BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 12 2006, pages 1477–1485
doi:10.1093/bioinformatics/btl110

Gene expression

Identifying gene expression changes in breast cancer that distinguish early and late relapse among uncured patients

Philippe Broët[1,*], Vladimir A. Kuznetsov[2,*], Jonas Bergh[3], Edison T. Liu[2] and Lance D. Miller[2]

[1]Faculté de Médecine—Université, Paris-XI, IFR69, 16 Avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France, [2]Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore and [3]Department of Oncology and Pathology, Radiumhemmet, Karolinska Institute and Hospital, S-171 76 Stockholm, Sweden

Received on July 11, 2005; revised and accepted on March 20, 2006
Advance Access publication March 21, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: In recent years, microarray technology has revealed many tumor-expressed genes prognostic of clinical outcomes in early-stage breast cancer patients. However, in the presence of cured patients, evaluating gene effect on time to relapse is quite complex since it may affect either the probability of never experiencing a relapse (cure effect) or the time to relapse among the uncured patients (disease progression effect) or both. In this context, we propose a simple and an efficient method for identifying gene expression changes that characterize early and late recurrence for uncured patients.

Results: Simulation results show the good performance of the proposed statistic for detecting a disease progression effect. In a study of early-stage breast cancer, our results show that the proposed statistic provides a more powerful basis for gene selection than the classical Cox model-based statistic. From a biological perspective, many of the genes identified here as associated with the speed of disease recurrence have known roles in tumorigenesis.

Contact: broet@vjf.inserm.fr;kuznetsov@gis.a-star.edu.sg

1 INTRODUCTION

Since the inception of genome-wide transcript analysis technologies such as serial analysis of gene expression (SAGE) and DNA microarrays, there has been much interest in identifying gene expression changes in primary human tumors associated with survival outcomes (class comparison) to better understand the disease process and to develop so-called gene signatures (class prediction) to improve patient prognosis (Simon *et al*., 2004). Although these two issues are clearly different, they share a common key gene selection process step which may be more crucial than the gene signature modeling or the multiple comparison procedure considered.

For the analysis of censored survival times, the semiparametric Cox proportional hazard regression model is the favored choice and the statistic being considered is usually the Wald statistic derived from the corresponding univariate Cox partial likelihood function (Cox, 1972, 1975). In practice, the genes with the largest statistics are selected for further confirmatory analyses or for inclusion in a gene signature (van't Veer *et al*., 2002; Beer *et al*., 2002; Wang *et al*., 2005).

For early-stage cancer in which a fraction of the patients may be cured (sometimes referred to as long-term survivors) after the primary treatment, evaluating the association of gene expression changes to tumor relapse is quite complex since it may relate either to the probability of never experiencing a relapse (herein called cure or long-term effect) or to the time-to-relapse among the patients who are susceptible to relapse (herein called disease progression or short-term effect) or both. From a clinical point of view, prognostic factors with a cure effect are relevant for identifying non-susceptible patients who will not benefit from adjuvant systemic therapies but would otherwise sustain their side effects, whereas those factors with a disease progression effect would be useful for selecting patient with high risk for early relapse who may highly benefit from more aggressive therapeutic strategies. Such clinical problems arise for lymph-node negative primary breast cancer patients for whom it is well accepted that more than half of them are amenable to cure after the local-regional treatment alone (EBCTCG, 1998).

However, the classical proportional hazard semi-parametric Cox model, which does not explicitly modelize these two effects, is not suited for evaluating the association between prognostic factor and time-to-event in the presence of a heterogeneous clinical group with cured patients (Maller and Zhou, 1996). Thus, a selection process based on proportional hazard model-based statistics may lead to the discarding of genes whose expression changes reflect rapidly progressive disease in susceptible patients and as such should be considered valuable therapeutic targets.

For long-term disease-free survival analyses, semi-parametric cure models have been proposed that rely either on two-component mixture models or bounded cumulative hazard models (for a review, see Tsodikov *et al*., 2003). However, proposed methods for investigating prognostic factors in cure models from a frequentist or Bayesian framework usually require complex computations and are too cumbersome for practical use in genome-wide analysis. This latter problem prompted us to propose a simple and easy-to-use statistic tailored for identifying genes with disease progression effects which can also take into account other conventional prognostic markers. This statistic extends previous work on a cure model in the two-sample comparison setting (Broët *et al*., 2001).

The paper is organized as follows. In Section 2, we introduce the semi-parametric cure rate model that allows us to derive the proposed statistic for testing the lack of disease progression effect together with extensions for including additional independent

[*]To whom correspondence should be addressed.

variables. In Section 3, we present the results of simulation experiments. In Section 4, we illustrate the performance and usefulness of our approach using Affymetrix microarray data to identify gene expression changes associated with early relapse in a cohort of 130 early breast cancer patients. We conclude with a discussion of the new insights obtained from our approach.

2 TESTING FOR NO DISEASE PROGRESSION EFFECT

In the following, we introduce the semi-parametric cure model that allows us to derive a new statistic suited for testing for the lack of disease progression effect. We also propose extensions for taking into account clinical prognostic factors.

2.1 Cure model

Let G_{ij} denote the gene vector for the *i-th* subject ($i = 1, \ldots, n$) and the *j-th* gene ($j = 1, \ldots, p$). For each patient i, let the random variables T_i and C_i be the survival and censoring times which are assumed to satisfy the classical condition of independent censoring. We let $X_i = min(T_i,\ C_i)$ denote the observed time of follow-up, $\delta_i = 1_{\{X_i = T_i\}}$ the indicator of death, $Y_i(t) = 1_{\{t \leq X_i\}}$ the indicator of being at risk at time t. For each subject i and the gene j , the data consist of X_i, δ_i and G_{ij}. The hazard function of T_i corresponding to every subject i with gene vector G_{ij} is denoted by $\lambda(t/_{G_{ij}}) = f(t/_{G_{ij}})/S(t/_{G_{ij}})$, where $f(t/_{G_{ij}})$ and $S(t/_{G_{ij}})$ are the probability density function and the survival function, respectively. The corresponding cumulative hazard function is denoted by $\Lambda(t/_{G_{ij}}) = -\log[S(t/_{G_{ij}})]$.

