Informe PEC1 Análisis de datos ómicos

Mario Them Álvarez

2025-03-31

Contents

Resumen

En el presente trabajo se ha escogido un set de datos del repositorio Metabolomics Workbench que analiza la lipidomica y metabolomica de muestras de plasma EDTA de dos grupos de pacientes con neumonia interesticial y lobar, respectivamente, con igual presencia de sexos. En primer lugar, se ha creado un objeto de tipo SummarizedExperimment, que contiene los datos y metadatos del estudio. Luego, se ha realizado un análisis exploratorio de los datos del estudio, utilizando técnicas estadísticas como PCA, cluster annalysis y distintas visualizaciones gráficas, como boxplots o histogramas, con el objetivo de obtener una visión general del estudio. Los resultados de este análisis exploratorio muestran algunos metabolitos desregulados en ambos grupos, así como una ligera diferencia entre los grupos, lo cual da indicaciones sobre la dirección de los análisis estadísticos a realizar sobre este set de datos, que sirvan para determinar si estas diferencias son significativas. También muestra posibles outliers en los datos.

Objetivos

- Conocer y crear un objeto de tipo SummarizedExperiment de 0.
- Poner en práctica distintas técnicas estadísticas de análisis de datos multivariante sobre un conjunto de datos real.
- Determinar si existen grupos de observaciones en el set de datos elegido, y si corresponden con los grupos del estudio.
- Averiguar si hay outliers en los datos.

Métodos

Los datos del presente trabajo se han obtenido de la página Metabolomics Workbench. Los datos son mediciones de 114 marcadores lipoproteicos y 24 metabolitos de 21 individuos con neumonia lobar y 24 con neumonia intersticial. Se han medido mediante resonancia magnetica nuclear, y se encuentran expresados en unidades arbitrarias(AU).

Los datos se han descargado en forma de un archivo .txt en formato mwtab, que es un formato ampliamente utilizado en análisis metabolomicos, ya que permite agrupar los metadatos y datos del estudio en el mismo archivo. Los datos se han analizado utilizando el programa Rstudio, con versión de R 4.4.3.

Para facilitar el procesado de este archivo en R, se ha utilizado la libreria metabolomicsWorkbenchR de Bioconductor, que permite realizar querys a la base de datos Metabolomics Workbench, especificando el tipo de datos a obtener(factores,mwtab...) y un id de estudio.

Para la creación del SummarizedExperiment y el acceso a sus slots, se ha utilizado el constructor SummarizedExperiment() de la libreria del mismo nombre de Bioconductor.

Para el análisis exploratorio de los datos, se han usado funciones de gráficos básicas en R como boxplot o hist, asi como funciones del paquete ggplot2 y del paquete car. En cuanto a las técnicas estadísticas de este análisis, se ha realizado un análisis de componentes principales(PCA) y un cluster jerarquico. La visualizacion de los resultados de estos análisis se ha hecho mediante funciones del paquete factoextra.

Resultados

En primer lugar, se realizan las querys para obtener los datos de Metabolomics Workbench, descargandose por un lado el mwtab entero y por el otro cada uno de las secciones correspondientes, para facilitar el acceso a los datos de interés:

```
# Se descarga el mwtab del estudio seleccionado
mwtab <- do_query(context = "study", input_item = "study_id",
    input_value = "ST003674", output_item = "mwtab")
# Se descargan los datos experimentales del estudio
datos <- as.data.frame(do_query(context = "study", input_item = "study_id",
    input_value = "ST003674", output_item = "data"))
# Se descargan los datos de las muestras
datos_muestras <- as.data.frame(do_query(context = "study", input_item = "study_id",
    input_value = "ST003674", output_item = "factors"))
# Se descargan los datos de los metabolitos
datos_metabolitos <- as.data.frame(do_query(context = "study",
    input_item = "study_id", input_value = "ST003674", output_item = "metabolites"))
```

Después, se crean los distintos slots del objeto SummarizedExperiment y se construye el objeto:

```
# Se construye el slot assay
assay<-datos[ ,c(4,8:52)]
for(i in seq(1,45)){
  colnames<- c(colnames, mwtab$SUBJECT_SAMPLE_FACTORS[[i]][[2]])
}
colnames(assay)<-colnames[1:46]
rownames(assay)<-assay[ ,1]
assay <- assay[ ,-1]
# Se construye el slot colData
colData <- datos_muestras[ ,c(2:4,8:9)]
colData <- colData[order(colData$ST003674.local_sample_id),]
colData <- colData[ ,-1]
# Se construye el slot rowData
rowData<-data.frame(datos_metabolitos[ ,1:4])
rowData$metabolite_name<-rownames(assay)
# Se construye el objeto de tipo SummarizedExperiment
st003674<-SummarizedExperiment(
  assay = assay,
  rowData = rowData,
```

```
  colData = colData,
  metadata = mwtab$STUDY)
```

Tras la creación del objeto de tipo SummarizedExperiment, podemos proceder con la exploración de los datos.

