CANCERTOOL MANUAL

"Cancer" is the common name designated to a group of diseases, characterized by the uncontrolled proliferation and potential dissemination of cells that accumulate mutations. Nowadays, it is the main cause of death worldwide (https://www.iarc.fr). Among all histological types of cancer, prostate, breast, lung and colorectal are the most frequent (http://gco.iarc.fr/).
The socio-economic impact and complex biological nature of these diseases have spear-headed the work of many research groups, who are integrating multiple methodologies to get an overview of the nature of cancer occurrence. One of these methodologies is the study of gene expression alterations in tumor tissues. This strategy is hampered by the complexity of getting appropriate study populations, the associated cost and processing of samples, as well as the analysis and representation of obtained data.
CANCERTOOL is envisioned to address these limitations, as a webtool that integrates gene expression data from various publicly available studies, so that researchers can access quickly and easily to a summary of relevant information as well as perform a number of basic analysis and, importantly, visualize and represent the results in an output format suitable for publication in scientific journals. It is important to emphasize that CANCERTOOL is based on a selected set of studies based on the molecular and clinical data available, so that a higher impact of the studies on the pathogenesis and progression of the disease can be achieved.

Herein, we describe how to benefit from CANCERTOOL.

Mandatory input files' requirements

- Files with gene lists MUST be in plain text (E.g.: *.txt, *.csv, *.rtf)
- Files with gene lists MUST have one single column, with a gene ID per line
- If you prefer to use the enabled text areas to submit gene lists, the guidelines described for the input files MUST be followed
- The identifiers available to perform our analysis are:
    - Gene Symbol (examples: *PPARGC1A*, *PTEN*, *MITF*)
    - Ensembl Human Gene IDs (examples: ENSG00000000003)
    - EntrezGene IDs (examples: 7105, 64102, 8813)
    - EMBL (Genebank) IDs (examples: AY358825, AF291656, U84895)

  Any other identifier type is not allowed and will not be recognized by CANCERTOOL.

0.- General information

Homepage: to redirect to initial page

Datasets: to get all the information about the datasets considered by CANCERTOOL. In this option, the user can download an Excel with the general information for all the cancers' datasets as well as the extended clinical attributes of each cancer type (one Excel sheet per dataset).

Help: the manual of CANCERTOOL, which can be read there or downloaded in pdf.

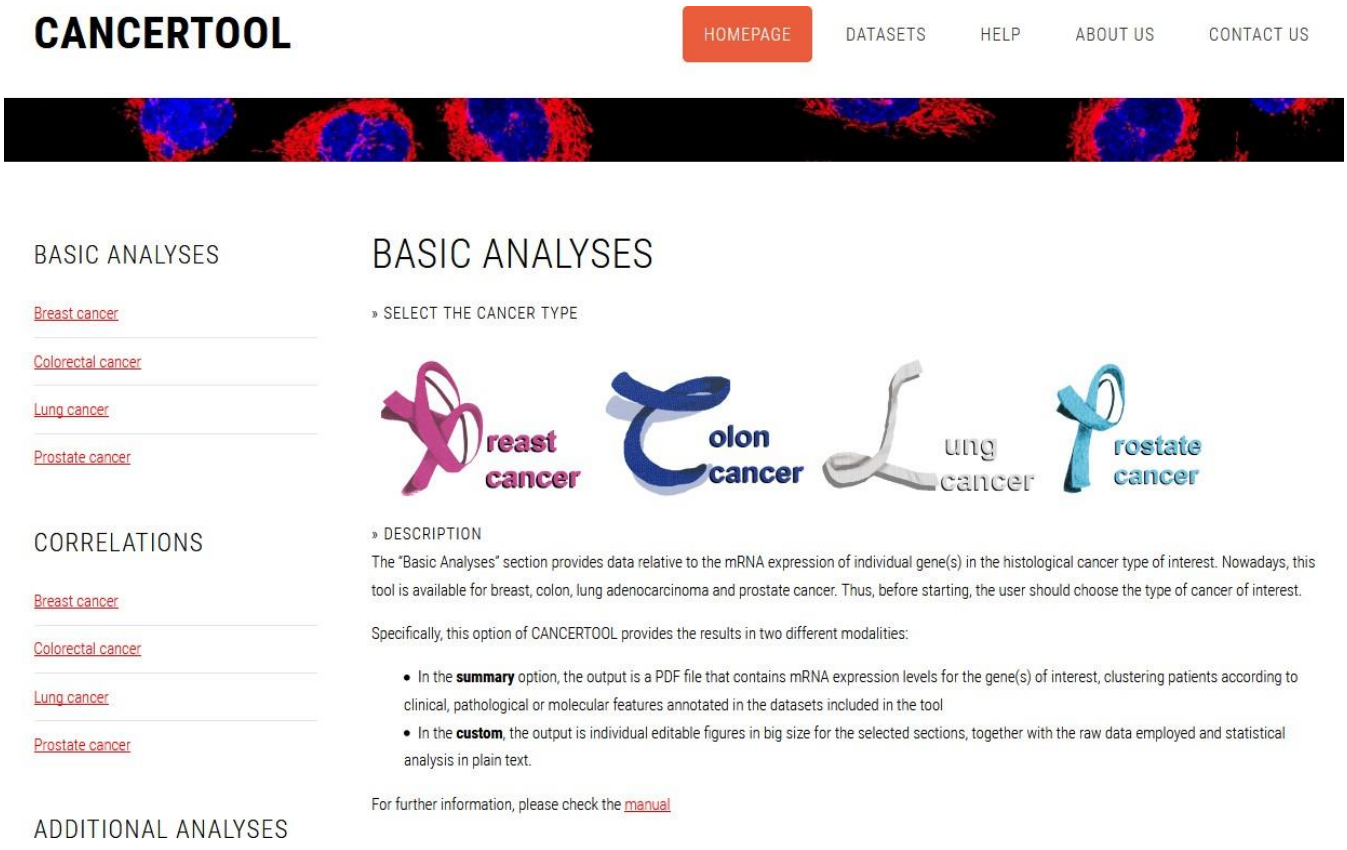
About us: description of the team that conceived and developed CANCERTOOL.

Contact us: a contact form to communicate with the developers of CANCERTOOL.

