BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 9 2013, pages 1149–1157
doi:10.1093/bioinformatics/btt124

Gene expression

Advance Access publication March 14, 2013

Modeling risk stratification in human cancer

Thierry Rème[1,*], Dirk Hose[2], Charles Theillet[3] and Bernard Klein[1]

[1]INSERM-UM1, U1040, Institut de Recherche en Biothérapie, 34295 Montpellier, France, [2]Medizinische Klinik V, Universitätsklinikum Heidelberg, Nationales Centrum für Tumorerkrankungen, 69120 Heidelberg, Germany and [3]INSERM, U896, Institut de Recherche en Cancérologie de Montpellier, 34295 Montpellier, France

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Despite huge prognostic promises, gene expression-based survival assessment is rarely used in clinical routine. Main reasons include difficulties in performing and reporting analyses and restriction in most methods to one high-risk group with the vast majority of patients being unassessed. The present study aims at limiting these difficulties by (i) mathematically defining the number of risk groups without any a priori assumption; (ii) computing the risk of an independent cohort by considering each patient as a new patient incorporated to the validation cohort and (iii) providing an open-access Web site to freely compute risk for every new patient.

Results: Using the gene expression profiles of 551 patients with multiple myeloma, 602 with breast-cancer and 460 with glioma, we developed a model combining running log-rank tests under controlled chi-square conditions and multiple testing corrections to build a risk score and a classification algorithm using simultaneous global and between-group log-rank chi-square maximization. For each cancer entity, we provide a statistically significant three-group risk prediction model, which is corroborated with publicly available validation cohorts.

Conclusion: In constraining between-group significances, the risk score compares favorably with previous risk classifications.

Availability: Risk assessment is freely available on the Web at https://gliserv.montp.inserm.fr/PrognoWeb/ for personal or test data files. Web site implementation in Perl, R and Apache.

Contact: thierry.reme@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 28, 2012; revised on February 13, 2013; accepted on March 7, 2013

1 INTRODUCTION

Prediction of cancer patients' survival is one of the most challenging tasks in daily practice in oncology. Gene expression profiling (GEP) has been used in cancer to look for cancer heterogeneity and predict for corresponding classification outcome. While the main clinical prognostic factor of multiple myeloma (MM) is the International Staging System (ISS) based on serum β2-microglobulin and serum albumin (Greipp *et al.*, 2005), two high-risk scores have combined GEP with overall survival (OS) as input data and outcome issue (Decaux *et al.*, 2008; Shaughnessy *et al.*, 2007). Several other GEP-based signatures using either tumor progression or cell cycle-associated indexes at diagnosis or relapse have been published (Chng *et al.*, 2008; Hose *et al.*, 2010; Kuiper *et al.*, 2012; Mulligan *et al.*, 2007). Gene expression microarrays were early used to identify breast carcinoma subtypes (Sørlie *et al.*, 2001) and outcome (van 't Veer *et al.*, 2002), intrinsic molecular groups (Guedj *et al.*, 2012; Kao *et al.*, 2011), recurrence, metastasis and survival (Reis-Filho and Pusztai, 2011). Gene expression-based profiling of malignant gliomas has been recognized to produce a more robust classification than the conventional histological diagnosis (Gravendeel *et al.*, 2009; Li *et al.*, 2009; Nutt *et al.*, 2003; Shirahata *et al.*, 2007) and also to directly predict for survival (Freije *et al.*, 2004; Petalidis *et al.*, 2008).

Despite a huge prognostic promise in a wide range of malignant diseases, GEP is rarely used in clinical routine to assess outcome. Main reasons include difficulties in performing the analysis and reporting of the data in clinical practice, inability of direct application of defined groups and frequently arbitrary assumptions in expression-based scores, e.g. risk-group size and number. Most GEP-based risk scores in MM delineate a group of patients with adverse prognosis, but using arbitrary quartile cutoffs on gene expression or index calculation. Most reports of breast cancer survival prediction still consider only two risk classes using semi-supervised predictive methods (Bair and Tibshirani, 2004; Naderi *et al.*, 2007).

We present here a novel survival prediction model (termed RS for risk score) built on gene expression profiles of patients suffering from MM ($n = 551$), breast cancer ($n = 602$) or glioma ($n = 460$). Prognostic genes were selected without arbitrary cutoffs. A score summing up the risk information of these prognostic genes identified the maximum number of patients' groups showing a statistically significant different OS, without a priori assumption. The RS score allowed successfully classification within three groups of two independent cohorts of patients with MM, breast cancer or glioma, and correlated with classification features like conventional prognostic factors, molecular subtypes or pathological grades in both training and validation cohorts. Finally, we provide an open-access Web site to compute risk (RS) for any novel patient.

2 METHODS

Computations were performed using R (http://www.R-project.org) and Bioconductor (Gentleman *et al.*, 2004).

*To whom correspondence should be addressed.

2.1 GEP of tumor samples

For MM, a training cohort (termed HM cohort) of 206 previously untreated patients with MM from the University Hospitals of Heidelberg (Germany) and Montpellier (France) was used. Patient characteristics have been published previously (Hose *et al.*, 2009; Sprynski *et al.*, 2009). Affymetrix U133 Plus 2.0 GEP is available at ArrayExpress under n° E-MTAB-362 for HM cohort and E-MTAB-363 for RHM cohort (similar to HM with novel agents in induction phase) as purified primary malignant plasma cells, n° E-MEXP-2360 for healthy bone marrow plasma cells and E-TABM-937 for human myeloma cell lines.

