BRIEF COMMUNICATIONS

Limits of Predictive Models Using Microarray Data for Breast Cancer Clinical Treatment Outcome

James F. Reid, Lara Lusa, Loris De Cecco, Danila Coradini, Silvia Veneroni, Maria Grazia Daidone, Manuela Gariboldi, Marco A. Pierotti

Data from microarray studies have been used to develop predictive models for treatment outcome in breast cancer, such as a recently proposed predictive model for antiestrogen response after tamoxifen treatment that was based on the expression ratio of two genes. We attempted to validate this model on an independent cohort of 58 patients with resectable estrogen receptor–positive breast cancer. We measured expression of the genes HOXB13 and IL17BR with real time–quantitative polymerase chain reaction and assessed the association between their expression and outcome by use of univariate logistic regression, area under the receiver-operating-characteristic curve (AUC), a two-sample *t* test, and a Mann–Whitney test. We also applied standard supervised methods to the original microarray dataset and to another independent dataset from similar patients to estimate the classification accuracy obtainable by using more than two genes in a microarray-based predictive model. We could not validate the performance of the two-gene predictor on our cohort of samples (relation between outcome and the following genes estimated by logistic regression: for HOXB13, odds ratio [OR] = 1.04, 95% confidence interval [CI] = 0.92 to 1.16, *P* = .54; for IL17BR, OR = 0.69, 95% CI = 0.40 to 1.20, *P* = .18; and for HOXB13/IL17BR, OR = 1.30, 95% CI = 0.88 to 1.93, *P* = .18). Similar results were obtained with the AUC, a two-sample two-sided *t* test, and a Mann–Whitney test. In addition, estimates of classification accuracies applied to two independent microarray datasets highlighted the poor performance of treatment-response predictive models that can be achieved with the sample sizes of patients and informative genes to date. [J Natl Cancer Inst 2005;97:927–30]

Several studies have demonstrated that breast cancers with distinct pathologic features can be recognized by their gene expression profile *(1–11)*. Microarrays have been used to identify expression patterns capable of predicting outcome or response after specific treatments such as tamoxifen, which is a standard adjuvant treatment for patients with primary, estrogen receptor–positive breast cancer *(12,13)*. Currently, many patients do not respond to treatment, and so additional biomarkers predictive of treatment failure within endocrine-responsive diseases are required.

Recently, a tamoxifen-response predictive model consisting of only two genes has been described *(14)*. By using microarray gene expression profiles of 60 tamoxifen-treated patients, HOXB13 and IL17BR were identified as the two genes whose expression ratio predicts clinical outcome. This finding was validated by use of real time–quantitative polymerase chain reaction (RT-QPCR) on an independent set of 20 formalin-fixed, paraffin-embedded samples by correctly classifying the outcomes of 16 patients ($P = .01$). However, by considering the data from relapsed and disease-free patients separately, although the probability of obtaining such a correct classification by chance remained low for disease-free patients (nine of 10 correctly classified, $P = .02$; 95% confidence interval [CI] for the proportion of correctly classified samples = 0.55 to 0.99), this estimate increased drastically for relapsed patients (seven of 10 correctly classified, $P = .34$; 95% CI = 0.35 to 0.93). Although the proposed predictive model is very appealing from clinical and practical points of view because of its potential straightforward application in many laboratories, the results of the validation set (i.e., the statistically nonsignificant results for the relapsed patients) indicate that a larger validation set is required.

For this reason, we applied this two-gene predictive model for relapse to a dataset derived from a cohort of 58 patients with early-stage, estrogen receptor–positive primary breast cancer who were treated at the Istituto Nazionale Tumori between March 1, 1991, and December 31, 1997, with radical or conservative surgery plus radiotherapy followed by adjuvant monotherapy with tamoxifen (median treatment duration = 60 months, range = 27–84 months). All patients signed an informed consent to donate any tissue leftover after diagnostic procedures to Istituto Nazionale Tumori. A tumor was classified as estrogen receptor positive if the ligand binding assay detected more than 10 fmol of estrogen bound per mg of total protein. Disease recurred with distant metastasis in 18 patients (16 patients as a first event and two as a second event after local-regional recurrence) of the 58 patients within a median time of 31 months (range = 14–43 months) from surgery. Forty of the 58 patients were disease free after a median time of 93 months (range = 70–125 months).

