Supplementary Material
SurvExpress Tutorial

SurvExpress: An online biomarker validation tool for cancer gene expression data using survival analysis.

by Raul Aguirre-Gamboa, Hugo Gomez-Rueda, Emmanuel Martinez-Ledesma, Antonio Martinez-Torteya, Rafael Chacolla-Huaringa, Alberto Rodriguez-Barrientos, José G. Tamez-Peña, Victor Treviño*

Tecnológico de Monterrey
Cátedra de Bioinformática
*vtrevino@itesm.mx

Table of Contents

Summary

SurvExpress is a comprehensive gene expression database and web-based tool that provides survival multivariate analysis and risk assessment from a list of genes (biomarker) in human cancer datasets.

SurvExpress web tool is free, open to all users, and does not need login or registration. The web address is http://bioinformatica.mty.itesm.mx/SurvExpress (which should automatically redirect to actual application page http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp).

This tutorial contains three sections: a SurvExpress Quick Tutorial for rapid access, a Detailed Tutorial to provide complete reference, and Examples.

SURVEXPRESS QUICK TUTORIAL

Referring to Figure 1 below, the quick steps are:

1. Input your list of genes in http://bioinformatica.mty.itesm.mx/SurvExpress
2. Select the tissue, dataset, and click GO (this will launch the Analysis Page).
3. Select your features and clinical outcome.
4. Click Go (this will call the server and respond with plots and tables).
5. Check your results. For details see detailed reference.

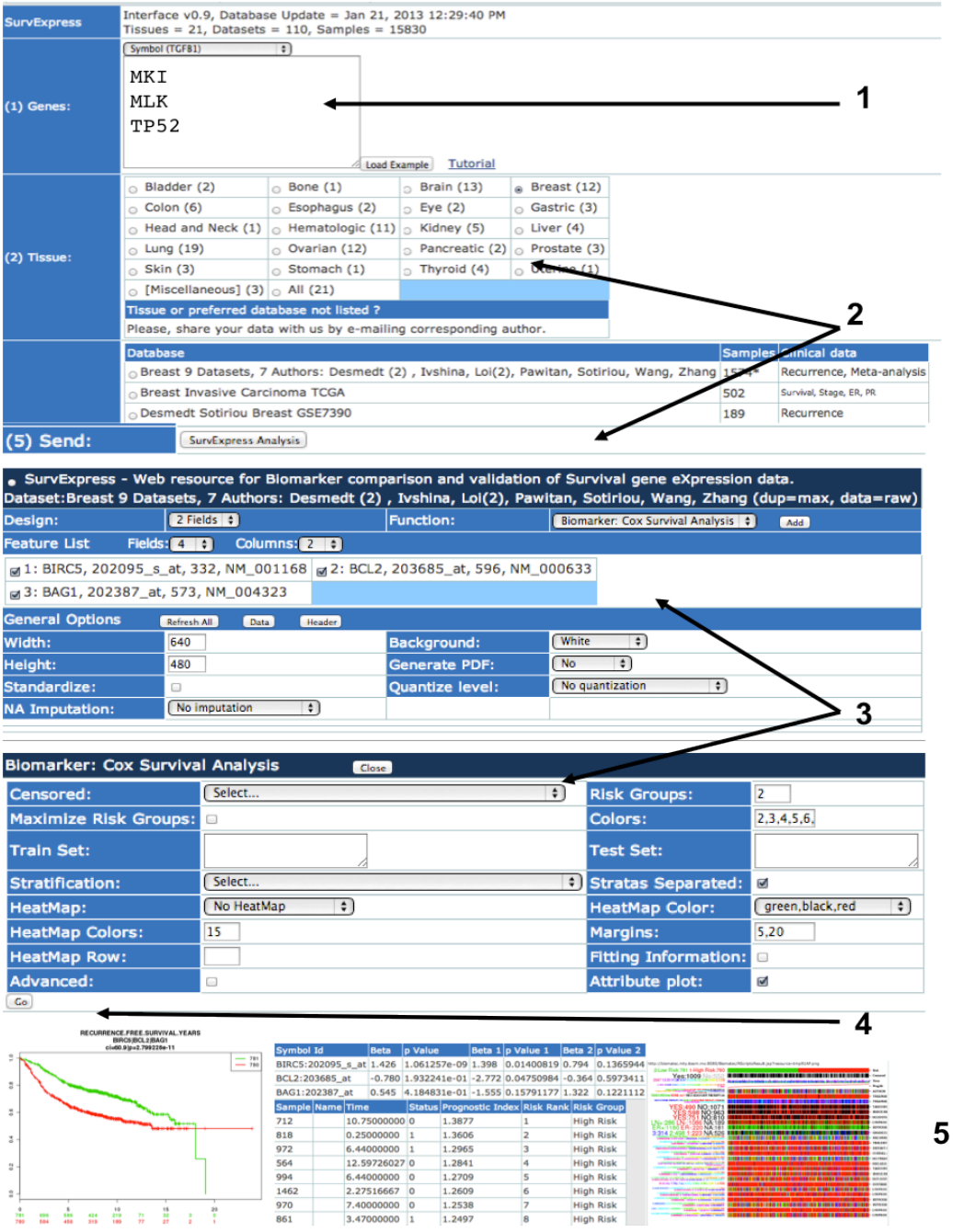


Figure 1. Quick reference of SurvExpress.

DETAILED TUTORIAL

SurvExpress consists of three pages: (i) the ***Input Page*** in which the genes and the dataset to work with are specified; (ii) the ***Analysis Page*** where the output is customized; and (iii) the ***Results Page***, which displays the corresponding plots and tables and is shown below the Analysis Page.

Key Concepts

To better interpret these pages for non-statistician minded users, please refer to Box 1 for key concepts.

Box 1: Key concepts in Survival Analysis useful to interpret SurvExpress outputs.

Survival Analysis. It is a branch of statistics to study time-to-event data of biological organisms. Includes a series of statistical tools as shown below.

Kaplan-Meier Plot. A graphical representation of the survival probability (vertical axis) versus time (horizontal axis) estimated with data using $S(t_i)=S(t_{i-1})*(n_i-d_i)/n_i$, where $S(t_0)=1$, t_i is *i-th* observed time, n_i is the number of events at time t_i (deaths) and n_i is the number of individual not having the event (alive) just before t_i (assuming ordered times t_i). This function generates a staggered curve, which represents the fraction of deaths in every stage known as instantaneous **hazard**.

Log-rank test. It is common to represent few Kaplan-Meier curves in the same plot to be visually compared. For example, assuming low- and high-risk groups, if both curves occasionally cross each other or decline similarly, it can be interpreted as no difference in survival times since the fraction of survivors is comparable in both curves across time. On the other hand, if both curves do not falloff similarly, it can be interpreted that there is a difference in survival times. Instead of a visual inspection, the Log-rank test has been proposed to evaluate statistically the equality of survival curves. The statistical test can be defined as the difference between the observed and expected events in a group.

Risk Factor. Any random variable that can be related to the survival time. For example, age, mutations, diet, treatment, or gene expression.

Cox Proportional Hazard regression. A mathematical model that relates survival time to data of a suspected risk factor. This allows us to model the survival time of a certain patient using its gene expression providing an estimation of the relationship of such genes with the survival time. This can be expressed as $h_i(t)/h_0(t) = exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots)$ where h_i is the hazard of the *i-th* individual, h_0 is the baseline hazard function, *exp* is the exponent function, and the term within the *exp* function is the prognostic index or risk score. There are three common statistical tests to evaluate the goodness-of-fit of the Cox model, that is, whether all β_i values are zero. These are the Log-Likelihood-Ratio test, the Wald test, and the Score test. In general, all tests produce similar results. There are also statistical tests for each particular β_i coefficient testing whether $\beta_i=0$. The most common is the Wald test.

Prognostic index. Also known as risk score, this is the linear part of the Cox model, and it is commonly used to generate the risk groups. For example, the simplest method splits the samples at the median after ranking the samples by their prognostic index, which generates low-risk and high-risk groups.

Concordance Index. This is a summary indicator whether subjects with higher risk prediction will experience the event after subjects of lower risk. This is a generalization of the AUROC used in classification problems. Values close to 0.5 are putatively random. This index is one of the most important indicators of prediction power. As it depends on the ranks (e.g. using the prognostic index from a Cox prediction), it is independent of the model used.

Hazard ratio. Can be defined as the ratio of the hazard between risk groups, and can be interpreted as the chance of an event occurring in the higher-risk population divided by the chance of an event occurring in the lower-risk population.

For more information please refer to these interesting books [1], [2], [3].

Input Page

The SurvExpress web-based tool input stage consists on simple steps to select the genes, datasets and multiple probe options in which the survival analysis and risk assessment will be performed.

SurvExpress mainly requires a list of genes, which can be provided in: Symbol, Entrez/GeneID, Ensembl, HGNC, MIM, HPRD or Vega Identifiers. For illustrative purposes, in this tutorial we will use the following list of genes: MKI67, ESR1, MMP11, GRB7, GSTM1, AURKA, PGR, CTSL2, ERBB2, CD68, BIRC5, BCL2, BAG1, CCNB1, SCUBE2, and MYBL2 as shown in Figure 2.

Figure 2: SurvExpress input stage, step 1. A shows the text box for the input of genes. B shows the *load example* button that will load a random list of genes from a commercial biomarker, and select randomly a tissue and dataset to analyze. C shows the options of available IDs.

In addition, SurvExpress includes a link that includes this tutorial and a function to provide example genes. The last is activated using the button “Load Example” (Figure 2 - B).

After a list of genes has been specified, SurvExpress needs the dataset where expression data will be extracted for specified genes. Since the number of datasets is more than 100, we classified them depending on the source tissue or organ following the conventions from Disease Ontologies (http://www.disease-ontology.org). The option *Miscellaneous* includes datasets that are pending or hard to classify. Therefore, the second step is the selection of tissue or type of cancer (Figure 3-A). In addition, we provide an option labeled *All* in which all datasets are shown at once. This facilitates the search of a specific dataset. Moreover, by keeping the mouse over a tissue, a globe window with information about datasets will be shown (Figure 3-C). By selecting the tissue, the section 3 of the page will be expanded with a table containing associated datasets (Figure 3-B). For this tutorial, we will use the TCGA dataset “Breast Invasive Cancer TCGA” as shown in the Figure 3.

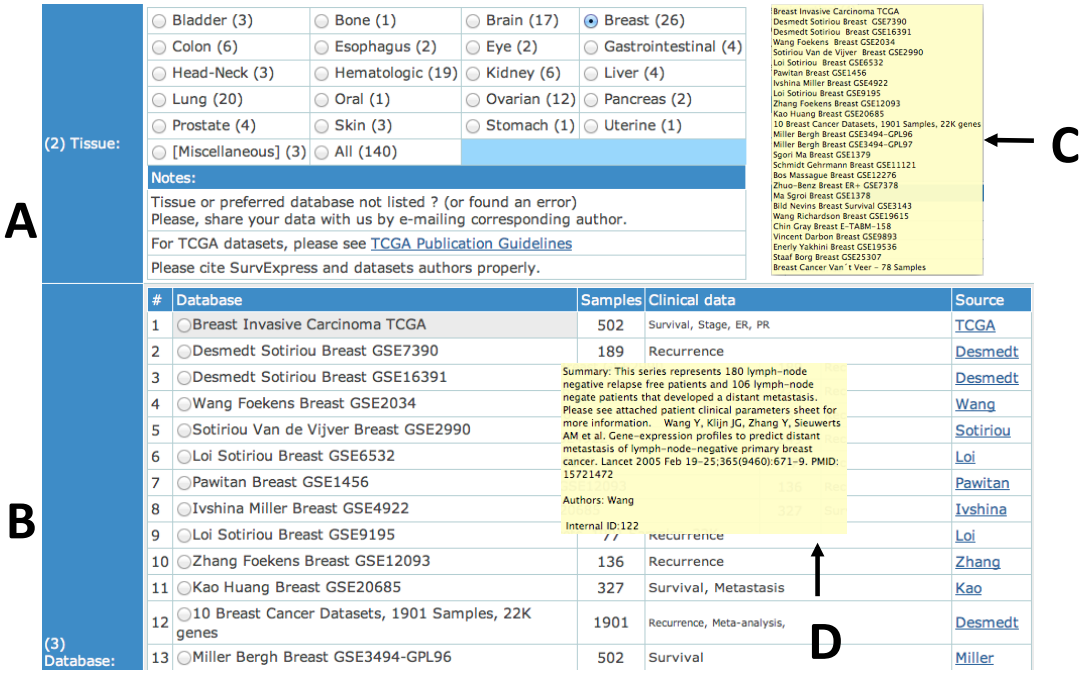


Figure 3. Selecting dataset. A shows the list of tissues or organs available. B shows the datasets available for selected tissue. C shows the information of the tissue by keeping the mouse over an option. D shows a short description of the dataset (mostly overall design and data source) that will appear when the mouse is kept over a dataset option. The "Source" column show a link to the data source or related publication.