A cohort (TT2 cohort) of 345 previously untreated patients with MM treated with the Total Therapy 2 protocol (Barlogie *et al.*, 2006) at the University of Arkansas for Medical Sciences (UAMS, Little Rock, AR, USA) was used for validation. Affymetrix U133 Plus 2.0 GEP of the TT2 cohort and the TT3 cohort including novel agents in induction treatment are available at Gene Expression Omnibus (GEO) under n° GSE24080. For breast cancer, a training cohort of 285 patients whose survival data were kindly provided by the Ligue National Contre le Cancer as part of the Cartes d'Identité des Tumeurs program (http://cit.ligue-cancer.net/) was used (CIT cohort). The corresponding molecular taxonomy is reported (Guedj *et al.*, 2012), and Affymetrix U133 Plus 2.0 GEP is available as E-MTAB-365 ArrayExpress dataset. A second breast cancer cohort of 317 patients (FOO cohort) with published molecular subtyping (Kao *et al.*, 2011) was used for validation. Affymetrix U133 Plus 2.0 GEP is available at GEO (GSE20685).

For glioma, a training cohort (NL cohort) of 272 outcome-documented patients with reported histological staging (Gravendeel *et al.*, 2009) was used. Affymetrix U133 Plus 2.0 GEP is available at GEO database (GSE16011). The validation cohort (NIH cohort) was obtained from the Rembrandt database (https://caintegrator.nci.nih.gov/rembrandt/) and includes 188 clinically documented patients of 507 patients published on GEO under n°GSE4290 using Affymetrix U133 Plus 2.0 chips. The clinical characteristics of patients with MM, breast cancer and glioma are depicted in Supplementary Table S1.

2.2 Normalization of training datasets

For each cancer type, the training samples were normalized together using the GCRMA preprocessing (Wu *et al.*, 2004). Recorded preprocessing parameters were then used to normalize each individual sample of the validation cohort of a given cancer type. The training and validation samples were then adjusted together for batch effect using the ComBat Empirical Bayes method (Johnson *et al.*, 2007).

2.3 Prognostic genes selection and score calculation in training datasets

2.3.1 Selection of prognostic genes When applied to our myeloma training set, the stringency of methods using maximally selected rank statistics ['maxstat', (Hothorn and Lausen, 2003)] led to a short and unusable probe set list at a false discovery rate level of 5% (Benjamini and Hochberg, 1995). Probe sets were therefore selected for prognostic significance using a two-step multiple testing correction as follows:

(1) About half the 54 675 initial probe sets was filtered out using the 'nsFilter' function of the 'genefilter' R-package, eliminating Affymetrix arrays control probes, probes without Entrez Gene identifiers and the less variable genes across samples by setting the variance cutoff at 0.15, approximately corresponding to a 65% quantile (Supplementary Fig. S1A).

(2) For each conserved probe set, the training samples were ordered from lowest to highest expression signals. A 'running' log-rank test was performed using each of these ordered signals as a cutoff point to assess the OS difference between patient groups with a lower or equal versus a higher signal (Supplementary Fig. S1B). The number of expected events per group, i.e. the number of deaths distributed in groups in the case of the null hypothesis, allowed determining the prognostic trend. A gene is of poor prognosis if the observed number of deaths in the high-expression group is higher than the expected one (Supplementary Fig. S2A). Conversely, it is considered of good prognosis if expected deaths exceed observed ones (Supplementary Fig. S2B). The minimal number of expected events per group was set to 2, as recommended for chi-square calculations (Conover, 1999; Fisher, 1990), eliminating log-rank tests on too small groups. All calculated positions of the expression scale result in a minimal log-rank *P*-value adjusted according to Benjamini and Yekutieli (Benjamini and Yekutieli, 2001) for multiple testing under dependency by dividing by the position number.

(3) The thousands of adjusted minimal log-rank *P*-values were then ordered for a second step of multiple testing correction according to their rank (Benjamini and Hochberg, 1995). The level of false discovery rate was set to a maximum level of 5%. Using this double multiple testing correction, 19 risk probe sets were selected with a 0.05 adjusted *P*-value for patients with MM, 68 probe sets with a 0.01 adjusted *P*-value for patients with breast cancer and 39 probe sets with almost no error (adjusted *P*-value <1.0E-20) for patients with glioma.

2.3.2 Calculation of the scorers A continuous RS 's' was calculated for each cohort sample as the difference between the sum of risk (poor prognosis) and the sum of protective (good prognosis) GCRMA-normalized expression signals of the probe sets.

2.4 Building risk groups in training datasets

Patients were ordered according to increasing s score $s = (s_1, s_2, \ldots, s_k, \ldots, s_n)$ with $s_k \leq s_{k+1}$. Two cutoff points s_i and s_j, with $i \in [1, n-2]$ and $j \in [i+1, n-1]$, were moved together in all possible score values.

The total number of possible cutoff pairs is $t = \frac{(n-1)(n-2)}{2}$.

s_i and s_j create three patient groups RS1, RS2 and RS3 (highest *s*). The global X_{ij} log-rank chi-square considering all groups, the between-consecutive group X_{ij}^{12} and X_{ij}^{23} log-rank chi-squares (excluding, respectively, the group RS3 then group RS1 patients) and the calculated number of deaths per group c_{ij}^1, c_{ij}^2 and c_{ij}^3 in the case of the null hypothesis were simultaneously recorded. The *P*-values of the *t* global chi-squares were adjusted according to their ranks (Benjamini and Hochberg, 1995). Possible s_i, s_j cutoff pairs were the score pairs for which simultaneously (i) the adjusted global chi-square was significant at a 5% risk, (ii) the between-consecutive group chi-squares were both significant at a 5% risk and (iii) the expected events in all groups c_{ij}^1, c_{ij}^2 and c_{ij}^3 were greater or equal to 2, as previously defined (Conover, 1999; Fisher, 1990). The elevated number of cutoff pairs is highly understandable, as moving a patient from one group to the next one produced an additional cutoff point with a close statistics. To reduce this high number, for each lower cutoff value s_i of the preselected cutoff pairs, the resulting global log-rank and the two between-group log-rank chi-squares calculated for all available positions of the upper cutoff were reduced to their maximums in recording $\max_{i+1 \leq j \leq n-1}(X_{ij})$, $\max_{i+1 \leq j \leq n-1}(X_{ij}^{12})$, $\max_{i+1 \leq j \leq n-1}(X_{ij}^{23})$ (Supplementary Animation). The between-group log-rank chi-squares were made comparable by centering on mean and reducing by standard deviation division using the 'scale' R function. The smallest between-group log-rank chi-square:

$$\min_i\left(\max_{i+1 \leq j \leq n-1}(X_{ij}^{12}), \max_{i+1 \leq j \leq n-1}(X_{ij}^{23})\right)$$

was recorded along s_i. Finally, the best lower score cutoff was selected as containing the maximum of these values:

$$\max(\min_{i}(\max_{i+1\le j\le n-1}(X_{ij}^{12}), \max_{i+1\le j\le n-1}(X_{ij}^{23})))$$