Here, we introduce the following semi-parametric bounded cumulative hazard model (Broët *et al*., 2001; Tsodikov *et al*., 2003) which is defined by the general survival function:

$$S(t/_{G_{ij}}) = \exp[-\theta e^{\beta_{1j} G_{ij}}(1 - e^{-H(t)e^{\beta_{2j} G_{ij}}})], \tag{1}$$

where $H(t)$ is an arbitrary function increasing with time from zero to infinity, which can be considered as a pseudo-cumulative hazard function and θ is a positive parameter. Here, β_{1j} and β_{2j} are parameters, belonging to $\mathbb{R}$, for the cure and disease progression effects of the gene j, respectively. This model is a semi-parametric model since a parametric form is assumed only for the genes effects, the function $H(t)$ being treated non-parametrically. Moreover, it is a cure model since the function $S(t/_{G_{ij}})$ is improper with its limiting value $e^{-\theta e^{\beta_{1j} G_{ij}}}$ representing the probability of not experiencing the event of interest. The cumulative hazard $\Lambda(t/_{G_{ij}})$ is bounded, being $\leq \theta e^{\beta_{1j} G_{ij}}$. In this model, the parameter vector β_{1j} quantifies the genes' cure (or long-term) effect and the β_{2j} quantifies the genes' disease progression (or short-term) effect on the pseudo-survival function $e^{-H(t)}$ through a proportional hazard relationship. This model can be written in terms of the hazard functions $\lambda(t/_{G_{ij}})$ as follows:

$$\lambda(t/_{G_{ij}}) = \theta h(t) e^{\beta_{1j} G_{ij}} e^{\beta_{2j} G_{ij} - H(t) e^{\beta_{2j} G_{ij}}}, \tag{2}$$

where $h(t) = [\partial H(t)/\partial t]$ is an arbitrary baseline hazard function. As seen in (2), the cure effect acts in multiplying the hazard rate by a quantity which is constant over time whereas for the disease progression effect this quantity is changing over time. This latter time-varying effect is related to the changes in composition of the population since the susceptible patients group is progressively exhausted as time goes on.

2.2 Test statistic

2.2.1 Score statistic We derive a score statistic for testing the hypothesis ($H_{0j} : \beta_{2j} = 0$) of no disease progression effect for the *j-th* gene. Based on the previous model, the corresponding partial log-likelihood is

$$LL(\beta_{1j}, \beta_{2j}) = \sum_{i=1}^{n} \left[\begin{array}{l} \delta_i(\beta_{1_j} G_{ij} + \beta_{2_j} G_{ij} - H(t_i) e^{\beta_{2j} G_{ij}}) \\ -\ \ln\left(\sum_{k=1}^{n} Y_k(t_k) e^{\beta_{1j} G_{kj}} e^{\beta_{2j} G_{kj}} e^{-H(t_i) e^{\beta_{2j} G_{kj}}} \right) \end{array} \right].$$

When there are ties among the events, we consider the modified partial likelihood as proposed by Breslow (1974).

Thus, the components of the score vector deduced from the partial likelihood under H_{0j} can be written as follows:

$$\begin{aligned} \hat{V}_1^{H_{0j}} &= \frac{\partial LL(\beta_{1j}, \beta_{2j})}{\partial \beta_{1j}} \mid_{\beta_{2j}=0} = 0 \\ \hat{V}_2^{H_{0j}} &= \frac{\partial LL(\beta_{1j}, \beta_{2j})}{\partial \beta_{2j}} \mid_{\beta_{2j}=0} \\ &= \sum_{i=1}^{n} \delta_i \hat{\Phi}(t_i) \left\{ G_j - \frac{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj}}{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}}} \right\}, \end{aligned}$$

Where $\hat{\Phi}(t) = 1 - \hat{H}(t) = 1 + \ln[1 - (\hat{\Lambda}_0(t)/\hat{\theta})]$.

Here, $\hat{\Lambda}_0(t)$ is the left-continuous version of the Breslow's estimator (Breslow, 1972, 1974) for the cumulative hazard function under H_{0j}, $\hat{\theta}$ is its value computed at the last observed failure time and $\hat{\beta}_{1j}$ is the maximum partial likelihood estimator of β_{1j} under H_{0j}. In practice, $\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \delta_i [\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}}]^{-1}$.

The corresponding observed information matrix $\hat{I}_{H_{0j}}$ under H_{0j} is obtained from the second derivatives and is given as follows:

$$\frac{\partial^2 LL(\beta_{1j}, \beta_{2j})}{\partial^2 \beta_{1j}} = \sum_{i=1}^{n} \delta_i \left[\begin{array}{c} \left(\frac{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj}}{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}}} \right)^2 \\ - \frac{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj}^2}{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}}} \end{array} \right]$$

$$\frac{\partial^2 LL(\beta_{1j}, \beta_{2j})}{\partial^2 \beta_{2j}} = \sum_{i=1}^{n} \delta_i \left[\begin{array}{c} \left(\frac{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj} \hat{\Phi}(t_i)}{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}}} \right)^2 \\ - \frac{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj}^2 \hat{\Phi}^2(t_i)}{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}}} \end{array} \right]$$

$$\frac{\partial^2 LL(\beta_{1j}, \beta_{2j})}{\partial \beta_{1j} \partial \beta_{2j}} = \sum_{i=1}^{n} \delta_i \left[\begin{array}{c} \frac{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj} \hat{\Phi}(t_i) \sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj}^2 \hat{\Phi}^2(t_i)}{\left(\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} \right)^2} \\ - \frac{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}} G_{kj} \hat{\Phi}^2(t_i)}{\sum_{k=1}^{n} Y_k(t_k) e^{\hat{\beta}_{1j} G_{kj}}} \end{array} \right]$$

Under H_{0j}, the statistic of no disease progression effect $S_j^{\mathrm{DPE}} = (0, \hat{V}_2^{H_{0j}}) \hat{I}_{H_{0j}}^{-1} (0, \hat{V}_2^{H_{0j}})'$ is asymptotically distributed as a χ^2 with one degree of freedom (Cox and Hinkley, 1974).

2.2.2 Extension for taking into account conventional clinical prognostic factors In order to adjust for conventional clinical factors, we propose a simple strategy which depends on the existence of a disease progression effect for the clinical factor. For the following, denote Z the clinical covariate and Z^l its discretized counterpart (with L stratums; $l = 1, \ldots, L$).

First, we test for the hypothesis of no disease progression effect for the clinical covariate using the score statistic introduced above.

- If this latter hypothesis is rejected, we propose to consider a stratified (from Z^l) version of the cure rate model introduced in the previous section. Thus, the components of the statistic are found by summing the previous first and second derivatives across each stratum.
- If this latter hypothesis is not rejected, the following extended cure model is considered :

$$S(t/_{Z_i, G_{ij}}) = \exp\left[-\theta e^{\beta_0 Z_i} e^{\beta_{1j} G_{ij}} (1 - e^{-H(t) e^{\beta_{2j} G_{ij}}})\right], \qquad (3)$$

where Z_i is the clinical covariate for the *i-th* subject and β_0 the corresponding regression coefficient.

