

SURCOMED User's Guide: Quick Start

SURCOMED is a method to identify cancer-driving mechanisms based on the search of gene expression survival signatures using tumor transcriptomic and clinical data from cancer patients. The usage of the SURCOMED web-based interface can be described in 5 steps:

1. Select the search space (see Figure 1)
2. Add constraints (see Figure 2)
3. Configure optimization (see Figure 3)
4. Run the search (see Figure 4)
5. Identify the underlying biological mechanisms (see Figure 5)

The screenshot shows the SURCOMED web interface. The main heading is "Search for optimal combinations of survival markers". The interface is divided into several sections:

- Cancer type:** Includes a dropdown menu with "breast" selected.
- Entity type:** Includes a dropdown menu with "symbols" selected.
- Search space:** Includes a list of genes (FASLG, GDF10, IFNA17, CNTF, LEFTY2, GREM1, CSF1, TNFSF9, IL2, CER1, IFNA7, INHA, BMP6, TNFSF13, IFNA4, IL23A, IFNA14) and an "Upload" button. A red box highlights this section, and a red arrow points from a text box to it.
- Forced states:** Includes a dropdown menu with "Examinar..." and an "Upload" button.
- Censored states:** Includes a dropdown menu with "Examinar..." and an "Upload" button.
- Other parameters:** Includes input fields for Population size (30), Selection number (15), Elite size (5), Iteration number (5), and Combination size (3).

A red box highlights the "Search space" section, and a red arrow points from a text box to it. The text box contains the following text:

1. Select the search space
The search space can be defined by copy-paste in this entry box or by browsing a local file

At the bottom of the form, there is a "send results to:" field with an "email" input and a "Search" button.

Figure 1. Select the search space. The search space is the list of genes eligible for the combinations. They can be copied and pasted in the entry box or loaded from a file. For a genome-wide search, click the radio button with the label "genome". Each gene symbol should be specified in a different line.

The screenshot shows the SURCOMED web interface. At the top, there are tabs for 'Search for combinations' and 'Test combinations', and a 'SURCOMED' logo. Below the tabs, there is a 'load example' button. The main section is titled 'Search for optimal combinations of survival markers'. It is divided into several panels:

- Cancer type:** A dropdown menu with 'breast' selected.
- Entity type:** Radio buttons for 'symbols' and 'gene list'.
- Search space:** A list of genes including FASLG, GDF10, IFNA17, CNTF, LEFTY2, GREM1, CSF1, TNFSF9, IL2, CER1, IFNA7, INHA, BMP6, TNFSF13, IFNA4, IL23A, and ICN1A.1A. Below the list is an 'Examinar...' input field and an 'Upload' button.
- Forced states:** A section with an 'Examinar...' input field, a 'forc.txt' file upload button, and a text input 'or enter gene name'. Below this is a dropdown for 'SPP1' and a 'x' button.
- Censored states:** A section with an 'Examinar...' input field, a 'cens.txt' file upload button, and a text input 'or enter gene name'. Below this is a dropdown for 'BMP3' and a 'x' button.
- Other parameters:** Input fields for 'Population size' (30), 'Selection number' (15), 'Elite size' (5), 'Iteration number' (5), and 'Combination size' (3).

A red box highlights the 'Forced states' and 'Censored states' sections. A callout points to them with the text: '2. Add constraints Forced and Censored states can be added one by one or in block from a file'. At the bottom, there is a 'send results to:' input field with 'email' entered, and a 'Search' button.

Figure 2. Add constraints. These constraints refer to gene states that will be avoided or included by the algorithm in every combination. Constraints can be manually added individually through the interface or loaded as a set from a file. In the latter case, the state can be either "high" or "low", and the syntax should be as follows: *gene_name::state*

Example:
 TNFSF9::low
 IL2::high
 CER1::low

load example

Search for optimal combinations of survival markers

Cancer type
☒ breast ⓘ

Entity type
☒ symbols

Search space
☒ gene list
☐ genome

Search space
 FASLG
 GDF10
 IFNA17
 CNTF
 LEFTY2
 GREM1
 CSF1
 TNFSF9
 IL2
 CER1
 IFNA7
 INHA
 BMP6
 TNFSF13
 IFNA4
 IL23A
 IFNA1A
 list.txt

Forced states
 Examiner... forc.txt

 or enter gene name

Censored states
 Examiner... cens.txt

 or enter gene name

Other parameters
 Population size 30
 Selection number 15
 Elite size 5
 Iteration number 5
 Combination size 3

3. Configure optimization
 The default iteration number of 5 is just for exploratory purposes; we suggest at least 20 iterations for an actual search

send results to: email

© SIB Swiss Institute of Bioinformatics 2016 Vital-IT

Figure 3. Configure optimization. There are 5 optimization parameters that can be changed, namely, population size, selection number, elite size, iteration number, and combination size.

