Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in Non-Small-Cell Lung Cancer

Balázs Györfy1*, Pawel Surowiak2, Jan Budczies3, András Lánczy1

1 Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of Sciences, Budapest, Hungary, 2 Department of Histology and Embryology, Wroclaw Medical University, Wroclaw, Poland, 3 Institut für Pathologie, Charité – Universitätsmedizin Berlin, Berlin, Germany

Abstract

In the last decade, optimized treatment for non-small cell lung cancer had lead to improved prognosis, but the overall survival is still very short. To further understand the molecular basis of the disease we have to identify biomarkers related to survival. Here we present the development of an online tool suitable for the real-time meta-analysis of published lung cancer microarray datasets to identify biomarkers related to survival. We searched the caBIG, GEO and TCGA repositories to identify samples with published gene expression data and survival information. Univariate and multivariate Cox regression analyses were performed to pinpoint those which could serve as indicators of the tumor's biological state. In essence there are two possible scenarios for this: either a single gene or a signature. Of these, survival was best predicted by CDK1 (p = 1E-16), CD24 (p = 1E-16) and CADM1 (p = 7E-12) in adenocarcinomas and by CCNE1 (p = 2.3E-09) and VEGF (p = 3.3E-10) in all NSCLC patients. Additional genes significantly correlated to survival include RAD51, CDKKN2A, OPN, EZH2, ANXA3, ADAM28 and ERCC1. In summary, we established an integrated database and an online tool capable of uni- and multivariate analysis for in silico validation of new biomarker candidates in non-small cell lung cancer.

Introduction

Although lung cancer treatment options have improved significantly in the last decade leading to better survival for patients with every stage of the disease, it is still leading cancer related deaths in the United States with 160 thousand deaths each year [1]. With approximately 85% of all cases the most common type of lung cancer is non-small cell lung cancer (NSCLC), which includes adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and bronchioloalveolar carcinoma [2]. Similarly to other cancer entities we can expect new molecular subtypes to emerge in the future, as it is now well accepted that the light microscopy based histologic subdivision uses only one of many phenotypic manifestations of the genetic changes that underlie lung cancer development [2].

The identification of genes whose altered expression is associated with survival differences might enclose the knowledge to pinpoint those which could serve as indicators of the tumor's biological state. In essence there are two possible scenarios for this: such biomarker can either be an individual gene or a signature comprising a set of genes. While numerous individual genes associated with survival have been published in the last thirty years, new microarray-based multigene molecular prognostic models using genomic signatures have only emerged in the last ten years [3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19]. A prerequisite for the reproducibility of such genomic signatures is the availability of raw data, which was only ensured by publications of the last six years [9,10,11,12,13,14,15,16,17,18]. Remarkably, in two cases not the signature as a whole, but genes as each individually important prognostic markers have been identified [15,19].

The initial discovery of a prognostic marker must be followed by several validation studies. Then, the results of these are usually synthesized in a meta-analysis including a large number of preferably more than thousand patients. Here, by uniting relevant data from several studies, statistical power is increased and more accurate estimates can be achieved. Several previous meta-analyses endeavored to perform such a meta-analysis of previous studies for solitary gene candidates including VEGF [20], MMP9 [21], cyclin E [22], survivin [23] and CDK1 [24].

Here, we integrated available genome-level transcriptomic datasets and then used this database to perform a meta-analysis of previously suggested survival associated biomarker-candidates. We also set up a global portal for such meta-analysis enabling express validation of new candidates without large-scale bioinformatic effort in an automated framework.


Competing Interests: The authors have declared that no competing interests exist.

