Enliven Archive
Intensifying Thoughts

Research Article www.enlivenarchive.org Enliven: Journal of Genetic, Molecular and Cellular Biology

ISSN: 2379-5700

# Genomic and Transcriptomic Data Integration in American Patients with Uterine Carcinosarcoma

Cristobal Ricardo De Leon Garcia*

*Servicio Nacional de Aprendizaje SENA, Colombia*

*Corresponding author: Cristobal Ricardo De Leon García, Servicio Nacional de Aprendizaje SENA, Colombia, Tel: +57 1 3133697899; E-mail: deleong1@outlook.com

Doi: https://doi-ds.org/doilink/05.2022-89188986/9.3-004

Received Date: 13th August 2021
Accepted Date: 08th September 2021
Published Date: 15th September 2021

Citation: De Leon Garcia CR. Genomic and Transcriptomic Data Integration in American Patients with Uterine Carcinosarcoma. Enliven: J Genet Mol Cell Biol. 2021; 9(3): 004.



## Abstract

Uterine carcinosarcoma (UCS) are aggressive neoplasms consisting of high-grade malignant epithelial and mesenchymal elements. UCS represent less than 5% of all uterine malignancies. In the US, approximately two in 100,000 women develop UCS annually. At the time of diagnosis, approximately one-third of patients have disease that has spread beyond the uterus. The survival percentage for patients with UCS projected at 5 years from the time of diagnosis is 40% to 75% for neoplasia confined to the uterus. Uterine carcinosarcoma is a cancer with high frequency of mutations specifically insertion and deletion polymorphisms such as Copy Number Alterations (CNA) linked to messenger Ribonucleic Acid (mRNA) transcripts. In this work was carried out omic data integration using CNA genomic data and transcripts from mRNA sequence counts from 57 American patients with different levels of infiltration and invasiveness of UCS analyzing 16383 genes and 60488 transcripts separately. For analyzing CNA genes, Component Principal Analysis (PCA) was carried out and for analyzing mRNA sequences counts, Differential Expression Analysis was carried out. After CNA and mRNA separately analysis, 36 genes and 96 transcripts highly significant were found, which were used in the integration analysis. Integrative analysis was carried out using *Sparse Least Square (sPLS)* methodology using *mixOmics* package in R software. Integrative analysis was based on graphical analysis from two output plots. Samples graphical representation, from RNAseq and CNA data show the clustering between samples. On RNAseq, samples showed clustering around central zero of all types of tumors, without clear separation between them. This indicates variance of different samples is not explained by the transcripts (genes). Clusters top and bottom of central zero especially tumor with most infiltration and invasiveness explained the most proportion of variance. On CNA genes, samples showed clear separated clustering's according with types of tumors. Tumor of less infiltration and invasiveness were clustered more closely near of central zero and tumor with most infiltration and invasiveness were clustered more closely away from central zero. Many samples were clustered very closely at central zero especially samples belonging tumors with less infiltration and invasiveness indicating some CNA genes have a weak influence on tumors with less infiltration and invasiveness. Samples from both RNAseq and CNA genes showed a strong negative correlation between them. Tumors with more infiltration and invasiveness showed high dispersion under the central zero, while tumors with less infiltration and invasiveness shows moderate dispersion above the central zero. This indicates both types of genes mRNA transcripts and CNA genes are highly expressed in aggressive tumors. According to genes graphical representation, three important CNA genes were highly expressed, *TPM3*, tropomyosin 3, *RPS27* ribosomal protein S27 genes, both located on chromosome 1 and ACTR1A, Alpha-centractin gene located on chromosome 10 was seen keeping direct positive correlation with the transcript *ENG00000122145* human transcript located on Chromosome 16. On RNAseq genes, four genes5 were highly expressed, *ENSG00000143028 (SYPL2,* Synaptophysin-like protein 2) human gene located on Chromosome 1; *ENSG00000077522 (ACTN2*, alpha actinin-2) human gene also located on Chromosome 1; *ENSG00000086967 (MYBPC2,* Myosin-binding protein C) human gene located on Chromosome 19 and *ENSG00000253767 (PCDHGA8*, Protocadherin gamma-A8) human gene located on Chromosome 5. The results showed a high correlation between CNA and mRNA genes, indicating that copy number alteration also results in differential gene expression in uterine carcinosarcoma.

Keywords: Data integration; Correlation; Copy number alteration; RNAseq counts

## Introduction

Uterine carcinosarcoma (UCS) is a cancer developing in uterus. Carcinosarcoma means that, tumor shows both histologic features endometrial carcinoma and sarcoma. Endometrial carcinoma begins into endometrium (inner uterus layer) while sarcoma begins in uterus outer muscle layer [1]. Uterine carcinosarcoma (UCS) are rare but aggressive neoplasms consisting of high-grade malignant epithelial and mesenchymal elements, representing less than 5% of all uterine malignancies. In the United States of America, approximately two in 100,000 women develop UCS annually [2].

In different studies, influence of genes on uterine carcinosarcoma expression has been demonstrated. Mutations in the TP53 gene (Tumor Protein 53, acts as a tumor suppressor: tumor cell antigen) and low frequency in the FBXW7 genes (F-box / WD repeat-containing protein 7, mediates ubiquitination and subsequent proteosomal degradation of target proteins), PIK3CA (Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform, participates in cell signaling in response to various growth factors.) And PPP2R1A (Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform, regulates transcription and intervenes in RNA splicing) [3]. In a more recent publication, Zhao et al. [4] found a lower frequency of mutations in the TP53 and FBXW7 genes of 24 normal non-hypermuted UCS tumor exomes.

Genomic data refer to mutations into genome such as Copy Number Variation (CNV) that are short nucleotide's chain (2 – 4 nt) into genomic DNA known as Indel (insertions, deletions, or amplifications), they are polymorphisms located in genomic DNA of germinal cells, oocytes, and sperm. But when these mutations are present in somatic cells, they are called Copy Number Alterations (CNA) [5]. CNV have been associated in humans with different diseases, especially with psychiatric disorders such as schizophrenia [6] Xu et al., while CNA have been associated with various types of cancer [7].

Transcriptomic data refer to transcript elements or Ribonucleic Acid (RNA) universe (i.e., micro-RNA (miRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA)), it´s deeply complex, and new isoforms of known genes and short RNA species continue to be discovered, such as microRNAs (miRNAs) and enhancer RNAs. Previous and de novo identified transcripts of many genes have been associated with expression of certain types of cancer Cherniack et al. [8] and have shed light based on complex traits such as personality or physiological response to catastrophic events [9].

Developing data integration models of data described above it can explore the dynamic interconnectivity of biological systems during pathophysiological relevant processes in order to cover as much information as possible that explains better functioning at the molecular level. Integration of genomic and transcriptomic data approach enables understanding of disease processes such as cancer in a "biological pathway" rather than a "single molecule" level and accelerate progress toward disease-modifying interventions [10].

The modern trend towards personalization medicine explains the importance of prevention and personalized treatment of diseases that occurs because of molecular integration of genomic, transcriptomic, metabolomic and proteomic factors, leaving behind the idea of individual metabolic actions are causing of disease [11].

## Methods and Material

### Patients and Study Location

Samples were extracted from uterine tumor cells and from normal surrounding cells of 57 women with UCS as well as from normal blood cells for CNAs, however, for mRNA only samples were taken uterine cell, from the Nationwide Children's Hospital and research center in Ohio, United States of America [12].

## Data Source

Data were obtained from The Cancer Genome Atlas (TCGA) repository, Project ID: TCGA-UCEC; dbGAPStudy accession: phs000178 (http://gdac.broadinstitute.org/), from the Broad Institute Genomic Data Analysis Center, (2016). All samples were processed in this center to obtain significant genomic events such as Copy Number Alterations (CNA) and transcriptome events such as RNAseq (mRNA) counts, for determining statistical association with important variables (genes) inherent to uterine carcinosarcoma [13].

## Genomic Data: Copy Number Alterations (CNA)

Copy number variation (CNV) uses data from the Affymetrix SNP 6.0 matrix to identify genomic repeating regions and infer the copy number of these repeats. The chip outputs are processed from the TCGA using the DNAcopy R package to perform an analysis of circular binary segmentation (CBS) [14]. CBS translates noisy intensity measurements into chromosomal regions of equal copy number. The final output files are segmented into genomic regions with the estimated copy number for each region. Then, Genomic Data Commons (GDC) further transforms these copy number values into segment mean values, which are equal to log2 (copy number / 2). Diploid regions will have a segment mean of zero, amplified regions will have positive values, and deletions will have negative values (Genomic Data Commons, https://gdc.cancer.gov/).

CNA data are collected in 3 tables: First Table, Copy Number Segment associates to the contiguous chromosomal segments with genomic coordinates, average intensity of the array and the number of probes that bind to each segment (see Table 1). Second table, Masked Copy Number Segment with the same information as copy number segment, except segment with probes known to contain germline mutations are removed. Third, Copy Number Estimated table shows gains and losses at the gene level, generated from the table above.
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/

Numerical copy number variation (CNV) values at the focal level, were generated from "copy number segment files from which germline mutations were removed" from tumor aliquots using GISTIC2 [15,16] at the project level. Only genes encoding proteins were kept, and their numerical CNV values were further limited by a limit of noise of 0.3: Genes with focal CNV values less than -0.3 are classified as "loss" (-1); genes with focal CNV values greater than 0.3 are classified as "gain" (+1) and genes with values of Focal CNVs between -0.3 and 0.3 are classified as "neutral" (0) (see Table 2).