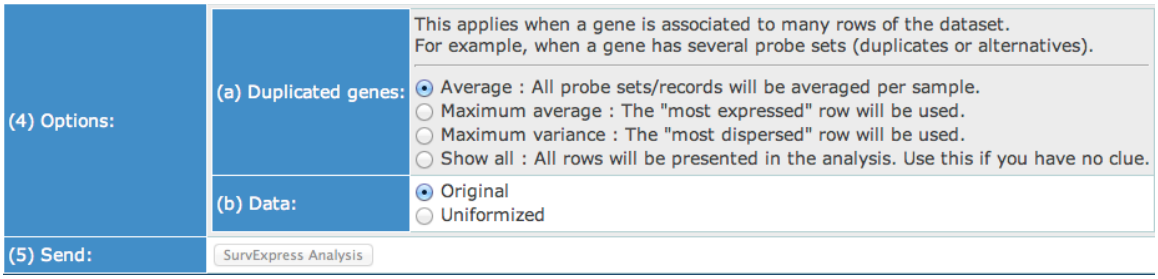


Figure 4: Specifying how duplicated genes will be handled and the SurvExpress Analysis button.

SurvExpress also requires that the user specify how duplicated genes will be treated (e.g. several probes of the same gene in one dataset). Figure 4 shows the four options. The *average* will use all probesets of a gene to compute an average per sample (additional gene information such as the id is taken from the first found probeset). The *maximum row average* or *variance* will extract only the row whose mean or variance is highest. The *show all* option will show all available probesets, then the user may activate and deactivate specific genes or probesets in the analysis page.

In addition, the original data has been transformed to a uniform distribution between 0 and 1 where 0 stands for the lowest level (no expressed) and 1 for the highest expression. This is employed for internal research purposes, so **we recommend using the Original quantile-normalized data**.

Finally, the analysis page is launched after pressing the "*SurvExpress Analysis*" button. This action submits the genes for searching and extracting expression data. The process may last few seconds (approximately 1 second per gene per 200 samples).

Analysis Page

To illustrate the analysis, we used in (1) the 20 genes shown in Input Page, breast in (2), Breast Invasive Carcinoma TCGA in (3), and "Average" and Original in (4). After 12 seconds, the Analysis Page starts by showing how many rows were found for each input gene and how were found. The algorithm for searching the expression values for a gene is shown in Figure 5-A. When the input is the gene symbol, SurvExpress first search the symbol in the original annotation file associated with the database. If it is not found, SurvExpress attempts to find the official symbol searching in the file Homo_sapiens.gene_info provided by NCBI. If an official symbol was found, the dataset is re-sought. If the input genes are not symbols, a similar process is performed searching first in Homo_sapiens.gene_info. An example of the genes found, and the searching process is shown in Figure 5. This information can also be shown when keeping the mouse over the title of the analysis page as shown in Figure 6.

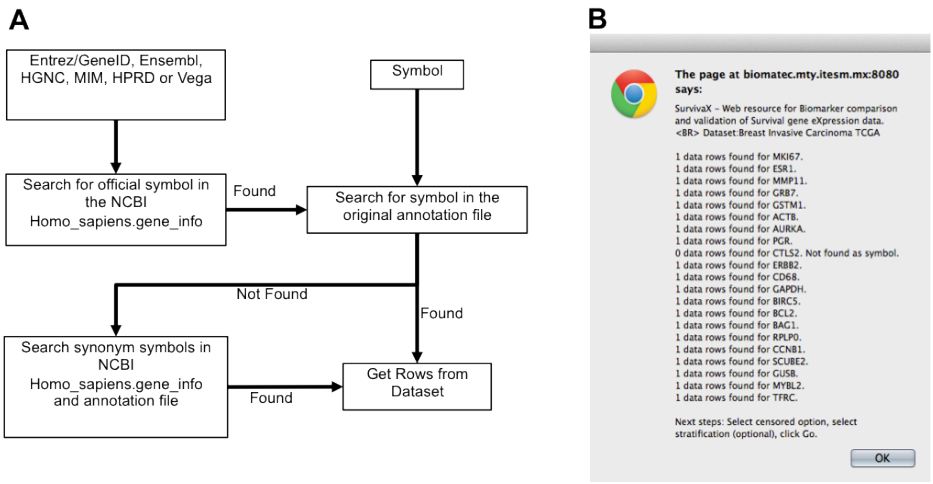


Figure 5: Process of finding genes in SurvExpress. A, the search algorithm performed by SurvExpress. B, the messages box reporting the genes found.

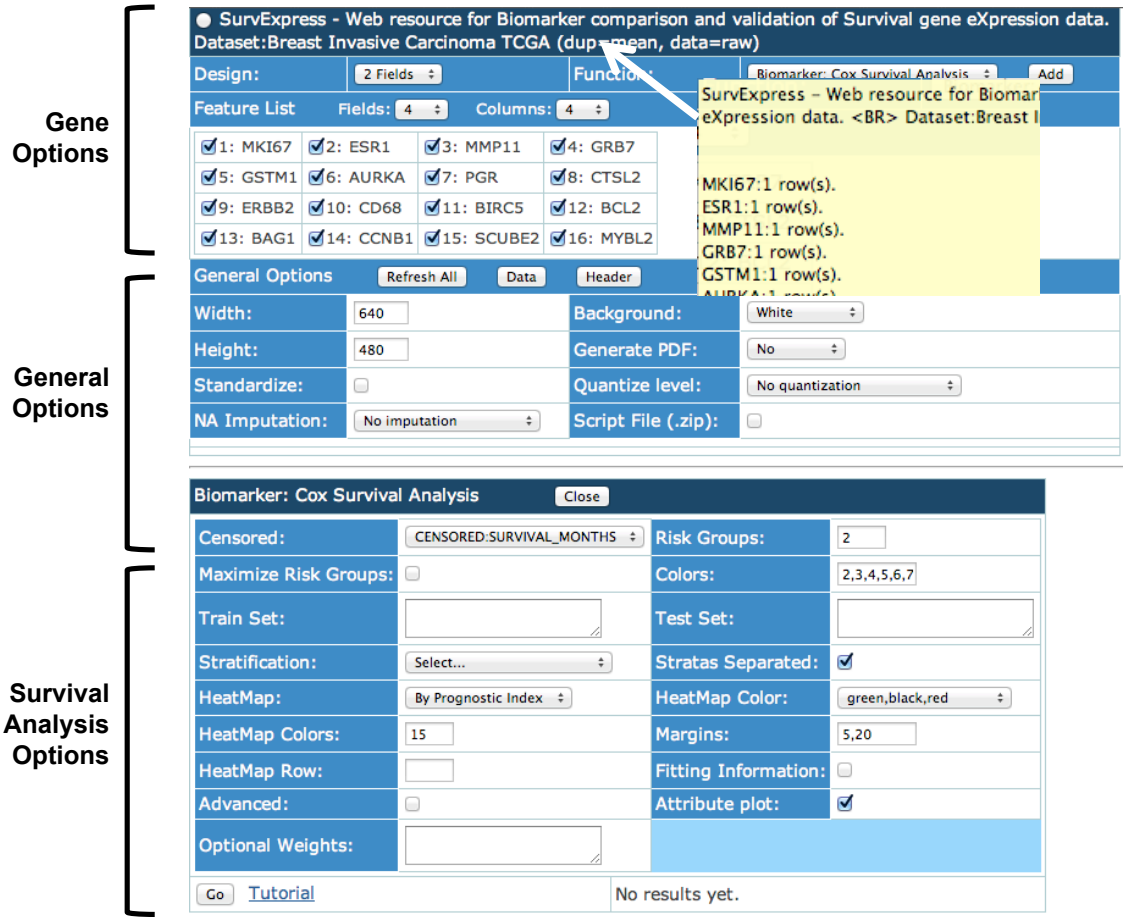


Figure 6: Complete page of the analysis page and its 3 sections.

Figure 6 shows the analysis page that consists of three option sections (Gene, General, and Survival Analysis) to specify the survival analysis and risk assessment that will be performed to selected data.

In the *Gene Options*, the user can select which genes or probes will be used in the survival analysis. This can become handy in the evaluation of single genes within a certain biomarker or a gene signature turning the genes on and off. For our example, we will leave all genes activated. The user can also change the appearance of the gene list by selecting the number of informative fields per gene and the number of gene columns per row. This setting has only visual effects of the page used for documenting purposes. The design is also an appearance option controlling the number of parameters of the survival analysis that will be shown per row.

In the *General Options*, the user can specify the size of the resulted images and whether the output is saved on a PDF file or in traditional PNG images. The PDF file can be the visualized within the browser or downloaded by clicking the provided link. In the *General Options,* the user can also specify how the data will be manipulated. For example, if the *Standardize* option is checked, the data is transformed to mean = 0 and standard deviation = 1 per gene. The NA imputation is used in cases where there is missing data. Since most of the datasets were downloaded from GEO and we ran a NA imputation procedure to all data, this option may be not needed. The quantize level option is used to transform gene expression values to specific levels. Since some biomarkers will be tested on RT-PCR and measurements of microarray data is noisy this option may help in showing whether the association of genes is sensitive to small changes in gene expression values. For example, if three levels are used with uniform data (values from 0 to 1), all values lower than 0.333 will be set to 0, values between 0.333 and 0.666 will be set to 0.5, and values larger than 0.666 will be set to 1. If original data is used, the 95% of the centered data is equivalently transformed to the specified number of levels, the 2.5% lowest values are set to lowest level and the 2.5% highest values are set to the highest level. The *Script File* option is used to generate a zip file containing the data and R scripts at the end of the output.

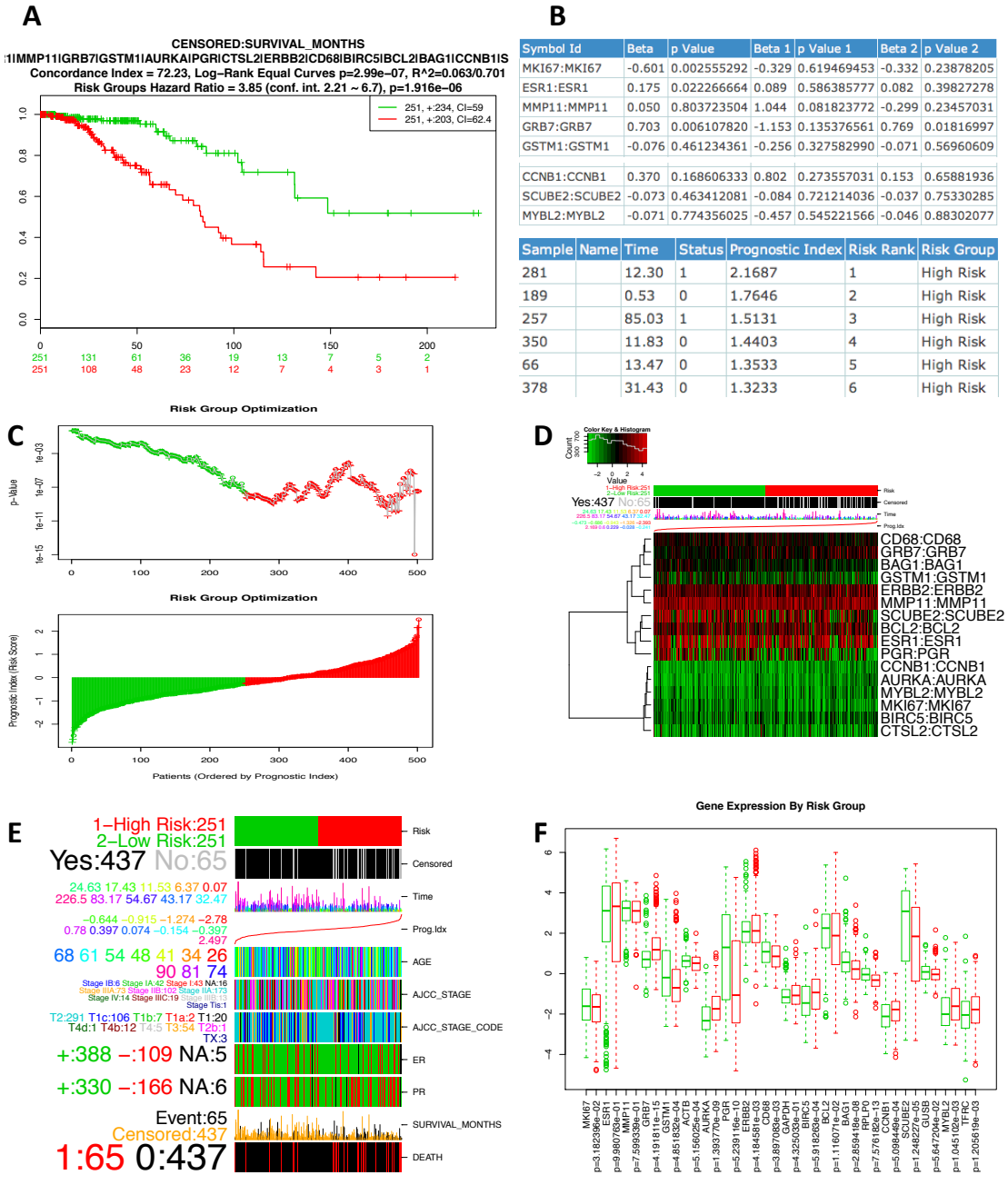
The *Survival Analysis Options* detail permits the users to select the specific parameters on which the survival evaluation will be performed. Only the selected features from the *Gene Options* will be used. A short explanation of each parameter is included in Table 1. In our example, we select the variable "SUVIVAL_MONTHS" in the "CENSORED" parameter leaving as default the rest of the parameters as shown in Figure 6.