Clinical and pathobiologic details of these 58 patients are presented in supplemental Table 1 (Available at: http://jncicancerspectrum.oupjournals.org/jnci/content/vol97/issue12). Most patients were older than 50 years of age (93.1%) and had lymph node–positive disease (77.5%; 53.5% had one to three positive lymph nodes and 24.0% had more than three positive lymph nodes). Their

Affiliations of authors: Department of Experimental Oncology, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano, Italy (JFR, LL, LDC, DC, SV, MGD, MG, MAP); Molecular Cancer Genetics Group, Fondazione Istituto FIRC di Oncologia Molecolare (IFOM), Milano, Italy (JFR, LL, LDC, MG, MAP).

Correspondence to: James F. Reid, Fondazione Istituto FIRC di Oncologia Molecolare (IFOM), Milano, Italy (e-mail: james.reid@ifom-ieo-campus.it); Manuela Gariboldi, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano, Italy (e-mail: manuela.gariboldi@istitutotumori.mi.it); Marco A. Pierotti, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano, Italy (e-mail: marco.pierotti@istitutotumori.mi.it).

See "Notes" following "References."

DOI: 10.1093/jnci/dji153

Journal of the National Cancer Institute, Vol. 97, No. 12,

tumors were larger than 2 cm (62.1% of tumors), were progesterone receptor positive (79.3% of tumors; i.e., more than 25 fmol of progesterone bound per mg of total protein by ligand binding assay), and were HER-2/neu negative (77.6% of tumors). HER-2/neu status was immunohistochemically assessed with polyclonal antibody against $p185^{HER2}$ protein (1:2000 dilution, DAKO, Milan, Italy) and defined as positive when strong membrane labeling was observed. A limitation of any validation study on independent cohorts can be related to having a different mixture of case patients than that of the original study. Compared with the previously described cohort *(14)*, our cohort had a prevalence of tumors that were lymph node positive (77.5% vs. 47.2%), HER-2/neu positive (20.7% vs. 5.4%), and larger than 2 cm (62.1% vs. 47.2%).

RT-QPCR used TaqMan gene expression assays for the following genes: HOXB13 labeled with FAM-MGB (a 6-carboxyfluorescein fluorescent dye and a minor groove binding [MGB] molecule attached to the 3′ end, which stabilizes the probe annealing; product Hs00197189), IL17BR labeled with FAM-MGB (product Hs00218889), and human GAPDH VIC-MGB (VIC is a proprietary fluorescent dye; product 4326317E), a housekeeping gene used for normalization (Applied Biosystems, Foster City, CA). Gene expression data were quantified as described by the manufacturer and log-transformed (Fig. 1, A, and raw data in supplemental Table 2; available at http://jncicancerspectrum.oupjournals.org/jnci/content/vol97/issue12).

We followed the procedures as previously outlined *(14)* to evaluate the association between the expression of the two genes and outcome with a two-sided *t* test with unequal variances, with an area under the receiver-operating-characteristic curve (AUC) analysis, and with re-estimated univariate logistic models because the original models were not reported. In addition to a *t* test, the nonparametric Mann–Whitney test was also considered to avoid making assumptions on the distribution of expression data, which departed from normality for HOXB13.

However, our analyses of this independent set of samples did not find any statistically significant association between the gene expression of HOXB13, IL17BR or their ratio and outcome after tamoxifen treatment (e.g., from univariate logistic regression, for HOXB13 odds ratio [OR] = 1.04, 95% confidence interval [CI] = 0.92 to 1.16, $P = .54$; for IL17BR, OR = 0.69, 95% CI = 0.40 to 1.20, $P = .18$; and for HOXB13/IL17BR, OR = 1.30, 95% CI = 0.88 to 1.93, $P = .18$). Results of the latter model, with the overlapping estimated probabilities of recurrence for disease-free patients and relapsed patients are shown in Fig. 1, B. Similar *P* values were obtained from *t* tests, Mann–Whitney tests, and AUC analyses (Table 1).

Table 1. Association and discrimination of real time–quantitative polymerase chain reaction expression data from 58 primary estrogen receptor–positive, lymph node–positive breast cancers from patients treated with adjuvant monotherapy with tamoxifen

Analysis	HOXB13	IL17BR	HOXB13/IL17BR
Mean comparison*: mean (DF) − mean (R)	−0.85	0.42	−0.55
95% CI	(−3.74 to 2.05)	(−0.22 to 1.06)	(−1.42 to 0.31)
t test *P*	.56	.19	.20
Mann–Whitney *P*	.49	.21	.23
AUC†			
Coefficient	0.55	0.59	0.58
95% CI	(0.40 to 0.71)	(0.43 to 0.75)	(0.41 to 0.74)
P	.51	.27	.20
Logistic regression‡			
Odds ratio	1.04	0.69	1.30
95% CI	(0.92 to 1.16)	(0.40 to 1.20)	(0.88 to 1.93)
P	.54	.18	.18

*Difference between means of gene expression between the disease-free group and the relapsed group (disease-free [DF] – relapsed [R]). CI = confidence interval based on normality assumption. *t* test *P* = *P* value from two-sided *t* test with unequal variances; Mann–Whitney *P* = *P* value from two-sided Mann–Whitney test.

