Actividad Integradora

Bruno Yánez, Javier Lizárraga, Maximiliano Martínez, Pedro Escoboza

```r
rm(list=ls());
options(stringAsFactors = FALSE);
```

```r
library("gplots"); # heatmap.2()
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
# Función para cálculo de diferencia de prueba t student.
t_student_diff <- function(df, index_list_a, index_list_b, col_names =  c("Tumor", "Normal", "Diff")) {
  res <- t(apply(df, 1,
                 function(x) {
                   m_1 <- mean(x[index_list_a], na.rm = TRUE);
                   m_2 <- mean(x[index_list_b], na.rm = TRUE);
                   m_diff <- abs(m_1 - m_2);
                   c(m_1, m_2, m_diff);
                 }));
  colnames(res) <- col_names;
  return(res);
};

# Función para cálculo de diferencia de prueba t student para dataframes con esquema de clases.
t_student_classes <- function(df, classes, cr_a, cr_b,
                              col_names = c("A", "B", "p_value", "fold_change")){
  samples_a <- which(classes == cr_a);
  samples_b <- which(classes == cr_b);
  t_res <- t(apply(df, 1,
                   function(x){
                     t_test <- t.test(x[samples_a], x[samples_b]);
                     c(t_test$estimate[1], t_test$estimate[2], t_test$p.value, t_test$estimate[1] - t_te
                   }));
  colnames(t_res) <- col_names;
  return(t_res);
};

# Regresa un dataframe con los primeros n resultados ordenados por la columna col.
get_top_n <- function(df, col, n, decreasing = FALSE){
  return(head(df[order(col, decreasing=decreasing),],n));
};

# Normalización de datos.
```

```
normalize <- function(x, min, max){
  return((x-min)/(max-min));
};

# División de datos en grupos por rangos de valores.
freq_groups <- function(vec, bounds){
  num_bounds <- length(bounds);
  freqs <- integer(num_bounds);
  for (i in 2:num_bounds){
    for (j in 1:length(vec)){
      if (vec[j] >= bounds[i-1] & vec[j] < bounds[i]){
        freqs[i] = freqs[i] + 1;
      }
    }
  }
  return(freqs);
};
```

Análisis de Multi_Cancer_Data

```
load("Multi_Cancer_Data.Rdata");
df <- multi_cancer_data;
rm(multi_cancer_data);
```