Table 1 Copy Number Segment

| Sample | Chromosome | Start | End | Num_Probes | Segment-Mean |
|---|---|---|---|---|---|
| TCGA-N5-A4R8-10A-01D-A28T-01 | 1 | 61735 | 3003488 | 666 | 0.0141 |
| TCGA-N5-A4R8-10A-01D-A28T-01 | 1 | 3004357 | 3026260 | 6 | -1.807 |
| TCGA-N5-A4R8-10A-01D-A28T-01 | 3 | 75578858 | 84702413 | 5009 | -0.0024 |
| TCGA-N5-A4R8-10A-01D-A28T-01 | 3 | 84702463 | 84702600 | 2 | -1.5993 |
| TCGA-N5-A4R8-10A-01D-A28T-01 | 3 | 84706284 | 98944632 | 5989 | 0.005 |
| TCGA-N5-A4R8-10A-01D-A28T-01 | 3 | 98944763 | 9894472306 | 44 | 0.3475 |

Rows represents samples (patients' identification) according to The Cancer genome Atlas (TCGA). Also, chromosomes with start and end of regions where segment mean was identified

Table 2 Copy Number Estimated

| Gene Symbol | Gene ID | Cytoband | TCGA-N9-A4C | TCGA-N9-A4G | TCGA-NA-A5T | TCGA-NF-A4X | TCGA-N5-A4RX |
|---|---|---|---|---|---|---|---|
| ENSG00000008128.21 | 0 | 1p36.33 | 0 | -1 | -1 | 0 | 1 |
| ENSG00000008130.14 | 0 | 1p36.33 | 0 | -1 | -1 | 0 | 1 |
| ENSG00000007606.14 | 0 | 1p36.33 | 0 | -1 | -1 | 0 | 1 |
| ENSG00000078369.16 | 0 | 1p36.33 | 0 | -1 | -1 | 0 | 1 |
| ENSG00000078808.15 | 0 | 1p36.33 | 0 | -1 | -1 | 0 | 1 |
| ENSG00000107404.16 | 0 | 1p36.33 | 0 | -1 | -1 | 0 | 1 |
| ENSG00000116151.12 | 0 | 1p36.32 | 0 | -1 | -1 | 0 | 1 |
| ENSG00000127054.17 | 0 | 1p36.33 | 0 | -1 | -1 | 0 | 1 |

Copy Number Alterations. Rows represents genes and Colom's represents samples (patients' identification) according to The Cancer genome Atlas (TCGA). Rows represents genes. Columns represent aliquots, which are associated with categorizations of Copy Number Alterations values (insertion=1, deletion=-4 and neutral=0) for each gene.

Data used in the present study were derived from the Copy Number Estimated table, extracted from TCGA.

## Coding of CNA Samples According TCGA

CNA samples of this study from the TCGA repository generally have the following encoding (see Figure 1).

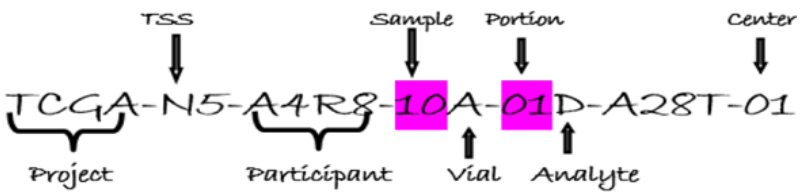



Figure 1. Sample's coding. TCGA: The Cancer Genome Atlas; N5: Sample of patients with uterine carcinosarcoma from MSKCC (Memorial Sloan Kettering Cancer Center); A4R8: Patient identification; 10A: Samples with code "10" correspond to samples derived from normal blood from patients with cancer, letter "A" means that sample is labeled on the first container or vial A; 01D: Portion in milligrams of sample or first portion, letter "D" means Deoxy Ribonucleic Acid (DNA); A28T: Wells Chip DNA identification; 01: Code of the center that received the sample for analysis, in this case 01: corresponds to the Broad Institute of MIT and Harvard.

## Transcriptome Data: RNAseq (mRNA Counts)

mRNA quantification analysis measures expression at the gene level. Samples were processed by TCGA institute using Illumina HiSeq_2000 chip technique, subsequently, HTSeq program (High Throughput Sequencing, high-performance sequencer) is used to generate two files of raw reads: "raw read count": the reads of fragments per kilobase of transcripts (Fragment per Kilobase of transcript per Million mapped read, FPKM ) and the upper quartile normalization readings (FPKM-UQ, upper quantile) [17, Anders S and Zanini F]. But before generating these values, reads are aligned with tGRCh38 reference human genome and then mapped reads are quantified. Alignment is performed following the methods used by the International Cancer Genome Consortium (ICGC) (https://icgc.org/). Finally, mapped reads of each gene are enumerated using the HT-Seq-Count program, thus generating "counts" file as described in Table 3.

Table 3. RNA-seq counts. Patients with Uterine Carcinosarcoma

| X1<br><chr> | TCGA-N5-A4R8-01A-11R-A28V-07<br><dbl> | TCGA-N5-A4RA-01A-11R-A28V-07<br><dbl> |
|---|---|---|
| ENSG00000000003.13 | 2610 | 1673 |
| ENSG00000000005.5 | 552 | 5 |
| ENSG00000000419.11 | 6718 | 2284 |
| ENSG00000000457.12 | 1034 | 567 |
| ENSG00000000460.15 | 802 | 663 |
| ENSG00000000938.11 | 252 | 96 |

Columns indicate samples with their respective count of the mRNA sequences identified in each gene (rows). Data downloaded from TCGA. Project ID: TCGA-UCEC; dbGAPStudy accession: phs000178

## Data Preparation

Previously to the integrative analysis, was carried out analysis in both omics separately.

## Copy Number Alterations (CNA)

CNA data were directly downloaded from TCGA repository using the function GDCquery from TCGAbiolinks and TCGAutils libraries in R.

```
query <- GDCquery (project = "TCGA-UCS",
        data. category = "Copy Number Variation",
        data. type = "Gene Level Copy Number Scores",
        access = "open")
```

Table 4. Copy Number Alterations. Genes and Patients with original code.

| Gene Symbol<br><chr> | Gene ID<br><dbl> | Cytoband<br><chr> | TCGA-N9-A4Q7-01A-11D-A28Q-01<br><dbl> | TCGA-N9-A4Q8-01A-31D-A28Q-01<br><dbl> |
|---|---|---|---|---|
| ENSG00000008128.21 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000008130.14 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000067606.14 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000078369.16 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000078808.15 | 0 | 1p36.33 | 0 | -1 |

Rows represents genes and columns from three onwards, represents patients (samples).

Table 4 First five rows from downloaded data (scores):

```
scores [1:5,1:5]
```

A Heatmap was carried out to understand better the original data:

```
library (ComplexHeatmap
library(dplyr)
scores. matrix <- scores %>%
  dplyr::select(-c(1:3)) %>%  # Removes metadata from the first 3 columns
  as. matrix
rownames (scores. matrix) <- paste0(scores$`Gene Symbol`,"_", scores$Cytoband)
# gain in more than 100 samples
gain. more.than.hundred.samples <- which(rowSums(scores.matrix == 1) > 100)
# loss in more than 100 samples
loss. more.than.hundred.samples <- which(rowSums(scores.matrix == -1) > 100)
lines. selected <- c (gain. more.than.hundred.samples,loss.more.than.hundred.samples)
Heatmap (scores. matrix [lines. selected,],
      show_column_names = FALSE,
      show_row_names = TRUE,
       row_names_gp = gpar (fontsize = 8),
      col = circlize::colorRamp2(c(-1,0,1), colors = c("red","white","blue")))
```

*ENSG00000146648 .14* gene on the first top line of the heatmap it is located on chromosome 7 in region 55,019,017-55,211,628 bp (Ensembl database) identified like *EGFR* (Epidermal growth factor receptor Epidermal growth factor receptor) in UNIPROT (Uniprot database) with code P00533. This gene encodes tyrosine kinase receptors. Mutations of this gene have been associated with lung cancer (https://www.newsmedical.net/health/LungCancer Genetics).

For developing a better analysis, were manually modified rows and columns in *scores* file. On Table 4 numbers after dot in column "Gene Symbol" were deleted and the column of patients with original code was replaced by short participant's code (see Table 5).

Table 4. Copy Number Alterations. Genes and Patients with original code.

| Gene Symbol <chr> | Gene ID <dbl> | Cytoband <chr> | TCGA-N9-A4Q7-01A-11D-A28Q-01 <dbl> | TCGA-N9-A4Q8-01A-31D-A28Q-01 <dbl> |
|---|---|---|---|---|
| ENSG00000008128.21 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000008130.14 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000067606.14 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000078369.16 | 0 | 1p36.33 | 0 | -1 |
| ENSG00000078808.15 | 0 | 1p36.33 | 0 | -1 |

Rows represents genes and columns from three onwards, represents patients (samples).

