PRECLINICAL STUDY

An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients

Balazs Györffy · Andras Lanczky · Aron C. Eklund ·
Carsten Denkert · Jan Budczies · Qiyuan Li ·
Zoltan Szallasi

Abstract Validating prognostic or predictive candidate genes in appropriately powered breast cancer cohorts are of utmost interest. Our aim was to develop an online tool to draw survival plots, which can be used to assess the relevance of the expression levels of various genes on the clinical outcome both in untreated and treated breast cancer patients. A background database was established using gene expression data and survival information of 1,809 patients downloaded from GEO (Affymetrix HGU133A and HGU133+2 microarrays). The median relapse free survival is 6.43 years, 968/1,231 patients are estrogen-receptor (ER) positive, and 190/1,369 are lymph-node positive. After quality control and normalization only probes present on both Affymetrix platforms were retained ($n = 22,277$). In order to analyze the prognostic value of a particular gene, the cohorts are divided into two groups according to the median (or upper/lower quartile) expression of the gene. The two groups can be compared in terms of relapse free survival, overall survival, and distant metastasis free survival. A survival curve is displayed, and the hazard ratio with 95% confidence intervals and logrank P value are calculated and displayed. Additionally, three subgroups of patients can be assessed: systematically untreated patients, endocrine-treated ER positive patients, and patients with a distribution of clinical characteristics representative of those seen in general clinical practice in the US. Web address: www.kmplot.com. We used this integrative data analysis tool to confirm the prognostic power of the proliferation-related genes TOP2A and TOP2B, MKI67, CCND2, CCND3, CCNDE2, as well as CDKN1A, and TK2. We also validated the capability of microarrays to determine estrogen receptor status in 1,231 patients. The tool is highly valuable for the preliminary assessment of biomarkers, especially for research groups with limited bioinformatic resources.

Keywords Survival analysis · Breast cancer · Prognosis

B. Györffy (✉) · A. Lanczky
Joint Research Laboratory of the Hungarian Academy
of Sciences and the Semmelweis University, Semmelweis
University 1st Department of Pediatrics, Bokay u. 53-54,
1083 Budapest, Hungary
e-mail: zsalab2@yahoo.com

A. Lanczky
Pazmany Peter University, Budapest, Hungary

A. C. Eklund · Q. Li · Z. Szallasi
Center for Biological Sequence Analysis, Technical University
of Denmark, Lyngby, Denmark

C. Denkert · J. Budczies
Charité Universitaetsmedizin, Berlin, Germany

Z. Szallasi
Children's Hospital Informatics Program at the Harvard-MIT
Division of Health Sciences and Technology (CHIP@HST),
Harvard Medical School, Boston, MA, USA

Background

Biomarkers are a readily measurable set of parameters with directly applicable information on the clinical course of cancer. The first biomarkers were established at the cellular, histological, or whole organism level. For example, tumor grade has traditionally been regarded as an important indicator of breast cancer prognosis [1]. Also, Adjuvant! Online, a SEER (Surveillance Epidemiology and End Results—an authoritative source of information on cancer incidence and survival in the United States) data-based algorithm, integrates various clinical (age, nodal status)

and histopathological parameters (estrogen receptor, size, grade) to predict 10-year mortality rate in breast cancer [2, 3]. With the introduction of biomarkers such as estrogen receptor and HER2 in evaluating the clinical course of breast cancer, biomarker discovery has shifted toward a more molecular level with a large number of individual gene or protein expression levels being tested. To date numerous additional genes have been suggested as being capable to predict prognosis in breast cancer [4]. This shift has also been further investigated by the fact that qualitative biomarkers are usually difficult to assess in a consistent fashion; e.g., the concordance of tumor grade assessments by three independent pathologists is less than 50% [5].

Following the identification of new gene expression-based biomarkers various steps of independent validations must be completed. While direct measurement of gene expression levels, e.g., by QRT–PCR, is the most reliable method to do this; it is often desirable to test few candidate genes without major further investment in order to choose the most promising candidates and eliminate those that are most likely to fail. Microarray cohorts combined with appropriate clinical data offer exactly such a cost effective tool to prescreen potential new biomarkers.

