Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients

Balázs Győrffy¹,², András Lánczy³ and Zoltán Szállásí⁴,⁵

¹Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of Sciences, Budapest, Hungary
²Laboratory of Functional Genomics, Institute of Pathology, Charité, Berlin, Germany
³2nd Department of Pathology, Semmelweis University Budapest, Budapest, Hungary
⁴Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark
⁵Children’s Hospital, Harvard Medical School, Boston, Massachusetts, USA

(Correspondence should be addressed to B Győrffy who is now at First Department of Pediatrics, Semmelweis University, Bókay János u. 53-54, H-1083 Budapest, Hungary; Email: gyorffy@kmplot.com)

Abstract

The validation of prognostic biomarkers in large independent patient cohorts is a major bottleneck in ovarian cancer research. We implemented an online tool to assess the prognostic value of the expression levels of all microarray-quantified genes in ovarian cancer patients. First, a database was set up using gene expression data and survival information of 1287 ovarian cancer patients downloaded from Gene Expression Omnibus and The Cancer Genome Atlas (Affymetrix HG-U133A, HG-U133A 2.0, and HG-U133 Plus 2.0 microarrays). After quality control and normalization, only probes present on all three Affymetrix platforms were retained (n=22,277). To analyze the prognostic value of the selected gene, we divided the patients into two groups according to various quantile expressions of the gene. These groups were then compared using progression-free survival (n=1090) or overall survival (n=1287). A Kaplan–Meier survival plot was generated and significance was computed. The tool can be accessed online at www.kmplot.com/ovar. We used this integrative data analysis tool to validate the prognostic power of 37 previously published biomarkers on ovarian cancer prognosis.

Endocrine-Related Cancer (2012) 19 197–208

Introduction

With a mortality of 8.4 per 100,000 women, ovarian cancer is the most common cause of death among gynecological malignancies (http://seer.cancer.gov) with a 5-year survival rate of 10–30%. Relative to breast cancer, the molecular characteristics of epithelial ovarian cancer (EOC) are more heterogeneous. Despite extensive research, clinical–pathological factors including tumor stage, residual disease after surgery, histological type, and tumor grade are still the most important features related to patient outcome. To date, only two biomarkers have been approved by the Food and Drug Administration (FDA) for monitoring patients with EOC: CA125 (MUC16; Gadducci et al. 1995, 2004, Cooper et al. 2002, Riedinger et al. 2006) and HE4 (WFDC2; Huhtinen et al. 2009, Moore et al. 2009, 2010).

Several additional genes have been suggested as potential biomarkers for the progression of EOC. Low expression of p21 (Ferrandina et al. 2000, Plisiecka-Halasa et al. 2003, Bali et al. 2004), bax...


Given the large number of potential biomarkers for EOC, the immediate challenge is to validate the most robust candidates eligible for further investigation. Recent advances in genomic technologies together with powerful bioinformatic tools can enable us to deliver this prerequisite. We recently developed an online biomarker validation tool using microarray data of 2000 breast cancer patients (Gyorffy et al. 2010). In this, the expression of a selected gene can be used to split patients into groups, and the proportional survival of these groups is compared to each other.

In this study, our aim was to implement an online survival analysis tool for the rapid assessment of prognosis-related genes in ovarian cancer and to test the validity of previously proposed biomarkers. Furthermore, we also developed additional analysis options including the computation of multigenic prognosis predictors and the option of grouping patients based on applied treatment protocols.

Materials and methods

Collection of ovarian cancer microarray data sets

We searched Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) and The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov) to identify data sets suitable for the analysis. In this, the keywords ‘ovarian’, ‘cancer’, ‘survival’, ‘gpl96’, ‘gpl570’, and ‘gpl571’ were used. Only publications with available raw microarray gene expression data, clinical survival information, and at least 20 patients were included. Only three microarray platforms, GPL96 (Affymetrix HG-U133A), GPL570 (Affymetrix HG-U133 Plus 2.0), and GPL571/GPL3921 (Affymetrix HG-U133A 2.0), were considered because they are frequently used and because these particular arrays have 22 277 probe sets (representing 13 435 unique genes) in common. The use of almost identical platforms and identical probe sets is vital because different platforms for gene expression profiling measure expression of the same gene with varying accuracy, on different relative scales, and with diverse dynamic ranges (Tan et al. 2003). Finally, we controlled all samples using the ranked expression of all genes to identify repeatedly published microarrays.