1.- Basic Analyses section

The "Basic Analyses" section provides data relative to the mRNA expression of individual gene(s) in the histological cancer type of interest. Nowadays, this tool contains available for breast, colon, lung adenocarcinoma and prostate cancer. Thus, before starting, the user should choose the type of cancer of interest.

Specifically, this option of CANCERTOOL provides the results in two different modalities. In the **summary** option, the output is a PDF file that contains mRNA expression levels for the gene(s) of interest, clustering patients according to clinical, pathological or molecular features annotated in the datasets included in the tool. In the **custom** analysis, the output is individual editable figures in big size for the selected sections, together with the raw data employed and statistical analysis in plain text.



Screenshot of the webpage where to choose the type of cancer

Once the type of cancer is selected, CANCERTOOL will redirect the user to a screen where the following information is requested:

1. *Optional*: To specify a name for the test. If a study name is not provided, a default name will be assigned. Only alphanumeric characters are allowed, so please, avoid the following characters: ~ # % & * { } \ : < > ¿ / + | " . ( ) = ? / , ; ' `
2. ***Mandatory***: The user must type or upload the gene(s) of interest. Uploaded files should follow input files guidelines (see above). If the list is typed into the window provided in the screen only

one gene per line should be inserted.

3. ***Mandatory***: The user must select the type of gene identifier that has been uploaded or typed into the text box.
4. ***Mandatory*** (Note: Summary request is selected by default): The user must select the type of analysis to be performed among the following:
    - Summary: Default option. No additional choices required.
    - Custom analysis: The user must choose among the following options:
        - The datasets of interest to be analyzed (at least one)
        - The required clinical, pathological or molecular features to be compared (at least one).

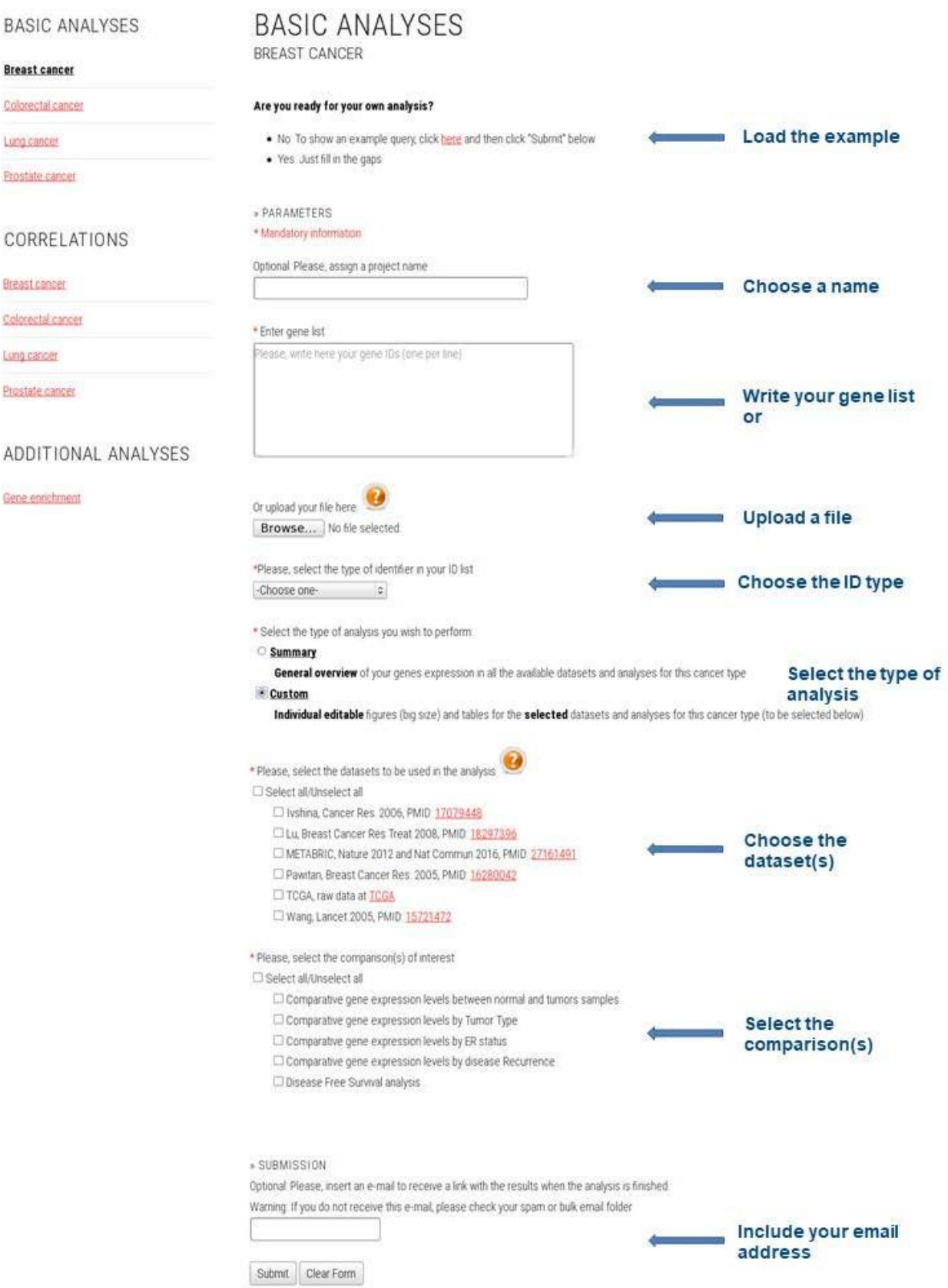
    The available options will vary in each histological cancer type, based on data availability. For more information, the user can check the information contained in each dataset.
5. *Optional*: The user can enter a valid e-mail address in the corresponding field. This will result in the submission of results to the indicated e-mail address (with a link to the webtool and to the ZIP file download page) once the analysis is finalized. If an email is not provided, the results will be made available in the webtool site once the analysis if finalized, with the option of downloading a ZIP file or visualizing them on the website.

Warning: If you do not receive this e-mail, please check your spam or bulk email folder.