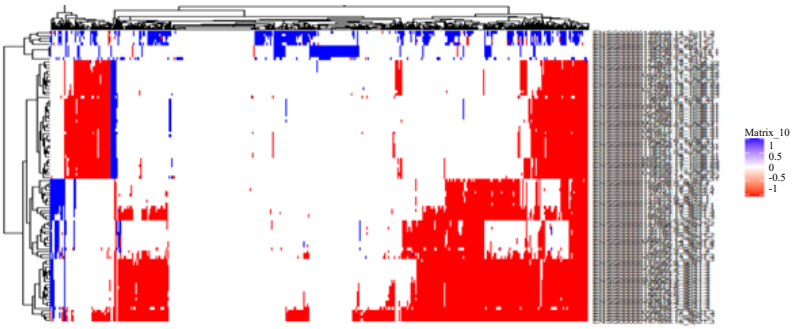



Figure 2. Upper blue bands indicate cluster of genes with insertions (+ 1). Red bands indicate cluster of with deletions (-1) and blank bands represent genes without deletions or insertions. Most of them are genes located in chromosome 7 and 12: ENSG00000146648.14, ENSG00000132432.12, ENSG000001324348, ENSG0000015497811, ENSG00000170419 9, ENSG000001392665, ENSG0000013 5446.15, EN SG0 0000135439.10.

Table 5. Copy Number Alterations. Genes and short patients' identification.

| | Gene_ID <chr> | A4PO <chr> | A4Q1 <chr> | A4X2 <chr> | A4QY <chr> |
|---|---|---|---|---|---|
| 1 | ENSG00000008128 | -1 | -1 | 0 | 0 |
| 2 | ENSG00000008130 | -1 | -1 | 0 | 0 |
| 3 | ENSG00000067606 | -1 | -1 | 0 | 0 |
| 4 | ENSG00000078369 | -1 | -1 | 0 | 0 |
| 5 | ENSG00000078808 | -1 | -1 | 0 | 0 |

Numbers of gene after dot were deleted. Patients ID was cropped.

```
myscores<-read.csv ("C:/Users/CRISTOBAL DE LEON/Documents/
GDCdata/myscores_p.csv", header = TRUE, sep = ";")
myscores<-as.data. frame(myscores)
head(myscores)
```

An important characteristic of these data is that each patient is categorized according to infiltration degree and tumor invasiveness. This classification is performed by FIGO (International Federation of Gynecology and Obstetrics) [18]. Table 6 show FIGO´classification according to infiltration degree and tumor invasiveness in each patient.

FIGO classification is used to replace the columns, Patiens_id, by infiltration and invasiveness degree of tumor (Table 7).

Columns and rows on Table 7 were manually rotated. FIGO classification is now in raws. This transposition will later facilitate the Principal Component Analysis.

Table 6 Samples classification according to infiltration degree and tumor invasiveness.

| patient_ID | figo_stage | patient_ID | figo_stage | patient_ID | figo_stage | patient_ID | figo_stage | patient_ID | figo_stage |
|---|---|---|---|---|---|---|---|---|---|
| A4PO | II | A59B | IA | A5NN | IB | A4VG | IVB | A4R1 | IIIC1 |
| A4Q1 | II | A4Q7 | IA | A4WC | IB | A4RM | IVB | A4PN | IIIC1 |
| A4X2 | II | A4Q3 | IA | A56S | IB | A4QX | IVB | A4RT | IIIC1 |
| A4QY | III | A5NM | IA | A4PM | IB | A4Y8 | IVB | A4PQ | IIIC2 |
| A4VC | III | A4PI | IA | A4QV | IB | A4R0 | IVB | A4RA | IIIC2 |
| A4WX | IV | A4PL | IA | A4Y5 | IIB | A4WA | IC | A4RO | IIIC2 |
| A4RV | IA | A4QW | IIA | A5I1 | IIIB | A4VW | IC | A4PZ | IIIC2 |
| A4WU | IA | A4VD | IIIA | A4VU | IIIB | A4Q4 | IIIC | A5CP | IIIC2 |
| A4RF | IA | A4WF | IIIA | A4RD | IVB | A4W6 | IIIC | A4RU | IIIC2 |
| A4R8 | IA | A4VF | IB | A59E | IVB | A4VE | IIIC | | |
| A4V9 | IA | A4RS | IB | A4RN | IVB | A4Y0 | IIIC | | |
| A4Q8 | IA | A59F | IB | A4RJ | IVB | A4PP | IIIC1 | | |

Stage I, the tumor is in the organ where it originally formed; Stage II to IV, the tumor spreads to tissues beyond the organ of origin; A, B and C indicate infiltration degree and tumor invasiveness: A, tumor is loca
in a part of the organ, B, the tumor is infiltrated in healthy neighboring cells of the organ of origin; C, tumor invades neighboring tissues or organs.
Source: Clinical Data (NHI, 2012):
(https://training.seer.cancer.gov/staging/systems/schemes/figo.html)

Table 7. Copy Number Alterations. Patients replaced by FIGO stages classification

| | Gene ID<br><chr> | II<br><int> | II<br><int> | II<br><int> | III<br><int> | III<br><int> | IV<br><int> | IA<br><int> | IA<br><int> |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ENSG00000008128 | -1 | -1 | 0 | 0 | -1 | 0 | 1 | 0 |
| 2 | ENSG00000008130 | -1 | -1 | 0 | 0 | -1 | 0 | 1 | 0 |
| 3 | ENSG00000067606 | -1 | -1 | 0 | 0 | -1 | 0 | 1 | 0 |
| 4 | ENSG00000078369 | -1 | -1 | 0 | 0 | -1 | 0 | 1 | 0 |
| 5 | ENSG00000078808 | -1 | -1 | 0 | 0 | -1 | 0 | 1 | 0 |
| 6 | ENSG00000107404 | -1 | -1 | 0 | 0 | -1 | 0 | 1 | 0 |

6 rows | 1-10 of 56 columns

```
myscores<-read.csv ("C:/Users/CRISTOBAL DE LEON/Documents/GDCdata/myscores_t.csv", header = TRUE, sep = ";")
head(myscores)
```

Table 8. Copy Number Alterations. Columns are variables (genes) and rows are Patients.

| | Patients_ ID<br><chr> | ENSG0000<br>0008128<br><int> | ENSG00000<br>008130<br><int> | ENSG0000<br>0067606<br><int> | ENSG0000<br>0078369<br><int> | ENSG000<br>00078808<br><int> | ENSG000<br>00107404<br><int> | ENSG000<br>00116151<br><int> |
|---|---|---|---|---|---|---|---|---|
| 1 | II | -1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 2 | II | -1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 3 | II | 0 | 0 | 0 | | 0 | 0 | 0 |
| 4 | III | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | III | -1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 6 | IV | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

6 rows | 1-9 of 16384 columns

## Principal component analysis (PCA) for Copy Number Alterations data:

PCA were carried out to obtain the most significant genes that will be used later for integrating with RNAseq data.

First, it has been built the matrix M from myscores file in R:

```
M<-read.csv ("C:/Users/CRISTOBAL DE LEON/Documents/GDCdata/myscores_t.csv", header = TRUE, sep = ";")
M<-as.data. frame(M)
dim(M)
```

```
## [1]    55 16384
```

Matrix M contains 55 rows (patients according with tumor FIGO classification) and 16384 genes (variables).

```
set. seed (9091)
na.row<-sample (1: nrow(M), replace = TRUE)
na.col<-sample (1: ncol(M), replace = TRUE)
M.na<-as. matrix(M)
M.na<-M.na [, -1]
```

After some application codes on matrix M, it obtains matrix M.na for PCA analysis

```
dim(M.na)
## [1]    55 16383
```

The goal of PCA is to identify directions (or principal components) along which the variation in the data is maximum. In other words, PCA reduces the dimensionality in multivariate data in two or three principal components with minimal loss information.

```
pca.CNA<-PCA (M.na, scale. unit = TRUE, ncp = 56, graph = F)
pca.CNA
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 55 individuals, described by 16383 variables
```

From this analysis it can see the most variance explained by first 6 components:

```
head(pca.CNA$eig)
```

| | eigenvalue | percentage of variance | cumulative percentage of variance |
|---|---|---|---|
| comp 1 | 1067.5406 | 6.531298 | 6.531298 |
| comp 2 | 1000.5367 | 6.121362 | 12.652660 |
| comp 3 | 844.4412 | 5.166358 | 17.819018 |
| comp 4 | 798.3566 | 4.884409 | 22.703426 |
| comp 5 | 732.2390 | 4.479896 | 27.183322 |
| comp 6 | 664.5490 | 4.065764 | 31.249086 |

First and second component corresponds to the addresses with the maximum amount of variation in data set. It is seen that first component retains the highest variation, (6.53%) and the second component 6.12%. Graphic representation on Figure 3:

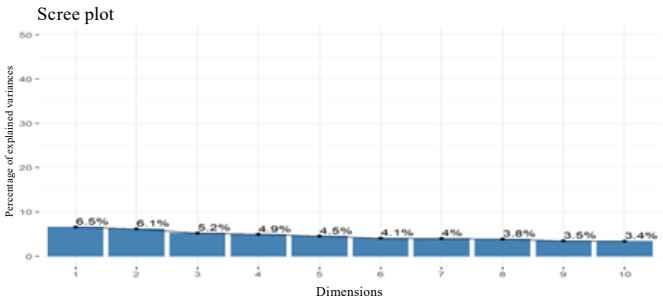



Figure 3. Principal Component Analysis. The highest percentage of variance is explained by the first and second components (11.6%)

```
fviz_eig (pca.CNA, addlabels = TRUE, ylim = c (0, 50))
```

The most significative genes from the first and second component were obtained (Tables 9 and 10 respectively), for that, is used the function dimdesc from factoextra and FactoMineR libraries in R

```
res. desc <- dimdesc (pca.CNA, axes = c (1,2), proba = 0.05)
res. desc$Dim.1
```

```
res. desc <- dimdesc (pca.CNA, axes = c (1,2), proba = 0.05)
res. desc$Dim.2
```

Table 9. Significative genes by first component

| Gene_ID | Correlation | p-value |
|---|---|---|
| ENSG00000204365 | 0.76539 | 1.01e-11 |
| ENSG00000156398 | 0.7150537 | 8.62e-10 |
| ENSG00000138175 | 0.7150537 | 8.62e-10 |
| ENSG00000171206 | 0.7150537 | 8.62e-10 |
| ENSG00000107882 | 0.7150537 | 8.62e-10 |
| ENSG00000138107 | 0.7150537 | 8.62e-10 |
| ENSG00000138111 | 0.7150537 | 8.62e-10 |
| ENSG00000151532 | 0.6992983 | 2.87e-09 |
| ENSG00000109452 | 0.6583856 | 4.67e-08 |
| ENSG00000170153 | 0.6583856 | 4.67e-08 |
| ENSG00000109436 | 0.6583856 | 4.67e-08 |
| ENSG00000189184 | 0.6583856 | 4.67e-08 |
| ENSG00000254535 | 0.6583856 | 4.67e-08 |
| ENSG00000138650 | 0.6583856 | 4.67e-08 |
| ENSG00000151470 | 0.6583856 | 4.67e-08 |

Table 10. Significative genes by second component

| Gene_ID | Correlation | p-value |
|---|---|---|
| ENSG00000143515 | 0.8450245 | 5.00e-16 |
| ENSG00000143575 | 0.8450245 | 5.00e-16 |
| ENSG00000143569 | 0.8450245 | 5.00e-16 |
| ENSG00000143612 | 0.8450245 | 5.00e-16 |
| ENSG00000163263 | 0.8450245 | 5.00e-16 |
| ENSG00000143549 | 0.8450245 | 5.00e-16 |
| ENSG00000177954 | 0.8450245 | 5.00e-16 |
| ENSG00000118217 | 0.8176596 | 2.58e-14 |
| ENSG00000160716 | 0.7882102 | 9.16e-13 |
| ENSG00000160714 | 0.7882102 | 9.16e-13 |
| ENSG00000163239 | 0.7882102 | 9.16e-13 |
| ENSG00000169291 | 0.7882102 | 9.16e-13 |
| ENSG00000160712 | 0.7882102 | 9.16e-13 |
| ENSG00000143595 | 0.7882102 | 9.16e-13 |
| ENSG00000215853 | 0.7882102 | 9.16e-13 |
| ENSG00000159450 | 0.7882102 | 9.16e-13 |
| ENSG00000182898 | 0.7882102 | 9.16e-13 |
| ENSG00000163191 | 0.7882102 | 9.16e-13 |
| ENSG00000014914 | 0.7882102 | 9.16e-13 |
| ENSG00000143368 | 0.7882102 | 9.16e-13 |
| ENSG00000280649 | 0.7882102 | 9.16e-13 |

A final file of most significative (CNA genes) is manually built (Table 11) by further integration with RNAseq data.

Gene's symbol and its function was sought after in ENSEMBL and UNIPROT database (Table 12) On Table 13 gene identifications are represented by gene symbol

Table 11. Copy Number Alterations of most significative genes

| | Patients_ ID <chr> | ENSG00000 0204365 <int> | ENSG0000 00156398 <int> | ENSG00000 0138175 <int> | ENSG00000 0171206 <int> | ENSG0000 00107882 <int> | ENSG0000 00138107 <int> | ENSG0000 00138111 <int> |
|---|---|---|---|---|---|---|---|---|
| 1 | II | 0 | -1 | -1 | -1 | -1 | -1 | -1 |
| 2 | II | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | II | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | III | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | III | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | IV | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

6 rows | 1-9 of 37 columns

The most CNA significative genes (36 genes) from first (15 genes) and second (21 genes) principal component respectively.

Table 12. Genes and chromosome positions, function and diseases of some genes.

| Gene_ID | Gene_Symbol | Chromo-some | start | end | Function | Diseases |
|---|---|---|---|---|---|---|
| ENSG00000204365 | C10orf126 | 10 | 28,846,408 | 28,881,898 | Uncharacterized protein | |
| ENSG00000156398 | SFXN2 | 10 | 102,714,538 | 102,743,492 | Mitochondrial amino-acid transporter that mediates transport of serine into mitochondria. | |
| ENSG00000138175 | ARL3 | 10 | 102,673,731 | 102,714,397 | Small GTP-binding protein which cycles between an inactive GDP-bound and an active GTP-bound form | Joubert syndrome (cerebellar ataxia) and Retinitis pigmentosa |
| ENSG00000171206 | TRIM8 | 10 | 102,644,479 | 102,658,318 | E3 ubiquitin-protein ligase that participates in multiple biological processes including cell survival, differentiation, apoptosis, and in particular, the innate immune response | TRIM8 deficiency leads to increased polyinosinic-polycytidylic acid– and LPS-triggered induction of downstream anti-microbial genes including TNF |
| ENSG00000107882 | SUFU | 10 | 102,503,972 | 102,633,535 | It is a sulfotransferase rather than a scaffold assembly protein | |
| ENSG00000138107 | ACTR1A | 10 | 102,461,881 | 102,502,712 | Component of a multi-subunit complex involved in microtubule-based vesicle motility. ATP binding | |
| ENSG00000138111 | MFSD13A | 10 | 102,461,395 | 102,477,045 | Transmembrane protein | |
| ENSG00000151532 | VTI1A | 10 | 112,446,998 | 112,818,744 | Vesicle trafficking and to promote fusion of the lipid bilayers | |
| ENSG00000109452 | INPP4B | 4 | 142,023,160 | 142,847,432 | Plays a role in the late stages of micropinocytosis by dephosphorylating phosphatidylinositol 3,4-bisphosphate in membrane ruffles. Antagonizes the PI3K-AKT/PKB signaling pathway by dephosphorylating phosphoinositide's and thereby modulating cell cycle progression and cell survival | Reduced INPP4B expression is associated with poor outcomes for breast, prostate, and ovarian cancer patients. |
| ENSG00000170153 | RNF150 | 4 | 140,859,807 | 141,212,877 | Ubiquitin protein ligase activity | |

Table 13 Genes symbol

| | Patients_ID | C10orf126 <int> | SFXN2 <int> | ARL3 <int> | TRIM8 <int> | SUFU <int> | ATR1A <int> | MFSD13A <int> | VTI1A <int> |
|---|---|---|---|---|---|---|---|---|---|
| 1 | II | 0 | -1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | II | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | III | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 |
| 5 | III | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | IV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

6 rows | 1-10 of 37 columns

## Data Preparation RNAseq counts

RNAseq counts data were directly downloaded from TCGA repository using the function GDCquery from TCGAbiolinks and TCGAutils libraries in R.

```
query.exp. hg38 <- GDCquery (project = "TCGA-UCS",
            data. category = "Transcriptome Profiling",
            data. type = "Gene Expression Quantification",
            workflow. type = "HTSeq - Counts",
            access = "open")
```

```
GDCdownload (query.exp. hg38)
```

Thus, is obtained the file named "raw. counts" with genes and their counts for each patient (Table 14).

Final file with gene symbol (CNA genes) that will be used for integration with RNAseq data

Table 14 Original RNAseq counts data

| **X1**<br><chr> | **TCGA-N5-A4RT-01A-11R-A28V-07**<br><dbl> | **TCGA-ND-A4W-01A-21R-A28V-07**<br><dbl> | **TCGA-N5-A4RA-01A-11R-A28V-07**<br><dbl> |
|---|---|---|---|
| ENSG00000000003.13 | 4765 | 6181 | 1673 |
| ENSG00000000005.5 | 878 | 1 | 5 |
| ENSG00000000419.11 | 4614 | 2081 | 2284 |
| ENSG00000000457.12 | 913 | 406 | 567 |
| ENSG00000000460.15 | 1549 | 478 | 663 |
| ENSG00000000938.11 | 105 | 133 | 96 |

6 rows | 1-4 of 57 columns

```
raw. counts <- GDCprepare (query = query.exp. hg38, summarizedExperiment = FALSE)
head (raw. counts)
```

For developing a better analysis, rows, and columns in raw. counts file (Table 14) were manually modified. Column" "X1" is changed by Gene_ID and numbers after dot in column "X1 were deleted, on patient's column code was replaced by short participant's code (see Table 15).