Setup of server for online survival calculation

The raw.CEL files were MAS5 normalized in the R statistical environment (www.r-project.org) using the affy Bioconductor library (Gautier et al. 2004). MAS5 can be applied to individual chips, making future extensions of the database uncomplicated. Furthermore, MAS5 ranked among the best normalization methods when compared with the results of RT-PCR measurements in our recent study (Gyorffy et al. 2009). For the analysis, only probes measured on GPL96, GPL570, and GPL571/GPL3921 were retained (n = 22 277). At this stage, we performed a second scaling normalization to set the average expression on each chip to 1000. Although this technique cannot remove all, but it can significantly reduce batch effects (Sims et al. 2008). We integrated the gene expression and clinical data using PostgreSQL, an open-source object-relational database system (www.postgresql.org). Data security is ensured through PostgreSQL...
permissions that are imposed on individual tables in the project databases.

The KMplot web application can be reached in a platform-independent user interface. The interactivity of the service is increased by the usage of JavaScript and Ajax technologies. The server is hosted on Debian Linux (www.debian.org) and is powered by Apache (www.apache.org). The server-side scripts were developed in hypertext preprocessor (PHP), which controls the analysis requests and delivers the results. Open Database Connectivity is used as a middleware layer between the R and the PostgreSQL database via the RODBC package (cran.r-project.org/package=RODBC). 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 (HR; and 95% confidence intervals) and logrank $P$ are calculated and displayed. The central server for the Kaplan–Meier plotter for ovarian cancer can be reached at www.kmplot.com/ovar.

**Probe set options**

We also implemented a set of probe-set-related options, including the option to use all probe sets available for a given gene on the microarray simultaneously and to use a combined expression of several probe sets. Using this option, it is possible to assess the effect of the mean expression of gene combinations on survival.

In addition, bee swarm plot can be drawn using the beeswarm package (www.cbs.dtu.dk/~eklund/beeswarm/). The bee swarm plot is capable of visualizing gene expression as nonoverlapping points in a one-dimensional scatter plot. A bee swarm plot can be used to quickly identify outlier samples and genes with bimodal distribution.

**Validation of previously published EOC biomarkers**

A PubMed search was performed using the keywords ‘ovarian cancer’, ‘survival’, ‘biomarker’, and ‘gene expression’ to identify genes described in the literature as potential EOC biomarkers. Then, using PubMed gene, we added a unique gene symbol for each of the genes and identified the corresponding Affymetrix probe set IDs. The capability of these genes to predict survival was measured by using the probe set IDs in the online analysis tool.

In the combination of several markers, their mean expression is first computed for each sample. Then, the median of these is used for splitting the patients into cohorts during the analysis.