Table 1. Parameters used to generate the Cox Survival Analysis.

Parameter	Description
Censored	Selection of the censored variable, e.g. overall survival, relapse, response. The options in this parameter are dependent on clinical information available for the dataset used.
Risk Groups	Number of risk groups in which the data will be split. At least a value of 2 must be used. In practice, an upper limit of 5 is sensible unless thousands of samples are available.
Maximize Risk Groups	By default, risk groups will be split in risk groups of the same size depending on the Prognostic Index (risk score) estimated by beta coefficients (or provided weights) multiplied by gene expression values. However, if *Maximize Risk Groups* is checked, risk group splitting is optimized using a simple algorithm shown in the section Risk Groups Plots using the inner-group p-value.
Colors	The colors used to identify risk groups. The default is set to green for low risk and red for high risk. Color numbers correspond to R-like color numbering (1=black, 2=red, 3=green, 4=blue, 5=cyan, 6=magenta, 7=orange, 8=gray, 9..16=like 1..8 but darker). To allow colorless figures for monochromatic publications, if all colors are 1, a gray scale will be used and line styles will be set in the Kaplan-Meier plots.
Train Set	Samples to be used in the estimation of beta coefficients using the Cox model. Samples numbers must be separated by comma. If not specified, all samples are used.
Tests Set	Samples to be used to estimate the prognostic index. Samples must be separated by comma. If not specified, all samples are used.
Stratification	Select the variable used to stratify the samples, e.g. age, gender. This option is populated with the clinical information available for each dataset.
Strata separated	If checked, a plot of the survival curve for each stratum will be shown. This is useful to compare the prediction along tumor sub-types, TNM, stages, histology, or any other clinical information.
Heatmap	If other than No Heatmap is selected, a heatmap showing gene expression values for every gene/probe will be shown. The selected option will direct how samples within the heatmap will be sorted. The most common option is the ordering by prognostic index.
Heatmap colors	Colors used to generate the gene expression colors of the heatmap.
Margins	Margin specifications for the heatmap plot. The format is bottom then right separated by a comma. This is useful when samples or gene names are large and want to be shown.
Heatmap row	Samples within the Heatmap can be ordered by the values of a specific gene/probe (using the option By Heatmap Row in the Heatmap parameter). This option specifies which of the input rows is used for sorting.
Fitting Information	If checked, the report will provide the fitting information used to generate the plots within them.
Advanced	Advance plots generate additional plots such as two survival ROC CURVES plots (using Kaplan-Meier (KM) and Nearest Neighbor Estimation (NNE) methods), as shown by Heagerty *et al.* [4]. The ROC curves are generated using the same times "ticks" applied in the Kaplan-Meier plot. These survival ROC curves may take a long time to compute.
Attribute plot	If checked, the estimated prognostic index is shown along with clinical information available for that dataset. This is useful to relate visually such variables as mutation status or hormone indicators to risk groups.
Optional Weights	By default, weights are estimated as the beta coefficients from the Cox model. However, the user can specify weight values to every gene. This can be used to approximate prognostic indicators that employ methods other than the Cox model. This parameter must be provided by one coefficient per gene/probe separated by commas.

Results Page

This page is generated after the specification of parameters and clicking the button *Go* in Figure 6. The results are shown just below this button. The basic scheme of results provided by SurvExpress consists in about seven plots as shown in Figure 7, which will be detailed in next sections.



Symbol Id	Beta	p Value	Beta 1	p Value 1	Beta 2	p Value 2
MKI67:MKI67	-0.601	0.002555292	-0.329	0.619469453	-0.332	0.23878205
ESR1:ESR1	0.175	0.022266664	0.089	0.586385777	0.082	0.39827278
MMP11:MMP11	0.050	0.803723504	1.044	0.081823772	-0.299	0.23457031
GRB7:GRB7	0.703	0.006107820	-1.153	0.135376561	0.769	0.01816997
GSTM1:GSTM1	-0.076	0.461234361	-0.256	0.327582990	-0.071	0.56960609
CCNB1:CCNB1	0.370	0.168606333	0.802	0.273557031	0.153	0.65881936
SCUBE2:SCUBE2	-0.073	0.463412081	-0.084	0.721214036	-0.037	0.75330285
MYBL2:MYBL2	-0.071	0.774356025	-0.457	0.545221566	-0.046	0.88302077

Sample	Name	Time	Status	Prognostic Index	Risk Rank	Risk Group
281		12.30	1	2.1687	1	High Risk
189		0.53	0	1.7646	2	High Risk
257		85.03	1	1.5131	3	High Risk
350		11.83	0	1.4403	4	High Risk
66		13.47	0	1.3533	5	High Risk
378		31.43	0	1.3233	6	High Risk

Figure 7: Default plots performed by SurvExpress. See text for each image description.

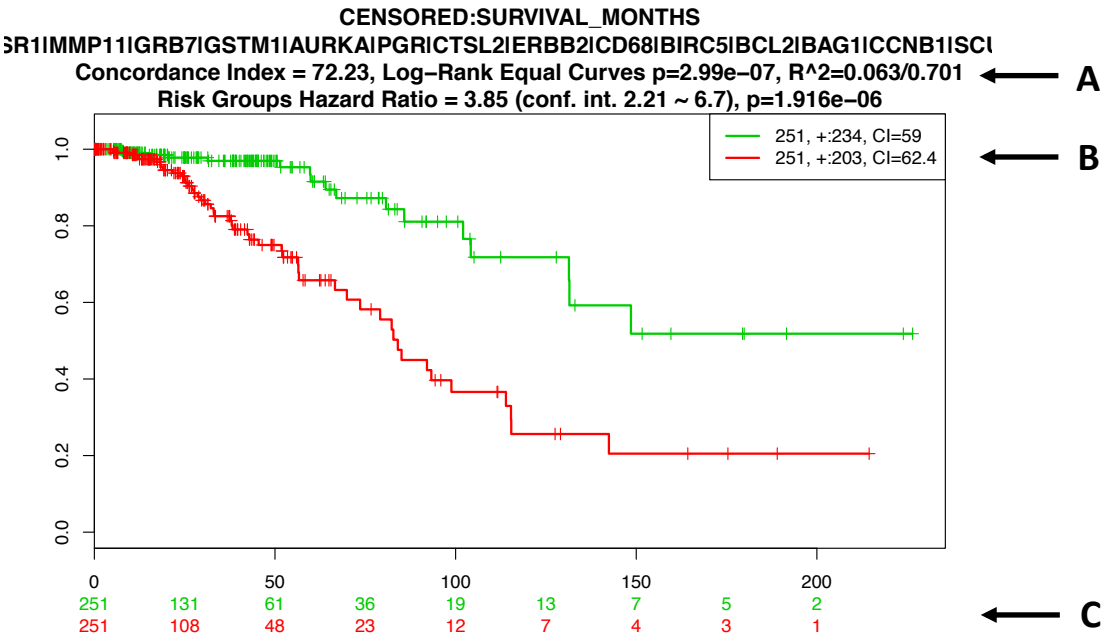


Figure 8: Kaplan-Meier Plot.

Kaplan-Meier plot

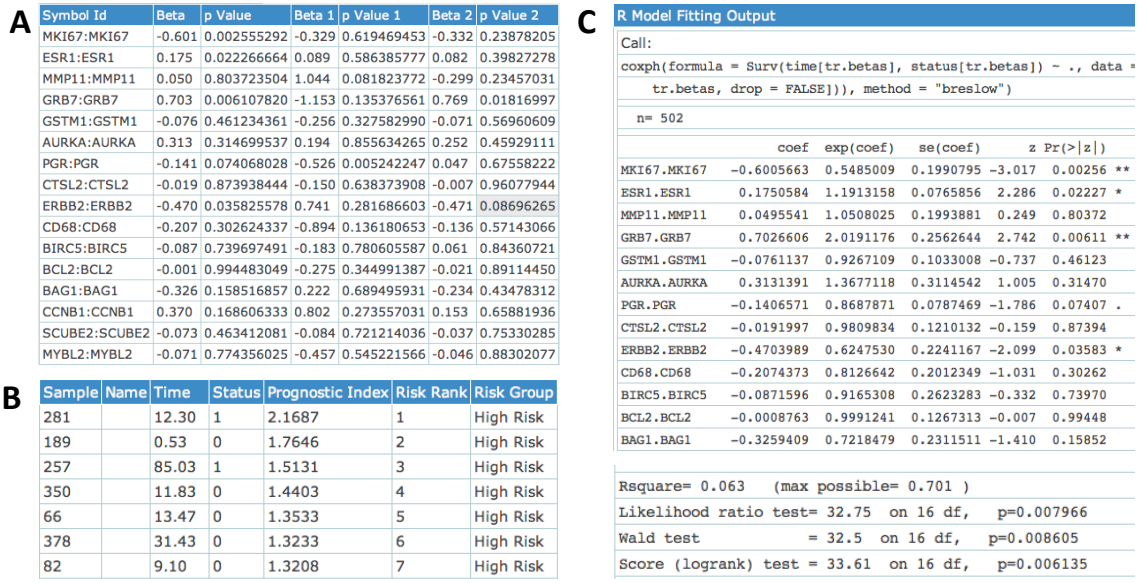
This a basic Kaplan-Meier plot (Figure 7A and Figure 8) which includes the following content:

A. The title of the plot shows the selected censored data and genes used. It also shows the Concordance Index (CI) and the p-value testing for equality of survival curves using a log-rank test, and the correlation coefficient estimated from deviance residuals. The CI estimates the probability that subjects with higher-risk prediction will experience the event after subjects of lower risk. *CI* is a generalization of the AUROC used in classification problems [5]. The *CI* is expressed as

$$CI = \frac{1}{|\Omega|} \sum_{i,j \in \Omega} \begin{Bmatrix} 1 \; if \; r_i > r_j \\ 0 \; otherwise \end{Bmatrix}$$

where r_i and r_j represent the risk predictors given by the corresponding prognostic index for subjects i and j respectively and Ω represent all subjects pairs (i, j) where $t_i < t_j$ and subject i is not censored. As in AUROC, CI values close to 0.5 are putatively 'random' whereas higher values are associated to better prediction. In addition, an estimation of the hazard ratio between the groups is shown. This is estimated by another Cox model using the risk group prediction as the covariate. p-Value and confidence interval is also shown.