Diferenia entre muestras normales y muestras de cáncer color

```
# Selección de muestras normales.
normal_samples_indexes <- grep("Normal", colnames(df));
print(normal_samples_indexes);
```

```
##  [1] 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209
## [20] 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228
## [39] 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247
## [58] 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266
## [77] 267 268 269 270 271 272 273 274 275 276 277 278 279 280
```

```
# Selección de muestras de cáncer colorrectal.
colorectal_cancer_indexes <- grep("Tumor__Colorectal", colnames(df));
print(colorectal_cancer_indexes);
```

```
##  [1] 33 34 35 36 37 38 39 40 41 42 43
```

```
# Prueba t student.
tstudent_normal_with_colorectal <- data.frame(t_student_diff(df, normal_samples_indexes, colorectal_canc

# Seleccionar 10 entradas con mayor diferencia.
tstudent_normal_with_colorectal <- get_top_n(tstudent_normal_with_colorectal, tstudent_normal_with_colo

print(tstudent_normal_with_colorectal);
```

```
##                                                                                    Tumor
## MMP12 Matrix metalloproteinase 12 (macrophage elastase)_L23808_at           -0.203500000
## CARCINOEMBRYONIC ANTIGEN PRECURSOR_M29540_at                                 0.036511111
## MMP1 Matrix metalloproteinase 1 (interstitial collagenase)_X54925_at        -0.143000000
## CDX1 Caudal type homeo box transcription factor 1_U51095_at                  0.078966667
## Transforming growth factor-beta induced gene product (BIGH3) mRNA_M77349_at -0.200255556
## TUMOR-ASSOCIATED ANTIGEN CO-029_M35252_at                                    0.095433333
## Homeobox protein Cdx2 mRNA_U51096_at                                        -0.293233333
## Gamma-glutamyl hydrolase (hGH) mRNA_U55206_at                               -0.089533333
## GC-Box binding protein BTEB2_D14520_at                                       0.007233333
## NF-E2-related factor 3_RC_AA132523_at                                       -0.088766667
##                                                                                 Normal
## MMP12 Matrix metalloproteinase 12 (macrophage elastase)_L23808_at           2.307545
## CARCINOEMBRYONIC ANTIGEN PRECURSOR_M29540_at                                2.538545
## MMP1 Matrix metalloproteinase 1 (interstitial collagenase)_X54925_at        2.088818
## CDX1 Caudal type homeo box transcription factor 1_U51095_at                 2.144182
## Transforming growth factor-beta induced gene product (BIGH3) mRNA_M77349_at 1.732727
## TUMOR-ASSOCIATED ANTIGEN CO-029_M35252_at                                   2.008545
## Homeobox protein Cdx2 mRNA_U51096_at                                        1.581818
## Gamma-glutamyl hydrolase (hGH) mRNA_U55206_at                               1.710818
## GC-Box binding protein BTEB2_D14520_at                                      1.775000
## NF-E2-related factor 3_RC_AA132523_at                                       1.659273
##                                                                                   Diff
## MMP12 Matrix metalloproteinase 12 (macrophage elastase)_L23808_at           2.511045
## CARCINOEMBRYONIC ANTIGEN PRECURSOR_M29540_at                                2.502034
## MMP1 Matrix metalloproteinase 1 (interstitial collagenase)_X54925_at        2.231818
## CDX1 Caudal type homeo box transcription factor 1_U51095_at                 2.065215
## Transforming growth factor-beta induced gene product (BIGH3) mRNA_M77349_at 1.932983
## TUMOR-ASSOCIATED ANTIGEN CO-029_M35252_at                                   1.913112
## Homeobox protein Cdx2 mRNA_U51096_at                                        1.875052
## Gamma-glutamyl hydrolase (hGH) mRNA_U55206_at                               1.800352
## GC-Box binding protein BTEB2_D14520_at                                      1.767767
## NF-E2-related factor 3_RC_AA132523_at                                       1.748039
```

Análisis de TCGA_COADREAD_comp_data

```r
rm(list=setdiff(ls(), lsf.str()));
load("TCGA_COADREAD_comp_data.RData");
df <- tcga_coadread;
rm(tcga_coadread);
```

Diferencia entre jóvenes y adultos

```r
# Prueba de t student para TCGA COARDREAD por las clases Young  y Old.
tcga_t_test <- t_student_classes(df,tcga_coadread_class,"Young","Old",c("Young", "Old", "p_value", "Fol

# Filtración de datos para eliminar entradas no significativas.
tcga_t_test_filter <- apply(tcga_t_test[,1:2],1,function(x){all(x<1)});
tcga_t_test <- tcga_t_test[-which(tcga_t_test_filter),];

# Ordenar por diferencia.
tcga_t_test <- tcga_t_test[order(tcga_t_test[,4], decreasing=TRUE),];

# Genes con mayor diferencia de expresión entre jóvenes y ancianos.
print("# Genes con mayor diferencia de expresión entre jóvenes y ancianos:");
```

```
## [1] "# Genes con mayor diferencia de expresión entre jóvenes y ancianos:"
```