†Area under the receiver-operating characteristic (AUC) curve. Standard errors (SEs) were obtained by a bootstrap procedure ($B = 200$), allowing AUC to be less than 0.5. 95% CIs were obtained as AUC ± 1.96 SE. *P* values were calculated for the null hypothesis AUC = 0.5; alternative AUC ≠ 0.5.

‡Regression coefficient from univariate logistic regression (coding: 0 = disease-free; 1 = relapsed). *P* is the *P* value from two-sided Wald test.

Because of this contradictory result, we investigated the feasibility of predicting recurrence after tamoxifen treatment by making thorough use of published microarray data. We applied standard supervised analysis on the laser-capture microdissections dataset from Ma et al. *(14)* (hereafter dataset TAM1) and on a subset of tamoxifen-treated patients with estrogen receptor–positive tumors who had not undergone chemotherapy and who had a known recurrence status from another published cohort of patients *(6)* (hereafter dataset TAM2) that reflects the clinical characteristics of patients in TAM1. Both of these datasets used dual-labeling competitive-hybridization technologies (cDNAs and oligonucleotides) and the Universal Human Reference RNA (Stratagene, La Jolla, CA). Because we lacked independent validation sets, estimates of prediction error rates were obtained by use of 10-fold cross-validation, a sampling method that divides the data into 10 parts, each of which is set aside to test the accuracy achieved by using the prediction rule built on the remaining data *(7,15,16)*. As the prediction algorithm, we used diagonal linear discriminant analysis, a method that showed good performance for microarray data *(17,18)*. The number of genes included in the model was estimated by cross-validation, repeatedly performing feature selection for each training set *(19)* by use of the highest univariate two-sided pooled variance *t* statistic. This procedure enabled us to assess the sensitivity of the classifier to the number of selected genes (Fig. 1, C) and to obtain an estimate of 30 genes for TAM1 and six genes for TAM2. To obtain a correct estimate of the misclassification error rate associated with these predictive models, we performed a full cross-validation study that took into account the fact that the number of genes included in the model was not specified a priori. The resulting error rates were 39% for TAM1 and 46% for TAM2. Notice how these error rates would have been misleadingly underestimated if this full cross-validation had not been considered (25% for TAM1 and 24% for TAM2; Fig. 1, C). When only two genes were selected in TAM1, the misclassification error rate was 37% [$P = .325$, for the permutation test on the cross-validated