The accuracy of microarray-based gene expression measurements has been evaluated by a wide array of diverse studies [6–8], leading to the general conclusion that it is a powerful surveyor of gene expression changes when its limitations are considered properly. While absolute gene expression levels are hard to estimate, relative gene expression levels can be measured in a consistent fashion; therefore, a preliminary test to evaluate prognostic biomarkers based on their relative gene expression levels is a prudent exploitation of already existing clinical microarray cohorts.

The Kaplan–Meier estimator (also known as the product limit estimator) estimates the survival function from lifetime data. An important benefit of the Kaplan–Meier curve is that the method takes into account "censored" data—losses from the cohort before the final outcome is observed (for instance, if a patient withdraws from a study). When no truncation or censoring occurs, the Kaplan–Meier curve is equivalent to the empirical distribution [9]. The association between a clinical parameter (or biomarker) and survival can be visualized by drawing a Kaplan–Meier plot in which patients are split into groups according to the parameter.

Our aim was to use the data generated in gene expression studies to develop an online survival analysis tool that can be used to assess the effect of single genes on breast cancer prognosis. Since many of the current ASCO proposed proliferation-related genes [10] do not hold sufficient evidence to be introduced in clinical practice, we also aimed to assess the effect of their expression on survival. Finally, we evaluated the capability of microarray data to predict estrogen receptor (ER) status.

Methods

A database was established using gene expression data downloaded from GEO. For this, the keywords "breast", "cancer", "gpl96", and "gpl570" were used in GEO (http://www.ncbi.nlm.nih.gov/geo/). Only publications with available raw data, clinical survival information, and at least 30 patients were included. Only Affymetrix HG-U133A (GPL96) and HG-U133 Plus 2.0 (GPL570) microarrays were considered, because they are frequently used and because these two particular arrays have 22,277 probe sets in common. The use of nearly identical platforms is important since different platforms for gene-expression profiling measure expression of the same gene with varying precision, on different relative scales, and with different dynamic ranges [11]. An overview of the clinical data is presented on Table 1.

After an initial quality control, redundant samples ($n = 384$) were excluded [12]. The raw CEL files were MAS5 normalized in the R statistical environment (www.r-project.org) using the affy Bioconductor library [13]. MAS5 can be applied to individual chips, making future extensions of the database uncomplicated. Moreover, MAS5 ranked among the best normalization methods when compared to the results of RT-PCR measurements in our recent study [8]. Then, only probes measured on both GPL96 and GPL570 were retained ($n = 22,277$). At this stage, we performed a second scaling normalization to set the average expression on each chip to 1,000 to avoid batch effects [14].

The Kaplan–Meier plotter is set up using a central server which can be reached over the internet. The background database is handled by a MySQL server, which integrates gene expression and clinical data simultaneously. Data is loaded into the R statistical environment, where calculations are performed. The package "survival" is used to calculate and plot Kaplan–Meier survival curves, and the number-at-risk is indicated below the main plot. Hazard ratio (and 95% confidence intervals) and logrank P are calculated and displayed. The user receives the feedback over the webpage. The system is summarized on Fig. 1.

In order to determine expression of the ER gene ESR1, we used the results from Gong et al. [15], who found that the probe set 205225_at had the highest mean and median expression values, the greatest range of expression values, and the strongest correlation with clinical ER status, and was therefore suggested for future ESR1 determinations. We also used their suggested threshold of 500 to determine ER status of the samples.