En primer lugar, se inspeccionan las dimensiones del slot assay, que contiene los datos experimentales:

```
dim(assay(st003674))
```

```
## [1] 139  45
```

Se trata de un data frame de 139 filas y 45 columnas. El siguiente paso sería ver de qué se tratan estas columnas y filas:

```
head(rownames(assay(st003674)))
```

```
## [1] "3-Hydroxybutyric_acid" "Acetic_acid"           "Acetoacetic_acid"
## [4] "Acetone"               "ApoA1(HDL)"            "ApoA1(HDL1)"
```

```
head(colnames(assay(st003674)))
```

```
## [1] "P_762" "P_763" "P_764" "P_765" "P_766" "P_767"
```

Las filas se corresponden a los metabolitos analizados, mientras que las columnas corresponden cada uno a un paciente, codificado como P_+3 números.

En cuanto a los datos, se realiza un resumen numérico de las 5 primeras columnas:

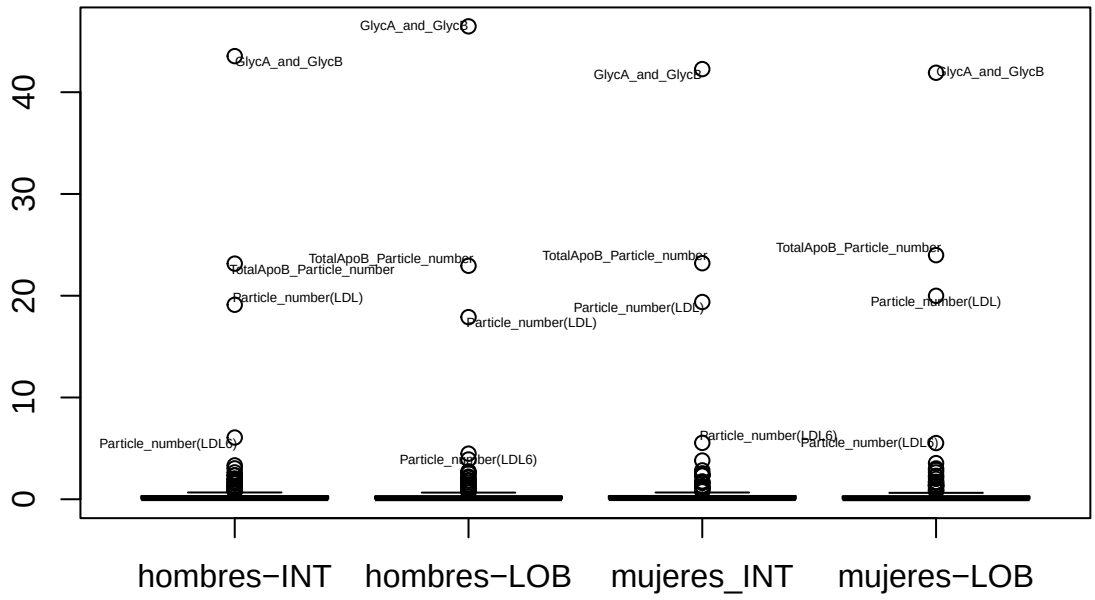
```
summary(assay(st003674)[1:5])
```

```
##      P_762              P_763              P_764              P_765
##  Min.   :   0.003   Min.   :   0.002   Min.   :   0.008   Min.   :   0.004
##  1st Qu.:   3.340   1st Qu.:   2.952   1st Qu.:   2.480   1st Qu.:   1.168
##  Median :   9.050   Median :   9.900   Median :  10.810   Median :   4.985
##  Mean   :  78.872   Mean   : 108.478   Mean   : 106.737   Mean   :  58.110
##  3rd Qu.:  24.840   3rd Qu.:  25.390   3rd Qu.:  26.137   3rd Qu.:  19.685
##  Max.   :2504.032   Max.   :5088.185   Max.   :4191.484   Max.   :3560.720
##  NA's   :2                             NA's   :1          NA's   :3
##      P_766
##  Min.   :   0.009
##  1st Qu.:   2.485
##  Median :   9.080
##  Mean   :  84.764
##  3rd Qu.:  27.795
##  Max.   :2930.434
##
```