The optimal upper cutoff contained the maximum global-log-rank chi-square for this best lower cutoff point:

$$\max_{i+1\le j\le n-1}(X_{(i\ optimal)j})$$

Survival of the RS groups was depicted using Kaplan–Meier curves and analyzed using log-rank statistics and Cox proportional hazard model. Extrapolation of survival curves was performed using a parametric regression model assuming a Weibull distribution fit (Therneau and Grambsch, 2000).

2.5 Validation on independent cohorts

The CEL file from each patient sample of the TT2, FOO or NIH validation cohorts was individually normalized with the processing parameters obtained with HM, CIT or NL cohorts, respectively, using incremental preprocessing by a modification (Meißner *et al.*, 2011) of the 'docval' R-package (Kostka and Spang, 2008). After batch effect correction, the expression values of the relevant signature were extracted and the s score was individually calculated. Patients were then assigned to their RS risk group according to the cutoffs from the relevant training cohort.

2.6 Open-access Web site for RS risk stratification of new patients

We provide here a simple open-access Web site (PrognoWeb, https://gliserv.montp.inserm.fr/PrognoWeb) automatically performing the computations required to determine RS classification for a new patient, i.e. GCRMA normalization of the Affymetrix.EL file, extraction of the predictive list of probe sets, computation of the score and classification according to the 3-group risk model. The user has to upload the Affymetrix U133 Plus 2.0.EL file into the Web site and RS and risk stratification is computed for any patient with MM, breast cancer or glioma within the indicated limitations.

3 RESULTS

3.1 Gene expression-based RS and survival

Filtering the probe sets with the highest variance in the training cohorts led to, respectively, 24 614, 27 731 and 28 821 probes for MM, breast cancers and gliomas (Supplementary Fig. S1A). For each probe set, the maximum significance of the running log-rank tests on the ordered expression scale was adjusted for the number of patients by multiple testing correction. These significances were then grouped, sorted and adjusted for the number of filtered probe sets according to their rank. This method yielded a selection of 19 prognostic probe sets distributing as 15 risk and 4 protective genes at an adjusted $P \le .05$ for MM, of 68 probe sets (37 risk and 28 protective genes, adjusted $P \le .01$) for breast cancer and of 39 probe sets (12 risk and 22 protective genes, adjusted $P \le 1.0\text{E-}20$) for gliomas (Supplementary Tables S2–S4). Each of the prognostic probe sets split patients' cohorts into two groups (Supplementary Fig. S1B), with a poor prognosis group comprising 7–27% of patients with MM, 4–31% of patients with breast cancer and 37–60% of patients with gliomas. For each cancer type, these prognostic probe sets were used to compute the s score (Supplementary Fig. S3A) and to define the number of patients' groups with a different OS as indicated in the Methods section.

3.2 Prognostic groups

Not surprisingly, the model could predict for two patients' risk groups with significant different survival in the training cohorts for the three cancer types (not shown). It could also predict for three prognostic risk groups but not for four risk groups likely owing to the size limitation of the training cohort.

To predict three patients' risk groups, 20 910, 40 186 and 36 585 combinations of cutoff pairs were possible for myeloma, breast cancer or glioma, respectively, of which 2756, 11 054 and 16 943 were accepted with the criteria of adjusted global chi-square significance, between-group chi-square significance and expected number of deaths per group (Supplementary Fig. S3B–D).

The best combination of cutoff pairs was selected by simultaneously optimizing between-group log-ranks as described in the Methods section and Supplementary Animation. Supplementary Figure S4 displays these maximized values plotted along the lower s score cutoff point for the different tumor types before and after centering and reduction (arrows). The optimal corresponding upper score cutoff is the one of the maximum global log-rank chi-square. In the HM cohort of patients with MM, this three-group model delineated a high-risk group (RS3, 9% of the patients) similar to that depicted in the two-group model, but split the low-risk group into a very good prognosis group RS1 (44% of the patients) and an intermediate-risk group RS2 (47% of the patients). Their difference in survival was likewise significant ($P \le 3.9 \times 10^{-3}$). The frequency of the s score distribution and the position of corresponding cutoffs are depicted in Figure 1B. The median survival was of 24 months for the high-risk RS3 group, 73% of the patients of RS2 intermediate-risk group were surviving at 4 years and 87% of the patients in the low-risk RS1 group at 7 years (Fig. 1A and Table 1).

Three prognostic groups were also identified for breast cancer with a patient distribution comparable with that of MM: a high-risk RS3 group comprising 11% of the patients and 68% of deceased patients at a median follow-up of 71 months and two comparably sized intermediate RS2 and low RS1 risk groups (36 and 53% of the patients, respectively). The RS2 group comprised 23% of deceased patients and RS1 9%. Ninety-eight percent (98%) and 90% of the patients were still alive at 4 years in RS1 and RS2, respectively.

The shape of the Kaplan–Meier curves and the s score frequency for breast cancer were comparable with those in MM (Fig. 1C and D). For patients with life-threatening glioma, the three-group model created a large poor prognosis RS3 group encompassing 53% of the patients with only two still alive, a small intermediate RS2 (15%) and a low risk RS1 group comprising only 32% of the patients. It should be noted that the rate of death remained high in both RS1 and RS2, with, respectively, 25 and 22% of surviving patients (Table 1). Expectedly, the s score distribution frequency shifted toward high s values (Fig. 1F).