Table 1 Clinical properties of the microarray datasets used in the analysis

GEO ID	Platform	ER+	Lymph node +	Relapse event	Average relapse free survival	Grade: 1/2/3	Age (years)	Size (cm)	# of CEL files after quality control	References
GSE12276	GPL570	NA	NA	204 (100%)	2.2 ± 1.8	NA	NA	NA	204	[21]
GSE16391	GPL570	55 (100%)	33 (60%)	55 (100%)	3.0 ± 1.2	2/35/18	61 ± 9	NA	55	[22]
GSE12093	GPL96	136 (100%)	0 (0%)	20 (15%)	7.7 ± 3.2	NA	NA	NA	136	[23]
GSE11121	GPL96	NA	0 (0%)	46 (23%)	7.8 ± 4.2	58/136/35	NA	2.1 ± 1	200	[24]
GSE9195	GPL570	77 (100%)	36 (47%)	13 (17%)	7.8 ± 2.5	14/20/24	64 ± 9	2.4 ± 1	77	[25]
GSE7390	GPL96	134 (68%)	NA	91 (46%)	9.3 ± 5.6	30/83/83	46 ± 7	2.2 ± 0.8	198	[26]
GSE6532	GPL96	70 (86%)	22 (27%)	19 (23%)	6.1 ± 3.1	0/54/1	64 ± 10	2.5 ± 1.2	82	[27]
GSE5327	GPL96	0 (0%)	NA	11 (19%)	6.8 ± 3.1	NA	NA	NA	58	[28]
GSE4922	GPL96	1	0	0	12.17	1	69	2.2	1	[29]
GSE3494	GPL96	213 (85%)	84 (33%)	NA	NA	67/128/54	62 ± 14	2.2 ± 1.3	251	[30]
GSE2990	GPL96	73 (72%)	15 (15%)	40 (39%)	6.6 ± 3.9	27/20/36	58 ± 12	2.3 ± 1.1	102	[31]
GSE2034	GPL96	209 (73%)	0	107 (37%)	6.5 ± 3.5	NA	NA	NA	286	[32]
GSE1456	GPL96	NA	NA	40 (25%)	6.2 ± 2.3	28/58/61	NA	NA	159	[33]
Total		968 (78%)	190 (15%)	689 (43%)	6.4 ± 4.1	198/534/312	57 ± 13	2.2 ± 1.1	1,809	

Parentheses: percentage of patients within the dataset

When comparing data from Surveillance, Epidemiology, and End Results (SEER), the population-based tumor registry program of the National Cancer Institute [16] to the overall characteristics of the patients used in our analysis (*only patients with all available clinical data*), some differences were observed. These differences could influence actual results when interpreting the resulting Kaplan–Meier plot. Therefore, a randomization algorithm-selected set of patients of similar, over-represented clinical characteristics were removed in making an additional filter for the analysis.
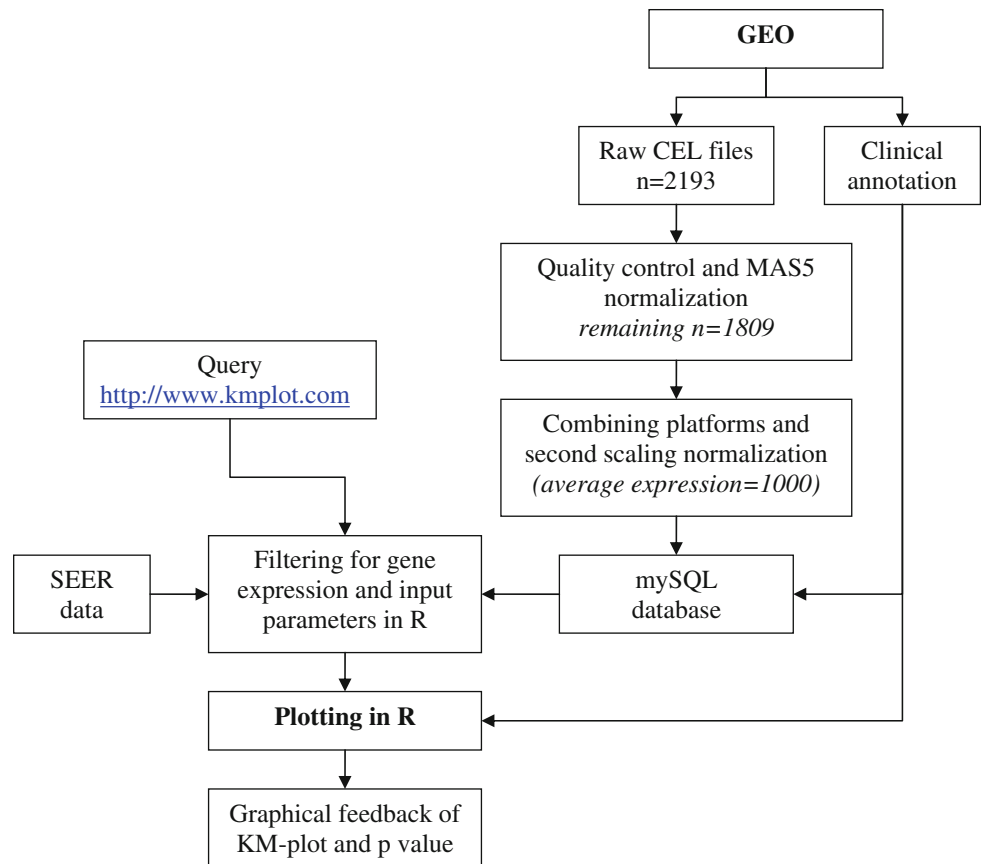
Results