Population size. This parameter refers to the number of combinations that are going to be evaluated at each iteration of the algorithm. Large populations help to prevent the optimization from becoming trapped in a local optimum, but they are more time consuming.

Selection number. This parameter refers to the number of combinations that are going to be selected from the population after the evaluation of these combinations in survival terms. We suggest to use a selection number of half the population size.

Elite size. This parameter refers to the number of the best historical combinations across iterations that will be directly transferred from one generation to the next to prevent loss of combinations with the best score. The elite number should be smaller than the selection number. We suggest an elite number not higher than half of the selection number in order to provide a certain freedom to the optimization process.

Iteration number. This parameter refers to the number of rounds of combinations that are going to be generated, evaluated, and selected by the algorithm. We suggest to start with some exploratory runs with 3-5 iterations, changing the population size before performing an actual analysis with 20 or more iterations. The greater the number of iterations, the longer it will take the algorithm to finish the search.

Combination size. This parameter refers to the number of gene states included in each combination. On the website, this number is limited to a maximum value of 5.

Search for optimal combinations of survival markers

Cancer type
☒ breast
☐ prostate

Entity type
☒ symbols
☐ genes

Search space
☒ gene list
☐ genome

Search space
 FASLG
 GDF10
 IFNA17
 CNTF
 LEFTY2
 GREM1
 CSF1
 TNFSF9
 IL2
 CER1
 IFNA7
 INHA
 BMP6
 TNFSF13
 IFNA4
 IL23A
 ICN1A14
 Examiner... list.txt
 Upload

Forced states
 Examiner... forc.txt
 Upload
 or enter gene name
 SPP1 x

Censored states
 Examiner... cens.txt
 Upload
 or enter gene name
 BMP3 x

Other parameters
 Population size: 30
 Selection number: 15
 Elite size: 5
 Iteration number: 5
 Combination size: 3

4. Run the search
 If an email is entered here, the user will receive a notification when the search is done with a link to the results

send results to: user@gmail.com
 Search

© SIB Swiss Institute of Bioinformatics 2016 Vital-IT

Figure 4. Run the search. Once the optimization parameters are set up, the user only has to click on the 'Search' button and wait. We strongly recommend to complete the dedicated box with an email address, so the search will be running on our server and the user will receive a notification when the search is done with a link to the results. By doing so, the user can leave the website and follow the link later on to access the results.