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>Reference</th>
<th>GEO platform</th>
<th>GEO number</th>
<th>No. of samples in data set</th>
<th>Median overall survival</th>
<th>Death event</th>
<th>Stage (1/2/3)</th>
<th>Grade (1/2/3)</th>
<th>Serous/ endometrioid</th>
<th>Treatment contains platin (out of)</th>
<th>Treatment contains Taxol (out of)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE41764</td>
<td>Denkert et al. 2009</td>
<td>GPL570</td>
<td>149</td>
<td>80</td>
<td>35 (2)</td>
<td>68 (7)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>27/29 (4)</td>
<td>79/79 (29)</td>
</tr>
<tr>
<td>GSE15622</td>
<td>Ahmed et al. 2007</td>
<td>GPL571</td>
<td>152</td>
<td>35</td>
<td>28 (7)</td>
<td>27 (0)</td>
<td>0/26/9</td>
<td>0/26/9</td>
<td>NA</td>
<td>20/35 (35)</td>
<td>15/35 (35)</td>
</tr>
<tr>
<td>GSE19829</td>
<td>Konstantinopoulos</td>
<td>GPL570</td>
<td>197</td>
<td>28</td>
<td>17 (5)</td>
<td>35 (0)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>GSE3149</td>
<td>Bild et al. 2006</td>
<td>GPL570</td>
<td>224</td>
<td>116</td>
<td>6 (9)</td>
<td>30 (4)</td>
<td>22 (0)</td>
<td>11 (0)</td>
<td>NA</td>
<td>53/22 (53)</td>
<td>NA (22)</td>
</tr>
<tr>
<td>GSE8931</td>
<td>Telli et al. 2008</td>
<td>GPL570</td>
<td>280</td>
<td>285</td>
<td>10 (0)</td>
<td>23 (0)</td>
<td>22 (2)</td>
<td>41 (0)</td>
<td>NA</td>
<td>53/22 (53)</td>
<td>NA (22)</td>
</tr>
<tr>
<td>GSE21040</td>
<td>Mok et al. 2009</td>
<td>GPL570</td>
<td>321</td>
<td>185</td>
<td>12 (8)</td>
<td>38 (3)</td>
<td>77 (0)</td>
<td>127 (0)</td>
<td>NA</td>
<td>90/85 (90)</td>
<td>NA (85)</td>
</tr>
<tr>
<td>TCGA_2011</td>
<td>TCGA_2011</td>
<td>GPL570</td>
<td>458</td>
<td>505</td>
<td>247 (147)</td>
<td>242 (142)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>233/452 (233)</td>
<td>458/473 (458)</td>
</tr>
<tr>
<td>NA</td>
<td>data not available</td>
<td>GPL570</td>
<td>285</td>
<td>1027</td>
<td>622 (285)</td>
<td>1287 (622)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>674/1121 (674)</td>
<td>11958 (674)</td>
</tr>
</tbody>
</table>

Table 1: Clinical properties of the ovarian cancer patients used in the analysis.