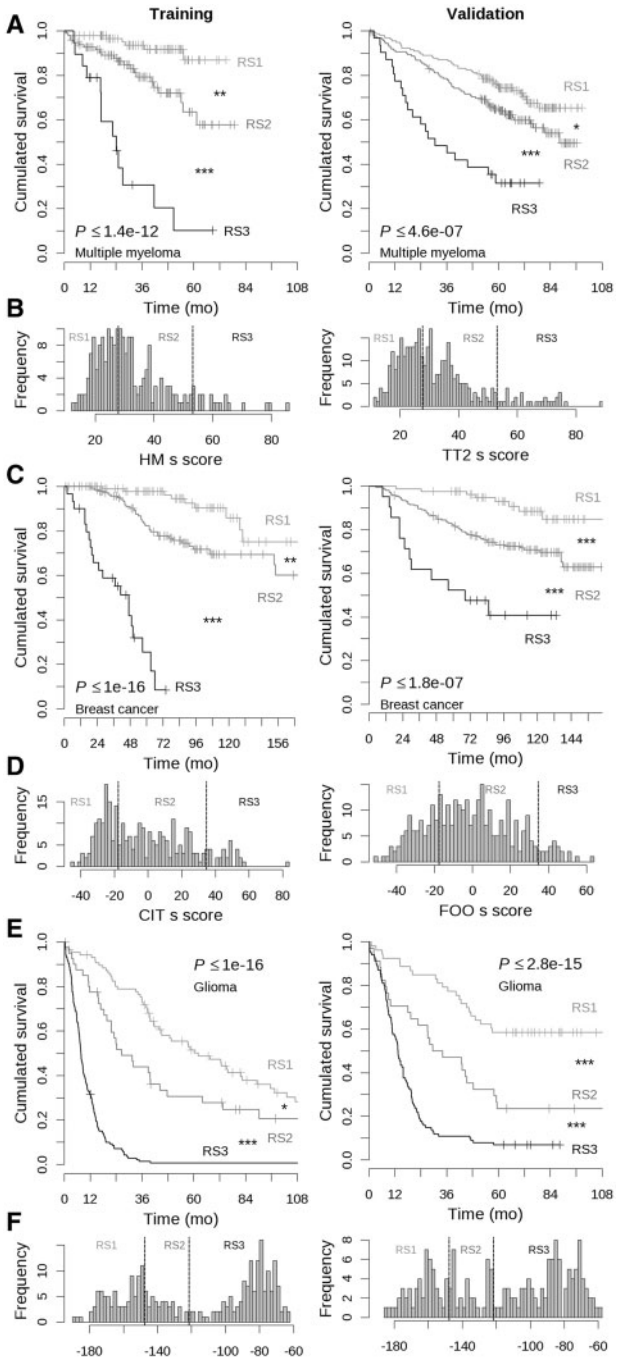

Fig. 1. Three-group RS molecular modeling of overall survival. Panels (**A**, **C** and **E**) display the Kaplan–Meier survival curves and log-rank tests using the threee-group survival model, respectively, in the training and validation cohorts. Global log-rank *P*-values and significance range for between-group log-rank *P*-values are indicated (*: $0.01 < P \leq 0.05$, **: $0.001 < P \leq 0.01$, ***: $P \leq 0.001$). Histograms of RS s score distribution and cutoffs between three groups of patients partitioned by increasing score for the training cohorts and the validation cohorts are displayed in panels (**B**, **D** and **F**)

3.3 Validation of the RS score on independent cohorts

To validate the relevance of RS classification, any patient of a validation cohort was considered as a single and new patient. This was achieved by GCRMA-normalizing the patient's raw data using training cohort parameters, computing the continuous s-score and assigning the patient to one of the prognostic RS groups using the cutoffs of the training cohort. Using this approach, s-score frequencies and RS group distributions observed in the three validation cohorts were similar to those found in the training sets (Fig. 1). Patients of the MM validation cohort were split into three groups accounting for 9, 46 and 45% of the patients, respectively. Survival difference was significant both between consecutive RS groups ($P \leq 0.029$ and $P \leq 1.4\text{E-}04$) and globally ($P \leq 4.6\text{E-}07$, Fig. 1A and Table 1).

The three-group models were also fully validated for patients with breast carcinoma or glioma (Fig. 1C and E and Table 1).

3.4 Independence of the risk model and conventional prognostic factors

The prognostic value of RS was compared with that of ISS (Greipp *et al.*, 2005) commonly used for MM in both the training and validation cohorts (Fig. 2A). Whereas ISS had a globally significant prognostic value as expected, it had difficulties to correctly identify high-risk patients. It detected only 36% of deaths in the ISS3 group (not shown), compared with 74% for RS3. The log-rank between intermediate- and high-risk groups is no longer significant for ISS in the HM cohort (not shown). When run together with RS in a multivariate Cox model analysis, ISS remained prognostic in the myeloma validation cohort only (Table 2), in agreement with ISS group distribution within RS classification (Fig. 2A).

For patients with breast cancer, the prognostic value of RS was compared with that of molecular subtypes. Table 2 shows that training and validation subtypes were individually significant with a lower hazard ratio than the corresponding RS. The distribution of molecular subtypes within RS groups showed that the RS3 high-risk group mostly comprised the worst prognosis basal-like (basL) and molecular apocrine (mApo) subtypes in the training cohort and their equivalent counterparts, subtypes I and II in validation (Fig. 2B). Intermediate subtypes luminal C (lumC) and luminal B (lumB), corresponding respectively to III and IV in validation, were largely represented in RS2 intermediate-risk group, whereas the RS1 low-risk group contains most of the good prognosis subtypes luminal A (lumA) and normal-like (normL), V and VI, respectively, in the CIT and FOO cohorts. When grouping the intrinsic gene classification populations by their published risk, i.e. basL and mApo, lumB and lumC, normL and lumA in the CIT cohort, and I and II, III and IV, V and VI in the FOO cohort, allowed comparing this three risk-level stratification with the three-group RS model. Low and intermediate populations almost superimpose, whereas the RS3 classification is more adapted at identifying the highest-risk patients (Supplementary Fig. S5). In univariate analysis, most subtypes were independently significant for OS (Table 2), whereas none remained independent when run in multivariate analysis with the RS categories (data not shown).