Thus, the components of the score vector for testing the null hypothesis of no disease progression effect for the *j-th* gene can be easily written as follows:

$$\hat{V}_1 = \frac{\partial LL(\beta_0, \beta_{1j}, \beta_{2j})}{\partial \beta_0} \mid_{\beta_{2j}=0} = 0$$

$$\hat{V}_2 = \frac{\partial LL(\beta_0, \beta_{1j}, \beta_{2j})}{\partial \beta_{1j}} \mid_{\beta_{2j}=0} = 0$$

$$\hat{V}_3 = \frac{\partial LL(\beta_0, \beta_{1j}, \beta_{2j})}{\partial \beta_{2j}} \mid_{\beta_{2j}=0}$$

$$= \sum_{i=1}^{n} \delta_i \hat{\Phi}(t_i) \left\{ \begin{array}{c} G_j - \\ \dfrac{\sum_{k=1}^{n} Y_k(t_i) e^{\hat{\beta}_0 Z_k} e^{\hat{\beta}_{1j} G_{kj}} G_{kj}}{\sum_{k=1}^{n} Y_k(t_i) e^{\hat{\beta}_0 Z_k} e^{\hat{\beta}_{1j} G_{kj}}} \end{array} \right\}$$

Here, $\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \delta_i [\sum_{k=1}^{n} Y_k(t_j) e^{\hat{\beta}_0 Z_k} e^{\hat{\beta}_{1j} G_{kj}}]^{-1}$ where $\hat{\beta}_0$ is the usual partial likelihood estimators of β_0 under the null hypothesis and $\hat{\theta}$ its value at the last observed failure time. The corresponding observed information matrix is obtained from the partial second derivatives in the same way as presented above. Thus, the corresponding statistic $S_j^{\mathrm{DPE}} = (0, 0, \hat{V}_3) \hat{I}_{H_{0j}}^{-1} (0, 0, \hat{V}_3)'$ is asymptotically distributed under the null hypothesis as a χ^2 with one degree of freedom.

3 SIMULATION

3.1 Method

A simulation study was performed to investigate the power properties of the proposed disease progression effect test statistic (denoted DPE) in comparison with the classical proportional hazard Cox model-based Wald test statistic (denoted PHM). Data were generated to mimic the disease progression and cure effects of a gene on the survival times according to the cure model described previously with $H(t) = t$. Censoring times were independently generated from a uniform distribution. For each gene, pseudo-expression values were independently sampled from a standard normal distribution. The number of subjects was chosen to be of 200. The following configurations were considered: plateau value ($e^{-\theta}$) of 50 and 75%; censoring of 0 and 25%; $\beta_1 = 0$, 0.25, 0.5 and $\beta_2 = 0$, 1, 1.25, 1.5, 2 that mimics realistic disease progression and cure effects. For each configuration, 200 replications were performed and the levels and powers of all tests were estimated at the nominal level 0.05.

Table 1. Simulation results for no censoring, cure fraction 50 and 75%

Cure effect (β1)	Disease progression effect (β2)					
	0	0.5	1	1.5	2	2.5
Cure fraction 50%						
0						
DPE	0.03	0.57	1.00	1.00	1.00	1.00
PHM	0.04	0.08	0.26	0.32	0.44	0.64
0.25						
DPE	0.03	0.63	1.00	1.00	1.00	1.00
PHM	0.75	0.82	0.99	0.98	0.97	0.99
0.5						
DPE	0.05	0.45	1.00	1.00	1.00	1.00
PHM	1.00	1.00	1.00	1.00	1.00	1.00
Cure fraction 75%						
0						
DPE	0.03	0.85	0.99	1.00	1.00	1.00
PHM	0.07	0.05	0.07	0.08	0.09	0.15
0.25						
DPE	0.06	0.83	1.00	1.00	1.00	1.00
PHM	0.47	0.55	0.54	0.62	0.61	0.72
0.5						
DPE	0.06	0.81	1.00	1.00	1.00	1.00
PHM	0.92	0.97	0.98	0.98	0.99	1.00

3.2 Results

Table 1 shows the simulation results for the uncensored case. As expected, the estimated level of the DPE test under its proper null hypothesis is within the binomial range [0.02–0.08]. In the presence of a disease progression effect without a cure effect, power gains of the proposed test are impressive as compared with the Cox-model-based Wald test. In the presence of a cure effect, power gains are lower but still interesting in comparison to the PHM test as soon as a non-negligible disease progression effect exists. It is worth noting that power gains increase with the plateau value. When the cure effect is important, the PHM test is more powerful than the DPE test. This is not surprising since the DPE statistic is devoted for detecting disease progression effect whereas the PHM statistic does not explicitly model these two effects. Moreover, if no disease progression exists, the cure model reduces to a proportional hazard model for which the PHM test is optimal.

Table 2 shows the simulation results for the censored case. With a 25% censoring rate, the observed levels of the proposed test statistic do not exceed the binomial bounds. Since the null hypothesis of no disease progression effect does not involve the plateau value estimate, it is not surprising that the DPE test maintains a correct type I error for censored cases. Concerning the power it appears that the trends observed in the uncensored case remain almost unchanged. Power gains for DPE are lower than in the uncensored case, but still remain impressive as compared with the PHM test as soon as no cure effect exists. When testing disease progression effect with a moderate cure effect, power values of the DPE and PHM tests are very close.