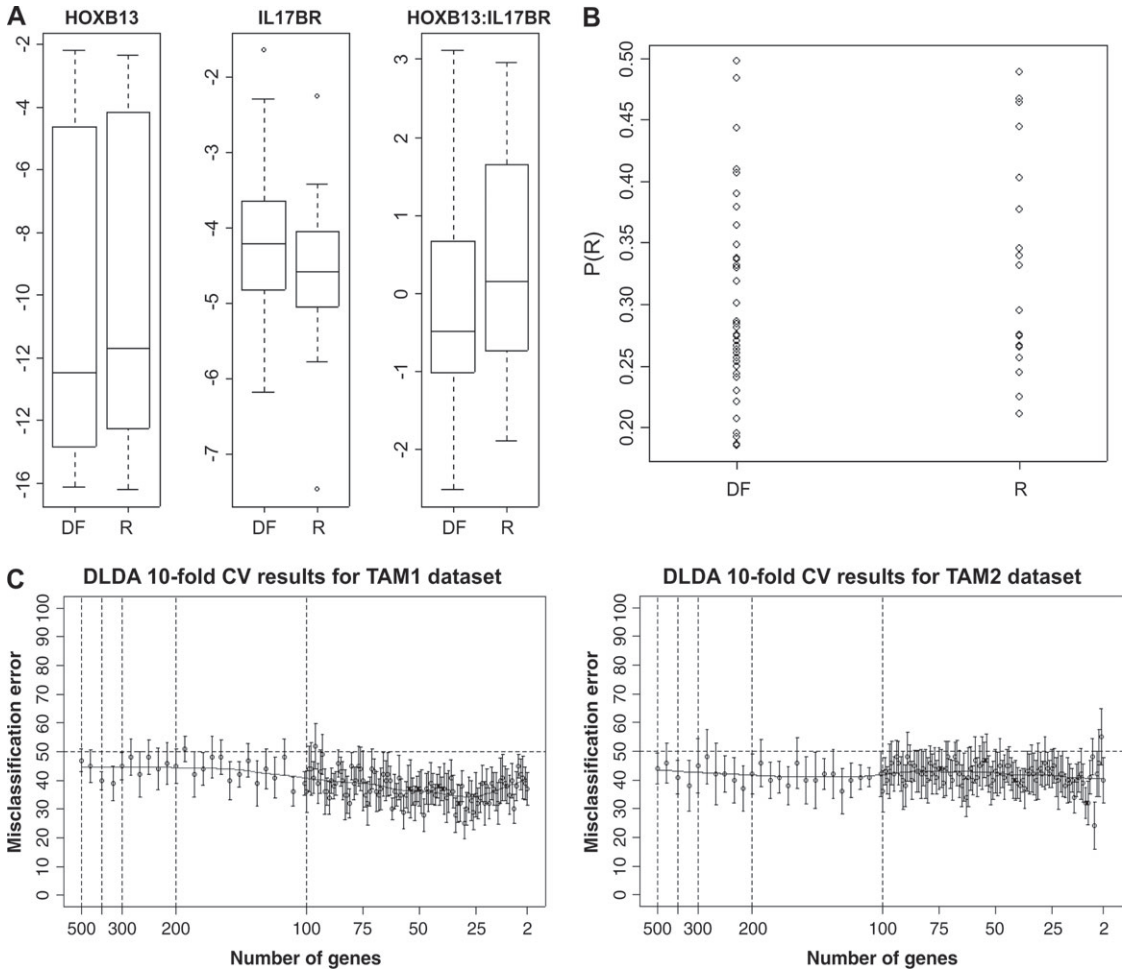

Fig. 1. Distributions of HOXB13 and IL17BR, associated predicted probabilities of recurrence estimated by logistic regression model, and average cross-validated misclassification error rates on two independent microarray data sets. **A)** Boxplots of the log-transformed amounts of targets for HOXB13 and IL17BR and their standardized ratio stratified by recurrence status on the Istituto Nazionale Tumori cohort. DF = disease free; R = recurrent. Total RNA was extracted from frozen samples with Trizol (Life Technologies, Frederick, MA) and treated with DNase. For cDNA synthesis, 2 μg of total RNA was reverse-transcribed and amplified in duplicate on an ABI PRISM 7700, according to the manufacturer's instructions. Relative quantitation of gene expression from real time–quantitative polymerase chain reaction (RT-QPCR) data were obtained by following the manufacturer's instructions. The ratio of the amounts of targets, HOXB13/IL17BR, was obtained after standardizing both genes separately (subtracting the mean and dividing by standard deviation). All the analyses were carried out with R, a language and environment for statistical computing *(25)* and receiver operating characteristic curves were fitted with the Bioconductor *(26)* roc library. The boxplots represent the first quartile (i.e., lower edge of the box), median (i.e., bar inside the box), third quartile (i.e., upper edge of the box), and minimum and maximum (i.e., horizontal lines). If any points are at a greater distance from the quartiles than 1.5 times the interquartile range (IQR), these are plotted individually and horizontal bars represent a distance of 1.5 times IQR from the upper or lower quartile. **B)** Probability of experiencing recurrence, stratified by recurrence status, estimated by logistic model including HOXB13/IL17BR as a covariate for the Istituto Nazionale Tumori cohort. **C)** Average 10-fold cross-validation (CV) misclassification error rates were determined by use of diagonal linear discriminant analysis (DLDA) of the TAM1 *(14)* and TAM2 *(6)* datasets on decreasing number of genes (from 500 to 250 at intervals of 50, from 250 to 150 at 10, from 150 to 100 at 5, and from 100 down to 2 the number was decreased one at a time). The superimposed curve was fitted by use of a lowess smoother, error bars represent average 10-fold cross-validation misclassification error rates plus or minus their standard error. Data sets were prepared as follows. Dataset TAM1 contained 60 samples (of which 59 were estrogen receptor-positive) from primary breast cancers from patients treated with tamoxifen monotherapy for 5 years (28 with recurrent disease and 32 who were disease free). Dataset TAM2 contained 99 breast cancer samples, 44 of which were estrogen receptor-positive by both a ligand binding assay and immunohistochemistry and were from patients who had not undergone chemotherapy and who had been treated with tamoxifen (16 with recurrent disease and 28 who were disease free). Raw data for dataset TAM1 were downloaded from the NCBI Gene Expression Omnibus (GEO) database at http://www.ncbi.nlm.nih.gov/geo/ (accession number GSE1378) and processed as previously described *(14)*, so that the raw background-corrected signals were normalized with a lowess transformation. The processed and normalized log ratios for dataset TAM2 were downloaded from the supporting information available from the publishers' web site at http://www.pnas.org/cgi/content/full/1732912100/DC1. Datasets were filtered according to the methods reported in the original publications. For the TAM1 dataset, only the filtered dataset that was based on overall variance of each gene (top 75th percentile) was used; this procedure resulted in a 5644 × 60 gene expression matrix. For the TAM2 data set, the whole set with no more than 50% missing values per gene was used; this procedure resulted in a 7525 × 44 gene expression matrix. Missing values in both data sets were imputed by the *k*-nearest neighbor method ($k = 10$), as previously described *(24)*, and finally data were standardized (all rows [genes] set to 0 = mean and 1 = standard deviation) before applying the cross-validation procedure. Microarray data preparation was done by use of the Bioconductor *(26)* Biobase library, multtest was used for the *t* tests, pamr was used for the missing value imputation and for creating balanced folds during the cross-validations, and sma [R (25) library] was used for the diagonal linear discriminant analysis.