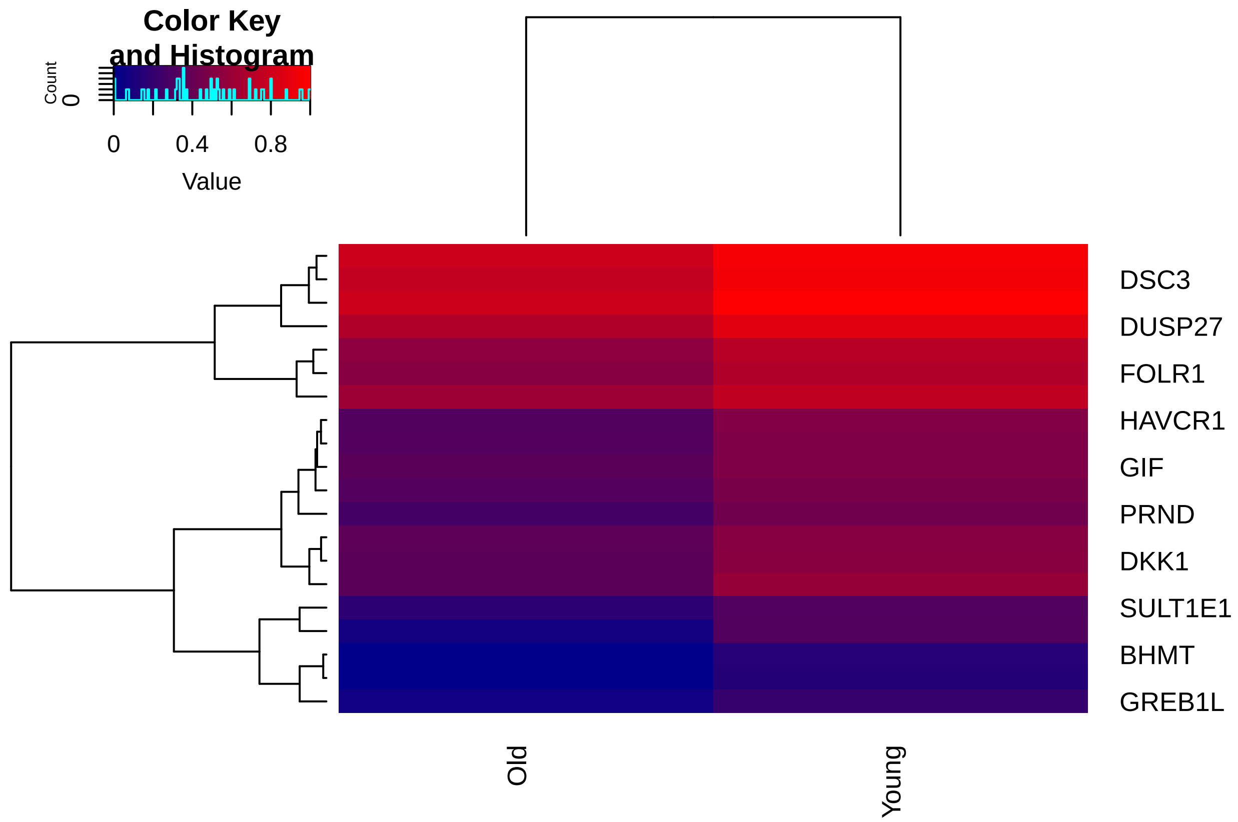
```r
write.table(rownames(tcga_t_test[which(tcga_t_test[,4] > 0),])[1:20], sep='\t', quote=F, row.names=F, c
```

```
## GATA4
## PCSK1N
## XIST
## DUSP27
## HAVCR1
## DSC3
## DKK1
## PRND
## FOLR1
## CPS1
## GAL
## FZD9
## GLDC
## GREB1L
## SULT1E1
## BHMT
## GIF
## PEG10
## NKX2-1
## LEFTY2
```

```r
# Generar matriz para mapa de calor.
hm_mat <- tcga_t_test[rownames(tcga_t_test)[1:20],];

# Remover columnas de p_value y fold_change.
hm_mat <- hm_mat[,-(3:4),drop=FALSE];
colnames(hm_mat) <- colnames(tcga_t_test)[1:2];

# Normalizar valores de expresión.
exp_values <- c(hm_mat[,1], hm_mat[,2]);
min_exp_values <- min(exp_values);
max_exp_values <- max(exp_values);

hm_mat[,1] <- normalize(hm_mat[,1], min_exp_values, max_exp_values);
hm_mat[,2] <- normalize(hm_mat[,2], min_exp_values, max_exp_values);


num_colors = 128;
```

Mapa de calor

```r
# Construcción de mapa de calor.
colors_h <- colorRampPalette(c("darkblue","red"))(num_colors);
h_breaks <- seq(from=0, to=1, length=num_colors+1);

heatmap.2(hm_mat, col=colors_h, trace="none", breaks=h_breaks, cexCol=1);
```