The various histological types of glioma were grouped into grades according to the WHO classification

Table 1. Overall survival (OS) analysis in the three RS groups of patients with multiple myeloma, breast cancer or glioma

Cohort	Prognostic group	% Patients per group	% Deceased per group[a]	Global log-rank $P\leq$	Between-group log-rank $P\leq$	OS at 48 months	Median survival (month)
Multiple myeloma HM	RS1	44	8		2.2E-03	92	NR[b]
	RS2	47	23	1.4E-12		72	NR
	RS3	9	68		2.2E-06	20	24.1
TT2	RS1	45	29		0.029	82	NR
	RS2	46	40	4.6E-07		70	88.3
	RS3	9	68		1.4E-04	39	30.6
Breast cancer CIT	RS1	36	9		2.3E-03	98	NR
	RS2	53	23	1.0E-16		90	NR
	RS3	11	68		1.0E-16	42	47
FOO	RS1	26	10		5.7E-04	98	NR
	RS2	68	29	1.8E-07		85	169.2
	RS3	7	57		4.0E-04	57	68.4
Glioma NL	RS1	32	75		0.013	58	62.0
	RS2	15	78	1.0E-16		31	26.5
	RS3	53	99		1.1E-12	1	7.7
NIH	RS1	28	42		4.0E-04	66	NR
	RS2	18	76	2.8E-15		35	31.8
	RS3	54	93		3.4E-04	9	13.4

[a]At the end of follow-up. [b]Not reached.

(Gravendeel *et al*., 2009). The differential survival between these grades was strongly significant ($P \leq 1.0E-16$) and revealed that extreme RS groups clearly reflected prognosis with all patients alive for the better grade I in RS1 and no survival, except one, for the worst grade IV in RS3 (not shown). When run in a Cox model analysis, RS and grades were both significant even in the validation cohort, and RS remained independent of conventional grading (Table 2).

The distribution of the grading into the RS groups is depicted in Figure 2C. Not surprisingly, most of RS3 patients are glioblastomas, grades II and III are similarly distributed in RS groups and most of the grade I pilocytic astrocytomas classified into the good prognosis RS1 group.

3.5 Open-access Web site for RS risk stratification of new patients

The different steps of the present risk assessment process can be computed online on our PrognoWeb Web site either in uploading personal Affymetrix U133 Plus 2.0 raw CEL files or using online available test files of patients with MM, breast cancer or glioma.

4 DISCUSSION

The aim of this study was to create an easy-to-use gene expression-based score allowing a multi-group stratification for OS while minimizing a priori assumptions. One mandatory assumption is that GEP of a tumor sample can predict for risk provoked by a whole tumor. In addition, tumor samples contain tumor cells but also cells of the tumor environment, except for cancers such as MM, for which tumor cells can be easily purified. This bias in using a sample partly representative of the whole tumor to assess risk cannot be easily cleared owing to ethical and economical reasons avoiding multiple tumor samples harvesting and introduces artifactual variation in GEP, weakening the risk information. Another assumption is that gene expression could feature risk stratification. It is clear that the best approach will be to monitor protein function, which is a consequence of gene transcription, translation and further protein modifications. Likely, the method to measure gene expression with microarrays will be shortly greatly improved using RNA sequencing and lead to a better model of risk stratification. But the principles developed here will still be applicable whatever the method used to generate the score scale.

Despite these biases, the current method made it possible to classify independent cohorts of patients with MM, breast cancer, or glioma into three risk groups defined by different OS between two consecutive groups. The number of risk groups that could be predicted is dependent on the number of patients in the training cohorts. In the current study, we used publicly available cohorts of 200–300 patients with GEP and prognostic information as well as a several year follow-up, leading to the finding of three prognostic groups. Both the cohort size and the homogeneity of the treatments per disease influenced the cutoffs stability. Larger series of previously untreated patients should provide more information about molecular heterogeneity and allow investigating how the RS methodology will translate this heterogeneity into additional groups of patients with different OS, particularly in

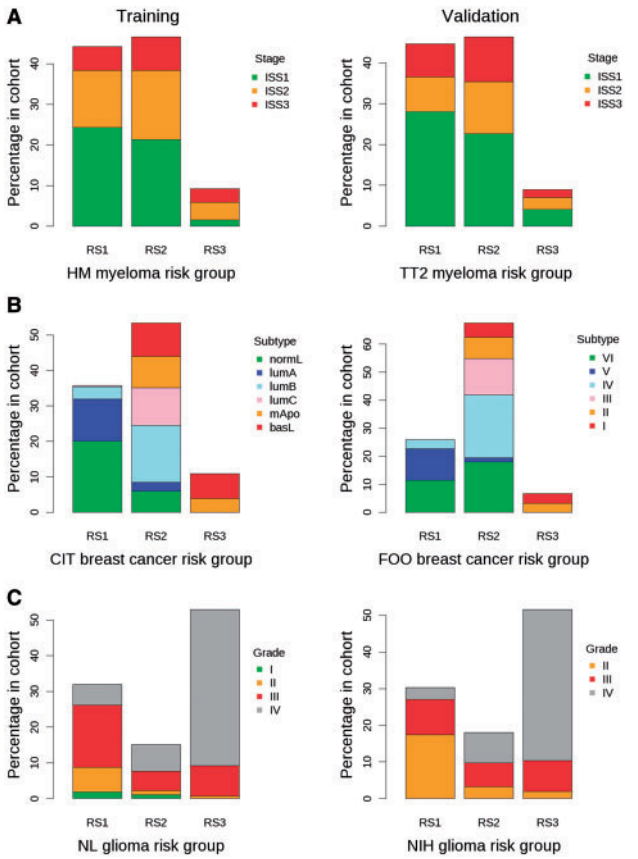

Fig. 2. Overlap analysis of clinical or molecular classifications with RS cancer risk groups. Distribution of multiple myeloma ISS stages (**A**), breast cancer subtypes (**B**) and glioma grades (**C**) are displayed within each corresponding RS group. Results are expressed as cohort percentages to reflect the relative RS distribution