A test example is provided in the website of the Basic Analyses pipeline that can be loaded by clicking in the hyperlink at the beginning of the page

The clinical, pathological and molecular features that can be selected in CANCERTOOL to compare the mRNA expression of the selected gene(s) are:

- **Status in cancer**: Available for all the cancer types. It shows the expression levels of each gene in non-tumoral specimens (defined as "Normal" or "N" and, in the case of Colonomics dataset, also "Normal Adjacent" or "N Adj" for normal tissue adjacent to the tumor) and cancer specimens (identified as "BCa" for Breast cancer, "CRC" for Colon Cancer, "LUAD" for Lung Cancer or "PCa" for Prostate Cancer). This analysis is presented in all datasets where expression values for the requested gene(s) are available. (Please, check available datasets). Statistical analysis: Student T-test or ANOVA, p-value is provided above each graph.



Screenshot of the Basic Analyses form, highlighting all the available sections

- **Status by tumor type/subtype**: Available for Breast and Prostate cancer datasets. It presents the mRNA expression levels for each gene in patients grouped by the typology of the tumor, with normal specimens if available. The available classification is:
  - Breast cancer:
    - Normal (N): Non-tumoral specimens
    - Normal-like (NL): Specimens from Normal-like cancer subtype
    - Basal-like (BL): Specimens from basal-like cancer subtype
    - HER2 Enriched (HE): Specimens from HER2-enriched cancer subtype
    - Luminal (L): Specimens from luminal cancer subtype. Some datasets provide an additional classification, differentiating Luminal A (LA) or Luminal B (LB).
  - Prostate cancer:
    - Normal (N): Non-tumoral specimens
    - PT: Specimens from primary tumors
    - M: Specimens from metastasis

This analysis is presented in all datasets that contain information about the queried gene (please, check the available datasets). Statistical analysis: ANOVA, p-value is provided above each graph.

- **Status by Gleason Score:** Only in prostate cancer. It presents mRNA expression levels of samples grouped by their Gleason grade. This analysis is provided for datasets with annotated Gleason grade (please, check the available datasets). Statistical analysis: ANOVA, p-value is provided above each graph.

- **Status by ER**: Only in Breast cancer. It presents mRNA expression levels of samples grouped by estrogen receptor status, ER positive (ER+) and ER negative (ER-). This analysis is provided for datasets with annotated ER status (please, check the available datasets). Statistical analysis: Student T-test, p-value is provided above each graph.

- **Status by Gender**: Only in Colorectal cancer. It presents mRNA expression levels of samples grouped by gender, males and females. This analysis is performed only in the datasets where this kind of data is available (please, check the available datasets). Statistical analysis: Student T-test, p-value is provided above each graph.

- **Status by Location**: Only in Colorectal cancer. It presents mRNA expression levels of samples grouped by the location of the tumor. This analysis is provided for datasets with annotated tumor location (please, check the available datasets). Statistical analysis: Student T-test or ANOVA, p-value is provided above each graph.

  - Ascending Colon (AC)
  - Cecum (CE)
  - Colon (CO)
  - Descending Colon (DC)
  - Distal (D)
  - Hepatic Flexure (HF)
  - Left (L)
  - Left Colon (LC)
  - Proximal (P)
  - Rectosigmoid Junction (RJ)
  - Rectum (RE)
  - Right (R)
  - Right colon (RC)
  - Sigmoid Colon (SC)
  - Splenic Flexure (SF)
  - Transverse Colon (TC)

- **Status by Stage**: Only in Colorectal cancer and lung cancer. It presents mRNA expression levels of samples grouped by disease stage. This analysis is provided for datasets with annotated tumor location (please, check the available datasets). Statistical analysis: Student T-test or ANOVA, p-value is provided above each graph.
- **EGFR mutant vs Non mutant**: Only in Lung cancer. It presents mRNA expression levels of samples grouped by EGFR status, EGFR mutant and non-mutant. This analysis is provided for datasets with annotated EGFR status (please, check the available datasets). Statistical analysis: Student T-test, p-value is provided above each graph.
- **KRAS mutant vs Non mutant**: Only in Lung cancer. It presents mRNA expression levels of samples grouped by KRAS status, KRAS mutant and non-mutant. This analysis is provided for datasets with annotated KRAS status (please, check the available datasets). Statistical analysis: Student T-test, p-value is provided above each graph.
- **Disease-Free Survival**: Available for cancer types. This type of analysis shows the differences into the relapse of the disease among different subgroups of the population. These subgroups are obtained by separating and comparing the patients in the four different quartiles. To estimate the survival function from available data, Kaplan-Meier Estimator is used and a Log-Rank test p-value is provided above each graph. The hazard ratio (HR) between two groups is calculated using Cox model and the resulting value is provided above each graph.
- **Overall Survival**: Available only for Lung cancer (please, check the available datasets). This type of analysis shows the mortality differences among different subgroups of the population. These subgroups are obtained by separating and comparing the patients in the four obtained quartiles. To estimate the survival function from available data, Kaplan-Meier Estimator is used and a Log-Rank test p-value is provided above each graph. The HR between two groups is calculated using Cox model and the resulting value is provided above each graph.
- **Metastasis Free Survival:** Available only for Lung cancer (please, check the available datasets). This type of analyses shows the metastasis relapse differences among different subgroups of the population. These subgroups are obtained by separating and comparing the patients in quartiles. To estimate the survival function from available data, Kaplan-Meier Estimator is used and a Log-Rank test p-value is provided above each graph. The HR between two groups is calculated using Cox model and the resulting value is provided above each graph.

Statistical Analyses description:

1. Outliers' selection. In all analyses, we check for potential outliers at each group. Outliers are defined as those samples whose expression values are smaller than Q1 - 3*IQR or larger than Q3 + 3*IQR, being (Q1: First quartile, splits off the lowest 25% of expression data from the highest 75%; Q3: Third quartile, splits off the highest 25% of data from the lowest 75, and IQR: interquartile range, the difference between 75th and 25th percentiles (IQR=Q3-Q1)). Outliers are eliminated by default in the Summary PDF, whereas in the Custom analysis two graphs are returned to the user per section and dataset: the first one with all the available samples, and the second one without the outliers.
2. Gene Expression Data distribution: Data normality distribution was confirmed for all gene expression datasets and, therefore, only parametric tests are considered on this section.
3. Student T-test is performed when comparing the mean values between two groups.
4. ANOVA is performed when comparing the mean values among more than two groups.
5. Posthoc analyses: In the custom analysis, the ANOVA statistics are accompanied by pairwise posthoc analyses using Bonferroni and Tukey HSD corrections.
6. Edgington or summatory method is performed when more than one dataset has been selected in the custom option. This test informs about the coherence among the selected datasets.
7. Log Rank test. When using the Kaplan-Meier Estimator, a Log-Rank test is calculated to check if there are significant differences among the resulting curves.
8. Hazard Ratio. HR is calculated in survival analyses between two groups using Cox proportional hazards regression model.
9. Adjusted p-value. All raw p-values are adjusted on custom part using the *Benjamini–Hochberg* (BH) procedure.