Esto nos muestra varias cosas:

- Por un lado, que hay valores faltantes (NAs), por lo que sería necesario decidir como se tratan esos valores, si imputarlos o ignorarlos a la hora de realizar los análisis estadisticos.
- Por el otro, que los valores tienen un rango bastante amplio (de apenas superior a 0 hasta poco más de 5000 AU), por lo que sería conveniente estandarizarlos o normalizarlos de alguna forma para facilitar el análisis.

Sabiendo esto último, se realiza un boxplot de los valores medios de los distintos metabolitos por grupo experimental. Se agrupan los datos por grupo experimental para mejorar la visualización de los mismos, ya que son muchos pacientes. Los valores medios se han estandarizado dividiendolos entre la media experimental por grupo, también con el objetivo de facilitar la visualización:

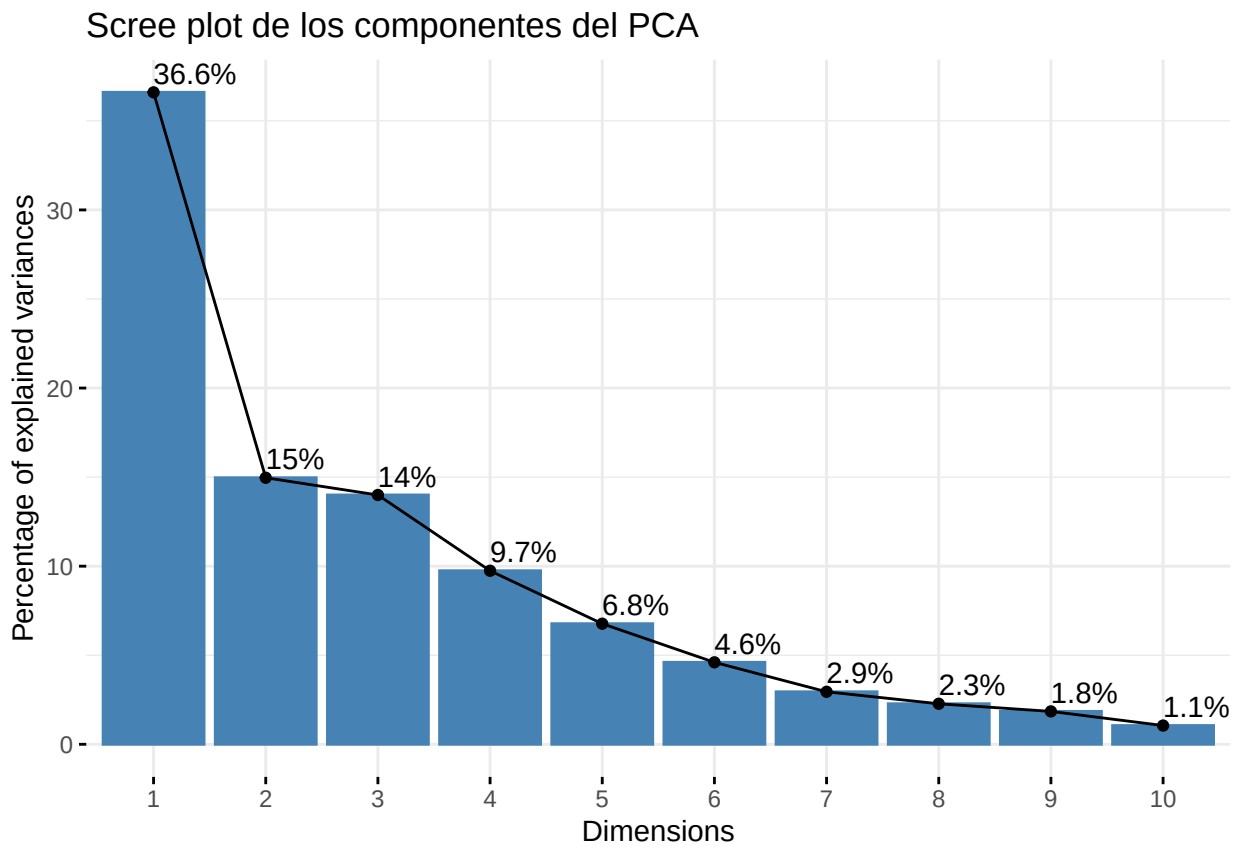



```
##  [1] "GlycA_and_GlycB"           "TotalApoB_Particle_number"
##  [3] "Particle_number(LDL)"      "Particle_number(LDL6)"
##  [5] "GlycA_and_GlycB"           "TotalApoB_Particle_number"
##  [7] "Particle_number(LDL)"      "Particle_number(LDL6)"
##  [9] "GlycA_and_GlycB"           "TotalApoB_Particle_number"
## [11] "Particle_number(LDL)"      "Particle_number(LDL6)"
## [13] "GlycA_and_GlycB"           "TotalApoB_Particle_number"
## [15] "Particle_number(LDL)"      "Particle_number(LDL6)"
```