Table 2. Simulation results for 25% censoring, cure fraction 50 and 75%

Cure effect (β1)	Disease progression effect (β2)					
	0	0.5	1	1.5	2	2.5
Cure fraction 50%						
0						
DPE	0.05	0.56	1.00	1.00	1.00	1.00
PHM	0.05	0.09	0.25	0.33	0.47	0.68
0.25						
DPE	0.05	0.50	1.00	1.00	1.00	1.00
PHM	0.68	0.80	0.96	0.98	0.99	0.99
0.5						
DPE	0.05	0.42	1.00	1.00	1.00	1.00
PHM	1.00	1.00	1.00	1.00	1.00	1.00
Cure fraction 75%						
0						
DPE	0.07	0.66	1.00	1.00	1.00	1.00
PHM	0.08	0.08	0.25	0.53	0.66	0.65
0.25						
DPE	0.03	0.64	1.00	1.00	1.00	1.00
PHM	0.66	0.78	0.85	0.99	0.99	1.00
0.5						
DPE	0.06	0.40	1.00	1.00	1.00	1.00
PHM	1.00	1.00	1.00	1.00	1.00	1.00

4 DISEASE PROGRESSION EFFECTS' GENES IN EARLY BREAST CANCER

4.1 Clinical and microarray datasets

The data come from an expression microarray study conducted jointly between the Genome Institute of Singapore and the Karolinska Institute (Stockholm, Sweden) and designed for investigating the prognostic effects of gene expression changes on the outcome of patients with primary invasive breast cancer (Miller *et al*., 2005). Here, we selected a homogeneous clinical group of 130 patients having lymph-node-negative breast cancer with positive steroid receptor (either estrogen or progesterone receptors), tumor size <50 mm, age between 35 and 80 years. All patients had been treated by modified radical mastectomy or breast-conserving surgery, followed by radiotherapy if indicated. None of these patients received chemotherapy and only a small fraction (15%) received hormonal therapy. Among the selected cases, 86 patients (66%) had a tumor <20 mm (stage T1) and 44 patients (34%) had a tumor between 20 and 50 mm (stage T2). The mean age at diagnosis was 62 years. According to the Elston–Ellis grading system (Elston and Ellis, 1991), 48 patients (37%) had tumor grade I, 66 (51%) grade II and 16 (12%) grade III. Protein levels of estrogen receptor (ER) and progesterone receptor (PR) were assessed by immunoassay (monoclonal 6F11 anti-ER and monoclonal NCL-PGR, respectively, Novocastra Laboratories Ltd, Newcastle upon Tyne, UK) and deemed positive if greater than 0.1 fmol/µg DNA. One hundred fifteen (88%) patients had tumors with positive ER and 123 (95%) with positive PR.

The clinical outcome considered in this study was the occurrence of any relapse from the disease (i.e. local or regional relapse, metastasis or disease-related death). Disease-free survival was calculated from the date of treatment to the time of relapse from the disease or last follow-up.

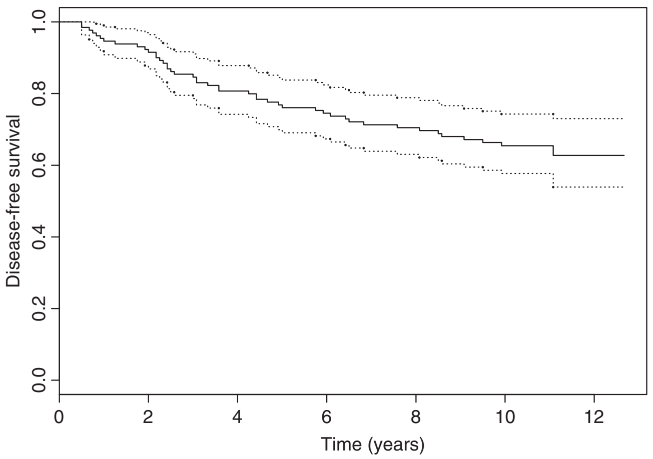

Fig. 1. Kaplan–Meier estimate of the disease-free survival. Dashed lines show the 95% confidence intervals.

For gene expression analyses, the Affymetrix Human U133 oligonucleotide arrays were used. Here, we considered U133A Chips with 22283 probe sets. Standardization and normalization of the data were carried out using the MAS5 procedure (Simon *et al*., 2004).

In our study, the median duration follow-up was 10.7 years. The five year disease-free survival was 76.1% [95%CI: 69.1–83.8] and the 10 year disease-free survival was 65.5% [95%CI: 57.7–74.3]. At the end of follow-up, 45 patients experienced a relapse from the disease. Figure 1 displays the Kaplan–Meier estimates of the disease-free survival (with the 95% confidence interval) for the entire cohort and shows a clear plateau value after ten years of follow-up.

From classical univariate Cox survival analysis, age was not significantly associated with the disease-free survival ($p = 0.61$), whereas high tumor size staging ($p = 0.005$) and histological grading ($p = 0.01$) were significantly associated with lower disease-free survival. Tumor size staging and histological grading were highly correlated ($p = 0.002$). When adjusting for these two factors in a multivariate Cox model, only tumor size staging showed a significant effect on the disease-free survival ($p = 0.02$). When we tested for a disease progression effect, the proposed test showed no statistical significance for the tumor size staging ($p = 0.3$). Thus, we decided to consider the statistic (denoted in the following S_j^{DPE}) derived from the extended cure model [Equation (3)] introduced in Section 2. We also calculated the corresponding Cox model-based Wald statistic adjusted for tumor size (denoted in the following S_j^{Cox}) and the corresponding p-values denoted $p^{S_j^{\mathrm{DPE}}}$, and $p^{S_j^{\mathrm{cox}}}$, respectively. The error criteria considered for the selection process was the classical false discovery rate (FDR) as introduced by Benjamini and Hochberg (1995). We estimated the FDR from the marginal distribution of the p-values without making any assumption on the distribution related to the modified genes according to the method proposed by Dalmasso *et al*. (2005).

4.2 Results

4.2.1 Results of the selection process Here we consider a typical situation where the investigator is interested in obtaining a list of top