NA, data not available; NA, total number of patients with available clinical data.
Table 2 The association between prognostic markers and progression-free survival. The patients were divided into two groups as having higher or lower expression as compared to the median. The markers were analyzed in subsets of patients with equivalent clinical characteristics to the cohorts in which the association has previously been described.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Gene</th>
<th>Reference</th>
<th>Survival</th>
<th>Analyzed in the cohort of</th>
<th>Affymetrix ID</th>
<th>Q</th>
<th>HR</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA(MUC 16)125</td>
<td>CA 125</td>
<td>Gadducci et al. 1995, Cooper et al. 2002, Gadducci et al. 2004, Riedinger et al. 2006</td>
<td>PFS</td>
<td>All patients</td>
<td>220196_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>201384_s_at</td>
<td>1.3</td>
<td>0.0003*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>201383_s_at</td>
<td>1.4</td>
<td>3.7 x 10^{-5},a</td>
<td></td>
</tr>
<tr>
<td>KRT19</td>
<td>Cytokeratin 19</td>
<td>Tempfer et al. 1998, Gadducci et al. 2001</td>
<td>PFS</td>
<td>Debulk = subopt.</td>
<td>201650_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>KLK6</td>
<td>Kallikrein 6</td>
<td>Diamandis et al. 2003</td>
<td>PFS</td>
<td>All patients</td>
<td>216699_s_at</td>
<td>2</td>
<td>0.79</td>
<td>0.002*</td>
</tr>
<tr>
<td>KLK10</td>
<td>Kallikrein 10</td>
<td>Luo et al. 2001</td>
<td>PFS</td>
<td>Stage = 3 + 4</td>
<td>209792_s_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>IL6</td>
<td>Interleukin 6</td>
<td>Scambia et al. 1995</td>
<td>OS</td>
<td>All patients</td>
<td>205207_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>IL7</td>
<td>Interleukin 7</td>
<td>Lambeck et al. 2007</td>
<td>OS</td>
<td>All patients</td>
<td>206693_at</td>
<td>3</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>IFNG</td>
<td>γ-Interferon</td>
<td>Marth et al. 2004</td>
<td>PFS</td>
<td>All patients</td>
<td>210354_at</td>
<td>3</td>
<td>0.81</td>
<td>0.004*</td>
</tr>
<tr>
<td>FAS</td>
<td>sFas</td>
<td>Hefler et al. 2000, Konno et al. 2000</td>
<td>PFS</td>
<td>All patients</td>
<td>207800_s_at</td>
<td>1</td>
<td>1.2</td>
<td>0.017</td>
</tr>
<tr>
<td>VEGFR</td>
<td>VEGFR</td>
<td>Hefler et al. 2006</td>
<td>OS</td>
<td>All patients</td>
<td>20834_at</td>
<td>2</td>
<td>1.2</td>
<td>0.064</td>
</tr>
<tr>
<td>CCND1</td>
<td>Cyclin D1</td>
<td>Bali et al. 2004, Barbieri et al. 2004</td>
<td>OS</td>
<td>Stage = 3 + 4</td>
<td>208711_s_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>CCND3</td>
<td>Cyclin D3</td>
<td>Levidou et al. 2007</td>
<td>OS</td>
<td>All patients</td>
<td>201700_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>P(CDK N2B)15</td>
<td>p15</td>
<td>Kudoh et al. 2002</td>
<td>PFS</td>
<td>All patients</td>
<td>209112_at</td>
<td>1</td>
<td>1.4</td>
<td>5.4 x 10^{-5,a}</td>
</tr>
<tr>
<td>P(CDK N2A)16</td>
<td>p16</td>
<td>Katsaros et al. 2004, Kommoss et al. 2007</td>
<td>PFS</td>
<td>Debulk = subopt.</td>
<td>207039_at</td>
<td>2</td>
<td>0.66</td>
<td>0.002*</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>p21</td>
<td>Ferrandina et al. 2000, Plisiecka-Halasa et al. 2003, Bali et al. 2004</td>
<td>PFS</td>
<td>Histology = serous</td>
<td>209644_x_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>CDKN1B</td>
<td>p27</td>
<td>Newcomb et al. 1999, Masciullo et al. 2000, Korkolopoulou et al. 2002, Schmider-Ross et al. 2006</td>
<td>PFS</td>
<td>All patients</td>
<td>205034_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>RB1</td>
<td>pRB</td>
<td>Dong et al. 1997, Konstantinidou et al. 2003</td>
<td>OS</td>
<td>Stage = 1</td>
<td>203132_at</td>
<td>1</td>
<td>1.4</td>
<td>5.4 x 10^{-5,a}</td>
</tr>
<tr>
<td>E2F1</td>
<td>E2F1</td>
<td>Suh et al. 2008</td>
<td>PFS</td>
<td>All patients</td>
<td>2028_s_at</td>
<td>1</td>
<td>0.83</td>
<td>0.017</td>
</tr>
<tr>
<td>E2F2</td>
<td>E2F2</td>
<td>Reimer et al. 2007</td>
<td>PFS</td>
<td>All patients</td>
<td>207042_at</td>
<td>3</td>
<td>0.86</td>
<td>0.037</td>
</tr>
<tr>
<td>E2F4</td>
<td>E2F4</td>
<td>Reimer et al. 2007</td>
<td>PFS</td>
<td>All patients</td>
<td>202248_at</td>
<td>3</td>
<td>0.85</td>
<td>0.034</td>
</tr>
<tr>
<td>TP53</td>
<td>p53</td>
<td>Buttitta et al. 1997, Reles et al. 2001</td>
<td>PFS</td>
<td>Stage = 3 + 4</td>
<td>211300_s_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>TP73</td>
<td>p73</td>
<td>Becker et al. 2006</td>
<td>OS</td>
<td>All patients</td>
<td>210746_at</td>
<td>1</td>
<td>0.84</td>
<td>0.075</td>
</tr>
<tr>
<td>BAX</td>
<td>bax</td>
<td>Tai et al. 1998, Skirnisdottir et al. 2001</td>
<td>PFS</td>
<td>Therapy = contains Taxol</td>
<td>208478_s_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

Downloaded from Bioscientifica.com at 01/08/2019 09:36:00AM via free access
Results