* E-mail: zsalab2@yahoo.com
Survival Analysis in Lung Cancer

Materials and Methods

Construction of lung cancer microarray database

We explored the Cancer Biomedical Informatics Grid (caBIG, http://cabig.cancer.gov/), microarray samples are published in the caArray project), the Gene Expression Omnibus ( GEO, http://www.ncbi.nlm.nih.gov/geo/) and The Cancer Genome Atlas (TCGA, http://cancergenenome.nih.gov) to identify lung cancer datasets using the keywords “lung”, “cancer”, “small-cell”, “NSCLC”, “survival”, “GPL96”, “GPL3921” and “GPL570” (and the alternative names of the microarray platforms). The search was restricted to publications with simultaneously available microarray gene expression data and published clinical characteristics including survival. To test randomness, a pairwise rank test was performed for the collected clinical data including age, sex, smoking history, histology, stage, grade, success of surgery, radiotherapy and applied chemotherapy for all patients in WinStat 2013. For the pairwise rank test, the samples were first sorted according to datasets. Then, each sample (“X”) in the series was compared with all values which occur later in the list of all samples (“Y”). Assuming randomness, the probability of X>Y compared with all values which occur later in the list of all samples is 0.5. The correlations between clinical variables and survival were investigated and Kaplan-Meier plots for these were plotted using WinStat 2013. Among the different microarray platforms, Affymetrix HG-U133A (GPL96), HG-U133 Plus 2.0 (GPL570) and HG-U133A 2.0 (GPL3921) were included, because these are regularly used and because these arrays have 22,277 probe sets in common. The use of the same probe sets enables to measure the same gene with similar accuracy, relative scale and dynamic range.

To avoid potential bias due to array errors, we have performed a quality check for all arrays. In this, the background (between 19 and 218), the raw Q (between 0.5 and 14), the percentage of present calls (over 30%), the presence of bioB/-C/-D- spikes, the GAPDH 3’ to 5’ ratio (below 4.5) and the beta-actin 3’ to 5’ ratio (below 18) were checked. The threshold values correspond to the 95% range of the arrays as described previously [25]. Quality control was not possible for GSE4573 as for this dataset only the MAS5 normalized data was available. A filtering was added to the database to exclude potentially biased arrays. Additionally, we compared all microarray files using the ranked expression of all genes to spot microarrays re-published in different studies.

Set-up of server for online survival calculation

The unprocessed CEL files were MAS5 normalized in the R environment (http://www.r-project.org) using the simpleaffy library (http://bioinformatics.pirc.man.ac.uk/simpleaffy/). We have selected MAS5 for normalization as it ranked among the best normalization methods when contrasted to the results of RT-PCR measurements in our previous study [26]. Moreover, MAS5 can be applied to single arrays, enabling seamless future extensions of the database. For the complete database, only the common probes measured in all three array platforms were retained (n = 22,277). Then, a second scaling normalization was performed to center the mean expression for each array to 1000 - this technique can significantly reduce batch effects. Gene expression and clinical data were integrated using PostgreSQL, an open source object-relational database system (http://www.postgresql.org/). To assess the prognostic value of a gene, each percentile (of expression) between the lower and upper quartiles were computed and the best performing threshold was used as the final cutoff in a univariate Cox regression analysis. Histology, grade, stage, gender and smoking history can be used in the multivariate analysis. However, the multivariate analysis uses less patients as the univariate analysis because not each patients has all clinical information. Kaplan-Meier survival plot and the hazard ratio with 95% confidence intervals and logrank P value were calculated and plotted in R using the “survplot” function of the “survival” Bioconductor package. The R script used by the software to perform the Kaplan-Meier analysis and to identify the best cutoff is available as R script S1.

The entire computational pathway is made accessible for re-analysis in a platform independent online available software running on a Debian Linux (http://www.debian.org) server powered by Apache (http://www.apache.org). The scripts on the server-side were developed in PHP, these control the user interface, the requests and the delivery of the results. The RODBC package provides a middleware layer between R and the PostgreSQL database. This platform can be reached over the internet via http://www.kmplot.com/lung.