misclassification error rate based on 2000 permutations *(20)*], indicating that two genes cannot predict a patient's response to tamoxifen treatment.

In conclusion, in our cohort of patients we failed to validate the predictive model proposed by Ma et al. *(14)*. Furthermore, building predictors for tamoxifen treatment by use of two independent microarray datasets did not provide promising results. These facts probably highlight the heterogeneous nature of the

underlying disease and, hence, the need for microarray data sets from much larger and/or more homogeneous cohort of samples to build more reliable predictive models. For the time being, because of the relatively small sample sizes of microarray experiments, this challenging task may only be circumvented by thoughtful experiment design *(21,22)* and by providing public access to published microarray data *(23)*, which will drive reproducible results and accelerate the design of appropriate meta-analytical techniques for integrating data from different studies. We believe that it is also crucial that microarray data be viewed as a valuable and rich source of additional information that can supplement information from clinical and pathobiologic markers toward the goal of developing efficient and subject-tailored treatment strategies.

References

(1) Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.

(2) Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869–74.

(3) Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 2003;100:8418–23.

(4) van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.

(5) van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.

(6) Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci U S A 2003;100:10393–8.

(7) West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 2001;98:11462–7.

(8) Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003;361:1590–6.

(9) Pittman J, Huang E, Dressman H, Horng CF, Cheng SH, Tsou MH, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proc Natl Acad Sci U S A 2004;101:8431–6.

(10) Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet 2003;362:362–9.

(11) Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, et al. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. J Clin Oncol 2004;22:2284–93.

(12) National Institutes of Health Consensus Development Panel. National Institutes of Health Consensus Development Conference statement: adjuvant therapy for breast cancer, November 1–3, 2000. J Natl Cancer Inst Monogr 2001;(30):5–15.

(13) Goldhirsch A, Wood WC, Gelber RD, Coates AS, Thurlimann B, Senn HJ. Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. J Clin Oncol 2003;21:3357–65.

(14) Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 2004;5:607–16.

(15) Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.

(16) Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. New York (NY): Springer-Verlag; 2004.

(17) Dudoit S, Fridly J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87.

(18) Speed T, editor. Statistical analysis of gene expression microarray data. Boca Raton (FL): Chapman & Hall/CRC; 2003.

(19) Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene expression data. Proc Natl Acad Sci U S A 2002;99:6562–6.

(20) Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002;9:505–11.

(21) Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002;32 Suppl:490–5.

(22) Gruvberger SK, Ringner M, Eden P, Borg A, Ferno M, Peterson C, et al. Expression profiling to predict outcome in breast cancer: the influence of sample selection. Breast Cancer Res 2003;5:23–6.

(23) Microarray Gene Expression Data (MGED). A guide to microarray experiments—an open letter to the scientific journals. Lancet 2002;360:1019.

(24) Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17:520–5.

(25) R Development Core Team. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2004. Available at: http://www.R-project.org/.

(26) Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80. Available at http://www.bioconductor.org/.

Notes

J. F. Reid and L. Lusa contributed equally as junior co-authors.

M. Gariboldi and M. A. Pierotti contributed equally as senior co-authors.

This work was supported by grants CNR/MIUR "Progetto Strategico Oncologia" (02.00385.ST97 to M.A. Pierotti), AIRC (Associazione Italiana per la Ricerca sul Cancro): individual grants to M. Gariboldi and M. A. Pierotti, and Sixth Framework Programme from the European Community: "Combating Cancer", TRANSFOG integrated project (proposal number 503438).

Manuscript received November 12, 2004; revised April 11, 2005; accepted April 13, 2005.