Cancertool

[CANCERTOOL (cicbiogune.es)](http://web.bioinformatics.cicbiogune.es/CANCERTOOL/manual.html)

Cancertool presents different sections, basic analysis, correlations and enrichment.

# Basic Analysis

Statistical analysis performed are ANOVA and t-Student (parametric test).

Status by tumor type or subtype

* Normal (N): Non-tumoral specimens, but patients with BPH (Benign Prostate Hyperplasia)
* PT: Specimens from primary tumors
* M: Specimens from metastasis

|  |
| --- |
| New Diagnosis Metastasis **NDM** are patients with really high lvls of PSA and metastasis in whose the illness did not show up until this stage because of the fast development of it.  50% of the deaths in PCa  Chemotherapy + surgery |

Status by Gleason score

|  |
| --- |
| **ISUP scale** is now use instead of the **Gleason score**. The scales go from 1 to 5, being considered tumoral tissue if the value is 3 |

Disease Free Survival. Differences into the relapse of the disease among different subgroups of the population. The survival function estimated by using Kaplan-Meier estimator and a long-rank test. The hazard ratio between two groups is calculated using Cox model.

**Statistical analysis**

1. Outliers detection below Q1 – 3\*IQR or above Q3 + 3\*IQR

**Outlier detection not consider because the database are curated.**

1. Gene express distribution to see if the data is normally distributed.
2. Parametric tests such as t-Student (N vs PCa) and ANOVA (N-PT-M)
3. Posthoc analysis (pairwise comparison using Bonferroni and Tukey)
4. Edgington or summatory method informs about the coherence among the selected datasets. Only performed when more than one dataset has been selected.
5. Log Rank test when using Kaplan-Meier estimator. It checks if there are significant differences among the resulting curves and it is based on X2 for each event time for group and sums the results.
6. Hazard ratio will show the increased rate of having an event in one curve versus the other.
7. Adjusted p-value using Benjamini-Hochberg

**Disease Free Survival** DFS

The disease-free survival gives information about when PTs per gene present biochemical recurrence, also known, as an increase in the PSA levels. These curves based on the distribution of the expression levels, which is distributed in three groups, Q1, Q2+Q3 and Q4. **ONLY PERFORMED IN PT** in order to study how long it takes to reappear.

|  |
| --- |
| **Kaplan-Meier Estimator**  Measures the survival time from a certain date to time of death or other significant event.  Product-limit estimator, non-parametric statistic used to step-wise estimate the survival function from lifetime data.  Where “d” are the number of death events at the time (t), “n” is the number of subjects at risk of death just prior to time.  The table should be sorted by ascending serial times beginning with the shortest times for each group.  The x-axis is the serial time (years) and y-axis is a cumulative probability of surviving given a time.  The mean, median and CI must be calculated per each group. |

|  |  |  |  |
| --- | --- | --- | --- |
| Subject | Serial time (years) | Status at serial time (1 or 0) | Group (1 or 2) |
|  | Var continuous | Var categorical or continuous  1 = event of interest, inc [PSA]  0 = censored, due to different reasons | Var categorical |

Table 1. Variables for the Kaplan-Meier estimator.

The probabilities of survival are two types and can be confusing, cumulative probabilities and an interval probability (= interval survival rate after the event). The cumulative probabilities are calculate from the interval survival rate. Censored subjects are not included in the probabilities. The Kaplan-Meier curves are constructed using the cumulative survival rates.

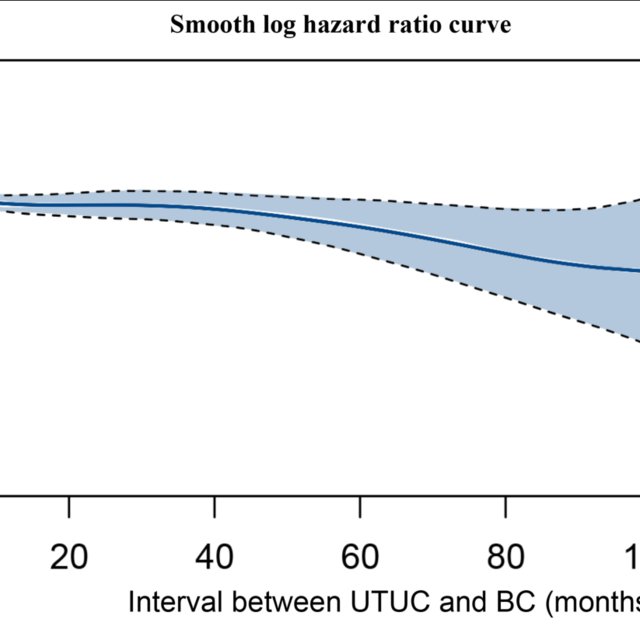
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | Serial time | Interval | No surviving at risk in the interval | Event | Censored | No surviving after the event | Interval survival rate | Cumulative survival rate |

Table 2. Production of the Kaplan-Meier curve. Number surviving at risk in the interval counting the subject of the interval and all those below

The **output** of the analysis is a compressed folder with the following files Datasets.xls, Leyends\_BasicAnalyses.pdf and LogFile.txt, and a summary report per gene in PDF format contrast with the information in the databases. In the Custom analysis, you get PDFs and PNGs of each graph per gene.

In the results, we can see a KM curve across quartiles (Q1, Q2+Q3, Q4), each represented by a line. In the lab, we can use different measurements in order to produce tables; we can use the mean, median and other statistical parameters to highlight different points.

**Smooth Hazard ratio** (Smooth HR)



This graph is an alternative to DFS, where we can see the log HR on the X-axis against the expression of each patient. All this together with a standard deviation measurement all along the curve. In the graph, we observe how the expression of the gen change in the patients from the database per time. Only performed in the PT patients.

# Correlations

Pairwise correlation of gene expression lvls of the two gene lists given. You can choose Spearman or Pearson correlations ¡, default version Pearson correlation. There are two boxes in which two gene lists are inserted and compared. Gene list 1 limited to 5 genes, and Gene list 2 (right) is limited to 10 genes.

**Heatmaps**

Representation of the correlation values between the gene expressions of one-gene vs another or more. *Se observa una correlacion entre la expression genica de un gen contra la expesion de los otros genes en las bases de datos seleccionadas. Basicamente, recoge la informaciondada por las correlaciones*.

# Enrichment analysis

In this section, you can find series of enrichment analysis. Those analysis are listed BioCarta, CMAP (Connectivity Map), HIPC (Human Immunology Project Consortium), Cancer, MIR, TFT, DOSE (Disease Ontology Enrichment), GO, KEEG, reactome and pathway. No further explanation will be given for are GO, KEEG, Reactome and pathway analysis.

BioCarta online maps of molecular pathways provides graphical models of molecular relationships, including proteomic and genomic information, as well as classical pathways and suggestions for new pathways.

DOSE analysis based on hypergeometric model and gene set enrichment analysis are also for discovering disease associations of high-throughput biological data. [DOSE (guangchuangyu.github.io)](https://guangchuangyu.github.io/software/DOSE/)