Table 15. RNAseq counts. Genes and short patients' identification.

| | Gene_ID<br><chr> | A4PO<br><int> | A4Q1<br><int> | A4X2<br><int> | A4QY<br><int> |
|---|---|---|---|---|---|
| 1 | ENSG0000000003 | 6217 | 3778 | 3211 | 5074 |
| 2 | ENSG0000000005 | 42 | 46 | 143 | 78 |
| 3 | ENSG0000000419 | 4035 | 1865 | 2022 | 3019 |
| 4 | ENSG0000000457 | 1233 | 556 | 968 | 557 |
| 5 | ENSG0000000460 | 1299 | 489 | 1269 | 1146 |

5 rows

```
mycounts_<-read.csv ("C:/Users/CRISTOBAL DE LEON/Documents/GDCdata/raw.counts_.csv", header = TRUE, sep = ";")
mycounts [1:5,1:5]
```

In the same way, that CNA data was carried out, patients in RNAseq counts data, were replaced by FIGO stages identification

```
Names <-c("II", "II","II",
"III","III","IV","IA","IA","IA","IA","IA","IA","IA","IA","IA","IA","IIIA","IIIA","IIIA","IB","IB","IB","IB","IB","IB","IB","IB","IIB","IIIB",
"IIIB","IVB","IVB","IVB","IVB","IVB","IVB","IVB","IVB","IVB","IC","IC","IIIC","IIIC","IIIC","IIIC","IIIC1","IIIC1","IIIC1","IIIC1","IIIC
2","IIIC2","IIIC2","IIIC2","IIIC2","IIIC2")
colnames(mycounts) [2:4] = paste0("II")
colnames(mycounts) [5:6] = paste0("III")
colnames(mycounts) [7:7] = paste0("IV")
colnames(mycounts) [8:17] = paste0("IA")
colnames(mycounts) [18:20] = paste0("IIIA")
colnames(mycounts) [21:28] = paste0("IB")
colnames(mycounts) [29:29] = paste0("IIB")
colnames(mycounts) [30:31] = paste0("IIIB")
colnames(mycounts) [32:40] = paste0("IVB")
colnames(mycounts) [41:42] = paste0("IC")
colnames(mycounts) [43:46] = paste0("IIIC")
colnames(mycounts) [47:50] = paste0("IIIC1")
colnames(mycounts) [51:56] = paste0("IIIC2")
```

```
Condition <-c ("II", "III", "IV", "IA", "IIA", "IIIA", "IB","IIB","IIIB","IVB","IC","IIIC","IIIC1","IIIC2")
colnames(mycounts) [2:4] = paste0("II")
colnames(mycounts) [5:6] = paste0("III")
colnames(mycounts) [7:7] = paste0("IV")
colnames(mycounts) [8:17] = paste0("IA")
colnames(mycounts) [18:20] = paste0("IIIA")
colnames(mycounts) [21:28] = paste0("IB")
colnames(mycounts) [29:29] = paste0("IIB")
colnames(mycounts) [30:31] = paste0("IIIB")
colnames(mycounts) [32:40] = paste0("IVB")
colnames(mycounts) [41:42] = paste0("IC")
colnames(mycounts) [43:46] = paste0("IIIC")
colnames(mycounts) [47:50] = paste0("IIIC1")
colnames(mycounts) [51:56] = paste0("IIIC2")
```

```
head(mycounts)
```

## *DESeq2* package analysis to RNAseq counts data:

*DESeq2* package provides methods for detection of differentially expressed genes with negative binomial generalized linear models, estimates the dispersion (quite wide in RNAseq counts data) and the logarithm with base two of the same with the Fold Change option to change the logarithmic base. The object class used by the *DESeq2* package to store read counts (mycounts) and study conditions (metadata) in addition to intermediate estimated quantities during statistical analysis is *DESeqDataSet*, which will normally be represented in code here as an object "dds" [19].

*DESeqDataSet* class extends the *RangedSummarizedExperiment* class from the *SummarizedExperimentpackage*. The "Ranged" part refers to rows (in this case, counts) can be associated with genomic ranks (exons of genes). This association facilitates the subsequent exploration of the results, making use of the range-based functionality of other Bioconductor packages (for example, finding the ChIP-seq peaks closest to the differentially expressed genes) [19].

A *DESeqDataSet* object must have an associated design formula. The design formula expresses variables that will be used in the model. Formula must be the sign (~) followed by the variables. The design can be changed later, however, all steps of the differential analysis must be repeated, as the design formula is used to estimate the dispersions and estimate the log2-fold changes of the model. *DESeqDataSetFromMatrix* function indicates that a matrix of counts has been started and the *DESeq* function estimates the size and dispersion factors of each gene and adaptation of a generalized linear model [19].

```
metadata<-read.csv ("C:/Users/CRISTOBAL DE LEON/Documents/GDCdata/metadata.csv", header = TRUE, sep = ";")
```

```
head(metadata,5)
```

| | Name <chr> | Condition <chr> |
|---|---|---|
| 1 | A4PO | II |
| 2 | A4Q1 | II |
| 3 | A4X2 | II |
| 4 | A4QY | III |
| 5 | A4VC | III |

5 rows

```
mycounts<-as.data. frame(mycounts)
metadata<-as.data. frame(metadata)
```

```
class(mycounts)
class(metadata)
names(mycounts) [-1]
metadata$Name
names(mycounts) [-1] ==metadata$Name
```

Start RNAseq data analysis by creating dds object in DESeq2:

```
dds<-DESeqDataSetFromMatrix (countData = mycounts,
                    colData = metadata,
                    design = ~Condition,
                    tidy = TRUE)
dds1<-DESeq(dds)
```

```
## estimating size factors
## gene-wise dispersion estimates
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing
## -- replacing outliers and refitting for 28283 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)
## estimating dispersions
## fitting model and testing
```

To display data or groupings, it is necessary to transform the counting data. DESeq2 program provides the regularized logarithmic transformation (rlog) which gives results similar to the transformation by the base 2 logarithm (log2) for high-count genes, with the rlog becoming a way to return homoscedastic data (equal variance).

```
logData<-rlog (dds1, blind = F)
```

```
## rlog () may take a long time with 50 or more samples,
## vst () is a much faster transformation
```

```
head(assay(logData),3)
DeseqMatrix<-estimateSizeFactors(dds1)
```

```
                   II      II.1      II.2       III     III.1        IV        IA      IA.1      IA.2      IA.3      IA.4      IA.5      IA.6
ENSG00000000003 12.132967 11.957122 11.710951 12.389775 12.320188 11.933550 11.93222 12.431889 12.628872 12.782627 12.477652 11.07928 11.827534
ENSG00000000005  5.742789  6.153295  7.249879  6.778646  4.643354  7.022518 12.65919  5.472742  8.953553  5.888208  4.032766  4.10015  7.080381
ENSG00000000419 11.499709 11.025479 11.044554 11.659637 11.296968 11.243428 11.46929 10.994355 11.200488 11.067666 10.475057 11.71546 11.276990
                    IA.7      IA.8      IA.9       IIIA     IIIA.1     IIIA.2        IB      IB.1      IB.2      IB.3      IB.4      IB.5      IB.6
ENSG00000000003 11.719135 12.145895 11.967158 11.38930 12.206753 12.14137 12.406208 11.997356 11.848244 11.774809 12.494961 12.349851 12.705659
ENSG00000000005  7.949658  5.718413  9.940843  9.56796  6.576937 13.25105  9.313064  8.058747  4.817829  9.628397  3.813405  8.372433  8.548669
ENSG00000000419 11.009176 11.535391 11.126248 11.07888 10.983980 11.03469 10.922323 10.493871 11.710448 10.781391 11.134840 11.265681 10.785738
                    IB.7       IIB      IIIB    IIIB.1       IVB     IVB.1     IVB.2     IVB.3     IVB.4     IVB.5     IVB.6     IVB.7     IVB.8
ENSG00000000003 11.945320 11.146325 12.314876 10.683122 11.700878 12.03425 12.37142 12.313926 11.904341 12.12782 11.606621 12.303906 11.972398
ENSG00000000005  8.128062  5.688107  8.850998  4.973758  6.943269 10.43910 11.54160  8.604392  5.017632 11.11294  7.925867  9.843125  8.425616
ENSG00000000419 11.146361 11.716827 11.061008 11.491143 11.632611 11.00627 11.35595 11.831702 11.328684 11.60100 11.982784 11.613702 11.046658
                     IC      IC.1      IIIC    IIIC.1    IIIC.2    IIIC.3     IIIC1   IIIC1.1   IIIC1.2   IIIC1.3     IIIC2   IIIC2.1   IIIC2.2
ENSG00000000003 11.812461 12.02538 11.307994 11.843575 12.612982 11.707683 11.284029 10.948684 11.430758 11.841208 11.74172 11.106690 11.224993
ENSG00000000005  5.222151  7.92608  6.483701  7.928393  7.978908  3.619135  8.341647  4.362315  6.034838  8.973218 13.17833  4.414491  6.572413
ENSG00000000419 11.724479 10.89850 11.949522 11.261786 11.340922 11.336472 11.288843 11.114552 11.355969 11.649690 11.43312 11.266206 11.246915
                   IIIC2.3   IIIC2.4   IIIC2.5
ENSG00000000003 11.754705 11.757899 13.13517
ENSG00000000005  5.781998  5.347933 14.51954
ENSG00000000419 11.227209 11.231242 11.18104
```