Construction of combined ovarian cancer microarray database

We identified 1287 unique patients in eight data sets meeting our criteria in GEO and TCGA. In the GSE3149 data set, we found two samples repeatedly published (GSM70546 = GSM70547 and GSM70511 = GSM70512). As for these samples the unique recognition of the appropriate clinical information was not possible, we removed them from the final database. Of the above, 72% have serous and 2% have endometrioid tumors. Patients are distributed across stage 1 ($n = 34, 3.6\%$), stage 2 ($n = 122, 13\%$), stage 3 ($n = 672, 71.7\%$), and stage 4 ($n = 109, 11.6\%$). Debulking was optimal (residual tumor $< 1 \text{ cm}$) in 674 out of 1119 patients.

The median overall survival is 31.0 months, 1090 patients have progression-free survival data, and 1287 have overall survival data. (note: some publications report 'disease-free survival' (Konstantinopoulos et al. 2010) or 'relapse-free survival' (Tothill et al. 2008) instead of 'progression-free survival'. These were merged as 'progression-free survival' to enable a meta-analysis of the complete database.). A summary of the clinical characteristics of the patients in each data set used in the analysis is shown in Table 1.

Table 2 continued

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Gene</th>
<th>Reference</th>
<th>Survival</th>
<th>Analyzed in the cohort of</th>
<th>Affymetrix ID</th>
<th>Q</th>
<th>HR</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCL2L1</td>
<td>Bcl-xl</td>
<td>Materna et al. 2007</td>
<td>PFS</td>
<td>All patients</td>
<td>212312_at</td>
<td>1</td>
<td>0.86</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>215037_s_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>206665_s_at</td>
<td>3</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>210334_x_at</td>
<td>2</td>
<td>0.75</td>
<td>0.00017*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>202076_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>202094_at</td>
<td>2</td>
<td>0.84</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>202095_s_at</td>
<td>1</td>
<td>0.84</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>207199_at</td>
<td>3</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>BIRC2</td>
<td>cIAP</td>
<td>Psysri et al. 2006</td>
<td>OS</td>
<td>Stage 3 + 4</td>
<td>201983_s_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>BIRC5</td>
<td>Survivin</td>
<td>Sui et al. 2002</td>
<td>PFS</td>
<td>All patients</td>
<td>216836_s_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>TERT</td>
<td>hTERT</td>
<td>Brustmann 2005</td>
<td>OS</td>
<td>Histology = serous</td>
<td>203510_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>EGFR</td>
<td>ERBB1</td>
<td>Skirimadottir et al. 2004, Psysri et al. 2005</td>
<td>PFS</td>
<td>Stage 1 + 2</td>
<td>203936_s_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>202076_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>211550_at</td>
<td>3</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>211551_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>210930_s_at</td>
<td>3</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>211607_x_at</td>
<td>3</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>217828_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>203892_at</td>
<td>1</td>
<td>0.33</td>
<td>0.05</td>
</tr>
<tr>
<td>ERBB2</td>
<td>ERBB2</td>
<td>Lassus et al. 2004</td>
<td>PFS</td>
<td>Histology = serous</td>
<td>201984_s_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>MET</td>
<td>c-Met</td>
<td>Sawada et al. 2007</td>
<td>OS</td>
<td>Stage 3 + 4</td>
<td>204855_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>MMP2</td>
<td>MMP-2</td>
<td>Torng et al. 2004</td>
<td>PFS</td>
<td>Histology = endom.</td>
<td>204531_s_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>MMP9</td>
<td>MMP-9</td>
<td>Sillanpaa et al. 2007</td>
<td>OS</td>
<td>Stage 1</td>
<td>203892_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>MMP14</td>
<td>MT1-MMP</td>
<td>Kamat et al. 2006</td>
<td>OS</td>
<td>Stage 2 + 3 + 4</td>
<td>203892_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>WFDC2 (HE4)</td>
<td>Epididymis protein 4</td>
<td>Huhtinen et al. 2009, Moore et al. 2009, 2010</td>
<td>PFS</td>
<td>All patients</td>
<td>203892_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>SERPINB5</td>
<td>Maspin</td>
<td>Secord et al. 2006</td>
<td>PFS</td>
<td>Debulk = subopt.</td>
<td>204855_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>BRCA1</td>
<td>BRCA1</td>
<td>Thrall et al. 2006</td>
<td>OS</td>
<td>All patients</td>
<td>211851_x_at</td>
<td>3</td>
<td>0.82</td>
<td>0.01</td>
</tr>
<tr>
<td>ERCC1</td>
<td>ERCC1</td>
<td>Darcy &amp; Tian 2007</td>
<td>PFS</td>
<td>Stage 3</td>
<td>204531_s_at</td>
<td>2</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Therapy = Tax + Plat</td>
<td>203719_s_at</td>
<td>1</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