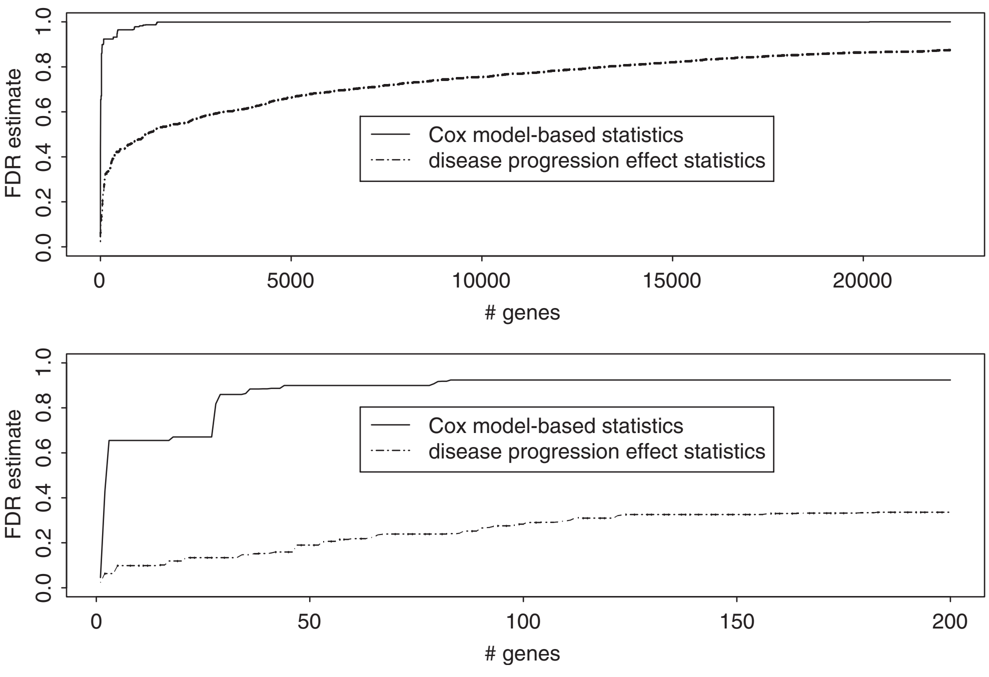

Fig. 2. FDR estimate as a function of the number (#) of probe sets (top gene selection strategy). Panel 1 corresponds to the entire set of probe sets. Panel 2 displays a zoom for the first 200 probe sets (panel 2).

probe sets for a defined FDR threshold based on the ordered *p*-values. Figure 2 displays the FDR estimate as a function of the number of probe sets selected. Panel 1 gives the results for all the probe sets whereas panel 2 displays a zoom on the first top 200 probe sets. As seen from these graphics, for a same FDR threshold, we can select a larger number of genes from our proposed statistic as compared with the Cox model-based statistics. When looking to the second panel, the FDR curve for the Cox model-based selection process shows a sharp increase with a value of 65% for the third probe set whereas for the proposed selection process the curve is slowly increasing. Choosing a classical 20% FDR cut-off for the statistical significance gives us a list of 52 probe sets with the proposed statistic and one probe set with the Cox model-based statistic. It is worth noting that the smallest observed *p*-value for our proposed statistic is 1.2×10^{-6} which is still significant when considering a restrictive criteria such as the familywise error rate (at a level of 5%) and using the Bonferroni procedure.

Figure 3 displays the survival curves for two representative probe sets (among the 52 selected ones) where gene expression measurements are dichotomized between those with high values (above the median) and those with low values (below the median). These figures show clearly the probe sets' time-varying effect with the two curves converging to a plateau value as time goes on. In the first case, an increase of the probe set expression is related to early relapse whereas for the other probe set it is the converse.

Among these 52 probe sets, one gene was selected with its three probe sets and one gene with its two probe sets leading to a subset of 49 different genes (Table 3).

Figure 4 displays results of a local smoothing procedure (Cleveland, 1979) of the time-dependent regression coefficient estimate, based on the Cox model-based Schoenfeld residuals, versus time (Marubini and Valsecchi, 2004). It clearly shows that genes' coefficients are not constant with their signs changing over time. As expected, testing for non-proportionality of the hazards leads to highly significant results for these genes (data not shown).

4.2.2 Disease progression effect principal component In order to explore the combined effect of the selected 49 disease progression effect genes on patient outcomes in a low-dimensional space, we performed a principal component analysis on the variance-covariance genes matrix and selected the largest principal component. This component, which may be viewed as a *super gene*, corresponds to the linear combination of the selected genes having maximal variation among tumor samples. We then calculated for each patient a corresponding *super gene* score and tested for a disease progression effect of the *super gene*. We observed a highly significant disease progression effect ($\chi^2 = 64.3$, $p < 10^{-15}$). Figure 5 displays the survival curves obtained when the *super gene* scores were dichotomized according to the median score. As seen from the graph, for patients with low *super gene* scores most of the relapse events occurred before five years, whereas for the other group relapses occurred after five years, despite the two groups having the same proportion of cured patients.

4.2.3 Biological insights of the disease progression effect genes For classifying selected genes by categories, we used the publicly available PANTHER (Protein Analysis Through Evolutionary Relationships, Version 6.0 2005, http://panther.appliedbiosystems.com) classification system software (Mi *et al.*, 2005). It includes interactive resources for analyzing gene expression data in relation to molecular functions, biological processes (GO ontology) and known pathways. In our selected subset of 49 disease progression effect genes, 46 were annotated. We

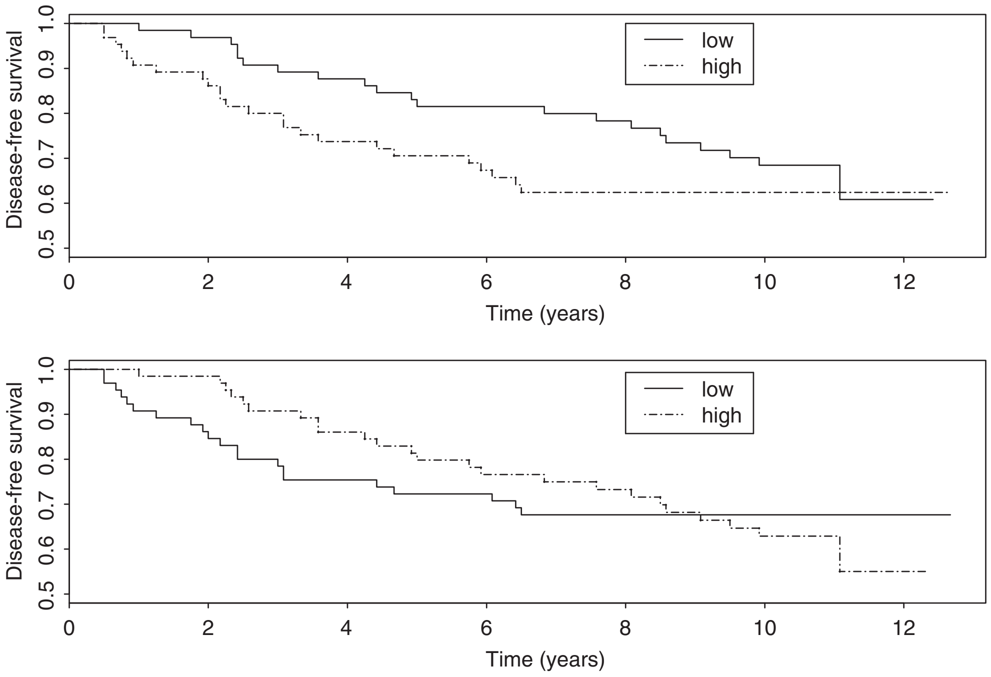

Fig. 3. Kaplan–Meier estimate of the disease-free survival according to the level of gene expression (below the median/above the median) for 2 representative probe sets (GGA1 for panel 1 and TXNIP for panel 2) having a disease progression effect.

investigated which categories (biological process, molecular function, pathway), if any, were statistically overrepresented in this 46 genes subset. We compared the number of genes observed in a specified category with the number that would be expected (based on the 23 481 annotated genes ID from the NCBI repertory) if there was no relationship between our selected subset and the specified category. We considered 243, 255 and 82 categories for biological process, molecular function and pathway, respectively. Table 4 displays the categories which are statistically overrepresented in our selected genes subset (at the 0.05 level).