Results provided in Basic Analyses

After Basic Analyses completion, the following files are obtained:

- Summary: If the **Summary option** has been chosen, the expression levels of every gene of your list are plotted for each status and phenotypic group accessible in the selected cancer type. In this case, the output is a compressed folder (*.zip) containing one summary report per gene in PDF format. These summaries are organized in different sections that correspond to the available clinical data, which can differ among the studied cancer types.
- The **Custom option** provides individual editable big size figures and tables for the queried gene(s) in the selected datasets and analyses for the chosen cancer type. CANCERTOOL will provide a compressed folder, with a sub-folder per selected analysis. Every folder has the same

structure: a subfolder for every uploaded gene and a plain text file with the statistical results. The folder for each gene contains: i) all the graphs (in PDF and PNG formats), ii) an Excel file with the raw data used in the analysis and the sample size (if the dataset allows providing transcript information, such data are also sent to the user on this table) and, iii) excel files with the results of pairwise posthoc analyses.

The provided folder also contains in its root an Excel file ("StatisticalResults.xlsx") with all the statistical results and an additional Excel file called "CompleteRAWdata.xlsx" containing the raw expression data of the requested genes for all samples in the selected datasets.

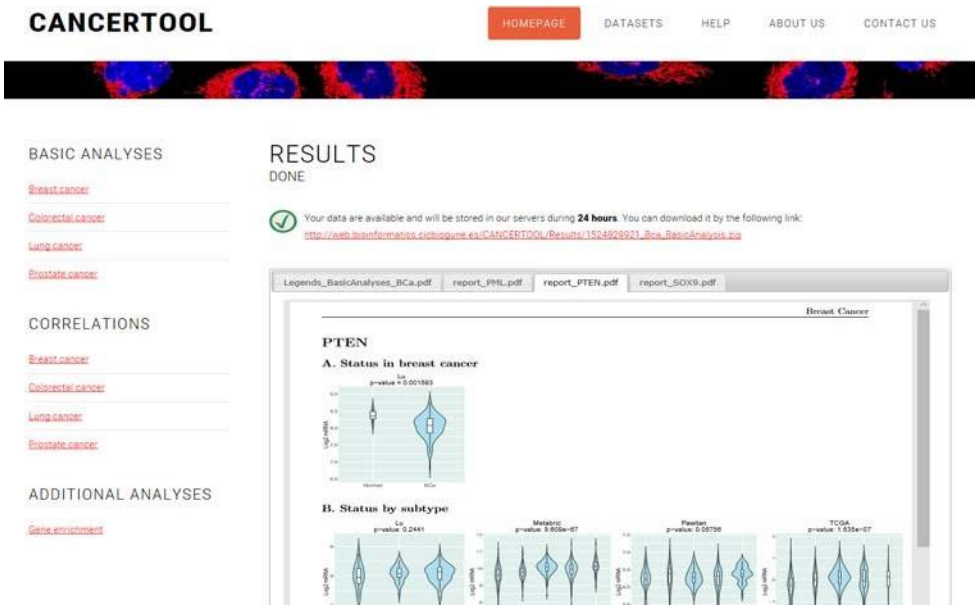
Warning: All the statistical results that are returned to the user in the excel files are calculated upon removal of outliers.

Both in the **Summary** and **Custom** analysis, the user is provided with:

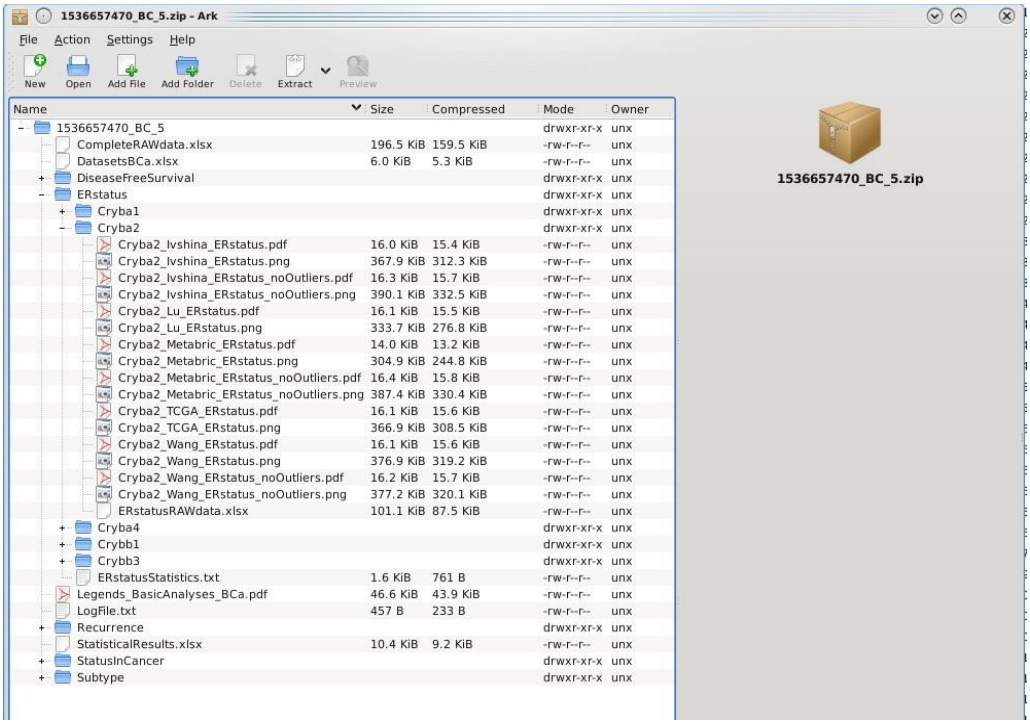
- **Datasets.xls**: An Excel file containing the information related with the datasets available for the chosen cancer type.
- **Legends_BasicAnalyses.pdf**: a PDF document with the legends associated to the Basic Analyses figures
- **LogFile.txt**: A plain text file which provides a short summary of the analysis performed.