PFS, progression-free survival; OS, overall survival, HR, hazard ratio; Q, quality score for the probe set as measured after normalization across the entire data set (1, average expression over 500 or maximal expression over 1000; 2, intermediate probes; 3, average expression below 100); NS, not significant (significance over 0.05). *bold indicates $P < 0.005$.

See Kaplan–Meier plots in Fig. 1.

Results

Construction of combined ovarian cancer microarray database

We identified 1287 unique patients in eight data sets meeting our criteria in GEO and TCGA. In the GSE3149 data set, we found two samples repeatedly published (GSM70546 = GSM70547 and GSM70511 = GSM70512). As for these samples the unique recognition of the appropriate clinical information was not possible, we removed them from the final database. Of the above, 72% have serous and 2% have endometrioid tumors. Patients are distributed across stage 1 ($n = 34, 3.6\%$), stage 2 ($n = 122, 13\%$), stage 3 ($n = 672, 71.7\%$), and stage 4 ($n = 109, 11.6\%$). Debulking was optimal (residual tumor $< 1 \text{ cm}$) in 674 out of 1119 patients. The median overall survival is 31.0 months, 1090 patients have progression-free survival data, and 1287 have overall survival data. (note: some publications report ‘disease-free survival’ (Konstantinopoulos et al. 2010) or ‘relapse-free survival’ (Tothill et al. 2008) instead of ‘progression-free survival’. These were merged as ‘progression-free survival’ to enable a meta-analysis of the complete database.). A summary of the clinical characteristics of the patients in each data set used in the analysis is shown in Table 1.

www.endocrinology-journals.org
Setup of online survival analysis platform

The Kaplan–Meier plot shows the association between the investigated marker and survival in which the samples are grouped according to the median (or upper or lower quartile) expression of the selected gene. Before running the analysis, the patients can be filtered using stage, histology, grade, and treatment parameters including debulking status and applied chemotherapy. In addition, as an alternative to progression-free survival, overall survival can also be investigated.

Since there is an already established biomarker (CA125), a clinician might be interested in a specific clinical cohort of patients having low CA125 levels. Therefore, we added an additional filtering option in which only patients having an average CA125 expression (average of the two reliable probe sets) below the lower quartile of all patients are included. We must note that while this study measured tissue levels of CA125, the FDA-approved test for ovarian cancer is serum based.

Validation of previously published EOC biomarkers

Markers of ovarian cancer prognosis have been identified using literature search. We computed Kaplan–Meier plots for 37 proposed biomarkers to assess their effect on prognosis (for the complete results see Table 2 and Fig. 1). All biomarkers were investigated in the same cohort in which they were discovered. High significance was achieved for CA125, KLK6, IFNG, P15, P16, CDKN1B, and BIRC5. In addition, we have also run the analysis for predicting progression-free survival in all patients.

In an effort to improve accuracy, a pair-wise combination of the three best performing probe sets was assessed independently. The combination of CA125 and CDKN1B with a cutoff at the lower quartile resulted in classification significance superior to the power of the markers independently (HR = 1.5 and \(P = 5.4 \times 10^{-6}\) vs HR = 1.4, \(P = 3.7 \times 10^{-5}\) and HR = 1.4, \(P = 5.4 \times 10^{-5}\) for CA125 and CDKN1B, respectively, see Table 2 and Fig. 1).