When looking to the biological process categories, apoptosis, oncogene, mRNA transcription and cell cycle genes were overrepresented. For the molecular function categories, genes having a protein kinase activity were overrepresented. For the pathway categories, we found an overrepresentation of genes having immune/inflammation functions, reflecting the well-known role of the microenvironment in cancer disease progression. Interestingly, we also showed an overrepresentation of genes related to the *Wnt* signaling pathway which is known to be implicated in oncogenesis of a wide range of human cancers including breast carcinomas (Howe and Brown, 2004).

Among the 49 disease progression effect genes, a high expression of 38 genes was related to early relapse whereas for 11 genes a low expression was related to early relapse. We also investigated if chromosome locations were statistically overrepresented in this 49 gene subset. We compared the number of genes we observed in a specified chromosome with the number that would be expected if there was no relationship between our selected subset and the chromosomal location. Here, chromosome 3 was significantly overrepresented ($p < 10^{-6}$) with 13 genes located on this chromosome. Moreover, it is worth noting that for the nine genes located on the 3*q* arm, an increase of the expression was related to early relapse.

4.2.4 Validation study For validating the disease progression effect of our selected subset of probe sets, we considered an independent breast cancer dataset from the study published by Wang *et al*. (2005). In this latter study, gene expression of 286 lymph-node-negative primary breast cancers was studied using Affymetrix Human U133 oligonucleotide arrays. The authors identified a gene signature of 60 probe sets for patients positive for ERs that is a prognostic factor for the development of metastasis.

Firstly, we tested the null hypothesis of no disease progression effect for our 52 selected probe sets on the 209 ER positive breast cancers from the Wang *et al*. (2005) study. Secondly, we tested the null hypothesis of no disease progression effect for the 60 probe sets selected by Wang *et al*. (2005) on our series (using both DPE and PHM statistics).

For these two groups of selected probe sets, we then compared the number of probe set statistics being significant at the classical 5% level with the number that would be expected if there was no relationship between the expression of the probe sets and the disease-free survival.

Among our 52 selected probe sets, 12 (23%) showed a disease progression effect in the Wang *et al*. (2005) series, this number being significantly higher than expected by chance alone ($p < 10^{-6}$). Among the 60 probes sets selected by Wang *et al*. (2005) four (6.7%) and six (10%) showed a relationship with the disease-free survival in our series, using DPE and PHM test statistics, respectively. When comparing these latter results with the numbers that would be expected if there was no relationship, we did not reach statistical significance with either the DPE ($p = 0.55$) or PHM ($p = 0.07$) test statistics.

Table 3. List of the selected 52 disease progression effect probe sets

AffyID	Gene symbol	Unigene name	Cytoband
Increased risk			
218114_at	GGA1	Golgi associated, gamma adaptin ear containing, ARF binding protein 1	22q13.31
222141_at	KELCHL	Kelch-like 22 (Drosophila)	22q11.21
222245_s_at	FER1L4	Fer-1-like 4 (*Caenorhabditis elegans*)	20q11.22
205930_at	GTF2E1	General transcription factor IIE, polypeptide 1, alpha 56 kDa	3q21–q24
209832_s_at	CDT1	DNA replication factor (Interim)	16q24.3
209361_s_at	PCBP4	Poly(rC) binding-protein 4	3p21
205977_s_at	EPHA1	EPH receptor A1	7q34
204094_s_at	KIAA0669	TSC22 domain family, member 2	3q25.1
213518_at	PRKCI	Protein kinase C, iota	3q26.3
200884_at	CKB	Creatine kinase, brain	14q32
218066_at	SLC12A7	Solute carrier family 12 (potassium/chloride transporters), member 7	5p15
211756_at	PTHLH	Parathyroid hormone-like hormone	12p12.1–p11.2
207493_x_at	SSX2	Synovial sarcoma, X breakpoint 2	Xp11.23–p11.22
201637_s_at	FXR1	Fragile X mental retardation, autosomal homolog 1	3q28
204701_s_at	STOML1	Stomatin (EPB72)-like 1	15q24-q25
208877_at	PAK2	P21 (CDKN1A)-activated kinase 2	3q29
210556_at	NFATC3	Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3	16q22.2
212782_x_at	POLR2J	Polymerase (RNA) II (DNA directed) polypeptide J, 13.3 kDa	7q11.2
206152_at	CENTG1	centaurin, gamma 1	12q14.1
210983_s_at	MCM7	MCM7 minichromosome maintenance deficient 7 (*Saccharomyces cerevisiae*)	7q21.3–q22.1
203486_s_at	ARMC8	Armadillo repeat containing 8	3q22.3
217926_at	HSPC023	HSPC023 protein	19p13.13
221219_s_at	KLHDC4	Kelch domain containing 4	16q24.3
204995_at	CDK5R1	Cyclin-dependent kinase 5, regulatory subunit 1 (p35)	17q11.2
47105_at	FLJ20399	Dihydrouridine synthase 2-like (SMM1, *S.cerevisiae*)	16q22.1
213420_at	DHX57	DEAH (Asp-Glu-Ala-Asp/His) box polypeptide 57	2p22.1
201908_at	DVL3	Dishevelled, dsh homolog 3 (Drosophila)	3q27
220388_at	FER1L4	Fer-1-like 4 (*C.elegans*)	20q11.22
221877_at	CDNA	FLJ38849 fis, clone MESAN2008936	19q13.31
216492_at	FLJ00060	hypothetical gene FLJ00060	19q13.42
219178_at	QTRTD1	Queuine tRNA-ribosyltransferase domain containing 1	3q13.31
201517_at	NCBP2	Nuclear cap binding-protein subunit 2, 20 kDa	3q29
210981_s_at	GRK6	G protein-coupled receptor kinase 6	5q35
204922_at	FLJ22531	hypothetical protein FLJ22531 (Interim)	11q13.2
201074_at	SMARCC1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	3p23–p21
215093_at	NSDHL	NAD(P) dependent steroid dehydrogenase-like	Xq28
212450_at	KIAA0256	KIAA0256 gene product (Interim)	15q21.1
215722_s_at	SNRPA1	small nuclear ribonucleoprotein polypeptide A'	15q26.3
209430_at	BTAF1	BTAF1 RNA polymerase II, B-TFIID transcription factor-associated	10q22–q23
Decreased risk			
214162_at	LOC284244	Hypothetical protein LOC284244	18q12.1
201008_s_at	TXNIP	Thioredoxin interacting protein	1q21.1
201009_s_at	TXNIP	Thioredoxin interacting protein	1q21.1
201010_s_at	TXNIP	Thioredoxin interacting protein	1q21.1
201041_s_at	DUSP1	Dual specificity phosphatase 1	5q34
212124_at	RAI17	Retinoic acid induced 17	10q22.3
219440_at	RAI2	Retinoic acid induced 2	Xp22
210832_x_at	PTGER3	Prostaglandin E receptor 3 (subtype EP3)	1p31.2
209189_at	FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	14q24.3
206286_s_at	TDGF1	Teratocarcinoma-derived growth factor 1	3p21.31
210218_s_at	SP100	Nuclear antigen Sp100	2q37.1
219689_at	LOC56920	Sema domain, immunoglobulin domain (Ig), short basic domain	3p21.1
203675_at	NUCB2	nucleobindin 2	11p15.1–p14