For a better understanding of data, prior to the analysis of differential expression, it can see some plots that explain some study conditions. In the Heatmap below (Figure 4) we can see the overexpressed genes represented with pink figures that stand out from the blue background of the Heatmap:

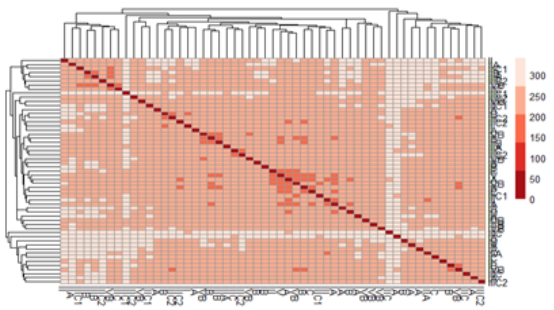



Figure 4. Strong red colors indicates values close to zero of greater similarity, while the paler colors, with values from 300 to 350, indicates greater distance between the samples, therefore less relationship between them. Very few genes are located near the strong red main diagonal. Most genes (pale pink and white colors) are very distant from the diagonal, only a few (strong pink colour) are close. This fact leads us to think that in the comparisons between the different conditions of the samples (between the different types of tumours) there will be few genes that will be differentially expressed.

## Descriptive and Graphical Analysis from RNAseq Data Processing

## Differential expression analysis:

Using *results* function from *DESeq2* and adding a contrast = c ("Condition", "II", "III", for example), comparison is made between groups, each group represents infiltration degree and invasiveness of the tumor, so, the first comparison is between II and III groups, the second comparison is between III and IV groups, the thirds comparison is between IV and IA groups and so on. This function also calculates the log2FoldChange (LFC) or estimate the effect indicating change in gene expression of one sample in relation to the other. Additionally, *DESeq2* performs a hypothesis test for each gene, thereby looking for technical variability of the experiment, for this it calculates the *p-value* of each gene. *p-value* is adjusted (padj) less than 0.05. The adjustment is obtained by the Benjamini-Hochberg method, indicating the false discovery rate [20]. After having made all comparisons, the differentially over-expressed genes are obtained, which will be used to build a file with the highly significant genes that will be used in the integration with the CNA genes.

*Comparison between II and III groups (samples=patients):*

```
res1<-results (dds1, contrast = c("Condition","II","III"), alpha = 0.05)
summary(res1)
```

```
## out of 53505 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up): 89, 0.17%
## LFC < 0 (down): 59, 0.11%
## outliers [1]: 365, 0.68%
## low counts [2]: 20601, 39%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of? results
## [2] see 'independentFiltering' argument of? results
```

Total genes comparison, 535051. Of these only 148 genes are differentially expressed, 89 are overexpressed (LFC> 0, up), with 17% expression of samples II in relation to samples III. 59 genes of lower expression.

## Adjustment of p-values

```
resadj1<-subset (res1, padj<0.05)
summary(resadj1)
```

```
## out of 148 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up): 89, 60%
## LFC < 0 (down): 59, 40%
## outliers [1]: 0, 0%
## low counts [2]: 0, 0%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of? results
## [2] see 'independentFiltering' argument of? results
```

Differentially expressed genes (148) are observed. The same 89 overexpressed but with a higher percentage of expression than the result without the adjusted p-values. 60% expression of samples II in relation to samples III.

Now let's look at the 30 differentially overexpressed genes according to the fitted p-values.

```
top_genes1<-resadj1[order(resadj1$log2FoldChange),]
top_genes_DESeq2_1<-rownames(top_genes1) [1:30]
top_genes_DESeq2_1
```

```
[1] "ENSG00000226317" "ENSG00000101441" "ENSG00000135220" "ENSG00000120211" "ENSG00000101292" "ENSG00000224817" "ENSG00000187172" "ENSG00000122145"
[9] "ENSG00000231213" "ENSG00000223784" "ENSG00000125533" "ENSG00000109132" "ENSG00000236956" "ENSG00000185640" "ENSG00000169164" "ENSG00000101076"
[17] "ENSG00000214381" "ENSG00000164825" "ENSG00000105388" "ENSG00000110680" "ENSG00000124253" "ENSG00000001626" "ENSG00000175311" "ENSG00000075891"
[25] "ENSG00000127831" "ENSG00000115844" "ENSG00000168878" "ENSG00000144355" "ENSG00000196510" "ENSG00000260877"
```

## Comparison between III and IV groups (samples):

```
res1A<-results (dds1, contrast = c ("Condition", "III,"IV"), alpha = 0.05)
summary(res1A)
```

```
## out of 53505 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up): 53, 0.099%
## LFC < 0 (down): 13, 0.024%
## outliers [1]: 365, 0.68%
## low counts [2]: 22654, 42%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of? results
## [2] see 'independentFiltering' argument of? results
```

```
top_genes2<-resadj2[order(resadj2$log2FoldChange),]
top_genes_DESeq2_2<-rownames(top_genes2) [1:6]
top_genes_DESeq2_2
```

```
[1] "ENSG00000147381" "ENSG00000046774" "ENSG00000211598" "ENSG00000211648" "ENSG00000276775" "ENSG00000211673"
```

After all the comparisons, a final file of most significative genes (RNAseq) is manually built (see Table 17) by further integration with CNA data

Table 17. RNAseq counts of most significative genes

| | Gene_ID <chr> | II <int> | II <int> | II <int> | III <int> | III <int> | IV <int> | IA <int> | IA |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ENSG00000046774 | 0 | 353 | 217 | 0 | 29 | 0 | 13 | 0 |
| 2 | ENSG00000048545 | 2 | 5 | 1 | 0 | 2 | 0 | 0 | 1 |
| 3 | ENSG00000077522 | 19 | 9846 | 162 | 19 | 10 | 108 | 594 | 101 |
| 4 | ENSG00000086967 | 3 | 2522 | 29 | 34 | 59 | 54 | 110 | 494 |
| 5 | ENSG00000101292 | 0 | 0 | 0 | 92 | 4 | 0 | 9 | 0 |
| 6 | ENSG00000101441 | 0 | 0 | 0 | 21 | 42 | 4 | 0 | 0 |

6 rows|1-10 of 56 columns

## Statistical data analysis

## CNA_36 and RNAseq_96 data integration

CNA_36 and RNAseq_96 data integration is carrying out by using *sparse partial least squares regression (sPLS)* of *mixOmics* R package. This is multivariate method that allows modeling multiple responses in data of high multicollinearity such as omic data (Wold et al., 2001), also, is not limited to correlated variables. The integration is achieved between 2 data matrices X and Y. *sPLS* is executed in the Bioconductor *mixOmics* package [21]. The aim of dispersed partial least squares (sPLS) methodology is to maximize the covariance between both data sets and to identify latent variables. In this work will be carried out correlation between two matrices: matrix X from RNAseq_96 counts (transcripts) data and matrix Y from CNA_36 data. Analysis involves the *perf* function that is used for cross-validation with 10-fold, with a few repetitions of 100 (nrepeat = 100).

## Matrix X creation

First, is created the *dds2* object with the RNAseq data of 96 differential expressed genes

```
dds2<-DESeqDataSetFromMatrix (countData = RNAseq_96,
                    colData = metadata,
                    design = ~Condition,
                    tidy = TRUE)
dds2
```

```
class: DESeqDataSet
dim: 96 55
metadata (1): version
assays (1): counts
rownames (96): ENSG00000046774 ENSG00000048545…… ENSG000000278532
ESNG00000280323
rowData names (0):
colnames (55): II II…. IIIC2 IIIC2
colData names (2): Name Condition
```

Then, is created the transposed X matrix (genes remain in columns like Y matrix of CNA data).

```
X <- assay(dds2)
XT<-t(X)
dim (XT)
```

```
## [1] 55 96
```

Matrix X transposed has 55 rows (patients) and 96 columns (transcripts)

## Matrix Y creation

Matrix Y is created with the CNA data of 36 most significant genes

```
Y <- CNA_36
Y<-as.data. frame(Y)
Y [] <-lapply (Y [2:36], as. numeric)
dim(Y)
```

```
1[] 55 37
```

```
na.row<-sample (1: nrow(Y), replace = TRUE)
na.col<-sample (1: ncol(Y), replace = TRUE)
Y<-as. matrix(Y)
Y<-Y [, -1]
dim(Y)
```

```
1[] 55 36
```

Matrix Y has 55 rows (patients) and 36 columns (genes)

## Integration using *sPLS*

Prior the integration, it´s necessary to include into model a larger number of principal component (ncomp = 5) because of perf function perform a deeper study to determine the real principal components must be selected to integrate into the model.

```
spls0<-spls (XT, Y, ncomp=5, mode = "regression")

tune. spls0<-perf (spls0, validation = "Mfold", folds = 9, progressBar = FALSE, nrepeat = 100)
```

Plotting this result, Figure 5 show the number of principal components selected by the model.