Some additional files might also be included in the results folder:

- **NotAvailableGenes.txt**: This file is provided when one or more of the submitted gene identifiers are not available in the chosen datasets, indicating the IDs to which this situation applies.
- **Annotation.txt**: If the user provides identifiers other than Gene symbol this file will be added to the results zip file. This file is a plain text with the Gene Symbol ID that corresponds to each identifier.



Screenshot of the results webpage (Summary option)

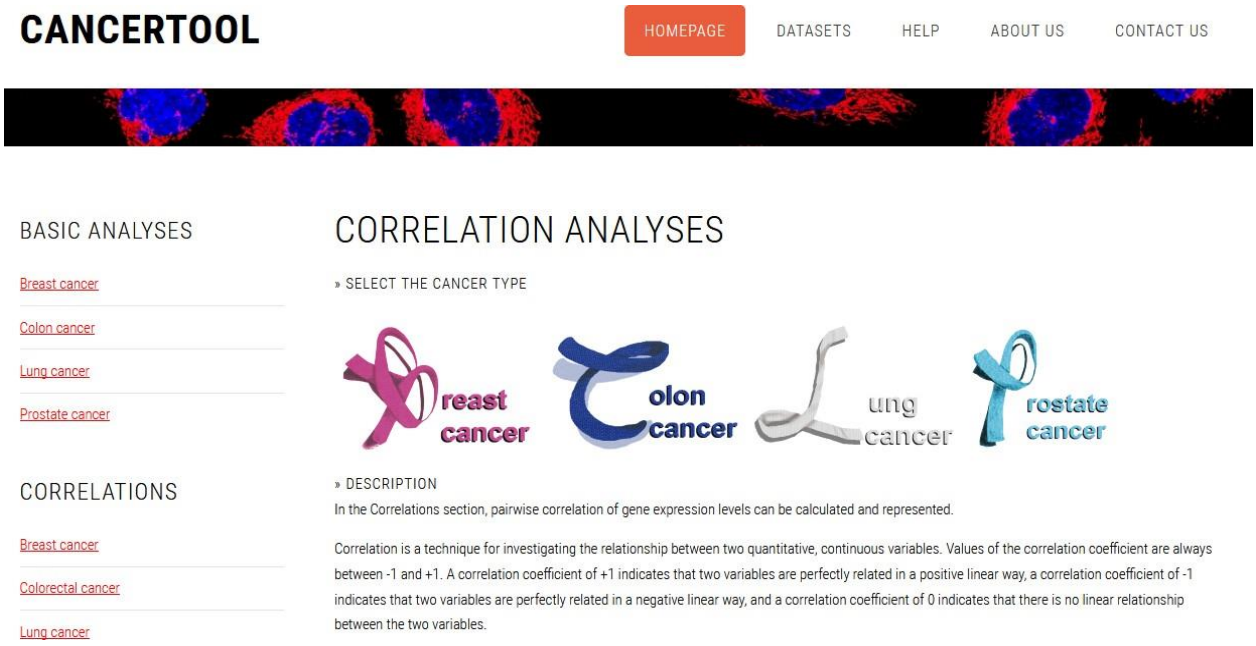


Screenshot of the resulting zip file (Custom option)

2.- Correlations section

In the **Correlations section**, pairwise correlation of gene expression levels can be calculated and represented.

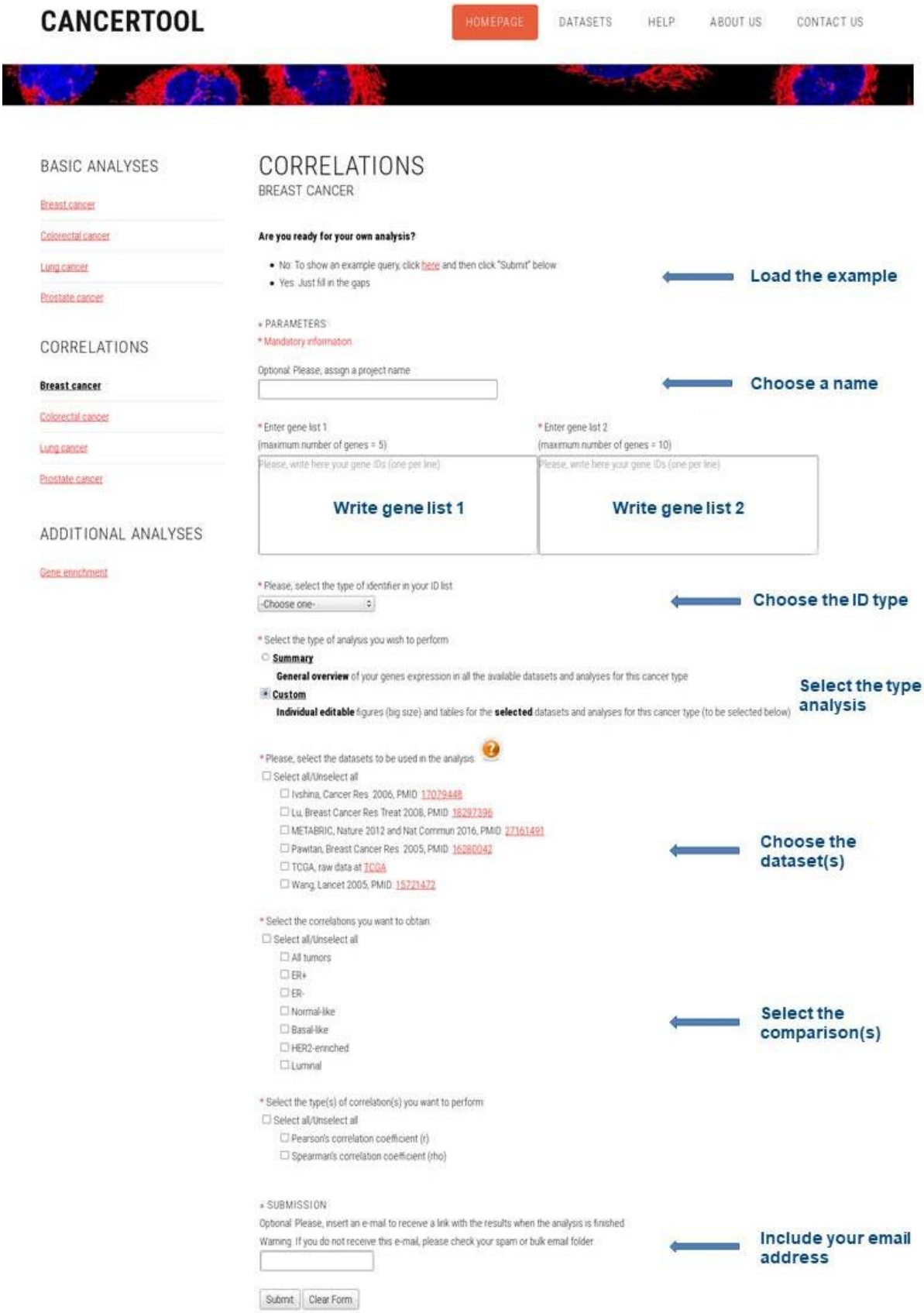
The user starts the analysis by selecting the histological cancer type of interest.