We identified 1,809 unique patients meeting our criteria in GEO. The median relapse free survival is 6.43 years, 968/1,231 patients are estrogen-receptor positive by histological or radioimmunoassay based evaluation, and 190/1,369 are lymph-node positive. Furthermore, 1,593 patients have relapsed free survival data, 594 have overall survival data and 767 have distant metastasis free survival data.

In order to analyze the association between a queried gene and survival, the samples are grouped according to the median (or upper or lower quartile) expression of the selected gene, and then the two groups are compared by a Kaplan–Meier plot. Before running the analysis, the patients can be filtered using ER status, lymph node status, and/or grade. Additionally, as an alternative to relapse free survival, overall survival and distant metastasis free survival can be employed. The web address is www.kmplot.com.

Many of the published microarray cohorts used patient selection criteria corresponding to the goals of the particular study. Therefore, the patients in our database may not be representative of breast cancer patients in general. Users of our service may be interested how a given gene is associated with outcome in a general "all comer" cohort, as might be seen in the everyday clinical practice. For this we established a patient cohort similar to SEER published prevalences. The eliminated samples were ER positive, node negative patients in all three grades from different datasets. The resulting reduced database includes 500 patients, and the prevalences of the individual breast cancer subtypes and clinical parameters are similar to the actual US prevalence numbers (Table 2).

A clinician might be interested in a specific clinical question related to the treatment of the patients. Therefore, we established two options for additional filtering: the first cohort represents a truly prognostic setting (e.g., systemically untreated patients, $n = 809$) and the second cohort the endocrine-treated ER positive patients ($n = 414$).

The ER status as determined by IHC was available for 1,231 patients, which we used to assess the efficacy of ER

Fig. 1 Flowchart of the
Kaplan–Meier plotter

determination on the microarray. The ER-positive samples ($n = 968$) had a markedly higher expression of the ESR1 gene than did the ER negative samples ($n = 263$). On Fig. 2, we illustrate the distribution of ER positive and ER negative samples as measured by microarray and IHC. 90.2% of the ER positive (945 out of 1,048), and 89.8% of ER negative (160 out of 183) predictions were correct.

Markers of cell proliferation have been proposed and evaluated as prognostic factors in breast cancer. We computed Kaplan–Meier plots for the markers Ki67, cyclin D, cyclin E, the cyclin inhibitors p27 and p21, thymidine kinase, and topoisomerase II to assess their effect on prognosis (Table 3; Fig. 3).