B. Survival risk curves are shown for each group; low and high risks are drawn in green and red respectively (or specified colors in parameters). The total number for each group is shown in the top right corner and the number of censoring samples are marked with +. The CI per curve is also included.
C. The x-axis represents the time (months or years) of the study. In rows and corresponding colors are shown the number of samples not presenting the event at the matching time.



A

Symbol Id	Beta	p Value	Beta 1	p Value 1	Beta 2	p Value 2
MKI67:MKI67	-0.601	0.002555292	-0.329	0.619469453	-0.332	0.23878205
ESR1:ESR1	0.175	0.022266664	0.089	0.586385777	0.082	0.39827278
MMP11:MMP11	0.050	0.803723504	1.044	0.081823772	-0.299	0.23457031
GRB7:GRB7	0.703	0.006107820	-1.153	0.135376561	0.769	0.01816997
GSTM1:GSTM1	-0.076	0.461234361	-0.256	0.327582990	-0.071	0.56960609
AURKA:AURKA	0.313	0.314699537	0.194	0.855634265	0.252	0.45929111
PGR:PGR	-0.141	0.074068028	-0.526	0.005242247	0.047	0.67558222
CTSL2:CTSL2	-0.019	0.873938444	-0.150	0.638373908	-0.007	0.96077944
ERBB2:ERBB2	-0.470	0.035825578	0.741	0.281686603	-0.471	0.08696265
CD68:CD68	-0.207	0.302624337	-0.894	0.136180653	-0.136	0.57143066
BIRC5:BIRC5	-0.087	0.739697491	-0.183	0.780605587	0.061	0.84360721
BCL2:BCL2	-0.001	0.994483049	-0.275	0.344991387	-0.021	0.89114450
BAG1:BAG1	-0.326	0.158516857	0.222	0.689495931	-0.234	0.43478312
CCNB1:CCNB1	0.370	0.168606333	0.802	0.273557031	0.153	0.65881936
SCUBE2:SCUBE2	-0.073	0.463412081	-0.084	0.721214036	-0.037	0.75330285
MYBL2:MYBL2	-0.071	0.774356025	-0.457	0.545221566	-0.046	0.88302077

B

Sample	Name	Time	Status	Prognostic Index	Risk Rank	Risk Group
281		12.30	1	2.1687	1	High Risk
189		0.53	0	1.7646	2	High Risk
257		85.03	1	1.5131	3	High Risk
350		11.83	0	1.4403	4	High Risk
66		13.47	0	1.3533	5	High Risk
378		31.43	0	1.3233	6	High Risk
82		9.10	0	1.3208	7	High Risk

C

R Model Fitting Output

```
Call:
coxph(formula = Surv(time[tr.betas], status[tr.betas]) ~ ., data =
    tr.betas, drop = FALSE])), method = "breslow")

  n= 502
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
MKI67.MKI67	-0.6005663	0.5485009	0.1990795	-3.017	0.00256	**
ESR1.ESR1	0.1750584	1.1913158	0.0765856	2.286	0.02227	*
MMP11.MMP11	0.0495541	1.0508025	0.1993881	0.249	0.80372	
GRB7.GRB7	0.7026606	2.0191176	0.2562644	2.742	0.00611	**
GSTM1.GSTM1	-0.0761137	0.9267109	0.1033008	-0.737	0.46123	
AURKA.AURKA	0.3131391	1.3677118	0.3114542	1.005	0.31470	
PGR.PGR	-0.1406571	0.8687871	0.0787469	-1.786	0.07407	.
CTSL2.CTSL2	-0.0191997	0.9809834	0.1210132	-0.159	0.87394	
ERBB2.ERBB2	-0.4703989	0.6247530	0.2241167	-2.099	0.03583	*
CD68.CD68	-0.2074373	0.8126642	0.2012349	-1.031	0.30262	
BIRC5.BIRC5	-0.0871596	0.9165308	0.2623283	-0.332	0.73970	
BCL2.BCL2	-0.0008763	0.9991241	0.1267313	-0.007	0.99448	
BAG1.BAG1	-0.3259409	0.7218479	0.2311511	-1.410	0.15852	