Screenshot of the webpage where to choose the type of cancer

Two boxes are provided for gene IDs of interest. All gene(s) inserted in the left box ("gene list 1") will be correlated with those present in the right box ("gene list 2"). The genes from the first list will be pairwise correlated with all the genes in the second one. This are the steps to carry out the correlation analysis:

1. *Optional*: To specify a name for the test. If a study name is not provided, a default name will be assigned. Only alphanumeric characters are allowed, so please, avoid the following characters: ~ # % & * { } \ : < > ¿ / + | " . ( ) = ? / , ; ' `
2. ***Mandatory***: Write on the left text box the gene IDs to be analyzed ("gene list 1"), following the guidelines indicated in the first point of the manual. This list is limited to **5 genes**.
3. ***Mandatory***: Write on the right text box the gene IDs to be correlated with gene list 1, following the guidelines indicated in the first point of the manual ("gene list 2"). This second list is limited to **10 genes**.
4. ***Mandatory***: Select the type of identifier that has been entered in the text boxes.
5. ***Mandatory***: Select the type of analysis to be performed. The possibilities are offered:
    A. **Summary** (Default option): No further choices are required.
    B. **Custom analysis**: Further choices:
        1. Select the datasets to be used for the correlation analysis. At least one dataset MUST be selected.
        2. Select the group or groups of patients to be considered for the correlations
        3. Select the type of statistical analysis to be performed in the correlation (Pearson and/or Spearman)



Screenshot of the Correlations form, highlighting all the available sections

6. *Optional*: The user can enter a valid e-mail address in the corresponding field. This will result in the submission of results to the indicated e-mail address (with a link to the webtool or to the ZIP file download page) once the analysis is finalized. If an email is not provided, the results will be made available in the webtool site once the analysis if finalized, with the option of downloading a ZIP file or visualizing them on the website.

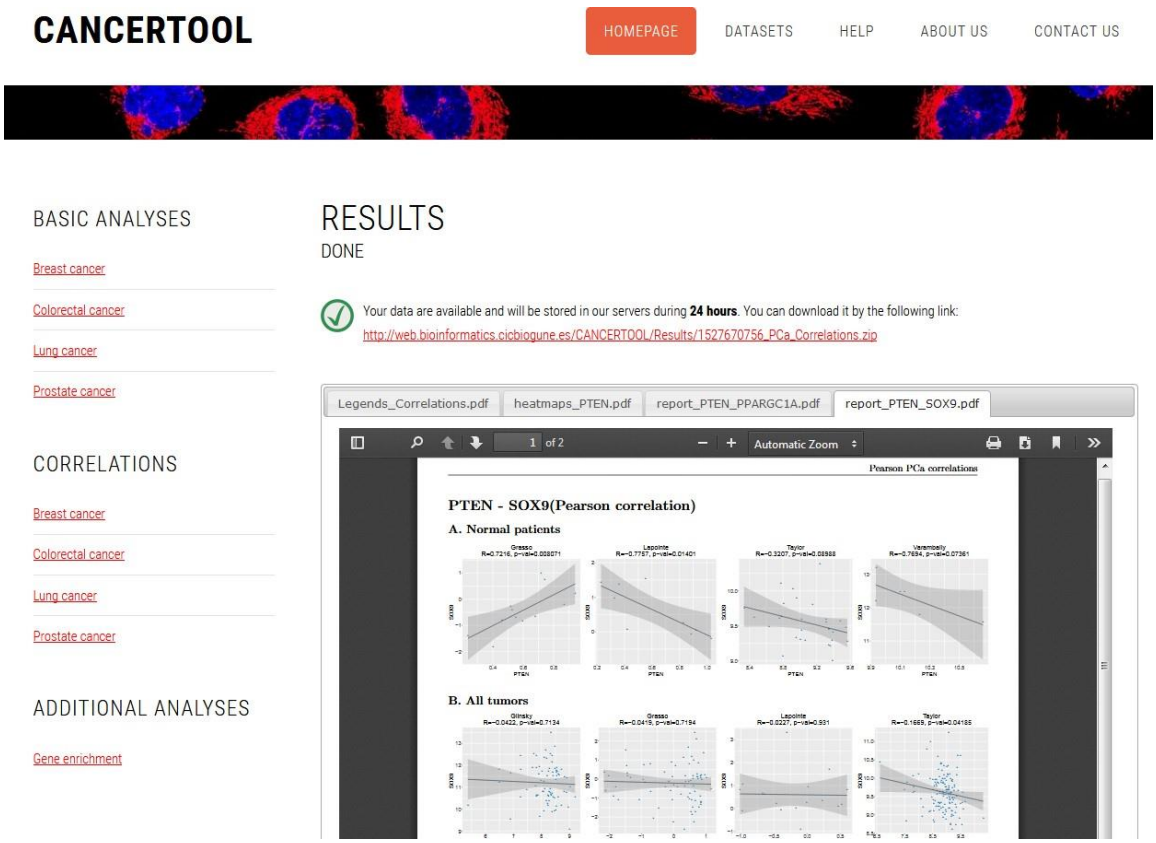
Warning: If you do not receive this e-mail, please check your spam or bulk email folder.

A test example is provided in the website of the Correlations pipeline, which can be loaded by clicking in the hyperlink at the beginning of the page.