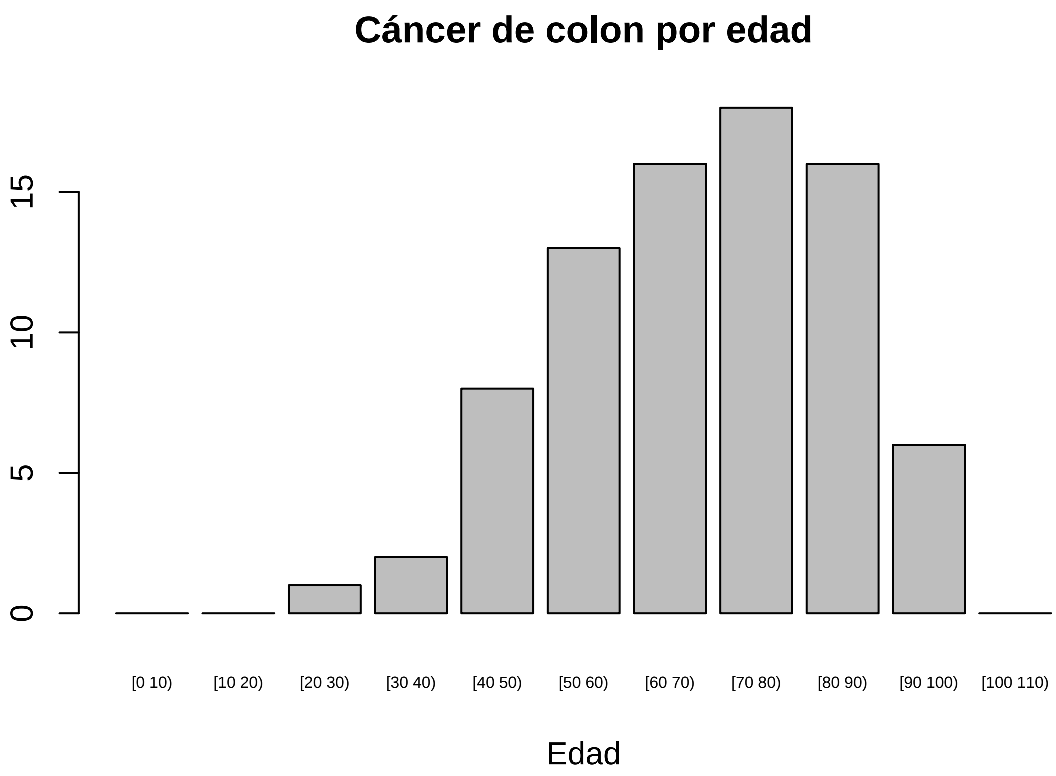


Análisis de 9_PACIENTES_DE_NUEVO_INGRESO.csv

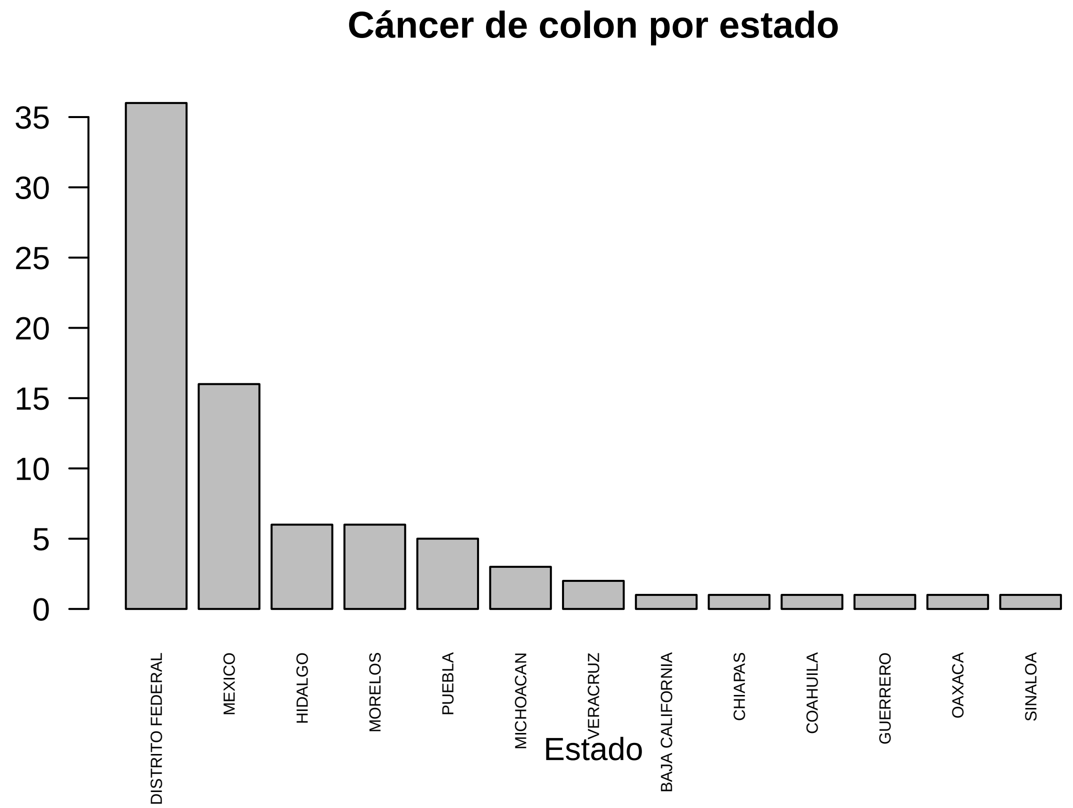
```
rm(list=setdiff(ls(), lsf.str()));
df <- read.csv("9_PACIENTES_DE_NUEVO_INGRESO.csv");

# Selección de entradas de tumores de colon.
colon_cancer <- df[grep("COLON", df$DESCRIPCION.DIAGNOSTICO),];
print(head(colon_cancer));
```

```
##     FOLIO EDAD      SEXO           ESTADO  MUNICIPIO DESCRIPCION.DIAGNOSTICO
## 85     85   76 Masculino          MORELOS XOCHITEPEC TUMOR MALIGNO DEL COLON
## 108   108   72 Masculino          HIDALGO    ACATLAN TUMOR MALIGNO DEL COLON
## 132   132   49  Femenino           MEXICO     CHALCO TUMOR MALIGNO DEL COLON
## 140   140   68  Femenino DISTRITO FEDERAL    TLALPAN TUMOR MALIGNO DEL COLON
## 145   145   54 Masculino           MEXICO   TULTEPEC TUMOR MALIGNO DEL COLON
## 180   180   35 Masculino DISTRITO FEDERAL XOCHIMILCO TUMOR MALIGNO DEL COLON
```

```
# Cáncer de colon por edad.
ranges <- c(0,10,20,30,40,50,60,70,80,90,100);
age_freq <- freq_groups(colon_cancer$EDAD, ranges);
label <- c("[0 10)", "[10 20)","[20 30)","[30 40)","[40 50)","[50 60)", "[60 70)", "[70 80)","[80 90)",
barplot(age_freq,  main="Cáncer de colon por edad", xlab="Edad", names.arg=label, cex.names=0.5);
```



```
# Cáncer de colon por estado.
state_freq <- as.data.frame(table(colon_cancer$ESTADO));
state_freq <- state_freq[which(state_freq$Freq != 0),];
state_freq <- state_freq[order(state_freq$Freq, decreasing=TRUE),];
barplot(state_freq$Freq, main="Cáncer de colon por estado", xlab="Estado", names.arg=state_freq$Var1, c
```


Cáncer de colon por estado
35
30
25
20
15
10
5
0
DISTRITO FEDERAL
MEXICO
HIDALGO
MORELOS
PUEBLA
MICHOACAN
VERACRUZ
BAJA CALIFORNIA
CHIAPAS
COAHUILA
GUERRERO
OAXACA
SINALOA
Estado