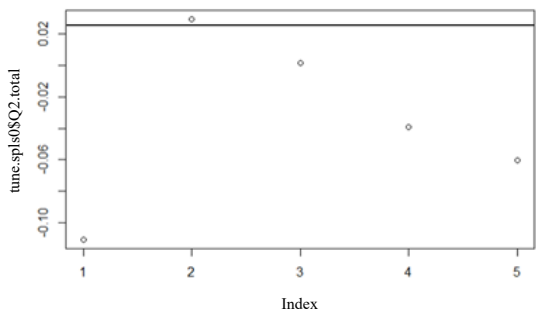
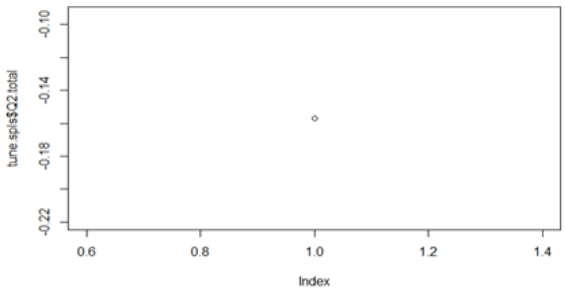



Figure 5. In this plotting it´s seen the number of principal components selected by the model (ncomp = 1) since there is only one component above the 0.0975 line, therefore it proceed to model both data sets with one component

```
plot (tune. spls0$Q2.total)
abline(h=0.0975)
```

```
spls<-spls (XT, Y, ncomp=1, mode = "regression")
tune. spls<-perf (spls, validation = "Mfold", folds = 9, progressBar = FALSE, nrepeat = 100)
```

```
plot (tune. spls$Q2.total)
abline(h=0.0975)
```




Now, it will model the entire 96 genes from matrix X (RNAseq) and the 36 genes from matrix Y (CNA) usin the *spls* function

```
MyResult.spls <- spls (XT, Y, keepX = c (96, 96), keepY = c (36,36))
```

From these results it can obtain two plots. The first plot (Graphical representation of the samples) shows, by clustering, the relationship between each block of samples separately, Block X (RNAseq data) and Block Y (CNA data) (see Figure 6).

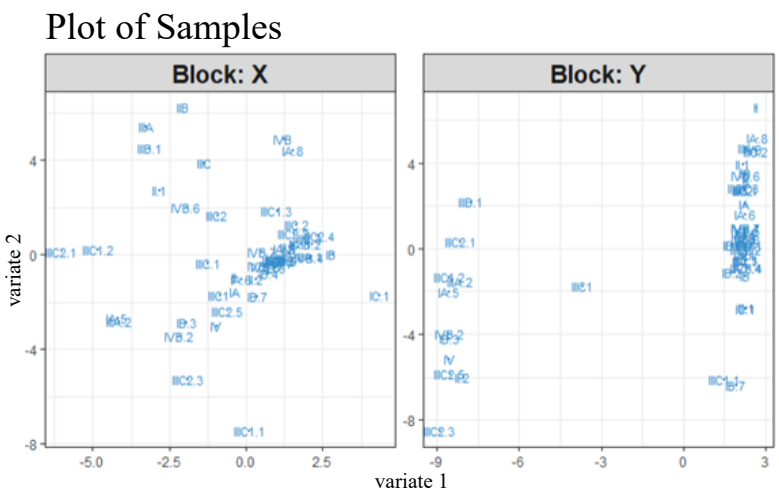



Figure 6. Plot of samples. Blocks X and Y shows clustering between samples. On Block X clustering is more remarkable between all of the samples, while on Block Y, separation between samples is more remarkable, thus, patients of most infiltration and invasiveness tumor (IIIB, IVB, IIIC) clustering togheter but more separated than patients of less infiltration and invasiveness tumor (IA, IB, IC). This means, transcripts (genes) on Block X are expressed in a similar way in the different types of tumors, while specifc CNA genes on Block Y are expressed in a different way according the types of tumor.

```
plotIndiv (MyResult.spls,
ind. names=TRUE,
title='Plot of Samples')
```

The second plot (Graphical representation of the variables) shows the correlation between genes (Figure 7).

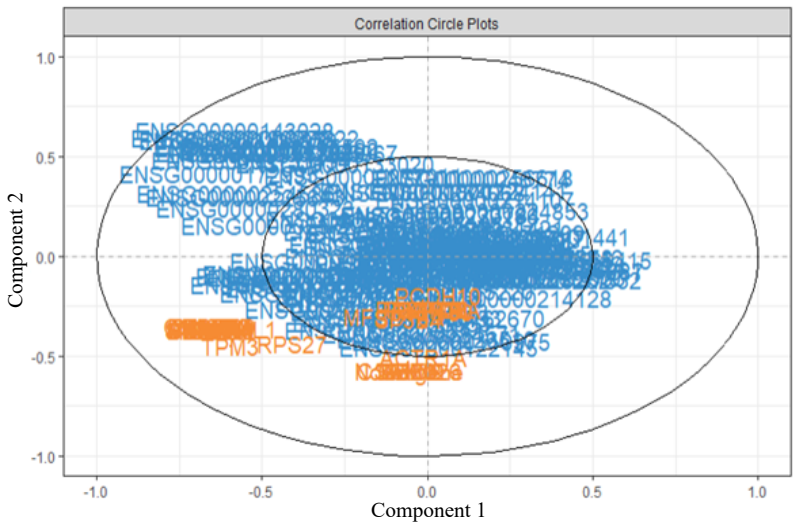



Figure 7. Correlation between genes. Orange genes are CNAs and blue genes are transcripts (RNAseq counts). All genes within the center circle (around zero) are not correlated while genes furthest from the central circle (far from zero) show high correlation. RNAseq and CNA genes located on the same plane (left side) are highly correlated either positively or negatively. For example, TPM3, RPS27 and ACTR1A genes (CNA) are positively correlated with each other but negatively correlated with blue genes (RNAseq) in the same plane above the central zero line.

```
plotVar (MyResult.spls,
var. names = TRUE, TRUE),
pch=c (5,5))
```

## Results

Graphic study on samples clustering in RNAseq and CNA data showed a clear clustering of samples. RNAseq samples showed clustering around central zero of all types of tumors, without clear separation between them. This indicates variance of different samples is not explained by the transcripts (genes). Clusters were also presented in all the samples without clear separation between them. Clusters top and bottom of central zero, especially of tumor samples with most infiltration and invasiveness (IVB, IC, IIIC, IIIC1, IIIC2, IIIC3) explained the most proportion of variance (Block X Figure 6).

CNA genes showed clearly separate clustering's according to the types of tumors. Samples of tumor of less infiltration and invasiveness (II, III, IV, IA, IIA, IIIA, IB, IIIB) were clustered more closely (right side Block Y Figure 7) near of central zero and tumor with most infiltration and invasiveness (IVB, IC, IIIC, IIIC1, IIIC2, IIIC3) were clustered more closely away from central zero (left side Block Y Figure 7). Many samples were clustered very closely at central zero especially samples belonging to tumors with less infiltration and invasiveness (II, III, IV, IA, IIA, IIIA, IB, IIIB), indicating some CNA genes have a weak influence on tumors with less infiltration and invasiveness.

Another important finding in both files (RNAseq and CNA data) according to graphic study on samples is that they showed a strong negative correlation between them. Tumors with more infiltration and invasiveness shows high dispersion under the central zero, while tumors with less infiltration and invasiveness shows moderate dispersion above the central zero. This indicates both types of genes (mRNA transcripts and CNA genes) are highly expressed in aggressive tumors.

According to graphical study on genes (Correlation Circle on Figure 7), most RNAseq genes are clustered in the central circle indicating that many of them don´t have influence in the expression of the different types of tumors. On the other hands very few CNA genes are within the central circle but most of them outside the central circle indicating that CNA genes have a most influence in the expression on the tumors than the RNAseq genes.

The variance explained by genes (RNAseq and CNA genes) located in the largest circle is similar but with opposite correlation in some of them. RNAseq genes located above the central horizontal line, strongly expressed, maintain a negative correlation with CNA genes below the

central horizontal line, while RNAseq genes located below the central horizontal line together with CNA genes, maintain a positive correlation, indicating both specific RNAseq and CNA genes are molecularly linked in the expression of this disease.

Among CNA genes located in the largest circle below the central horizontal line are, *TPM3* (tropomyosin 3) and *RPS27* (ribosomal protein S27) genes, both located on chromosome 1 and *ACTR1A* (Alpha-centractin: actin related protein 1A: regulation of G2/M transition of mitotic cell cycle. Spermatogenesis) gene is located on chromosome 10 also outside the central circle keeping direct positive correlation with the *TBX22* gene (T-box transcription factor, transcriptional regulator involved in developmental processes) located on chromosome X. [22,23].

