. Ma X, Wang Z, Ryan PD, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004;5:607–16.
5. Reid JF, Lusa L, De Cecco L, et al. Limits of predictive models using microarray data for breast cancer treatment clinical outcome. *J Natl Cancer Inst* 2005;97:927–30.
6. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817–26.
7. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* 2006;24: 1–12.
8. Lin LI. Assay validation using concordance correlation coefficient. *Biometrics* 1992;48:599–604.
9. Dobbin KK, Beer DG, Meyerson M, et al. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin Cancer Res* 2005;11: 565–72.
10. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002;12:111–39.
11. Simon RM, Korn EL, McShane LM, et al. *Design and Analysis of DNA Microarray Investigations*. New York: Springer, 1980.
12. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to ionizing radiation response. *Proc Natl Acad Sci USA* 2001;98:5116–21.
13. Liu JP, Chow SC. Statistical issues on the diagnostic multivariate index assay and targeted clinical trials. *J Biopharm Stat* 2008;18:167–82.
14. Liu JP, Liao CT, Chiu ST, Dai JY. A permutation two one-sided tests procedure to test minimal fold changes of gene expression levels. *J Biopharm Stat* 2008;18:802–26.
15. Liao CT, Lin CY, Liu JP. Non-inferiority tests based on concordance correlation coefficient for assessment of agreement for gene expression data from microarray experiments. *J Biopharm Stat* 2007;17:309–27.
16. Weerahandi S. Generalized confidence intervals. *J Am Stat Assoc* 1993;88:899–905.
17. Burdick RK, Graybill FA. *Confidence Intervals on Variance Components*. New York: Marcel Dekker, 1992.
18. Luo J, Duggan DJ, Chen Y, et al. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* 2001;61:4683–8.
19. Liu JP, Ma MC, Wu CY, Tai JY. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. *Stat Med* 2006;25:1219–39.
20. Li CR, Liao CT, Liu JP. On exact interval estimation for the difference in the paired areas under ROC curves. *Stat Med* 2008;27:224–42.
21. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6.
22. Van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;347:1999–2009.
23. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006;98: 1183–92.
24. US Food and Drug Administration. *Decision Summary k062694*. Rockville, MD: US FDA, 2007.