Discussion

The validation of prognostic biomarkers is a major bottleneck in ovarian cancer research. Here, we combined multiple large microarray data sets to increase the statistical power for a meta-analysis of 22 277 genes. We developed a freely accessible online tool to estimate the prognostic value of any selected gene in a large cohort of clinical patients. After dividing

![Figure 1](image_url)
the patients into two groups based on the expression of the selected gene, a Kaplan–Meier plot is generated. The implemented computations are performed in real time on our server. This enables seamless future extension using new data sets or new filtering options.

We have integrated data sets from GEO and TCGA – ~40% of samples used by www.kmplot.com/ovar are from the TCGA repository. For the TCGA samples alone, there is an option to perform analyses in the caIntegrator website (https://caintegrator.nci.nih.gov). The samples in TCGA are open access or restricted (access is granted to NIH staff and to eRA Commons principal investigators) – however, the Affymetrix HG-U133 microarray profiles for the ovarian cancer patients are publicly available. We plan to continuously incorporate new GEO data sets as well as new TCGA samples in www.kmplot.com.

In contrast to breast cancer, where several already approved markers are in clinical use, in ovarian cancer only minimal progress has been made in recent years. When investigating the previously proposed biomarkers, we found that only few genes are actually capable of predicting outcome in our combined data set: CA125, P15, KLK6, IFNG, P16, CDKN1B, and BIRC5. Of these, CA125 and CDKN1B resulted in very robust significance. These results may reflect the high genetic heterogeneity of ovarian cancer (Gyorffy et al. 2008) and emphasize the importance of potential improvements in prognosis.

The most extensively studied marker for EOC is CA125 (Gadducci et al. 1995, 2004, Cooper et al. 2002, Riedinger et al. 2006), and determining its concentration in serum is essential for monitoring ovarian cancer progression. Fifty percent increase in serum CA125 level has been correlated to progression, and present progression definition of the Gynecological Cancer Intergroup defines progression based on two elevated serum CA125 levels. According to our results, tumor level of CA125 gene was able to predict later clinical outcome. Notably, we observed two different probe sets representing CA125 as significant. A third probe set did not show significant prognostic power, but it has also displayed low quality in terms of average expression as compared to the probe sets with significant prognostic power.

The role of the cell cycle control gene p27 (CDKN1B) as a prognostic marker in ovarian cancer was suggested in several studies (Newcomb et al. 1999, Masciullo et al. 2000, Korkolopoulou et al. 2002, Schmider-Ross et al. 2006). In addition, numerous recent analyses also confirmed its role in 205 (Lee et al. 2011), 131 (Skirnisdottir et al. 2011), and 339 (Duncan et al. 2010) patients. p27 is measured by only one probe set on the microarrays, and this probe set delivered high prognostic power in our analysis.

P15 is a tumor suppressor gene previously associated with ovarian cancer progression in 45 patients (Kudoh et al. 2002). The methylation status of P15 has also been investigated but was not an independent prognostic factor in 145 patients (Tam et al. 2007). One of the probe sets measuring P15 (212857_x_at) delivered a high prognostic potential.

To this point, we have investigated the prognostic power of individual probe sets. However, recent reports based on genomic technologies use not only single selected genes, but also a combination of these. In addition, some of the markers are not related to ovarian cancer prognosis in general, but have discriminative potential in one of the subgroups, or are related to different treatment regimens. While the evaluation of all potential markers and all eligible combinations is beyond the scope of this study, our online tool was set up exactly to enable researchers to perform these tests on our database.

In summary, we reviewed previously reported biomarkers of ovarian cancer prognosis and assessed their performance in a meta-analysis of 1297 ovarian cancer patients. We also developed an online biomarker validation platform to mine all available microarray data to assess the prognostic power of 22 277 genes.

**Declaration of interest**

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

**Funding**

BGyörfy study was supported by the OTKA PD 83154, by the TAMOP-4.2.1.B-09/1/KMR-2010-0001, by the ETT 029/2009 grant and by the Alexander von Humboldt Stiftung. Z Szállási was supported by the Breast Cancer Research Foundation.

**References**


Received in final form 17 January 2012
Accepted 24 January 2012
Made available online as an Accepted Preprint 25 January 2012