the highly heterogeneous breast cancer (Sotiriou and Pusztai, 2009). Interestingly, considering homogeneous populations like healthy bone marrow plasma cells or human myeloma cell lines, the s score was as expected at low risk (RS1) for healthy plasma cells or at high risk (RS3) for myeloma cell lines (Supplementary Fig. S7).

For building RS score, prognostic probe sets were identified by splitting training cohorts into two prognostic groups maximizing a log-rank statistics. The selection of this probe set list could be further refined according to the final criteria, here three significant prognostic groups, using, for instance, backward selection. The fact that the proportion of patients to split the cohort was not fixed arbitrarily may explain why the currently identified prognostic RS genes poorly overlap with previously published prognostic genes. In MM, half of the 19 RS genes were relevant to cell cycle and mitosis (Supplementary Table S6). Only one of these genes was common to the 70 genes risk signature from UAMS GEP70 (Shaughnessy *et al.*, 2007) and none with the 15 Intergroupe Francais du Myelome (IFM) genes (Decaux *et al.*, 2008). Comparably, none of the 70 genes of the seminal signature of poor prognosis breast cancer (van 't Veer *et al.*, 2002) was found in the current 65 RS genes. This lack of overlap was confirmed with subsequent signatures (Finak *et al.*, 2008; Naderi *et al.*, 2007; Paik *et al.*, 2004; Parker *et al.*, 2009). In the same way, none to only one of the prognostic genes reported for glioma were shared with the present 34 RS genes (Cho *et al.*, 2009; de Tayrac *et al.*, 2011; Freije *et al.*, 2004). Moreover, training and validation cohorts were swapped to test for overfitting, and the whole process including initial normalization, batch correction and survival-significant probe set selection was run again. Noteworthy, new 74-, 98- and 28-gene signatures (Supplementary Tables S8–S10) were obtained, respectively, for MM, breast cancer and glioma. Although these signatures

Table 2. Categorical univariate and multivariate Cox model applied to multiple myeloma, breast cancer and glioma prognosis groups for OS

Multiple myeloma	HM		TT2	
	HR[a]	$P^b \leq$	HR	$P \leq$
RS2/RS1	3.5	3.6E-03	1.5	0.031
RS3/RS1	16.6	3.5E-09	3.9	3.0E-07
ISS2/ISS1	2.3	0.031	1.7	0.016
ISS3/ISS1	3.9	1.3E-03	2.8	4.3E-07
RS2	3.3	6.8E-03	1.4	NS[c] (0.12)
RS3	12.9	2.1E-07	3.7	1.2E-06
ISS2	1.8	NS (0.16)	1.7	0.026
ISS3	2.3	NS (0.06)	2.7	1.7E-06
Breast cancer	**CIT**		**FOO**	
	HR	$P \leq$	HR	$P \leq$
RS2/RS1	3.0	3.7E-03	3.3	1.2E-03
RS3/RS1	30.4	8.9E-16	10.1	4.8E-07
lumA/normL	1.2	NS (0.77)		
lumB/normL	2.8	0.024		
lumC/normL	2.7	0.054		
mApo/normL	4.7	1.0E-03		
basL/normL	3.9	3.4E-03		
Subtype V/VI			0.3	NS (0.08)
Subtype IV/VI			1.9	0.037
Subtype III/VI			1.1	NS (0.75)
Subtype II/VI			3.4	2.7E-04
Subtype I/VI			1.6	NS (0.27)
Glioma	**NL**		**NIH**	
	HR	$P \leq$	HR	$P \leq$
RS2/RS1	1.7	0.018	2.6	8.9E-04
RS3/RS1	9.0	2.0E-06	6.0	2.6E-13
Grade III/II[d]	1.2	NS (0.54)	2.1	0.026
Grade IV/II	3.6	2.0E-06	5.1	3.6E-08
RS2	1.6	0.043	2.2	0.020
RS3	7.0	1.0E-16	3.9	4.6E-05
Grade III	1.0	NS (0.86)	1.5	NS (0.23)
Grade IV	1.4	NS (0.27)	2.2	0.028

[a]Hazard ratio. [b]Wald test *P*-value. [c]Not significant at a 5% risk. [d]Excluding grade I gliomas absent in validation cohort.

shared no common gene, except one (*PHF15*) for breast cancer and three (*FKBP9*, *MTPAP* and *RAB18*) for glioma, with the initial ones, they could predict for three significantly different risk groups, close to the ones before swapping, both in the training and validation cohorts (Supplementary Table S5 and Fig. S6). This is in line with MicroArray Quality Control comments emphasizing that prognostic genes are mainly useful to build scores to identify patients with various risks, but are not relevant to understand tumor biology in vivo (MAQC Consortium, 2010). The prognostic genes, the computation of the RS and cutoffs are valuable only for patients treated with therapy lines, which have been validated in the current study. These parameters will likely change according to treatment improvement and should be updated regularly with publication of new data. However, RS score remains significant in MM treated with novel agents (Supplementary Table S7). It is important to stress that GEP raw files and patients' clinical data used in various publications should be publicly available according to MIAME recommendations (Minimum Information About a Microarray Experiment; http://www.mged.org/Workgroups/MIAME/miame_2.0.html).