AffyID, Affymetrix identification code for each probe set; risk, risk of early relapse.

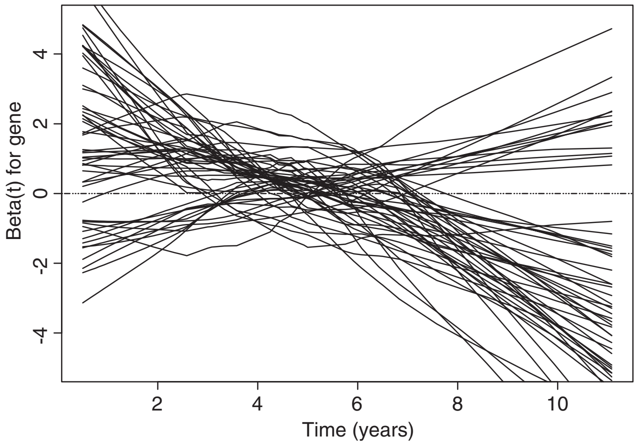

Fig. 4. Estimate of the time-dependent regression coefficient for the selected probe sets (Beta(t) for gene) versus time. If the proportional hazards assumption is true, Beta(t) will be a horizontal line. The dot-dashed horizontal line represents the value zero.

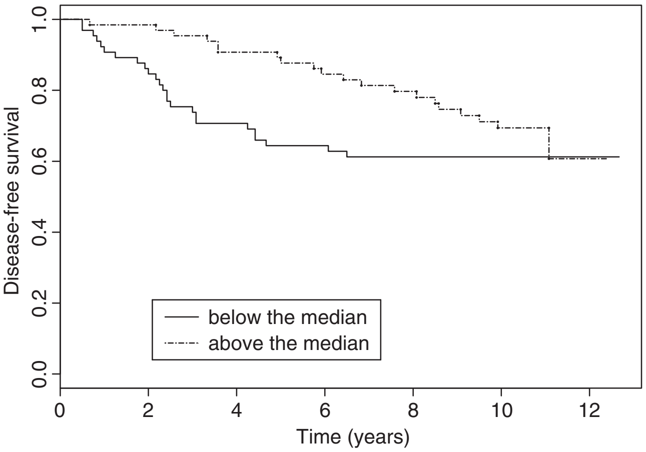

Fig. 5. Kaplan–Meier estimate of the disease-free survival according to the level of *super gene score* (below/above the median score). The *super gene* corresponds to the linear combination of the selected genes having maximal variation among our samples.

Table 4. Comparison between the number (#) of selected genes for a specified category (biological process, molecular function, pathway obtained from PANTHER software) with the number that would be expected if there was no relationship between these genes and the specified category, together with its corresponding *p*-values.

	# Genes NCBI	# observed	# expected	*p*-value
Biological Process				
Nucleoside, nucleotide and nucleic acid metabolism	3010	16	6.15	0.000203
Protein modification	1110	8	2.27	0.00172
Protein phosphorylation	644	6	1.32	0.00194
tRNA metabolism	42	2	0.09	0.00342
Mesoderm development	537	5	1.1	0.00473
mRNA transcription	1711	9	3.5	0.00719
Cell cycle control	381	4	0.78	0.00764
Pre-mRNA processing	269	3	0.55	0.0177
Oncogene	107	2	0.22	0.0204
Skeletal development	120	2	0.25	0.0252
Lactation, mammary development	13	1	0.03	0.0262
Cell cycle	886	5	1.81	0.0342
Induction of apoptosis	151	2	0.31	0.0384
Molecular function				
Nucleic acid binding	2534	14	5.18	0.000393
Helicase	161	3	0.33	0.00443
Annexin	63	2	0.13	0.00748
DNA helicase	67	2	0.14	0.00842
Kinase	658	5	1.35	0.0109
Protein kinase	512	4	1.05	0.0205
Other transcription factor	323	3	0.66	0.0284
Calmodulin related protein	149	2	0.3	0.0375
Other RNA-binding protein	157	2	0.32	0.0412
Kinase modulator	158	2	0.32	0.0416
Pathways				
Wnt signaling pathway	357	5	0.73	0.000807
Angiogenesis	218	4	0.45	0.00104
T cell activation	121	3	0.25	0.00199
Inflammation mediated by chemokine and cytokine signaling pathway	332	4	0.68	0.00474
B cell activation	102	2	0.21	0.0186
JAK/STAT signaling pathway	22	1	0.04	0.044

#, number; *p*-value, degree of significance.

5 DISCUSSION

The discovery of disease progression genes characterizing early and late relapse among uncured patients, which can only be accomplished by investigating survival data with long-term follow-up, advocates for the use of new statistics appropriate for identifying such genes. We propose in this paper a new statistic tailored for detecting disease progression effect genes which offers the investigator a powerful new and easy-to-use tool for the gene selection process. Furthermore, this statistic can be easily implemented using classical statistical software with survival analysis capabilities.