Among RNAseq genes located in the largest circle above the central horizontal line are, *ENSG00000143028* (SYPL2: Synaptophysin-like protein 2: heart development) human gene located on Chromosome 1; *ENSG00000077522* (*ACTN2*: alpha actinin-2, structural constituent of muscle) human gene also located on Chromosome 1; *ENSG00000086967* (*MYBPC2*: Myosin-binding protein C, fast-type: It may modulate muscle contraction or may play a more structural role, structural constituent of muscle) human gene located on Chromosome 19 and *ENSG00000253767* (*PCDHGA8* Protocadherin gamma-A8: Potential Calcium-dependent cell-adhesion protein. May be involved in the establishment and maintenance of specific neuronal connections in the brain.) human gene located on Chromosome 5, most of them involved in muscle disorders [22,23].

## Discussion

RNAseq and CNA genes showed a high expression and correlation between them in the different uterine cancer, especially the FIGO stages of most infiltration and invasiveness (IIIB, IVB, IIIC1, IIIC2), unlike gene fusion study findings by Chiang et al. [24], in uterine carcinosarcoma, which reported all tumors they found as FIGO stage IB. Approximately 16 of 96 RNAseq and 12 of CNA genes have relationship with this pathology: *TPM3, RPS27* and *ACTR1A CNA* genes are correlated to *ENSG00000122145* (TBX22), *ENSG00000143028* (SYPL2), *ENSG00000077522* (ACTN2), *ENSG00000086967* (*MYBPC2*) and *ENSG00000253767 (PCDHGA8)* RNAseq genes. *TPM3, RPS27 CNA* genes and *ENSG00000143028 (SYPL2), ENSG00000077522 (ACTN2)* RNAseq genes are located on chromosome 1. To date, TMP3 gene have been reported among tumors with NTRK1 (Neurotrophic Tyrosine Receptor Kinase) -related fusion-positive tumors [25]. *NTRK1* encodes for *TrK* (Tropomyosin receptor kinase) genes, *Trk* pathway aberrations, including gene fusions, is involved in many human cancers, with NTRK gene fusions [26,27]. *RPS27* gene has been reported by Xiong et al. [28] demonstrating that neddylation stabilizes *RPS27* gene "to confer the survival of cancer cells". Neddylation (ubiquitin-like protein NEDD8: neural-precursor-cell-expressed developmentally down-regulated 8 is conjugated to its target proteins) causes a structural change in the substrate [29]. *ACTR1A* gene has been reported by Wang et al. [30] which is correlated with metabolic enzyme phosphoglycerate mutase enzyme 1 (*PGAM1*), this is a key enzyme in the glycolysis pathway (glycolysis is related to cancer progression). The *SYPL2* gene (RNAseq gene) highly expressed here was identified in small cell lung cancer by combining affinity propagation clustering of selected genes from different cancer databases [31]. *ACTN2* gene was identified by Lo et al. [32] like responsible for hepatocellular carcinoma by using the transcriptome fusion genes methodology. *MYBPC2* gene has been found highly expressed in rhabdomyosarcoma (cancerous malignant tumor in the muscles attached to the bones) [33].

All these findings derived from omic data integration performed here, allow clear up interactions between large molecules at the metabolic level that can be used as diagnostic and treatment mechanisms in personalized medicine in patients who enroll in clinical trials.

## Conclusions

I found copy number alteration results in differential gene expression in uterine carcinosarcoma. High correlation between CNA (*TPM3, RPS27, ACTR1A*) and RNAseq (*SYPL2, ACTN2, MYBPC2*) genes may be involved in the pathogenesis of uterine carcinosarcoma.

## Acknowledgments

Data analysis was possible thanks to TCGA data open access necessary to carry out this study.

## Conflicts of Interest

The author declares no conflict of interest.

## References

1) American Cancer Society. Endometrial (uterine) Cancer. Am. Can Soc.
2) Kernochan LE, Garcia RL. Carcinosarcoma (malignant mixed Müllerian tumor) of the uterus: advances in elucidation of biologic and clinical characteristics. JJ Natl Compr Canc Netw. 2009, 7: 550-557.
3) Jones S, Stransky N, McCord CL, Cerami E, Lagowski J, Kelly D, et al. Genomic analyses of gynaecologic carcinosarcoma reveal frequent mutations in chromatin remodelling genes. Nat Commun. 2014, 5: 1-7.
4) Zhao S, Bellone S, Lopez S, Thakral D, Schwab C, English DP, et al. Mutational landscape of uterine and ovarian carcinosarcoma implicates histone genes in epithelial–mesenchymal transition. Proc Natl Acad Sci U S A. 2016, 113: 12238-12243.
5) Li W, Lee A, Gregersen PK (). Copy-number-variation and copy-number-alteration region detection by cumulative plots. BMC Bioinformatics. 2009, 10: S67.
6) Sutrala S, Goossens D, Williams N, Heyrman L, Adolfsson R, Norton N, et al. Gene copy number variation in schizophrenia. Schiz Res. 2007, 96: 1-3.
7) Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Gen. 2013, 45: 1134-1140.

8) Cherniack AD, Shen H, Walter V, Stewart C, Murray BA, Bowlby R, et al. Integrated molecular characterization of uterine carcinosarcoma. Cancer Cell. 2017, 31: 411-423.
9) Alvarez M, Schrey AW, Richards CL. Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Mol Ecol 2015, 24: 710-725.
10) Langley SR, Dwyer J, Drozdov I, Xiaoke Y, Mayr M. Proteomics: from single molecules to biological pathways. Review. Cardiovasc Res. 2013, 97: 612-622.
11) De Sanctis G, Colombo R, Damiani Ch, Sacco E, Vanon M. Omics and Clinical Data Integration. Chapter 15 from book: Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition. Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni. 2018.
12) Broad Institute TCGA Genome Data Analysis Center (2016). Aggregate Analysis Features. Broad Institute of MIT and Harvard.
13) Noble M, DeFreitas T, Heiman D. Final report: GDAC Firehose Integration with The Genomic Data Commons. Broad Institute of MIT & Harvard. 2016.
14) Olshen AB, Venkatraman R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004, 4: 557-572.
15) Beroukhim R, Mermel CH, Porter D, Guo W, Soumya R, Donovan J, et al. The landscape of somatic copy-number alteration across human cancer. Nature. 2010, 463: 899-905.
16) Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC 2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Gen Biol. 2011, 12: R41.
17) Anders S, Pyl PT. Wolfgang Huber HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015, 31: 166-169.
18) https://training.seer.cancer.gov/staging/systems/schemes/figo.html
19) Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Geno Biol. 2014, 15: 550.
20) Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics. 2001, 1165-1188.
21) Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017, 13: e1005752.
22) https://www.ensembl.org/index.html.
23) https://www.uniprot.org.
24) Chiang S, Cotzia P, Hyman DM, Drilon A, Ta WD, Zhang L, et al. NTRK Fusions Define a Novel Uterine Sarcoma Subtype with Features of Fibrosarcoma. Am J Surg Pathol. 2018, 42: 791-798.
25) Agaram NP, Zhang L, Sung YS, Chung CT, Antonescu CR, Fletcher CD. Recurrent NTRK1 gene fusions define a novel subset of locally aggressive lipofibromatosislike neural tumors. Am J Surg Pathol. 2016, 40: 1407-1416.
26) Kaplan DR, Martin-Zanca D, Parada LF. Tyrosine phosphorylation and tyrosine kinase activity of the TRK proto-oncogene product induced by NGF. Nature. 1991, 350: 158-160.
27) Barbacid M. Structural and functional properties of the TRK family of neurotrophin receptors. Ann N Y Acad Sci. 1995, 766: 442-458.
28) Xiong X, Cui D, Bi Y, Sun Y, Zhao Y. Neddylation modification of ribosomal protein RPS27L or RPS27 by MDM2 or NEDP1 regulates cancer cell survival. FASEB J. 2020, 34: 13419-13429.
29) Rabut G, Peter M. Function and regulation of protein neddylation. 'Protein modifications: beyond the usual suspects' review series. EMBO Rep. 2008, 9: 969-976.
30) Wang Y, Xiong X, Hua X, Liu W. Expression and Gene Regulation Network of Metabolic Enzyme Phosphoglycerate Mutase Enzyme 1 in Breast Cancer Based on Data Mining. BioMed. R.I. 2021, 39.
31) Li J, Chang M, Gao O, Song X. Gene Identification for Small Cell Lung Cancer via Combining Affinity Propagation Clustering and Conditional Mutual Information. IEEE 8th Data Driver Control and Learning System Conference, Dali China. 2019.
32) Lo LH, Lam CY, To JC, Chiu CH, Keng VW. Sleeping Beauty insertional mutagenesis screen identifies the pro-metastatic roles of CNPY2 and ACTN2 in hepatocellular carcinoma tumor progression. Elsevier. 2021, 541: 70-77.
33) Chen Z, Li XY, Guo P, Wang D-L. MYBPC2 and MYL1 as Significant Gene Markers for Rhabdomyosarcoma. Technol Cancer Res Treat. 2021, 20: 1-15.
34) Yu C, Qin N, Pu Z, Song C, Wang C, Chen J, et al. Integrating of genomic and transcriptomic profiles for the prognostic assessment of breast cancer. Breast Cancer Res Treat. 175: 691-699.
35) Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics Intell Lab Sys. 2001, 58: 109-130.