Another advantage of the present methodology is that RS score was validated without using gene expression variance or structure information from the validation cohorts, except for batch correction. Each patient of a validation cohort was considered as a single patient, data normalization computed and batch-adjusted using Affymetrix parameters from the training cohort and RS score and risk groups computed using cutoffs defined with the training cohort. This is another major difference with previously published transcriptome-based RS. In MM, the GEP70 score was delineated on the TT2 cohort using expression quartiles and was validated on their more recent, treatment-improved TT3 cohort using an adapted cutoff value (Shaughnessy *et al*., 2007). The same holds true for IFM score, which was validated arbitrarily using the top 25% of patient scores on TT2 cohort (Decaux *et al*., 2008). Finally, RS remained significant when compared with the currently used ISS RS in MM, both in terms of multivariate analysis (Table 2) and survival expression, especially for RS3 populations (Supplementary Fig. S8).

GEP has largely been used in breast cancer over the past decade in search for classification patterns and survival signatures (Reis-Filho and Pusztai, 2011). Most studies have sought a dichotomic risk in bulk patient populations without addressing disease complexity. Although built using different gene sets, most behave similarly in outcome prediction (Fan *et al*., 2006). A few used tumor subtypes to create survival indexes from clinical and molecular data, namely, the risk of relapse (Parker *et al*., 2009) and the recurrence (Paik *et al*., 2004). Both studies delineated a globally significant three-group model, but with pre-specified cutoff points and no evidence for between-group significance. Our RS score, based on mathematical criteria only, also allowed delineating three risk groups in both training and validation breast carcinoma cohorts: a high-risk group of up to 10% and two similarly sized intermediate- and low-risk groups with significant successive differential survival. Although a number of four-group models were delineated in MM and breast cancer training cohorts (not shown), the number of patients with simultaneously available clinical and microarray data has not permitted to validate survival models beyond three groups for these diseases. When affecting the molecular taxonomy to the various RS groups of the training cohort, we found a survival distribution perfectly consistent with the clinical outcome and respecting the disease heterogeneity (Fig. 2B). The poor outcome intrinsic basL and mApo subtypes constitute the RS3 high-risk group and part of the intermediate RS2 group. The latter encompasses most of the $ER^+/HER2^+$, poor prognosis lumC and ER^+ lumB, whereas the good prognosis normL and lumA subtypes constitute most of the low-risk RS1 group. Of note, this remarkable correlation between molecular classification and prognosis was achieved without a proliferative signature; our 65 gene list containing none of the supposedly necessary proliferation components (Reis-Filho and Pusztai, 2011) to the exception of *MKI67*. A gene enrichment analysis was unable to correlate with previously described pathways whatever the significance threshold (data not shown). This tight subtype–outcome association was propagated to the validation cohort just using the training s score cutoff points. The RS3 poor prognosis group contains the subtypes I ($ER^-/HER2^-$) and II ($ER^-/HER2^+$), equivalent, respectively, to the basL and mApo subtypes; the $HER2^+$ subtype III is included in RS2 and the ER^+ subtypes V and VI again constitute the good prognosis RS1 group. Thus, the present algorithm delineates a comprehensive OS model, and suggests that with a large enough number of documented patients, both the molecular heterogeneity and its outcome prediction could be simultaneously addressed.

Gliomas, the most common primary brain tumors, are currently classified according to largely variable interpretations of histopathological features, and graded following the WHO classification (Louis *et al*., 2007), from the grade I best prognosis astrocytic tumor through the anaplastic transformation between grades II and III to the highly malignant grade IV glioblastoma. In both training and validation glioma cohorts, our method perfectly depicts this hierarchy in detecting a majority of highly malignant glioblastomas in the RS3 group, glioblastoma and grade III anaplastic gliomas in RS2, but still some glioblastomas and a comparable number of grade II and III gliomas in RS1, reflecting the weakness of the current classification to clearly separate intermediate grade II/III gliomas. Thus our classification brings benefit to a fraction of grade III and IV gliomas, which may present a better outcome than expected. Contrarily to MM and breast carcinoma, a four-group survival model was significantly delineated in the training glioma cohort and applied to the validation cohort with marginal significance (not shown). In this case, the large high-risk, glioblastoma-containing group with a median OS of 15–20 months was further split into a small very poor prognostic group with an <12-month survival and a better one with a median OS of 2 years. Glioblastoma subtypes have been defined by gene expression signatures (Cooper *et al*., 2010; de Tayrac *et al*., 2011; Li *et al*., 2009; Verhaak *et al*., 2010). Implications like a differential expression in the MGMT enzyme (Esteller *et al*., 2000) repairing chemotherapy-induced DNA damage could be tested with larger data. Meanwhile, fitting a regression model on the survival of training and validation cohorts and their combination in every tumor type allow to approximate the life expectancy of patients in each risk group, namely, from high to low risk: 2, 8 and 13 years for myeloma; 4, 16 and >25 years for breast cancer and 1, 3 and 6 years for gliomas (Supplementary Fig. S9).

Finally, we think it is essential to provide easy-to-use tools that will allow clinicians to compute individual GEP-based RSs without heavy computation. A molecular classification of glioma is already provided with a similar input (Li *et al*., 2010). Thus, we provide the open-access 'Prognoweb' Web site to automatically compute the RS score for any single oncology patient with available Affymetrix files. Any report about a GEP-based RS should provide such an open access, helping to integrate, as currently developed (Meißner *et al*., 2011), the molecular prognosis assessment into the standard investigations at diagnostic or relapse. But the user has to be aware that the way to compute the RS and the cutoffs is valuable only for patients treated with therapy lines, which has been validated in the current study. The computation of RS and cutoffs will change likely according to treatment improvement. This recommendation is indicated in the Web site.

ACKNOWLEDGEMENTS

The Ligue Nationale Contre le Cancer and particularly its Cartes d'Identité des Tumeurs program (http://cit.ligue-cancer.net/) are gratefully acknowledged for providing data for the breast cancer training program. The authors are indebted to Drs John Shaughnessy and Bart Barlogie for the use of their transcriptome data of patients with MM. The authors thank Tobias Meißner for improving normalization scripts; Anja Seckinger, Uta Bertsch and Jennifer Klemmer for clinical data management and the Transcriptome Platform, Montpellier, France, for microarray experiments.