```
Rsquare= 0.063   (max possible= 0.701 )
Likelihood ratio test= 32.75  on 16 df,   p=0.007966
Wald test            = 32.5  on 16 df,   p=0.008605
Score (logrank) test = 33.61  on 16 df,   p=0.006135
```

Figure 9: Genes and Sample Tables from the SurvExpress results page.

Genes and Sample Tables

Three tables are displayed as a typical SurvExpress report (Figure 7B and Figure 9):

A. *Gene Table*: As shown in Figure 9-A, this table displays information about each gene, including betas' coefficients and corresponding Wald-test p-values (first Beta and p Value columns). If the advanced parameter is checked, additional beta and p-values are shown for each risk group estimated only that subpopulation. This can be used to explore whether coefficient or significance is specific or change between risk groups.
B. *Sample Table*: Reports the censored clinical data used in the analysis along with the estimated prognostic index and risk group assignment. This table is sorted by Prognostic Index as shown Figure 9-B.
C. *Fitting Table*: A table showing the information of the output generated by the *coxph* function from the *survival* package in R is also included. This is useful to observe all statistics provided.

Risk groups plots

The risk plots (Figure 7C) help the user to visualize how the risk groups partitions were made on SurvExpress to generate the Kaplan-Meier plots. This is convenient to observe variations on the p-values if the partition changes or to evaluate the relation between risk groups and prognostic index. By default, SurvExpress split the risk group by the median of the Prognostic Index generating risk groups of the similar number of samples.

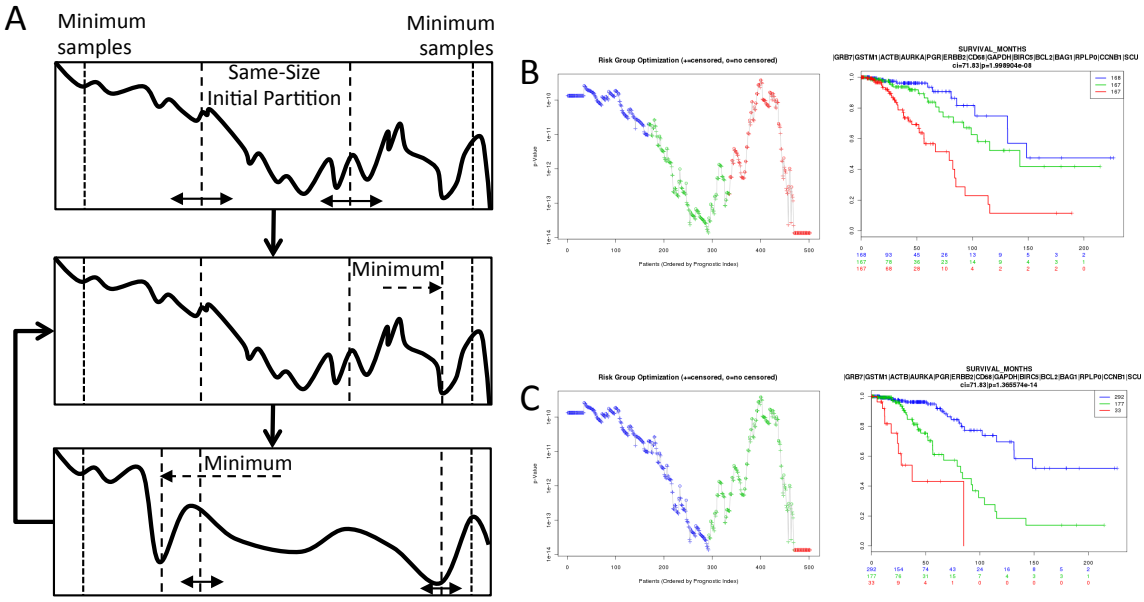


Figure 10. Algorithm and examples of generating partitions.

However, in SurvExpress, the partition can also be optimized by using the *Maximize Risk Groups* option. Using this option an algorithm decides where the partitions should be made to maximize the statistical significance of the separation of risk groups as shown in Figure 10-A for three risk groups. First, the algorithm start by partitioning samples by same-size risk groups. Then a p-value is estimated by changing the cut-off point one group at the time until a certain limit (five samples or L% of samples where L = 20/# risk groups). The new cut-off point is chosen so that the p-value is minimum. This process is repeated until no changes are needed. For example, Figure 10-B shows the default partition for three risk groups, and Figure 10-C shows the partition after maximization. Note that the p-values estimated initially may not correspond to those at the end for more than two groups.

A second example performed using the Hoshida Golub Liver GSE10143 is shown in Figure 11.

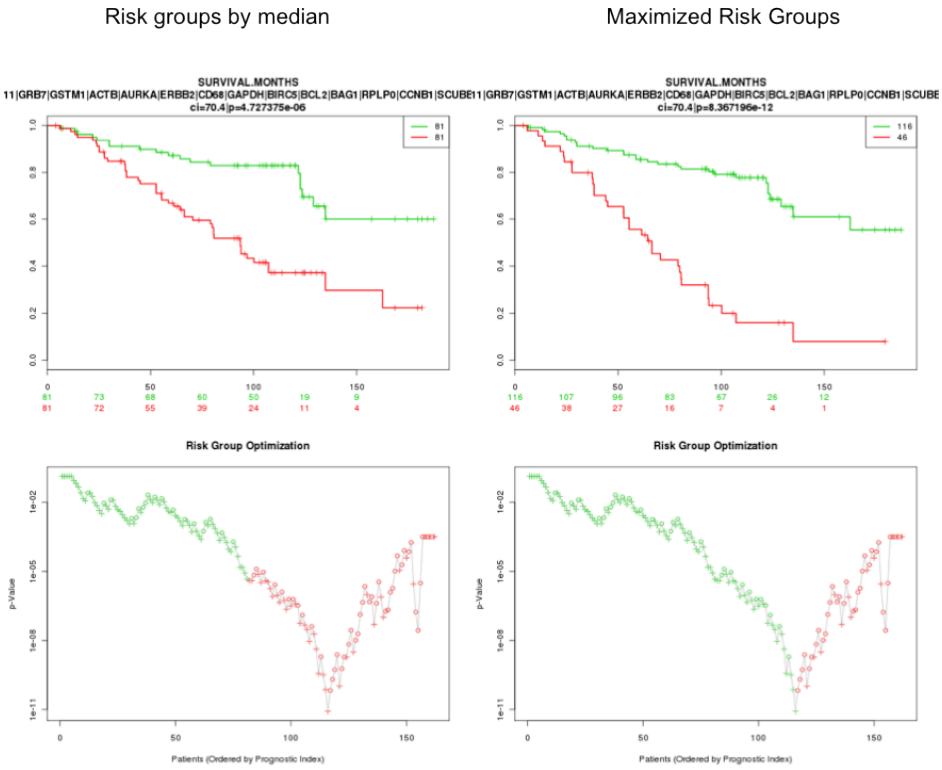


Figure 11: Effect of the group generation by maximizing the p-value in SurvExpress using the Hoshida Golub Liver GSE10143 dataset. Note that the number of samples in Kaplan-Meier plots also reflects the change.

Heatmap

The heatmap (Figure 7D and Figure 12) can be very helpful in the analysis and visual correlation of the survival analysis and gene expression. As well as in the previous graphic result, the heatmap enables the user to visualize the level of expression (by color) of each gene along samples ranked by their prognostic index. Samples are shown in x-axis while genes are shown in y-axis. Commonly, samples are sorted by prognostic index unless the user selects a different option for the corresponding option in *Heatmap* parameter. The genes are clustered by Euclidean distance (Figure 12).

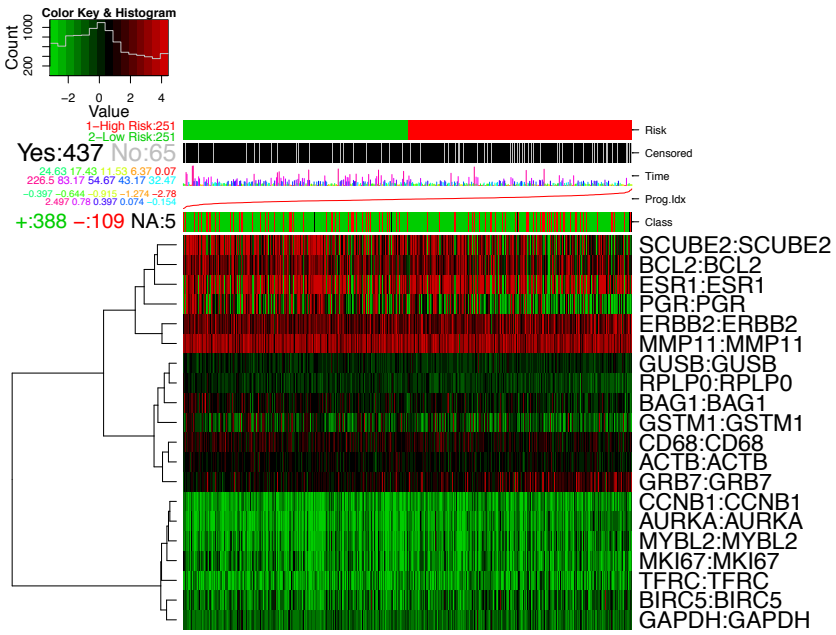


Figure 12: Expression profiles shown as a heatmap.

Clinical characteristics plot

The clinical characteristics plot (Figure 7E and Figure 13) is a data summary of the censored event selected together with other medical features of the dataset. This plot can be used to compare visually clinical features to risk estimations. For example, data in Figure 13 suggests that high risk can be associated with a negative result in the progesterone receptor assay (encouraging to perform a stratification within SurvExpress, a proper statistical test, literature search, or even a pilot study). By default, this plot uses the samples ranked by prognostic index but can be sorted by several other options available including a particular gene (see Table 1). This graph can easily be divided into two sections:

A. Information related to censoring event being analyzed (risk group assignment, censoring status, time related to event, and prognostic index).
B. Other clinical features contained in the analyzed dataset. This section depends on reported information. For the Breast Invasive Carcinoma from TCGA, there is information concerning about receptors, state, age, and AJCC state.

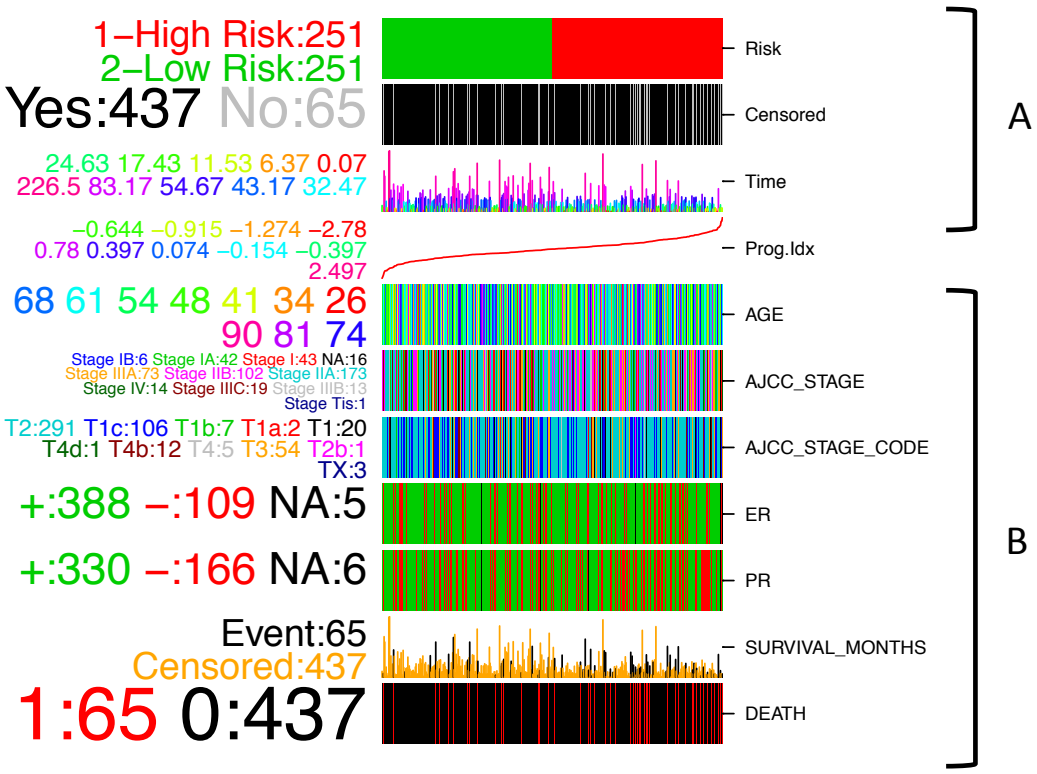


Figure 13: Clinical characteristics plot performed by SurvExpress analyzing the Breast Invasive Carcinoma from TCGA. Colored values in left indicate the frequency in the corresponding image at right.

Box plot of gene expression by risk groups

In this plot (Figure 7E and Figure 14) the gene expression of each gene is plotted along risk groups obtained in the analysis. The x-axis shows each gene and a p-value of the expression difference between risk groups. The p-value is obtained from a t-test for two risk groups or an f-test for more than two risk groups. The y-axis shows the expression levels. This plot is useful to visualize whether gene expression values are different between risk groups.

Figure 14: Box plot obtained in the results of SurvExpress, visualizing the expression levels of each feature in the risk groups generated.

Stratification

SurvExpress can also generate survival curves for types of samples by selecting a determined class present in the analyzed dataset. SurvExpress will generate two series of plots for every value of the selected class: (i) a series of plots using "overall betas", that is, using the Cox fitting from all samples in training, and (ii) a series of plots re-estimating betas for that specific value of the class. This is useful for comparisons of survival curves per class value and whether the fitting change between groups. Only class values greater than four times the number of risk groups are plotted. Figure 15 shows an example generated by the stratification of progesterone receptor (class PR whose values are "+" or "-"). It is clear that the general evaluation of risk groups is quite similar irrespective of the estimations of the betas. Nevertheless, due to re-estimation, the concordance index and the p-value of the risk group have changed.

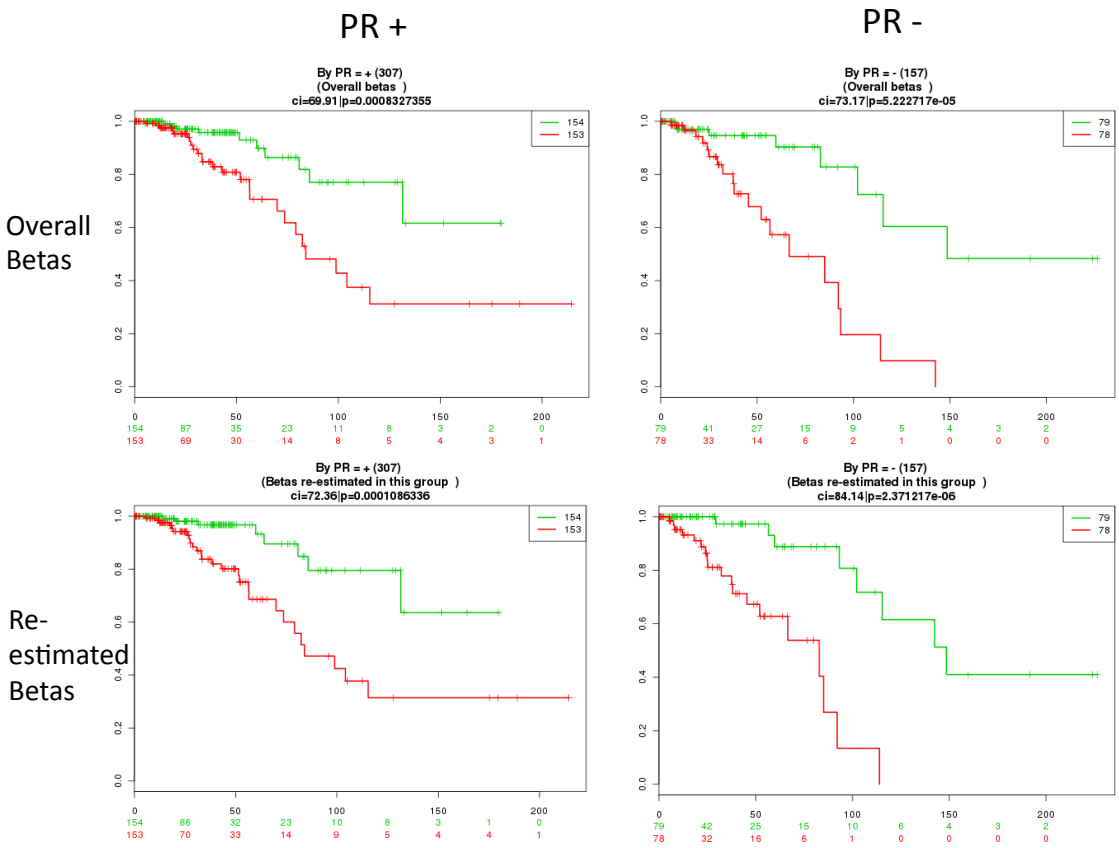


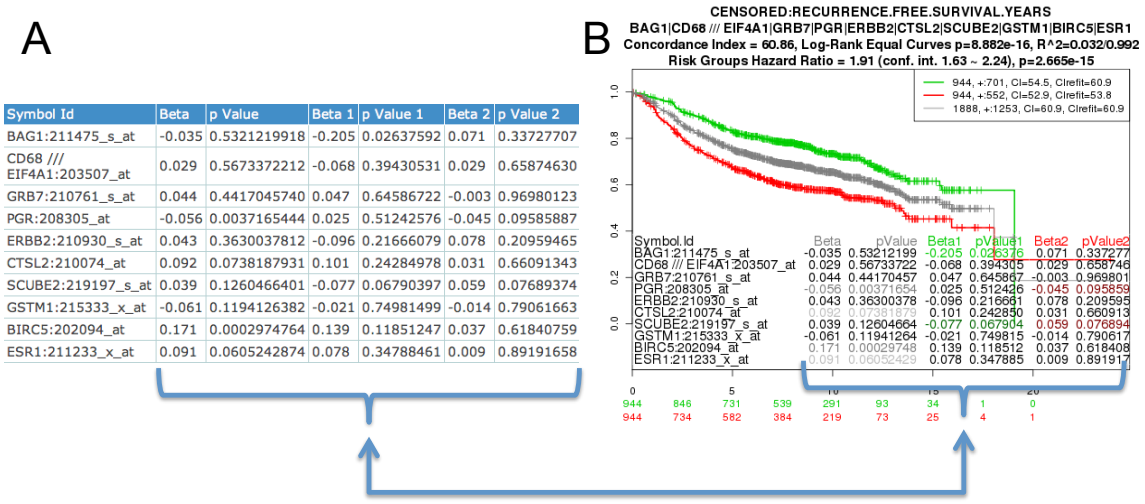
Figure 15: Stratification of the TCGA Breast Carcinoma dataset by progesterone receptor.

Fitting Information

The fitting information of the Cox model is always shown in tables. These tables include the gene symbol, the regression coefficient, and the p-value of the statistical significance of such coefficient in the columns “Symbol Id”, “Beta”, and “p Value” respectively. However, if the parameter *fitting information* is checked, additional fitting information estimated by risk groups is included in the table. For instance, if two risk groups are used in the field “Risk Groups”, four additional columns appear in the table corresponding to beta regression coefficients and p-values of each group as shown in Figure 16 A. This information is also included within the survival curves as shown in Figure 16 B.