Table 2 Overall clinical characteristics of the patients in our database, and the subset designed to match US prevalences are compared to SEER reported US prevalences

	All[a]		Prevalence-matched subset		SEER
	n	*%*	*n*	*%*	*%*
ER+	774	87.8	412	*82.4*	*76.3*
ER−	108	12.2	88	*17.6*	*23.7*
Node+	176	20.0	168	*33.6*	*36.5*
Node−	706	80.0	332	*66.4*	*63.5*
Grade 1	166	18.8	86	*17.2*	*17.1*
Grade 2	469	53.2	219	*43.8*	*44.0*
Grade 3	247	28.0	195	*39.0*	*38.9*
Total *n*	*882*		*500*		

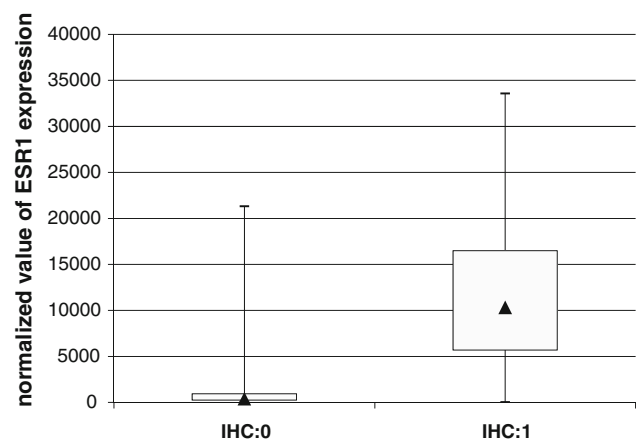[a] Only samples for which all clinical data was available simultaneously

Fig. 2 Box plot showing normalized expression of ESR1 (probe set 205225_at) in 1,231 tumors divided into two groups based on the IHC diagnosis of ER (1 = ER positive, $n = 968$; 0 = ER negative, $n = 263$)

Table 3 The association between proliferation genes and relapse-free survival

Marker	Gene name	Affymetrix ID	HR	RFS p
MKI67	Antigen identified by monoclonal antibody Ki-67	212020_s_at	0.95 (0.82–1.1)	1
		212021_s_at	1.13 (0.97–1.31)	1
		212022_s_at	*1.8 (1.5–2.1)*	*1.14E-12*
		212023_s_at	1.3 (1.1–1.5)	0.0352
CCND1	Cyclin D1	208711_s_at	1.3 (1.1–1.5)	0.0374
		208712_at	1.07 (0.93–1.25)	1[a]
CCND2	Cyclin D2	200951_s_at	1.2 (1.0–1.4)	0.946
		200952_s_at	*0.62 (0.53–0.72)*	*1.23E-08*
		200953_s_at	*0.68 (0.58–0.79)*	*9.02E-06*
CCND3	Cyclin D3	201700_at	*0.7 (0.6–0.82)*	*0.000114*
CCNE1	Cyclin E1	213523_at	1.2 (1.1–1.4)	0.1518
CCNE2	Cyclin E2	205034_at	*2.5 (2.1–2.9)*	*<1e-16*
		211814_s_at	1.2 (1.0–1.3)	1
CDKN1B	Cyclin-dependent kinase inhibitor 1B (p27, Kip1)	209112_at	1.3 (1.1–1.5)	0.0132
CDKN1A	Cyclin-dependent kinase inhibitor 1A (p21, Cip1)	202284_s_at	*0.68 (0.59–0.79)*	*1.21E-05*
TK1	Thymidine kinase 1, soluble	202338_at	1.2 (1.0–1.4)	0.506
TK2	Thymidine kinase 2, mitochondrial	204227_s_at	*0.53 (0.45–0.62)*	*7.26E-15*[a]
		204276_at	*0.67 (0.58–0.78)*	*4.18E-06*
		204277_s_at	0.81 (0.70–0.94)	0.1496
TOP2A	Topoisomerase (DNA) II alpha 170 kDa	201291_s_at	*2.3 (2.0–2.7)*	*<1e-16*
		201292_at	*1.8 (1.6–2.1)*	*2.05E-13*[a]
TOP2B	Topoisomerase (DNA) II beta 180 kDa	211987_at	*1.7 (1.5–2.0)*	*4.4E-11*