Validation of previously published survival associated biomarkers

A Pubmed search was performed to identify lung cancer survival associated biomarkers using all combinations of the keywords “lung cancer”, “NSCLC”, “adenocarcinoma”, “squamous cell carcinoma”, “survival”, “gene expression”, “signature” and “meta analysis”. Only studies published in English were included. Eligibility criteria also included the investigation of the biomarker in at least 50 patients - biomarkers described in experimental models only were omitted. For each gene/signature the exact conditions in which it was identified have been retrieved, and these have been used as filtering when selecting the patients for the survival analysis.

To visualize the performance of the various biomarkers in datasets including different number of patients, we have generated funnel plots depicting the hazard ratio (and confidence intervals) on the horizontal axis vs. the sample size on the vertical axis for each dataset. We also added an option to the online interface to simultaneously perform the analysis in each of the individual datasets. Finally, significance was set at p<0.01.

Results

Construction of combined lung cancer microarray database

We identified all together 1,715 patients, 1,120 in seven GEO datasets, 133 patients in TCGA and 462 patients in caArray. There were no samples repeatedly published. One sample (GSM370984) failed two parameters in the quality control - this array was excluded from all analyses. Additionally, in 215 arrays one parameter was out of the 95% range of all arrays - these arrays can be excluded from analyses by selecting the “exclude outlier arrays” in the online interface. Overall survival was published for 1,405 patients and time to first progression was published for 764 patients. We have collected age, sex, smoking history, histology, stage, grade, success of surgery, radiotherapy and applied chemotherapy for all patients - none of these parameters was significant in the pairwise rank test indicating random distribution of the data. A summary of these clinical properties for each dataset is presented in Table 1. The survival of the patients stratified by subtype, gender, smoking history and stage is presented in Figure 1.

Set-up of online survival analysis platform

We have employed Kaplan-Meier plots to visualize the association between the gene under investigation and survival. Before analysis, the patients were filtered using the available...
Validation of previously published NSCLC biomarkers

We identified 21 previously published survival associated individual genes and 7 gene expression signatures (listed in Table S1). Each of these biomarker candidates were investigated in a cohort having similar clinical characteristics as the patients in which they were originally described. For genes measured by several probe sets on the microarrays, those with the highest quality were used (high quality: average expression over 500 or maximal expression over 1000, low quality: average expression below 100, intermediate: all other probes). In case there were several high quality probes then the best performing was used. The analysis results are presented in Table 2 and Figure 2.

Discussion

The importance of cancer biomarkers is highlighted by the success of the HER2 gene in breast cancer. High HER2 expression was first a marker of worse survival, but the introduction of targeted anti-HER2 therapy changed the picture: today HER2 positive patients have an improved prognosis compared to women with HER2 negative disease [27].

Here, by using an integrated database of ten previously published transcriptomic datasets, we validated the association with survival for a set of genes in non-small-cell lung cancer. Generally, the strongest associations were found for those also investigated in a previous meta-analysis (VEGF, CCNE1 and CDK1). For all of these genes higher expression was associated with shorter survival. With over 5,000 patients, the meta-analysis for VEGF [20] employed the highest number of patients – our analysis also confirmed the correlation of VEGF expression and overall survival in NSCLC patients by both univariate and multivariate analyses. The importance of VEGF is due to the availability of targeted agents directly inhibiting its activation. Interestingly, for one of the genes (CDK1) a previous meta-analysis actually rejected a correlation between the gene and survival [24]. In contrast, our results represent a large-scale independent validation of the gene. In individual genes, only a few were associated with longer survival when displaying higher expression – these include CADM1, ANXA3, ADAM28, XIAP and XAF1. Future therapeutic targeting of these will only be possible using a different approach than for most genes in which higher expression actually results in shorter survival.