This data can be useful to observe the behavior of the coefficients across the risk groups. For instance, in the Figure 16 B the PGR gene is significant (p=0.0037) in the fitting irrespective of the risk group (in gray), and PGR seems to be related to high-risk group since p-value is more significant in the high-risk group (p=0.09) than in the low-risk group (p=0.512). A second example is the shown for the gene BAG1 in the same Figure 16 B. Although BAG1 seems to be not significant (p=0.532), it seems

related to low-risk group since its p-value is significant specifically in that group (p=0.026).



Symbol Id	Beta	p Value	Beta 1	p Value 1	Beta 2	p Value 2
BAG1:211475_s_at	-0.035	0.5321219918	-0.205	0.02637592	0.071	0.33727707
CD68 /// EIF4A1:203507_at	0.029	0.5673372212	-0.068	0.39430531	0.029	0.65874630
GRB7:210761_s_at	0.044	0.4417045740	0.047	0.64586722	-0.003	0.96980123
PGR:208305_at	-0.056	0.0037165444	0.025	0.51242576	-0.045	0.09585887
ERBB2:210930_s_at	0.043	0.3630037812	-0.096	0.21666079	0.078	0.20959465
CTSL2:210074_at	0.092	0.0738187931	0.101	0.24284978	0.031	0.66091343
SCUBE2:219197_s_at	0.039	0.1260466401	-0.077	0.06790397	0.059	0.07689374
GSTM1:215333_x_at	-0.061	0.1194126382	-0.021	0.74981499	-0.014	0.79061663
BIRC5:202094_at	0.171	0.0002974764	0.139	0.11851247	0.037	0.61840759
ESR1:211233_x_at	0.091	0.0605242874	0.078	0.34788461	0.009	0.89191658

Figure 16. Example of output generated when "Fitting Information" option is selected. Panel A shows the table format. Panel B shows the table drawn within the figure.

Advanced Option

The advanced option is experimental. At the moment, we are including estimation of survival curves using the R package survivalROC [4]. We estimate the ROC curves over different point in event times. SurvivalROC curves take a long time to compute (few minutes) and depend on the number of samples. Therefore, we have limited the estimation to 500 samples taken from 1 to the number of samples (e.g. for 1000 samples, samples 1, 3, 5, ... 999 will be used). This number of samples should be enough to have good estimates of the ROC curves. Examples of the output of two survivalROC methods are shown in Figure 17. For more information, please refer to survivalROC R package documentation.

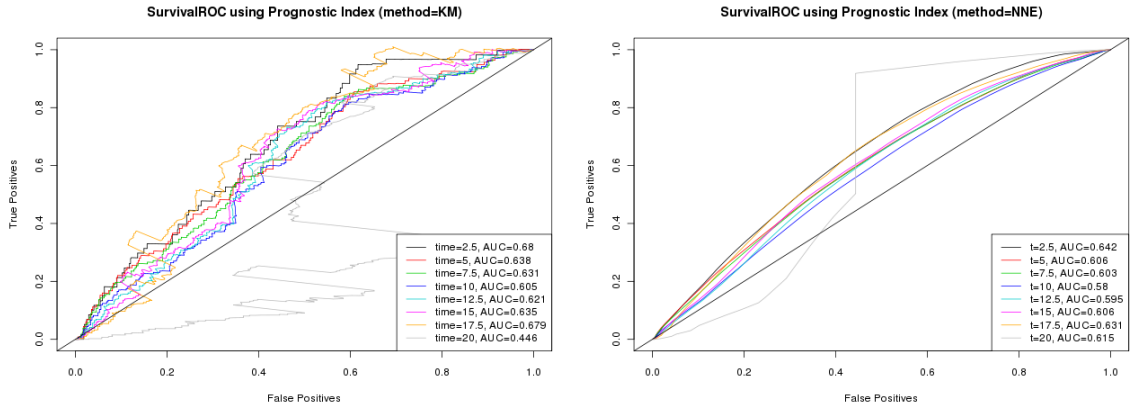


Figure 17. SurvivalROC plots generated when "Advanced" option is selected.

IMPLEMENTATION

SurvExpress, database and web tool, were implemented using MySQL for administration, java server pages (JSP) for the user front-end, R for the data analysis and plot generation, and Apache web server for file delivery (Figure 18). All this is running under a Linux-based server. We have tested our server under modest load; so we guess it will work on higher demands (although slower). If service is not available, please e-mail corresponding author. The service works under the University intranet. Therefore, it is affected by local traffic and Institutional Internet bandwidth.

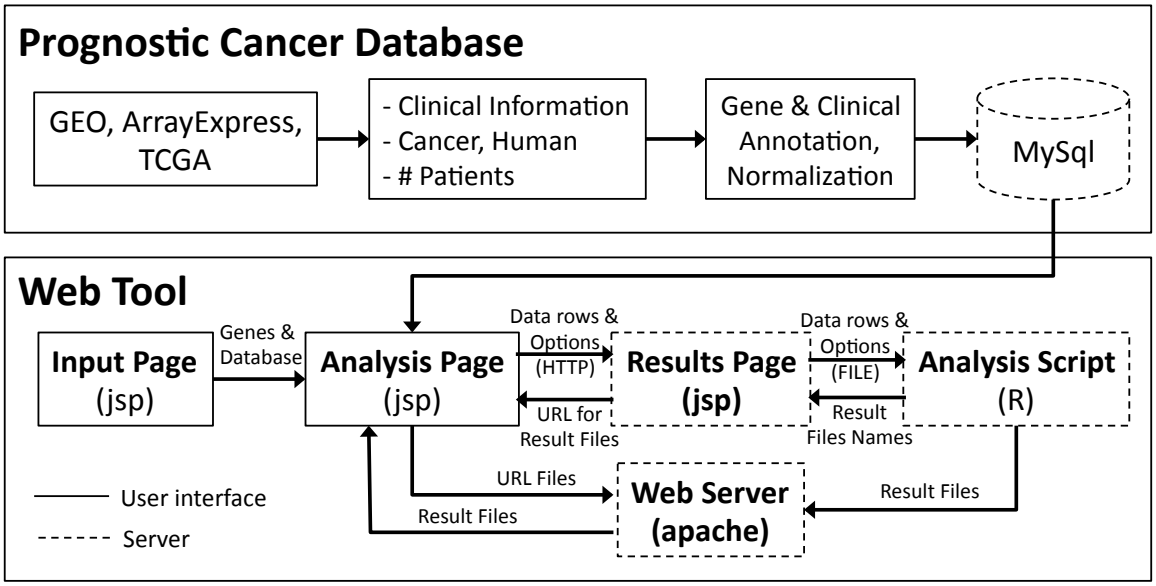


Figure 18. SurvExpress implementation.

Response Time

The response time depends on the number of requests by users (load), and the Institutional Internet bandwidth. Every request depends on the number of genes, the number of samples, the width and height of figure, and other options specified such as stratification, risk group maximization, and advanced (advanced takes longer times due to survivalROC curves). In preliminary tests over default options, the response time is approximately to 1 second per gene per 200 samples for a 640x480 image. For 20 genes and 1,900 samples, the average response time was 41 seconds. For larger datasets, the group maximization and concordance index takes particularly longer times since a two-dimensional matrix (of samples) needs to be

computed. For 20 genes, 1,900 samples, and stratification with two-levels factor (such as the estrogen receptor +/-), the response time increased to 64 seconds.

EXAMPLES

In this section we show two applications of SurvExpress: (a) One biomarker tested in several datasets; and (b) Two biomarkers tested on the same dataset.

OncotypeDX Biomarker for Breast Cancer

As an example for testing one biomarker in several datasets, we used the 16 OncotypeDX genes [6] not including reference genes. OncotypeDX estimates a recurrence score that is mainly offered to early-stage, estrogen positive, lymph node negative breast cancers. The genes included are AURKA, BAG1, BCL2, BIRC5, CCNB1, CD68, CTSL2, ERBB2, ESR1, GRB7, GSTM1, MKI67, MMP11, MYBL2, PGR, SCUBE2 (reference genes not used are ACTB, GAPDH, GUSB, RPLP0, TFRC). To estimate the score, OncotypeDX uses a weighting algorithm equivalent to a weight multiplied by corresponding gene expression normalized by a reference [6]. Here, in SurvExpress, we used Cox fitting (since gene expression data is not normalized to reference genes), the maximum row average for duplicated genes, two risk groups split at the median prognostic index in four breast cancer datasets (Table 2). These datasets reflect patient populations close to that suitable for the test (Wang and Ivshina), a dataset with partial information besides different event (TCGA), and a dataset with no other clinical information (Kao).

Table 2: Datasets and clinical for the Oncotype DX example.

Dataset	Platform	Samples/ Censored	ER+/-	LN+/-	Censored Data
Breast Invasive Carcinoma TCGA [7]	Agilent	502/437	388/109		Survival months
Kao Huang Breast GSE20685 [8]	Affymetrix	327/244			Recurrence years
Ivshina Miller Breast GSE4922 [9]	Affymetrix	249/160	211/34	81/159	Recurrence years
Wang Foekens Breast GSE2034 [10]	Affymetrix	286/179	209/77	0/286	Recurrence years

ER and LN stand for Estrogen Receptor and Lymph Node respectively.

The Kaplan-Meier plots shown in Figure 19 and summarized in Table 3 suggest that, overall, Oncotype DX can separate significantly low- and high-risk groups in the four datasets tested. Moreover, satisfactory indexes of concordance and ROC areas were obtained. These good results were obtained even though the majority of the coefficients of the 16 genes were not significant within the Cox fitting. This is not surprising because here we used a multivariate Cox while the provider use a different scoring algorithm. Nevertheless, to analyze the multivariate coefficients

further, we compared the corresponding coefficients of each gene across datasets as shown in Table 4. Only three genes had the four datasets coefficients of the same sign (the sign is associated to risk, positive to higher risks); 10 genes concurred in three datasets; and there was no majority in three genes. The significant coefficients are important supported by the fact that if all coefficients are replaced by 1 (using the weights option) no good prognosis is obtained in three of the four databases (data not shown).

Table 3: Results of the Oncotype DX in four breast cancer datasets.

Dataset	Genes Found	Response Time*	Significant Coefficients	Risk Groups p Value	CI	DEG between Risk Groups	Survival ROC**
TCGA	16	9.1s	4	2.9e-7	72.2	11	0.74
Kao GSE20685	16	6.1s	2	2.0e-5	69.1	16	0.69
Ivshina GSE4922	16	6.0s	2	5.0e-6	68.7	13	0.70