The patients were divided into two groups as having higher or lower expression as compared to the median. Bonferroni multiple testing correction was applied when generating the *P* value

RFS relapse free survival, *HR* hazard ratio

[a] See Kaplan–Meier plots on Fig. 3

Discussion

The discovery of prognostic markers is a high priority task in breast cancer biomarker research. In our study, we combined raw data from several studies; this enabled us to treat the data as a single dataset which makes the use of existing algorithms directly applicable. By combining multiple datasets the statistical power is dramatically increased. Prior to our study, no suitable tool was available which could help to estimate the prognostic value of any selected gene in a large cohort of clinical patients. In our service, after dividing the patients into two groups based on the expression of the selected gene, a Kaplan–Meier plot is generated. In this, 1,809 patient are used all together, of which 1,593 have relapse free survival data, 594 have overall survival data, and 767 have distant metastasis free survival data. As our service performs the requested analysis in real time on the original data, the extension of the analysis (e.g., the inclusion of additional samples or filtering for other clinical parameters) will be easily feasible in the future.

Since gene expression arrays might be used to confirm ER status, we implemented an estimation of ER status based on gene expression data. Previous studies have shown significant correlation between mRNA concentrations and routinely established (IHC based) clinical ER status [17–19]. In the study of Gong et al. [15] the same platform was used as in our study. They used immunohistochemistry to independently measure the ER status and to establish a statistical threshold for ESR1 mRNA level to assign ER status to tumor samples. They suggested using an ESR1 mRNA cutoff value of 500 to identify ER positive status with an overall accuracy of 90%. By using the above threshold in the 1,231 patients with available ER status data, we also achieved overall accuracy of 90%. Thus, we confirmed the capability to use microarrays to measure ER status. Because we performed a second scaling normalization, the original MAS5 expression values (as used in the study of Gong et al.) were slightly transformed. However, this transformation made it possible to compare gene expression measurements made on two different microarray platforms. On our webpage, the ER status for all
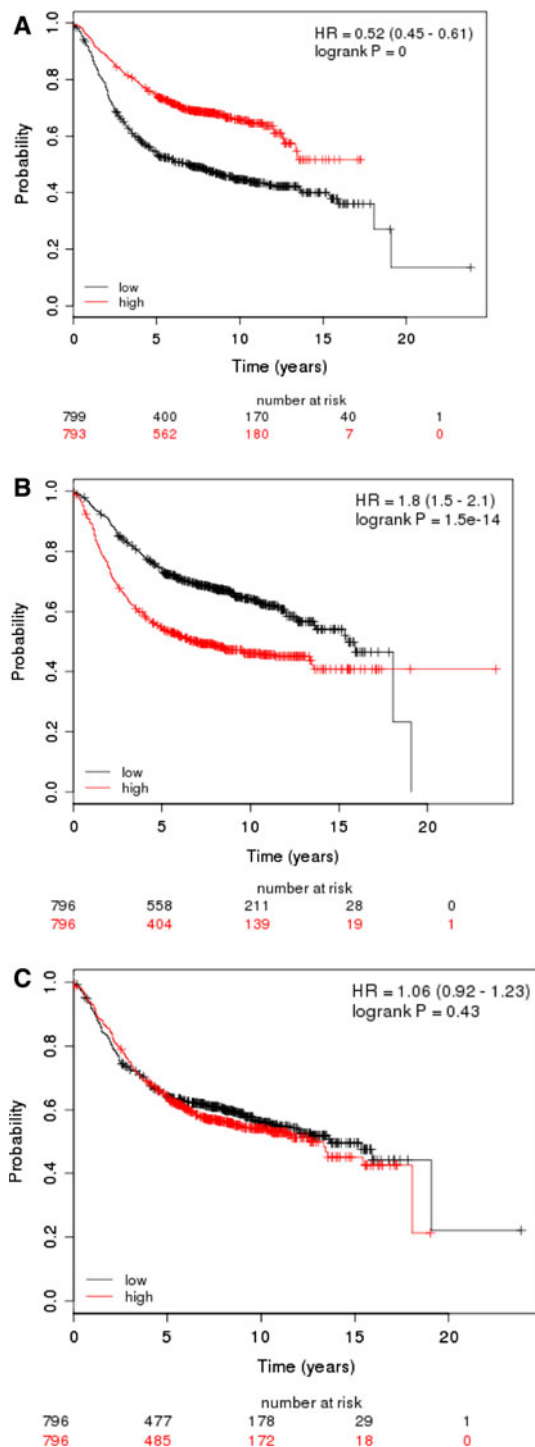
Fig. 3 The good prognostic effect of the over expression TK2 (a 204227_s_at) and the lower expression of TOP2A (b 201292_at) are highly significant. The expression of Cyclin D1 is not correlated to prognosis at all (c 208712_at)