Results provided in Correlations

After completion of the Correlations analysis, the following files are obtained:

- If the **Summary option** is selected, the output is a compressed folder with PDF files containing gene expression pairwise correlation analyses graphs in the datasets corresponding to the histological cancer selected. These summaries are organized in several sections that can differ for each cancer type depending on the clinical, pathological and molecular feature available for the datasets.

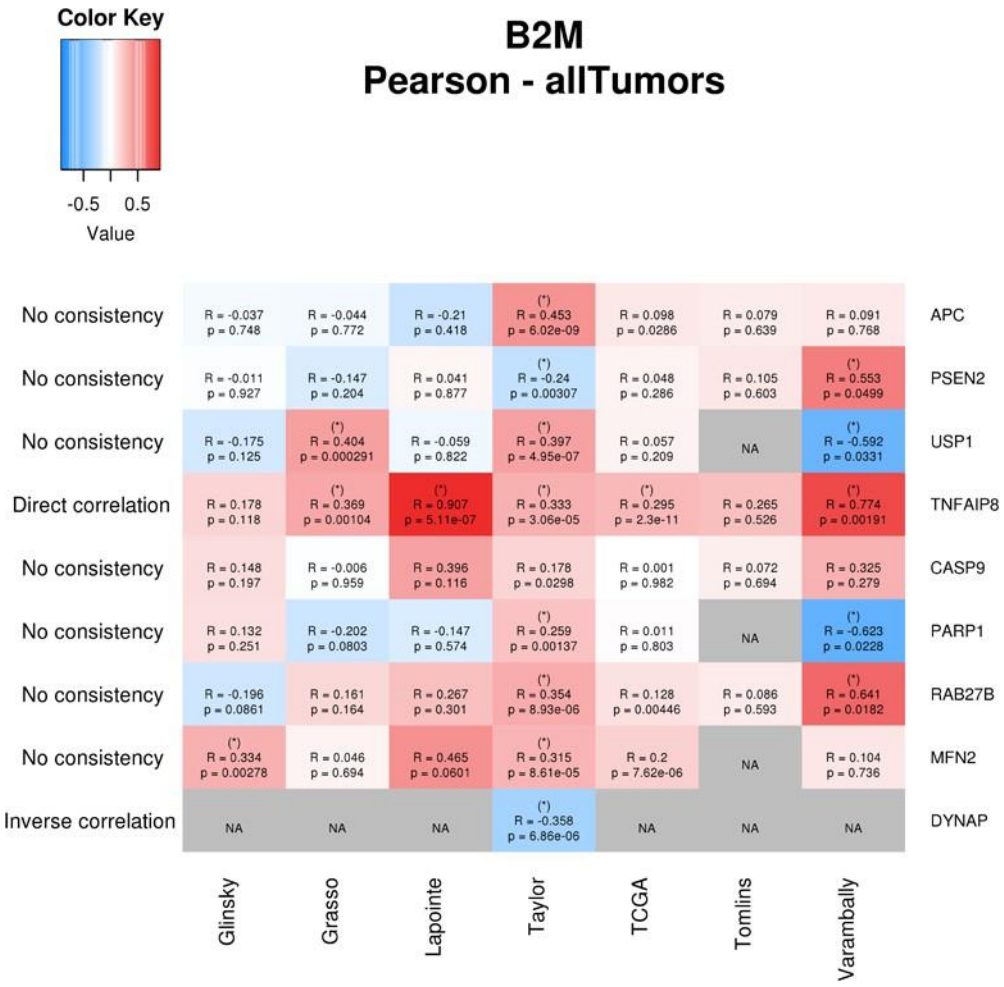


Screenshot of the results webpage (Summary option)

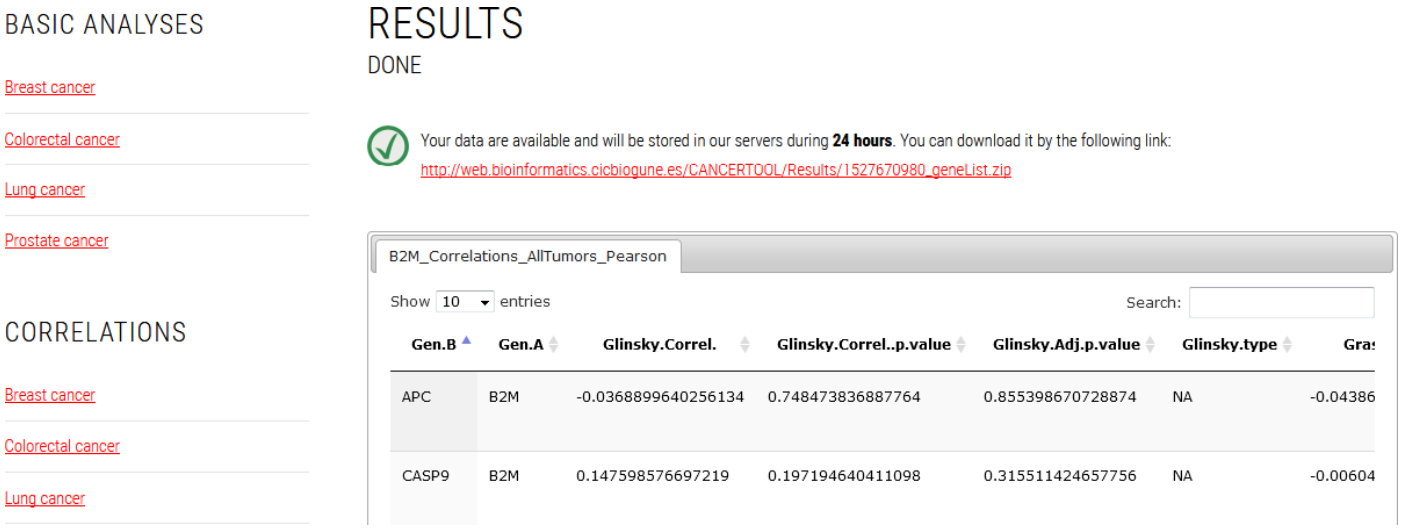
- If the **Custom Analysis option** is selected, a compressed folder will be provided with one subfolder per gene included in Gene List 1. In each subfolder you will find:

    - A table with statistical results for every correlation performed, in plain text and excel format (with results in different sheets). The table contains Pearson or Spearman correlation coefficient, p-value and adjusted p-value (all of them have been adjusted using *Benjamini–Hochberg* method), the directionality of the correlation (*Dataset.Type*; D: direct; I: inverse; NA: non Applicable) in Pearson and Spearman tests when significant ($p<0.05$). A consistency estimate (*Coherence*) is provided, that indicates that more than 50% of the datasets present a correlation with same directionality (direct or inverse; directional correlation). For this purpose, only datasets with a correlation coefficient greater than 20% ($-0.2<R<0.2$) and a p-value lower than 0.05 are considered. In the results table, additional information is provided: *AppearAt*, the number of datasets presenting data to calculate the correlation; *NumDirect*, number of datasets which present direct significant correlation; *NumInverse*, number of datasets which present direct significant correlation; *Coherence*, see above; *Sumatory*, the number of datasets presenting directional correlation; *avgSignifCorr (average significant correlation coefficient),* the mean correlation coefficient for all datasets with available data; *avgGeneralCorr (average correlation coefficient)*, the mean correlation coefficient for all datasets with available data; *percCorr (Percentage of correlation),* the percentage of datasets presenting directional correlation. For datasets that are not selected or that contain insufficient number of samples to correctly perform the analysis, the requested correlation will be annotated as *NA (Non Applicable)* in the corresponding position of the table.
    - A heatmap color coding the correlation coefficient (red towards +1 and blue towards -1) is provided for every gene entered in Gene List 1, with a grid that presents datasets in columns and Gene List 2 in lines. Each cell includes the correlation coefficient (R) and the p-value for the corresponding correlation analysis. Correlations with p-value<=0.05 and $|R| >= 0.2$ are indicated with (*). Cells with no correlation data are depicted with NA and colored in grey. On the left side, the coherence value among data sets is shown for correlation.
    - An Excel spreadsheet is provided containing the raw data used in each analysis and the sample size of the particular data set. If the dataset contains transcript information, this information is provided in the table.
    - A subfolder per selected dataset containing a big size correlation figure, in PDF and in PNG format, for every pair of genes queried.

In the results webpage, apart from the download link, the user can visualize the correlations tables obtained in the queried analysis. The tables are dynamic, sortable, and have a search engine enabled.



Example of a correlations heatmap



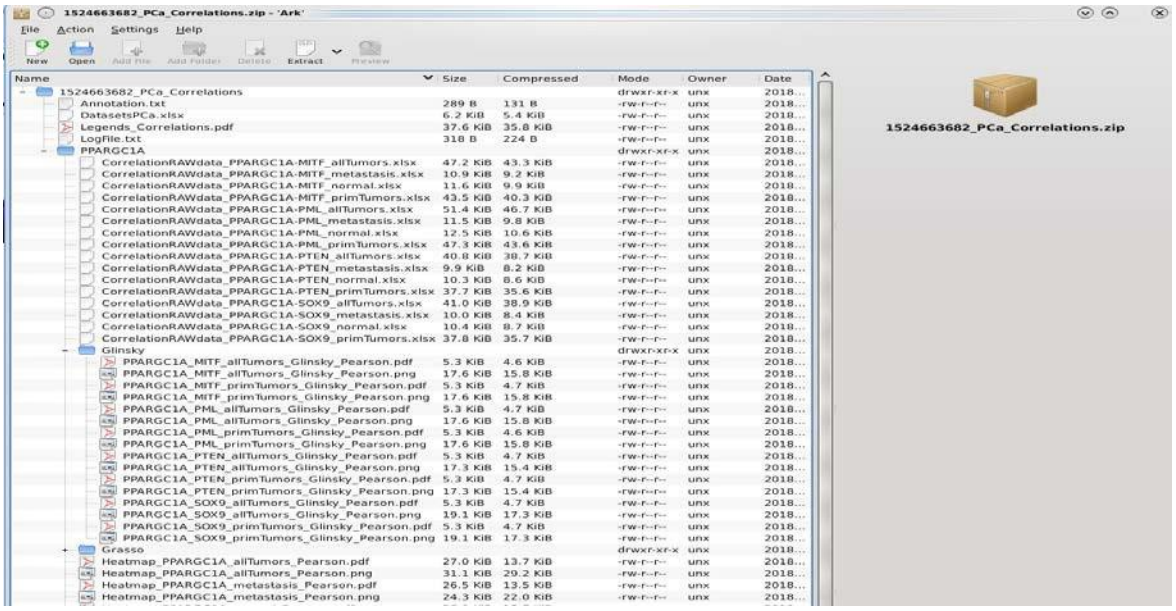
Example of the results webpage (Custom option)

The following additional files are provided regardless of the "Summary" or "Custom" analysis requested:

- **Datasets.xls**: An Excel file containing the information related with the datasets available for the chosen cancer type.
- **Legends_Correlations.pdf**: a PDF document with the legends associated to the correlations approach output figures
- **LogFile.txt**: A plain text file which provides a short summary about how the analysis has been performed.

Two files will be included in specific scenarios :

- **NotAvailableGenes.txt**: This file is provided when one or more gene identifiers are not available in the selected datasets, and will contain the ID of those genes.
- **Annotation.txt**: For Prostate cancer, this file will always provide the corresponding transcript identifiers fort each queried gene. In addition, for all cancer types, when the user inputs gene IDs other than Gene Symbol, this plain text includes the Gene Symbol that matches to each identifier queried.



Example of the resulting zip file obtained from the Custom option (Correlations)

Output remarks

1. All the graphs obtained with custom analyses of CANCERTOOL are delivered in PNG format (600 ppp) for escalar images and PDF format for vectorial ones, being both of them editable with programs such as the freeware Inkscape or GIMP.
2. All the results provided by CANCERTOOL are generated in plain text and Excel format. Please, **be aware of the local configuration of your PC, it MUST be in English format in order to avoid errors in the interpretation of the results. (Decimals must be separated by dots '.', and thousands by commas ',')**
3. All the output tables that are in plain text are tab delimited. Take it in account in order to export the resulting tables.
4. All the results obtained with CANCERTOOL are displayed directly on the website as a summary, and are downloadable via the provided link (available for 24 hours).
5. This tool offers the possibility (not mandatory) of entering an e-mail address in order to receive a message with a link to download the results after completion, avoiding waiting ties.