Wang GSE2034	16	6.0s	4	1.1e-7	69.1	13	0.73

CI stands for Concordance Index. DEG means differential expressed genes. *Response time of the results page. **SurvivalROC was estimated around time=6 years, curves took one order of magnitude more than the response time shown.

To demonstrate the analytical features of SurvExpress, we also performed the survival evaluation stratifying the samples using the provided tumor grades (AJCC Stage in the TCGA dataset and grade in the Invshina dataset) as shown in Figure 20. This figure shows that the performance of the estimated biomarker is different between sub-populations.

All this information is important because the performance of biomarkers in different populations, clinical information, probes per gene, gene expression technology, and conditions may give clues about possible applications in medical practice and about the biology of the tested biomarker.

Table 4: Gene coefficients within the Cox model for each dataset.

	TCGA		Kao		Ivshina		Wang			
Gene	**β**	**p**	**β**	**p**	**β**	**p**	**β**	**p**	**#Pos**	**#Neg**
AURKA	0.3130	0.3147	0.0060	0.9781	-1.8030	0.2156	4.0350	0.0100	3	1
BAG1	-0.3260	0.1585	0.0520	0.7755	3.5650	0.0465	0.0140	0.9932	3	1
BCL2	-0.0010	0.9945	-0.5840	0.0004	-0.0730	0.9550	-0.8520	0.4214	0	4
BIRC5	-0.0870	0.7397	-0.1110	0.4487	-0.0500	0.9765	-1.0540	0.3431	0	4
CCNB1	0.3700	0.1686	0.4430	0.0885	3.1070	0.0630	1.2130	0.4204	4	0
CD68	-0.2070	0.3026	0.0040	0.9792	-0.4990	0.6993	-1.4560	0.1350	1	3
CTSL2	-0.0190	0.8739	0.1700	0.3046	-0.0060	0.9955	-2.7350	0.0290	1	3

ERBB2	-0.4700	0.0358	-0.2400	0.3147	3.4180	0.2374	-1.1420	0.5878	1	3
ESR1	0.1750	0.0223	-0.0860	0.3554	1.3730	0.3301	-0.4740	0.6439	2	2
GRB7	0.7030	0.0061	0.3290	0.1421	0.2180	0.8742	-1.0310	0.2581	3	1
GSTM1	-0.0760	0.4612	-0.3020	0.0144	-0.5630	0.6843	0.0640	0.9449	1	3
MKI67	-0.6010	0.0026	-0.3190	0.1704	-0.7010	0.6891	0.7110	0.6680	1	3
MMP11	0.0500	0.8037	0.0370	0.6896	-2.4600	0.1166	4.7190	0.0079	3	1
MYBL2	-0.0710	0.7744	-0.0920	0.4525	3.2370	0.0011	0.3600	0.6563	2	2
PGR	-0.1410	0.0741	0.0320	0.6960	0.2320	0.6750	-1.7590	0.0011	2	2
SCUBE2	-0.0730	0.4634	0.1500	0.0757	0.3990	0.6919	1.4300	0.0836	3	1

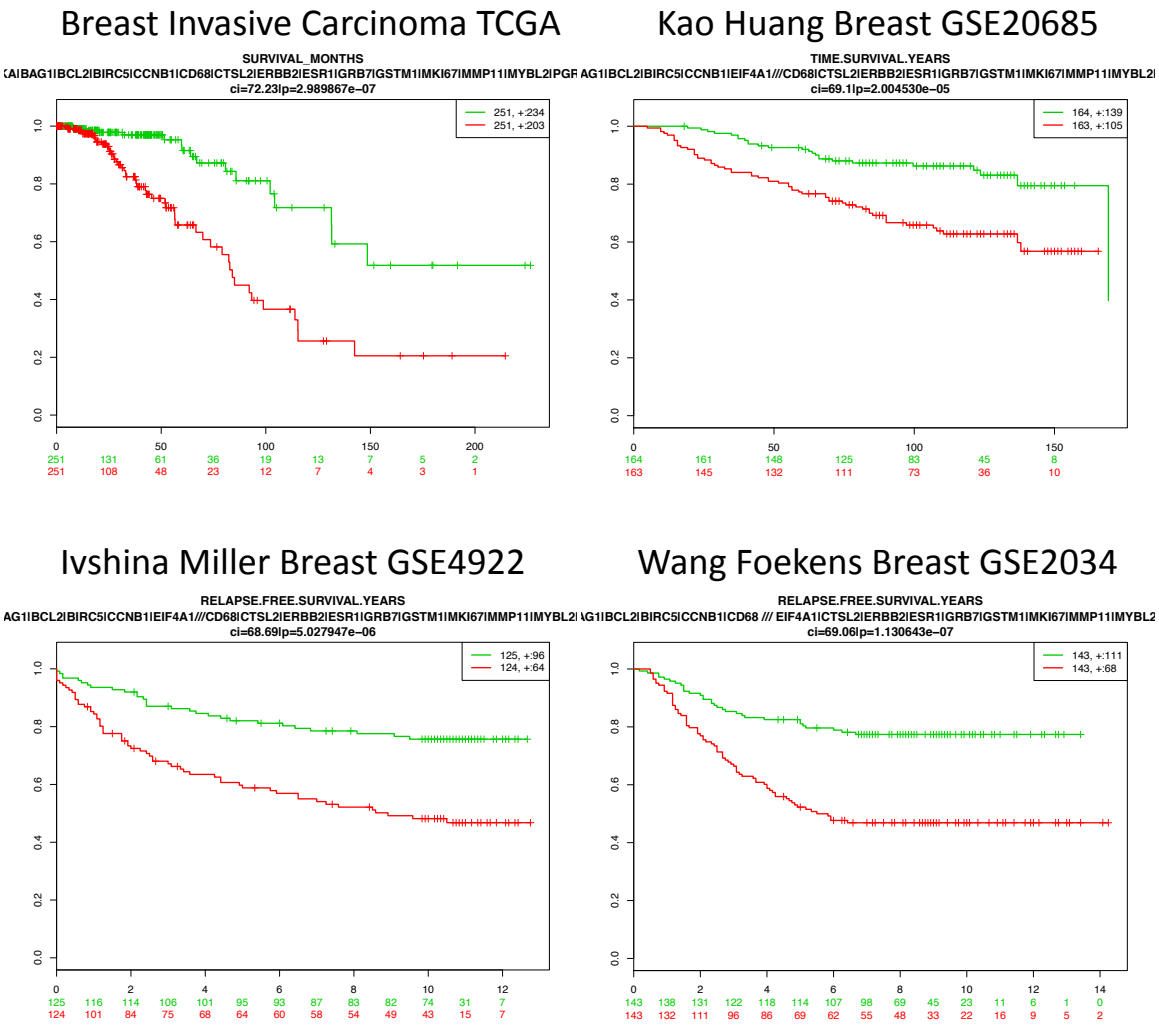


Figure 19: Evaluation of the breast cancer prognostic biomarker “OncotypeDX” in four breast cancer datasets.

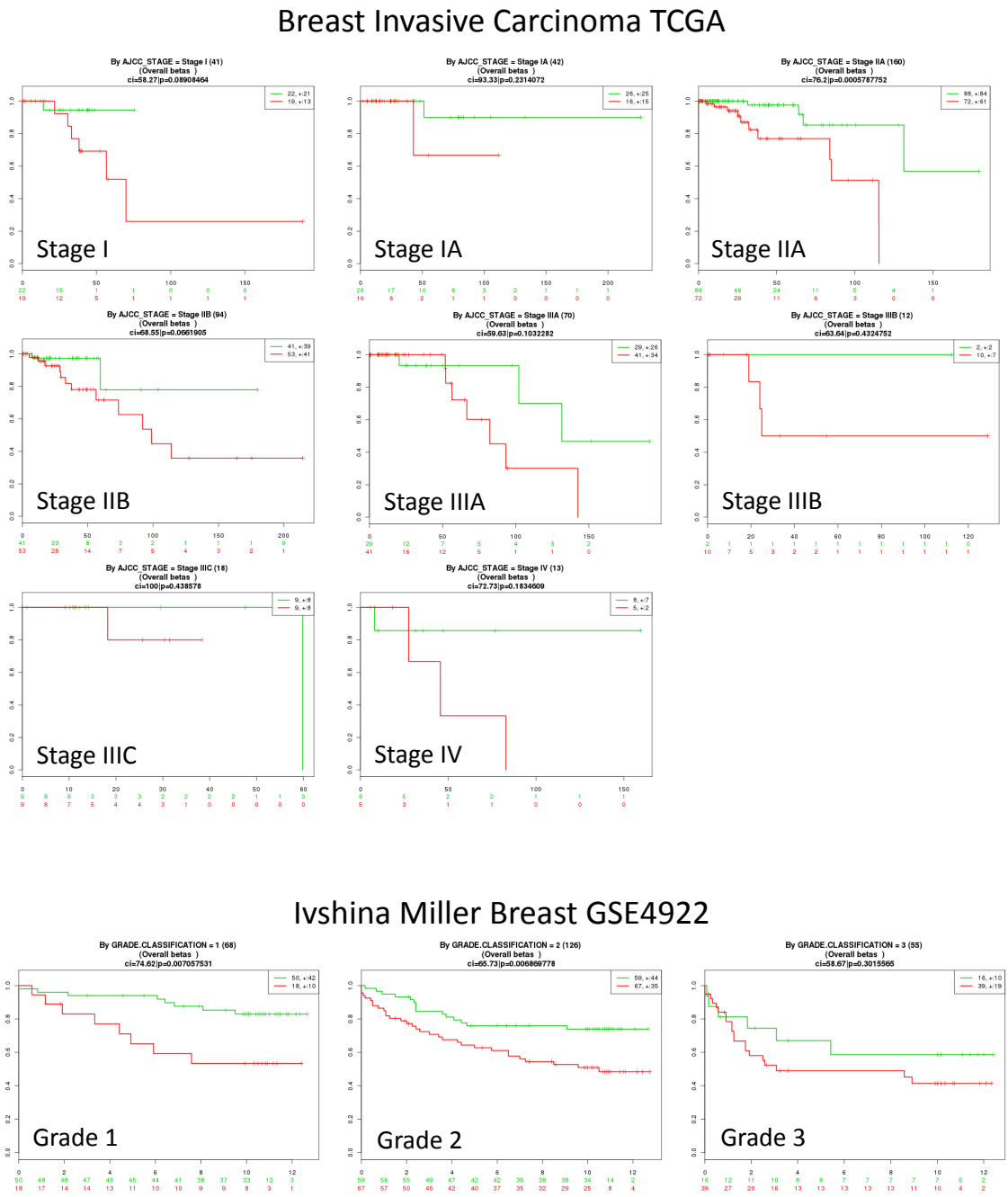


Figure 20: Evaluation of OncotypeDX in different breast cancer tumors of different grades or stages.

Comparing two Lung Cancer biomarkers

Here we will compare two biomarkers proposed for survival of non-small-cell lung cancer (NSCLC). The first lung biomarker was proposed by Boutros [11] and contains the following genes: STX1A, HIF1A, CCT3, HLA-DPB1, RNF5, and MAFK. The second lung biomarker was proposed by Chen [12] and contains the genes

DUSP6, MMD, STAT1, ERBB3, and LCK. These two lung biomarkers attempt to predict the same event (survival), use a similar number of genes, but none of the genes are the same. In SurvExpress, we used the maximum row average for duplicated genes, two risk groups, Cox fitting, and the prognostic median to generate risk groups. We used a special lung meta-base build in our research group, which is composed of more than 1,000 samples obtained from six authors (Bild [13], Raponi [14], Zhu [15], Hou [16], NCI [17], Okayama [18]), equivalent Affymetrix gene expression platform, and that contain all these genes.

Table 5: Datasets and results of the Boutros and Chen biomarkers for the lung cancer example.

Dataset	Samples/ Censored	Boutros p-Risk Groups	Boutros Overall CI	Chen p-Risk Groups	Chen Overall CI
Raponi Beer GSE4573 [14]	130/63	0.255	57.2	0.019	55.3
Bild Nevins GSE3141 [13]	108/50	0.027	59.4	0.023	57.6
Zhu Tsao GSE14814 [15]	72/49	0.401	59.0	0.009	63.6
Hou Philipsen GSE19188 [16]	64/23	0.771	54.5	0.028	62.4
Director´s Challenge Consortium NCI [17]	444/207	0.001	58.2	0.001	60.3
Okayama Kohno GSE31210 [18]	226/191	0.222	59.1	0.006	66.1

The results show that both biomarkers are able to separate risk groups characterized by differences in gene expression between them (see Kaplan-Meier and box plots respectively in Figure 21). Nonetheless, the p-value of the risk group separation, the concordance index, and the significance of the coefficients fitting in the Cox model were slightly better in the Chen biomarker. To analyze the biomarkers deeply, we tested the biomarker per author as summarized in Table 5. The results show that Boutros biomarker fails in four datasets (the log-rank test for difference in risk groups is not significant, see Figure 22), and that Chen biomarker works better in almost all databases (Figure 23). In summary, these results suggest that the performance of Chen biomarker is superior.

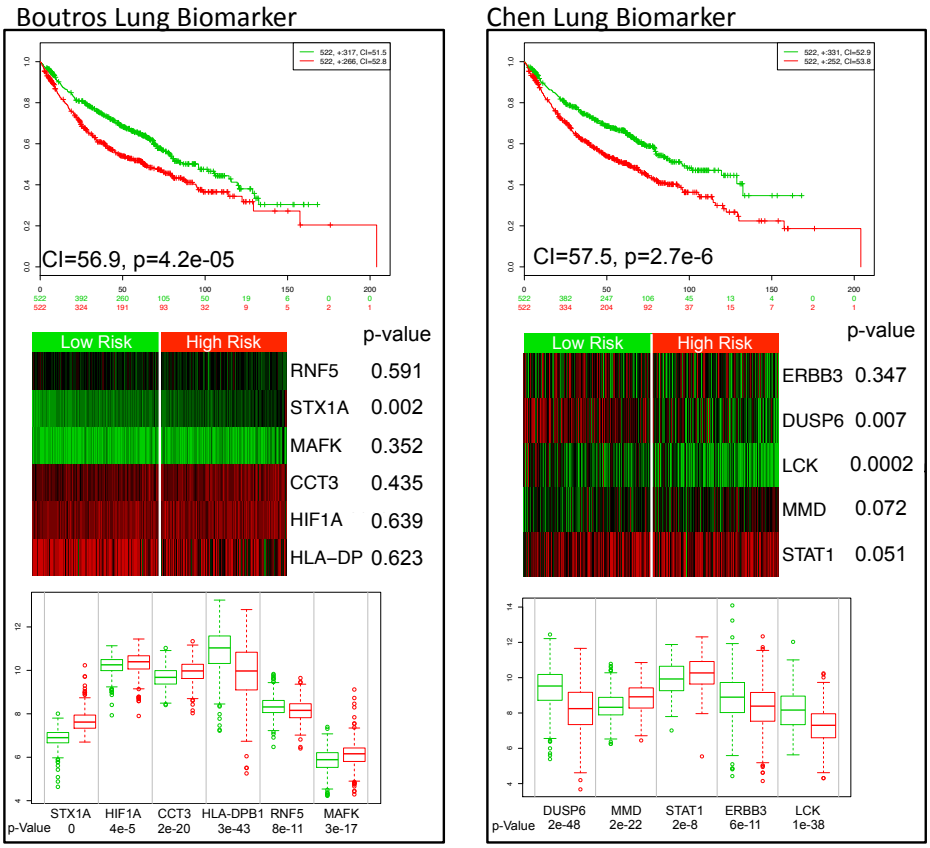


Figure 21. Comparison of two survival NSCLC biomarkers. Left panel shows the 6-gene Boutros biomarker while right panel shows the biomarker from Chen. P-values correspond to log-rank test, Chi-square/Wald test of Cox fitting, and t-test corresponding to the Kaplan-Meier, Heat Map, and Box plots respectively.

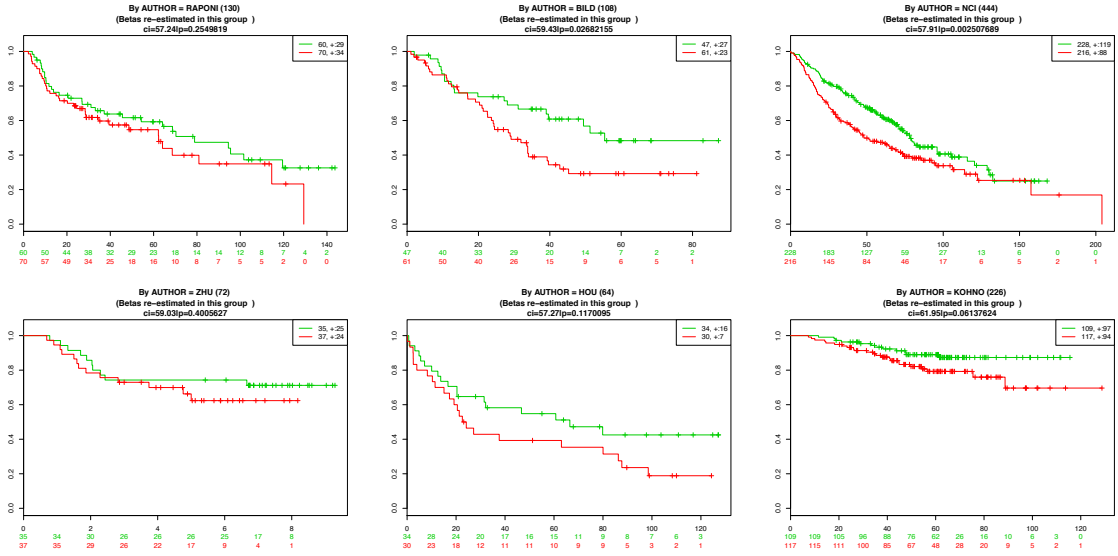


Figure 22. Performance of the Boutros biomarker among 6 authors.

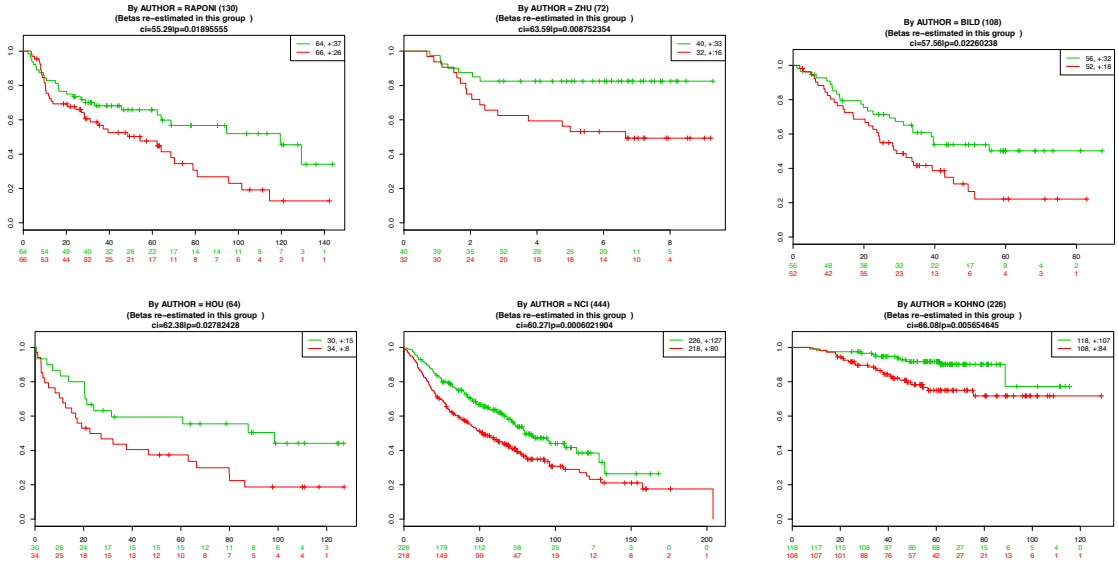


Figure 23. Performance of the Chen biomarker among 6 authors.

SOME APPROACHES FOR PROGNOSTIC BIOMARKER IDENTIFICATION

There have been plenty of statistical and computational efforts to evaluate the correlation or association of single genes or groups of genes with the clinical outcome for different types of cancer. Therefore, provide a comprehensive review of these techniques is out of the scope of SurvExpress. Nevertheless, in an effort to provide a starting point in which a prognostic biomarker can be identified, we will provide a brief representative summary of common methodologies. Independently of the source, biomarkers can be easily evaluated using the web tool SurvExpress. For this, we can use the coefficient values provided by their authors feeding the values as weights in the analysis page, or by pasting the genes filtered by any method and letting SurvExpress to estimate the β in the specific cohort. Some studies have suggested that the number of genes included in the prognostic signature is important [19]. Therefore, we recommend the use of a sensible small number of genes.

Univariate Feature Selection

The most straightforward and intuitive approach for prognosis biomarker discovery is to use the top-ranked genes from a univariate test. The univariate test is commonly the Cox model assessing one gene at the time. Such methodology is the first approach used to characterize any new cohort that is being published. Botling *et al.* performed one recent example of this methodology were they generated a cohort of 196 patients with clinical information and long-term follow-up information of non-small cell lung cancer [20]. They obtained 450 survival-related genes using a univariate Cox proportional hazard regression (significance level $p<0.01$). Then, genes were tested in a meta-base constructed out of five public datasets. Only 14 genes were found statistically significant ($p < 0.001$) with a false discovery rate < 1%. Afterwards, the gene cell adhesion molecule 1 (CADM1) was deeply analyzed using immunohistochemistry and *in-silico* validation on two independent non-small lung cancer cohorts. They concluded that such marker, CADM1, could be established as an immune-histochemical marker for survival prognosis. In SurvExpress, we could use either the 14 genes or the CADM1 gene. In the Lung Metabase provided by SurvExpress, CADM1 is significantly associated to risk.