patients can be assessed by gene expression, and this option increases the number of patients available for ER-status dependent classification from 1,231 to 1,809.

Another important clinical question is the use of proliferation-genes to predict prognosis in breast cancer patients. A previous analysis of 32 papers could not allow the inclusion of these markers on the list of ASCO-approved standard prognostic and predictive factors due to methodological problems [10]. In order to clarify these contradictory results, we computed Kaplan–Meier plots for the markers (*MKI67*, *CCND1*, *CCND2*, *CCND3*, *CCNE1*, *CCNE2*, *CDKN1B*, *CDKN1A*, *TK1*, *TK2*, *TOP2A*, and *TOP2B*) to assess their effect on prognosis. We can confirm the prognostic value of *MKI67*, *CCND2*, *CCND3*, *CCNE2*, and *CDKN1A*, as well as *TK2*. Both *TOP2A* and *TOP2B* had a very high predictive power. However, the results of *CCND1*, *CCNE1*, *CDKN1B*, and *TK1* were (although partially significant) not convincing. Therefore, we suggest the use of above prognostic genes as measured using microarrays.

We must note a limitation of our approach: the use of the median (or upper/lower quartile) sample for dividing the samples into high- and low-expression groups. In principle, a cutoff-free correlation analysis of gene expression and survival data is possible using Cox proportional hazard models. In this frame work, significance and hazard ratio could be assessed, but no survival curves for a good and a poor prognosis group could be drawn. The advantage of the use of the median for splitting is the negligible effect of outliers, which—due to the high dynamic range of the microarrays—could seriously skew the results when using the mean. Moreover, median enables to have high- and low-expression groups of virtually the same size which enables the drawing of robust Kaplan–Meier plot. The determination of an exact cutoff value for each transcript could improve the results. However, in this case the expression should be confirmed by independent methods like RT–PCR or immunohistochemistry to achieve a reliable correlation. Such a fine-tuning—as it has been done for the ESR1 gene in ovarian cancer [20]—must be performed for each gene individually and is therefore not in the scope of present study.

Finally, we added a visual enhancement to the service to help to select the probe sets with the best quality. The individual probes on the probe selection drop-down menu have background color, which represents the estimated quality of the probe set: green = excellent (average expression of the probe set in the 1,809 patients is over 500), red = not reliable (average expression < 100 and maximal expression < 500), yellow = intermediate (all other probes).

In summary, we developed a web service which is capable evaluating the prognostic utility of 22,277 genes in 1,809 breast cancer patients and specifically used this tool to evaluate the effect of the proliferation-related genes in breast cancer prognosis. The integrative genomic analysis

is still evolving; thus future integration of additional forms of data such as sequence, location, or copy number variations might potentially add vital additional information which will enable us to deliver higher accuracy in prognosis prediction.