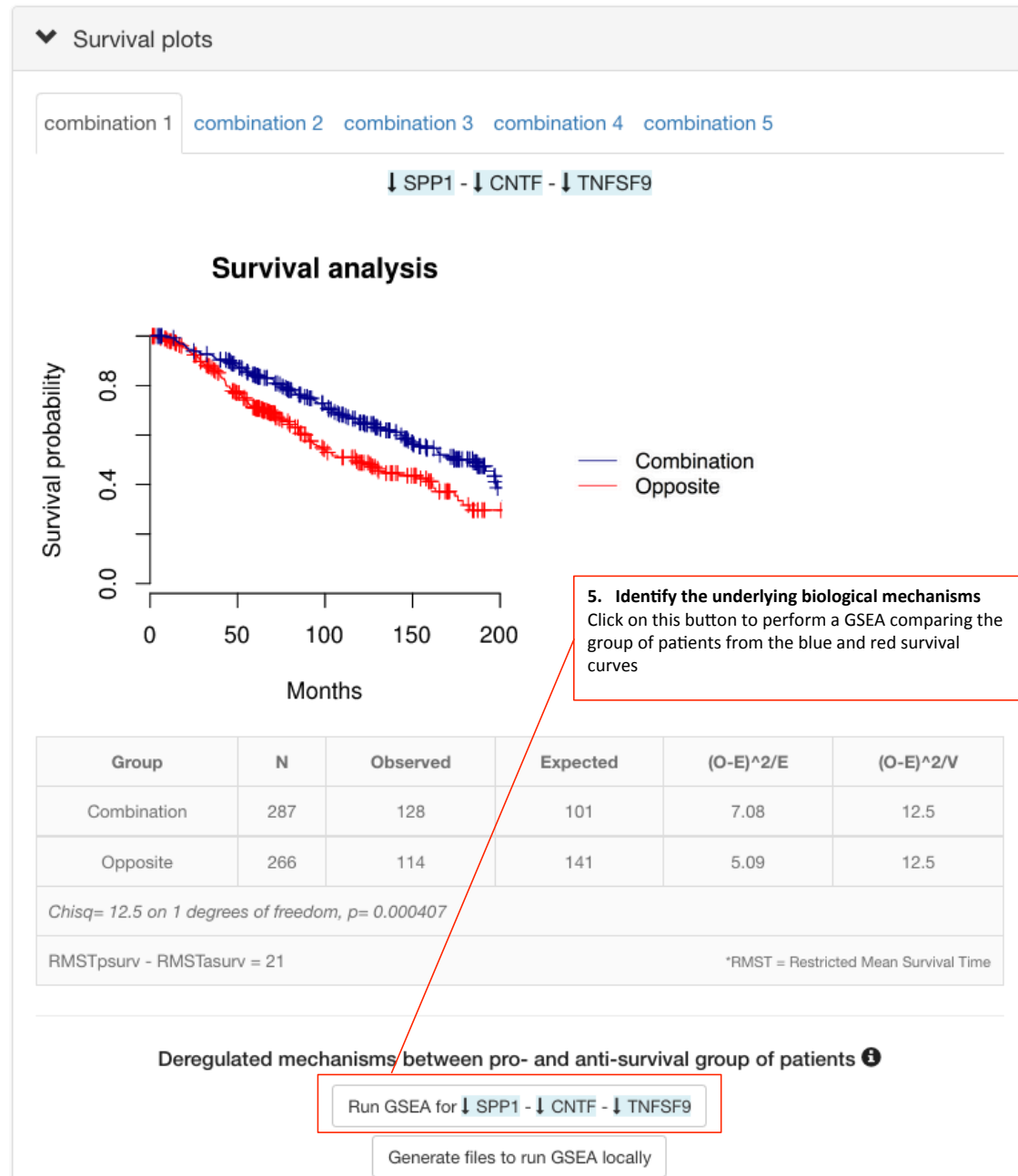


Figure 5. Identify the underlying biological mechanisms. Once the search is complete, the best combinations are displayed on the website with their corresponding survival curves. Just below the survival curves, there is a link to perform the Gene Set Enrichment Analysis (GSEA), or, alternatively, to generate and download the required files in order to run the GSEA locally with personalized settings.

Once the GSEA is done, the results can be explored online using the Internet browser or may be downloaded (see Figure 5). The main result is the list of up- and down-regulated biological mechanisms, which can be visualized online in the browser or downloaded.

GSEA Report for Dataset Expression

Enrichment in phenotype: **prosurvival** (235 samples)

- 741 / 2086 gene sets are upregulated in phenotype **prosurvival**
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 26 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot of enrichment results](#)
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to interpret results](#)

Enrichment in phenotype: **antisurvival** (242 samples)

- 1345 / 2086 gene sets are upregulated in phenotype **antisurvival**
- 5 gene sets are significantly enriched at FDR < 25%
- 90 gene sets are significantly enriched at nominal pvalue < 1%
- 195 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot of enrichment results](#)
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to interpret results](#)

The list of up- and down-regulated biological mechanisms can be visualized and downloaded as html or excel files

Dataset details

- The dataset has 25235 features (genes)
- No probe set => gene symbol collapsing was requested, so all 25235 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 698 / 2784 gene sets
- The remaining 2086 gene sets were used in the analysis
- List of [gene sets used and their sizes](#) (restricted to features in the specified dataset)

Gene markers for the **prosurvival versus antisurvival** comparison

- The dataset has 25235 features (genes)
- # of markers for phenotype **prosurvival**: 13180 (52.2%) with correlation area 51.0%
- # of markers for phenotype **antisurvival**: 12055 (47.8%) with correlation area 49.0%
- Detailed [rank ordered gene list](#) for all features in the dataset
- [Heat map and gene list correlation](#) profile for all features in the dataset
- [Butterfly plot](#) of significant genes

Global statistics and plots

- Plot of [p-values vs. NES](#)
- [Global ES](#) histogram

Other

- [Parameters](#) used for this analysis

Figure 5. GSEA results main page. Results from GSEA can be easily explored following the links from the main page, which can be accessed online or downloaded and visualized using your Internet browser. These results include the list of up- and down-regulated biological mechanisms associated with predefined gene sets, and other complementary information concerning the GSEA such as enrichment plots and statistical scores (p-values, normalized enrichment scores, or normalized enrichment score (NES)).