Funding: Ligue Nationale Contre le Cancer, Paris, France (Equipe labéllisée 2009) and Institut National du Cancer, Paris, France (2008-047 to B.K.); the Hopp-Foundation, Germany; the University of Heidelberg, Germany; the National Center for Tumor Diseases, Heidelberg, Germany; the Deutsche Forschungs-gemeinschaft, Bonn, Germany (TRR79 to D.H.) and the Deutsche Krebshilfe, Bonn, Germany.

Conflict of Interest: none declared.

REFERENCES

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*., **2**, E108.

Barlogie,B. *et al.* (2006) Total therapy 2 without thalidomide in comparison with total therapy 1: role of intensified induction and posttransplantation consolidation therapies. *Blood*, **107**, 2633–2638.

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

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.

Chng,W.J. *et al.* (2008) The centrosome index is a powerful prognostic marker in myeloma and identifies a cohort of patients that might benefit from aurora kinase inhibition. *Blood*, **111**, 1603–1609.

Cho,H. *et al.* (2009) Robust likelihood-based survival modeling with microarray Data. *J. Stat. Softw*., **29**, 1–16.

Conover,W.J. (1999) *Practical Nonparametric Statistics*. John Wiley and Sons, New York, p. 156.

Cooper,L.A. *et al.* (2010) The proneural molecular signature is enriched in oligodendrogliomas and predicts improved survival among diffuse gliomas. *PLoS One*, **5**, e12548.

de Tayrac,M. *et al.* (2011) A 4-gene signature associated with clinical outcome in high-grade gliomas. *Clin. Cancer Res*., **17**, 317–327.

Decaux,O. *et al.* (2008) Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myelome. *J. Clin. Oncol*., **26**, 4798–4805.

Esteller,M. *et al.* (2000) Inactivation of the DNA-repair gene mgmt and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med*., **343**, 1350–1354.

Fan,C. *et al.* (2006) Concordance among gene-expression–based predictors for breast cancer. *N. Engl. J. Med*., **355**, 560–569.

Finak,G. *et al.* (2008) Stromal gene expression predicts clinical outcome in breast cancer. *Nat. Med*., **14**, 518–527.

Fisher,R.A. (1990) *Statistical Methods, Experimental Design and Scientific Inference*. Oxford University Press, New York, p. 84.

Freije,W.A. *et al.* (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Res*., **64**, 6503–6510.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Gravendeel,L.A. *et al.* (2009) Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res*., **69**, 9065–9072.

Greipp,P.R. *et al.* (2005) International staging system for multiple myeloma. *J. Clin. Oncol*., **23**, 3412–3420.

Guedj,M. *et al.* (2012) A refined molecular taxonomy of breast cancer. *Oncogene*, **31**, 1196–1206.

Hose,D. *et al.* (2009) Inhibition of aurora kinases for tailored risk-adapted treatment of multiple myeloma. *Blood*, **113**, 4331–4340.

Hose,D. *et al.* (2010) Proliferation is a central independent prognostic factor and target for personalized and risk adapted treatment in multiple myeloma. *Haematologica*, **96**, 87–95.

Hothorn,T. and Lausen,B. (2003) On the exact distribution of maximally selected rank statistics. *Comput. Stat. Data Anal*., **43**, 121–137.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Kao,K.J. *et al.* (2011) Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer*, **11**, 143.

Kostka,D. and Spang,R. (2008) Microarray based diagnosis profits from better documentation of gene expression signatures. *PLoS Comput. Biol*., **4**, e22.

Kuiper,R. *et al.* (2012) A gene expression signature for high-risk multiple myeloma. *Leukemia*, **26**, 2406–2413.

Li,A. *et al.* (2009) Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Res*., **69**, 2091–2099.

Li,A. *et al.* (2010) GliomaPredict: a clinically useful tool for assigning glioma patients to specific molecular subtypes. *BMC Med. Inform. Decis. Mak*., **10**, 38.

Louis,D. *et al.* (2007) The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol*., **114**, 97–109.

MAQC Consortium. (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol*., **28**, 827–838.

Meißner,T. *et al.* (2011) Gene expression profiling in multiple myeloma—reporting of entities, risk, and targets in clinical routine. *Clin. Cancer Res*., **17**, 7240–7247.

Mulligan,G. *et al.* (2007) Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood*, **109**, 3177–3188.

Naderi,A. *et al.* (2007) A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, **26**, 1507–1516.

Nutt,C.L. *et al.* (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*., **63**, 1602–1607.

Paik,S. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl. J. Med*., **351**, 2817–2826.

Parker,J.S. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol*., **27**, 1160–1167.

Petalidis,L.P. *et al.* (2008) Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Mol. Cancer Ther*., **7**, 1013–1024.

Reis-Filho,J.S. and Pusztai,L. (2011) Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet*, **378**, 1812–1823.

Shaughnessy,J.D., Jr *et al.* (2007) A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood*, **109**, 2276–2284.

Shirahata,M. *et al*. (2007) Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis. *Clin. Cancer Res*., **13**, 7341–7356.

Sørlie,T. *et al*. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci*., **98**, 10869–10874.

Sotiriou,C. and Pusztai,L. (2009) Gene-expression signatures in breast cancer. *N. Engl. J. Med*., **360**, 790–800.

Sprynski,A.C. *et al*. (2009) The role of IGF-1 as a major growth factor for myeloma cell lines and the prognostic relevance of the expression of its receptor. *Blood*, **113**, 4614–4626.

Therneau,T. and Grambsch,P. (2000) *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York/Berlin, Heidelberg.

van 't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Verhaak,R.G. *et al*. (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.

Wu,Z. *et al*. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc*., **99**, 909–917.