References

1. Amat S, Penault-Llorca F, Cure H et al (2002) Scarff-Bloom-Richardson (SBR) grading: a pleiotropic marker of chemosensitivity in invasive ductal breast carcinomas treated by neoadjuvant chemotherapy. Int J Oncol 20:791–796
2. Ravdin PM, Siminoff LA, Davis GJ et al (2001) Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 19:980–991
3. Olivotto IA, Bajdik CD, Ravdin PM et al (2005) Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J Clin Oncol 23:2716–2725
4. Harris L, Fritsche H, Mennel R et al (2007) American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 25:5287–5312
5. Paik S, Shak S, Tang G et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817–2826
6. Draghici S, Khatri P, Eklund AC et al (2006) Reliability and reproducibility issues in DNA microarray measurements. Trends Genet 22:101–109
7. Shi L, Reid LH, Jones WD et al (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24:1151–1161
8. Gyorffy B, Molnar B, Lage H et al (2009) Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. PLoS One 4:e5645
9. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
10. Colozza M, Azambuja E, Cardoso F et al (2005) Proliferative markers as prognostic and predictive tools in early breast cancer: where are we now? Ann Oncol 16:1723–1739
11. Tan PK, Downey TJ, Spitznagel EL Jr et al (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 31:5676–5684
12. Győrffy B, Schäfer R (2009) Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. Breast Cancer Res Treat 118(3):433–441
13. Gautier L, Cope L, Bolstad BM et al (2004) affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315
14. Sims AH, Smethurst GJ, Hey Y et al (2008) The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets—improving meta-analysis and prediction of prognosis. BMC Med Genomics 1:42
15. Gong Y, Yan K, Lin F et al (2007) Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study. Lancet Oncol 8:203–211
16. Dunnwald LK, Rossing MA, Li CI (2007) Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Res 9:R6
17. Lacroix M, Querton G, Hennebert P et al (2001) Estrogen receptor analysis in primary breast tumors by ligand-binding assay, immunocytochemical assay, and northern blot: a comparison. Breast Cancer Res Treat 67:263–271
18. Pusztai L, Ayers M, Stec J et al (2003) Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. Clin Cancer Res 9:2406–2415
19. Paik S, Tang G, Shak S et al (2006) Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 24:3726–3734
20. Darb-Esfahani S, Wirtz RM, Sinn BV, Budczies J, Noske A, Weichert W, Faggad A, Scharff S, Sehouli J, Oskay-Ozcelik G, Zamagni C, De Iaco P, Martoni A, Dietel M, Denkert C (2009) Estrogen receptor 1 mRNA is a prognostic factor in ovarian carcinoma: determination by kinetic PCR in formalin-fixed paraffin-embedded tissue. Endocr Relat Cancer 16(4):1229–1239
21. Bos PD, Zhang XH, Nadal C et al (2009) Genes that mediate breast cancer metastasis to the brain. Nature 459:1005–1009
22. Desmedt C, Giobbie-Hurder A, Neven P et al (2009) The Gene expression Grade Index: a potential predictor of relapse for endocrine-treated breast cancer patients in the BIG 1–98 trial. BMC Med Genomics 2:40
23. Zhang Y, Sieuwerts AM, McGreevy M et al (2009) The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. Breast Cancer Res Treat 116:303–309
24. Schmidt M, Bohm D, von TC et al (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res 68:5405–5413
25. Loi S, Haibe-Kains B, Desmedt C et al (2008) Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. BMC Genomics 9:239
26. Desmedt C, Piette F, Loi S et al (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res 13:3207–3214
27. Loi S, Haibe-Kains B, Desmedt C et al (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. J Clin Oncol 25:1239–1246
28. Minn AJ, Gupta GP, Padua D et al (2007) Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci USA 104:6740–6745
29. Ivshina AV, George J, Senko O et al (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res 66:10292–10301
30. Miller LD, Smeds J, George J 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
31. Sotiriou C, Wirapati P, Loi S et al (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98:262–272
32. Wang YX, Klijn JGM, Zhang Y et al (2005) Gene-expression pro-files to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365:671–679
33. Pawitan Y, Bjohle J, Amler L et al (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7:R953–R964