Penalized regression

Another well-documented statistical approach to evaluate survival-related genes in multi-dimensional data is the use of the penalized regression such as Lasso (L1) and Ridge (L2). Lasso and Ridge perform a regression penalizing the β coefficients by a

constant λ resulting in many β coefficients close or equal to zero. The best value of λ is obtained by evaluating λ in a wide range of values. Lasso performs an absolute value penalization whereas Ridge performs a squared penalization. The work performed by Yoshihara is an example of gene selection obtained by such analysis [21]. To find a progression-free-survival (PFS) signature, they performed a univariate Cox proportional hazard regression and only the significant genes (P-value < 0.01) were taken into account. Then a prognostic index was generated using an adjustment in the regression coefficients using the ridge regression. The resulting 88-genes signature was independent of other clinical characteristics to predict PFS in 110 patients ovarian cancer datasets as well as in a different dataset. This signature was also tested to predict overall survival in another two cohorts showing promising performance.

Survival Principal Component Analysis

When using principal component analysis (PCA) to discover survival-associated genes, there is no guarantee that the genes selected will be correlated to survival time. Thus, Bair and Tibshirani proposed a supervised survival-oriented PCA in which the best-ranked variables from a Cox score in a training set are used to perform a PCA [22]. This method was successfully applied by Konstantinopoulos *et al.* to pre-select 650 genes [23]. Then, they generate a custom array, which was used to create a blind test cohort. Using additional filters, they generated a 19-genes model that discriminates between overall survival times in two validation sets. These genes can be easy pasted into SurvExpress.

Gene Ontology Associated Genes by Multiple Survival Screening

Previous approaches have used purely statistical information to select genes. Different procedures have been proposed that integrate biological information into the algorithm. In this regard, a gene ontology (GO) algorithm was developed under the assumption that biomarkers may not be robust in independent cohorts due to cancer heterogeneity. Multiple Survival Screening (MSS) starts by selecting a group of genes associated to survival ($p < 0.05$) to explore their related GO categories [24]. Then, they generated one million of 30 randomly selected genes sets that were tested for survival association on bootstrapped training groups. The top 30 most frequent genes among those highly predictive were used as potential signatures. Using this algorithm, they reported three robust signatures that stratify independent cohorts of breast cancer.

Protein-protein interaction network exploration

The integration of biological information into algorithms also includes protein-protein interaction (PPI) networks. This has the direct advantage of using sets of genes having functional biological meaning. In this subject, an algorithm was proposed to detect network modules having 'differential activity' [25], the equivalent of differential gene expression. However, this algorithm was developed for classification rather than for survival regression. Then, SurvNet adapted this algorithm to survival problems [26]. The procedure starts by using Cox survival associated genes as node network seeds. Then, an expansion step is performed to every seed. It follows a greedy algorithm that explores the PPI network expanding the module if a module score function increases. Finally, the modules obtained are clinically evaluated by performing a multivariable Cox proportional hazard regression. The results are displayed as best modules to predict survival. These modules can be easily pasted into SurvExpress to test the biomarker in several datasets.

References

1. Collett D (1993) Modelling Survival Data in Medical Research: Chapman and Hall/CRC. 368 p.
2. Therneau TM, Grambsch PM (2010) Modeling Survival Data: Extending the Cox Model (Statistics for Biology and Health): Springer. 350 p.
3. Kleinbaum DG, Klein M (2011) Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health): Springer. 715 p.
4. Heagerty PJ, Zheng Y (2005) Survival model predictive accuracy and ROC curves. Biometrics 61: 92-105.
5. Bovelstad HM, Borgan O (2011) Assessment of evaluation criteria for survival prediction from genomic data. Biom J 53: 202-216.
6. Paik S, Shak S, Tang G, Kim C, Baker J, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817-2826.
7. Network CGA (2012) Comprehensive molecular portraits of human breast tumours. Nature 490: 61-70.
8. Kao KJ, Chang KM, Hsu HC, Huang AT (2011) Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer 11: 143.
9. Ivshina AV, George J, Senko O, Mow B, Putti TC, et al. (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res 66: 10292-10301.
10. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671-679.
11. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, et al. (2009) Prognostic gene signatures for non-small-cell lung cancer.
12. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, et al. (2007) A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 356: 11-20.
13. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439: 353-357.
14. Raponi M, Zhang Y, Yu J, Chen G, Lee G, et al. (2006) Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 66: 7466-7472.
15. Zhu CQ, Ding K, Strumpf D, Weir BA, Meyerson M, et al. (2010) Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. J Clin Oncol 28: 4417-4424.
16. Hou J, Aerts J, den Hamer B, van Ijcken W, den Bakker M, et al. (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 5: e10312.

17. Shedden K, Taylor JMG, Enkemann SA, Tsao M-S, Yeatman TJ, et al. (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nature Medicine 14: 822-827.
18. Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, et al. (2012) Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res 72: 100-111.
19. Venet D, Dumont JE, Detours V (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol 7: e1002240.
20. Botling J, Edlund K, Lohr M, Hellwig B, Holmberg L, et al. (2013) Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. Clin Cancer Res 19: 194-204.
21. Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H, et al. (2010) Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. PLoS One 5: e9615.
22. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2: E108.
23. Konstantinopoulos PA, Cannistra SA, Fountzilas H, Culhane A, Pillay K, et al. (2011) Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer. PLoS One 6: e18202.
24. Li J, Lenferink AE, Deng Y, Collins C, Cui Q, et al. (2010) Identification of high-quality cancer prognostic markers and metastasis network modules. Nat Commun 1: 34.
25. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.
26. Li J, Roebuck P, Grunewald S, Liang H (2012) SurvNet: a web server for identifying network-based biomarkers that most correlate with patient survival data. Nucleic Acids Res 40: W123-126.