After surgery, about two-thirds of recurrences for early stage disease occur at distant sites. Therefore, the eradication of micrometastases must have a high priority as early as possible. A previous meta-analysis of all the trials investigating chemotherapy benefit demonstrated a 5% improvement in overall survival [28]. This survival advantage with chemotherapy was also maintained at 9 years of follow-up. For these reasons the use of adjuvant chemotherapy is the current standard of care for patients with early stage NSCLC. In our analysis system we have integrated the use of chemotheraphy to enable the validation of genes specifically related to survival in chemotherapy treated patients.

A major etiological factor for lung cancer is cigarette smoking which accounts for nearly 85% of all cases. Lung cancer...
HISTOLOGY

- Censors
- Adeno
- SCC
- Large

Overall survival (months)

n=440, p=1.3E-08

PFS (months)

n=161, p=0.0015

GENDER

- Censors
- Female
- Male

Overall survival (months)

n=788, p=9.2E-07

PFS (months)

n=636, p=0.16

STAGE

- Censors
- Stage 1
- Stage 2
- Stage 3

Overall survival (months)

n=300, p=6.4E-09

SMOKING HISTORY

- Censors
- Smokers
- Never smoked

Overall survival (months)

n=342, p=2.1E-07
Table 2. Performance of previously published biomarker candidates associated with survival in non-small-cell lung cancer.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Literature data</th>
<th>Meta-analysis results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>Ref.</td>
<td>n</td>
</tr>
<tr>
<td><strong>Genes examined in a meta-analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VEGF</td>
<td>[20] 5386</td>
<td>IHC, RT-PCR</td>
</tr>
<tr>
<td>MMP9</td>
<td>[21] 2029</td>
<td>IHC, RT-PCR</td>
</tr>
<tr>
<td>CCNE1 cyclin E</td>
<td>[22] 2606</td>
<td>IHC</td>
</tr>
<tr>
<td>BIRC5 survivin</td>
<td>[23] 2703</td>
<td>IHC, FISH, RT-PCR</td>
</tr>
<tr>
<td>CD2, CDK1</td>
<td>[24] 2731</td>
<td>IHC, RT-PCR</td>
</tr>
<tr>
<td><strong>Genes identified in original studies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CADM1</td>
<td>[15] 617 Array+IHC</td>
<td>ADE</td>
</tr>
<tr>
<td>CEA</td>
<td>[33] 97</td>
<td>IHC</td>
</tr>
<tr>
<td>RAD51</td>
<td>[34] 383</td>
<td>IHC</td>
</tr>
<tr>
<td>CDKN2A P16</td>
<td>[35] 106</td>
<td>IHC</td>
</tr>
<tr>
<td>OPN</td>
<td>[36] 25</td>
<td>IHC</td>
</tr>
<tr>
<td>CENK2A</td>
<td>[37] 82 RT-PCR</td>
<td>NSCLC surgical</td>
</tr>
<tr>
<td>EZH2</td>
<td>[38] 106</td>
<td>IHC</td>
</tr>
<tr>
<td>IFNAR2</td>
<td>[39] 113</td>
<td>IHC</td>
</tr>
<tr>
<td>ANXA3</td>
<td>[40] 125</td>
<td>MS, 2D-DIGE</td>
</tr>
<tr>
<td>S100A4</td>
<td>[41] 400</td>
<td>IHC</td>
</tr>
<tr>
<td>ADAM28</td>
<td>[42] 90</td>
<td>ELISA</td>
</tr>
<tr>
<td>XIAP</td>
<td>[43] 144</td>
<td>IHC</td>
</tr>
<tr>
<td>XAF1</td>
<td>[44] 51</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>CD24</td>
<td>[45] 267</td>
<td>IHC</td>
</tr>
<tr>
<td>ERCC1</td>
<td>[46] 51</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>HER2</td>
<td>[47] 83</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>CD82</td>
<td>[48] 151</td>
<td>RT-PCR</td>
</tr>
<tr>
<td><strong>Gene expression signatures identified using microarrays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>139-gene</td>
<td>[13] 253</td>
<td>Array</td>
</tr>
<tr>
<td>59-gene</td>
<td>[14] 100</td>
<td>Array</td>
</tr>
<tr>
<td>15-gene</td>
<td>[10] 133 Array+RT-PCR</td>
<td>see Table S1</td>
</tr>
<tr>
<td>50-gene</td>
<td>[9] 129</td>
<td>Array+RT-PCR+IHC</td>
</tr>
<tr>
<td>17-gene</td>
<td>[11] 91</td>
<td>Array</td>
</tr>
<tr>
<td>6-gene</td>
<td>[16] 138</td>
<td>Array+RT-PCR</td>
</tr>
<tr>
<td>38-gene</td>
<td>[17] 462</td>
<td>Array</td>
</tr>
</tbody>
</table>