Our test statistic stems from biological, clinical as well as statistical considerations. From a biological point of view, it is likely that a non-negligible fraction of genes measured at the time of the treatment is direct or indirect witness of the speed of the disease for the uncured patients. From a clinical perspective, the identification of genes that drive early relapse may not only improve patient prognosis, but also guide the discovery of new potential therapeutic targets appropriate for patients with rapidly progressive disease. From a statistical point of view, in such a mixed population (cured and uncured patients), the susceptible patients group which is progressively exhausted over time, leads to an observed time-varying effect, which advocates the use of cure rate models from which well-suited statistics can be derived.

As seen from the simulation study, the proposed statistic shows excellent power performances for assessing a disease progression effect as compared with the classical Wald statistic derived from the Cox model. Power gains are impressive for no, or small differences in the cure effect. In any case, our proposed score test maintains a correct type I error.

For the early breast cancer series considered in this work, a long-term survivor fraction exists, and having a large number of patients followed up to the first decade post surgery allows for an interpretable time sequence for tumor relapse. Based on the results, the proposed statistic leads to the selection of a larger subset of genes for a reasonable FDR as compared with the Cox model-based statistic. When looking to our selected genes, they exhibit a clear time-varying effect which explains why they are not selected using the classical Cox-based statistic. Moreover, we could easily understand that early evaluation (say at five years) may emphasize differences that will disappear with longer follow-up. This latter fact may explain recent findings regarding gene signatures for early breast cancer, where around two-thirds of patients have a negligible risk of tumor recurrence after 10-years (Bland and Copeland, 1998) and may be considered as cured (or long-term survivors). Recently, van't Veer *et al.* (2002) have identified a microarray-derived gene expression signature that predicts for distant metastasis. This 70-gene signature was derived from the probability of being free of metastasis at five years and later evaluated on the time-to-distant relapse with a longer follow-up (van de Vijver *et al.*, 2002). In this latter work, the authors reported that the hazard ratio for distant metastasis as a first event was estimated to be 8.8 (95%CI : 3.8–20) between the 'poor' versus the 'good' profile groups for the first five years and only 1.8 (95%CI : 0.69–4.5) after five years. Thus, we may hypothesize that this time-varying effect reflects the presence of disease progression effect genes in the signature.

In our study, we considered a classical *top-gene* selection strategy (based on the FDR criteria) even though *key genes* are not necessarily those with larger transcriptional variations. We also validated the prognostic potential of our selected subset of genes on an independent dataset published by Wang *et al.* (2005). Adjusting for tumor size, this latter variable being the classical clinical reflection of cell proliferation, leads us to explore different biological pathways involved in rapid progressive disease. Here, our study emphasizes the potential interest of genes involved in the *Wnt* signaling pathway (Howe and Brown, 2004).

Of particular interest is the overrepresentation of genes located in chromosome 3 and especially on the 3q arm. This finding is likely related to genomic amplification since it is consistent with recent comparative genomic hybridization results which show that gain of 3q is a strong predictor of recurrence in lymph node-negative invasive breast carcinomas (Janssen *et al.*, 2003). Notwithstanding that it was not the main purpose of our work, we investigated the interest of combining disease progression effect genes in a unique *super gene* component. As seen from our results, it clearly leads to a more powerful prognostic factor and thus warrants further investigations for prediction purposes.

In this work, we considered the same proportional hazard disease progression shape for each gene, however other disease progression effect shapes (e.g. accelerated life model) may also be considered and will require future exploration. We conclude that the proposed statistic is a powerful new approach for identifying genes with disease progression effects which could be valuable prognostic indicators useful in therapeutic decision making and for identifying candidate genes and pathways for future targeted therapies. Finally, this study emphasizes the need for deriving new statistics for genome-wide analysis where gains of power are a crucial issue.

ACKNOWLEDGEMENTS

The first author was supported by the Genome Institute of Singapore Sabbatical Program. We especially thank Sigrid Klaar, Hans Nordgren and Johanna Smeds of the Karolinska Institute for the clinical annotation and processing of the tumor specimens for microarray analysis.

Conflict of Interest: none declared.

REFERENCES

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B*, **57**, 289–300.

Bland,K.I. and Copeland,E.M. (1998) *The Breast: Comprehensive Management of Benign and Maligant Diseases*. 2nd edn.. Saunders, Philadelphia.

Breslow,N.E. (1972) Contribution to the discussion on the paper by D.R. Cox Regression and life tables. *J. R. Statist. Soc., Ser. B*, **34**, 216–217.

Breslow,N.E. (1974) Covariance analysis of censored survival data. *Biometrics*, **30**, 89–99.

Broët,P. *et al.* (2001) A semi-parametric approach for the two-sample comparison of survival times with long-term survivors. *Biometrics*, **57**, 844–852.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Cox,D.R. (1972) Regression models and life tables (with Discussion). *J. R. Statist. Soc. Ser. B*, **34**, 187–220.

Cox,D.R. (1975) Partial likelihood. *Biometrics*, **62**, 269–276.

Cox,D.R. and Hinkley,D. (1974) *Theoretical Statistics*. Chapman and Hall, London.

Dalmasso,C. *et al.* (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics*, **21**, 660–668.

EBCTCG: Early Breast Cancer Trialists' Collaborative Group. (1998), Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet*, **352**, 930–942.

Elston,C. and Ellis,I. (1991) Pathological prognostic factors in breast cancer. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, **19**, 403–410.

Howe,L.R. and Brown,A.M. (2004) Wnt signaling and breast cancer. *Cancer Biol. Ther.*, **3**, 36–41.

Janssen,E.A. *et al.* (2003) In lymph node-negative invasive breast carcinomas, specific chromosomal aberrations are strongly associated with high mitotic activity and predict outcome more accurately than grade, tumour diameter, and oestrogen receptor. *J. Pathol.*, **201**, 555–561.

Maller,R. and Zhou,X. (1996) *Survival Analysis with Long-Term Survivors*. John Wiley, New York.

Marubini,E. and Valsecchi,M.G. (2004) *Analysing Survival Data from Clinical Trials and Observation Studies*. John Wiley, New York.

Mi,H. *et al.* (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **33**, 284–288.

Miller,L.D. *et al.* (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA*, **102**, 13550–13555.

Simon,R., Korn,E.L., McShane,L.M., Radmacher,M.D., Wright,G. and Zhao,Y. (2004) *Design and Analysis of DNA Microarray Investigations*. Springer-Verlag.

Tsodikov,A.D. *et al.* (2003) Estimating cure rates from survival data: an alternative to two-component mixture models. *J. Am. Stat. Assoc.*, **98**, 1063–1068.

van t Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.