ADE: adenocarcinoma; SCC: squamous cell carcinoma; 2D-DIGE: two-dimensional difference gel electrophoresis; MS: mass spectrometry; n: number of tumor samples included in the study; *highest quality probe, when several high quality probes then the best performing; † see Figure 2, for the survival plots; $: p value of the univariate analysis; multivariate: using those two parameters where most data was available (histology and gender for NSCLC, gender and stage for adenocarcinoma and squamous cell carcinoma). Multivariate analysis was performed only for biomarker candidates significant at p<0.01 in the univariate analysis. doi:10.1371/journal.pone.0082241.t002
development is similar to other cancer types by involving a stepwise progression to a malignant transformation driven by the collective effect of genetic changes induced by inhaled carcinogens [29]. At the same time, the number of previously never-smoker lung cancer patients is also increasing [30]. Gathering new insights into the underlying mechanism and etiological factors in these patients is necessary to better understand the disease and to develop new treatment strategies [2]. In our database we had the smoking history for 1,042 patients (of these 187 never smoker) and the meta-analysis tool also includes the option to restrict to either smoker on nonsmoker cohorts of patients. Additional filtering options include the use of gender (data is available for 1,564 patients) and staging (697 patients). Combinations of these options enable to validate biomarker candidates in sub-hubs having a size not reached by any of the previous individual studies.

Previously, within the directors’ challenge project for lung adenocarcinoma, the combined use of clinical and gene expression information performed best for predicting prognosis [17]. The multivariate analysis in the online software enables to compare clinical and molecular variables. Unfortunately, not all clinical information is published for each patient - this significantly limits the potential of any multivariate analysis including both clinical and gene expression variables.

We must also mention some issues with meta-analyses that may undermine their validity - these include biases related to patient selection, to clinical heterogeneity, to different outcome measures, to methodological and statistical techniques [31]. One option the test for biases is plotting the sample size against the effect size as this is usually skewed and asymmetrical in the presence a bias [32]. Basically, without a bias, the largest variation should be observed most in the small studies and least in large studies. This is the concept of the original funnel plot which we employed to demonstrate the correlation between hazard rates and sample sizes for two selected genes. We added an analysis option to our tool to run the computations in each dataset separately to enable swift construction of such analyses for any gene.

Finally, we have also assessed previously published gene expression signatures to predict survival. Today, the clinical application of multigene signatures is still controversial, as many of them do not outperform prognostication using conventional parameters. Here, out of seven signatures, two were capable to predict survival in stage I [13], and in all NSCLC patients [14].

In summary, by utilizing genome-wide microarray datasets published in the last five years, we have successfully integrated a large scale database suitable for the in silico validation of biomarker candidates in non-small cell lung cancer.

Supporting Information
Table S1 List of genes involved in previously published gene sets.
(XLS)
R Script S1 R script used to generate Kaplan-Meier plots (R)

Author Contributions
Conceived and designed the experiments: BG. Performed the experiments: BG PS JB AL. Analyzed the data: BG PS JB AL. Contributed reagents/materials/analysis tools: BG PS JB AL. Wrote the paper: BG.

References