Como se puede observar, hay algunos metabolitos con unos niveles más altos que el resto en todos los grupos. Serían candidatos para su comparación con los niveles en personas sin neumonía.

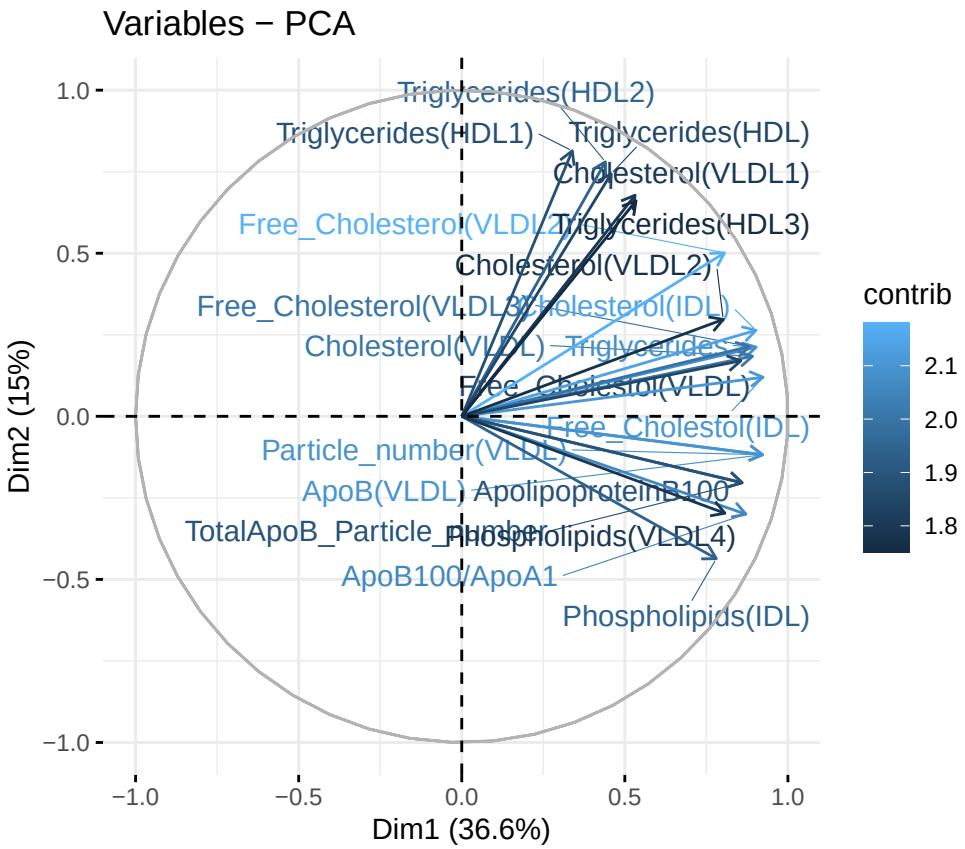
Para continuar, se realiza un análisis de componentes principales para poder determinar posibles grupos, y si estos coinciden con los grupos experimentales. A la hora de realizar el análisis de componentes principales, se han omitido los valores NA y se han escalado los datos, teniendo en cuenta lo observado en el resumen numérico del set de datos.

Se visualiza la contribución de cada una de las 10 componentes del análisis mediante un scree plot, lo que nos muestra la varianza explicada por cada una de las componentes.

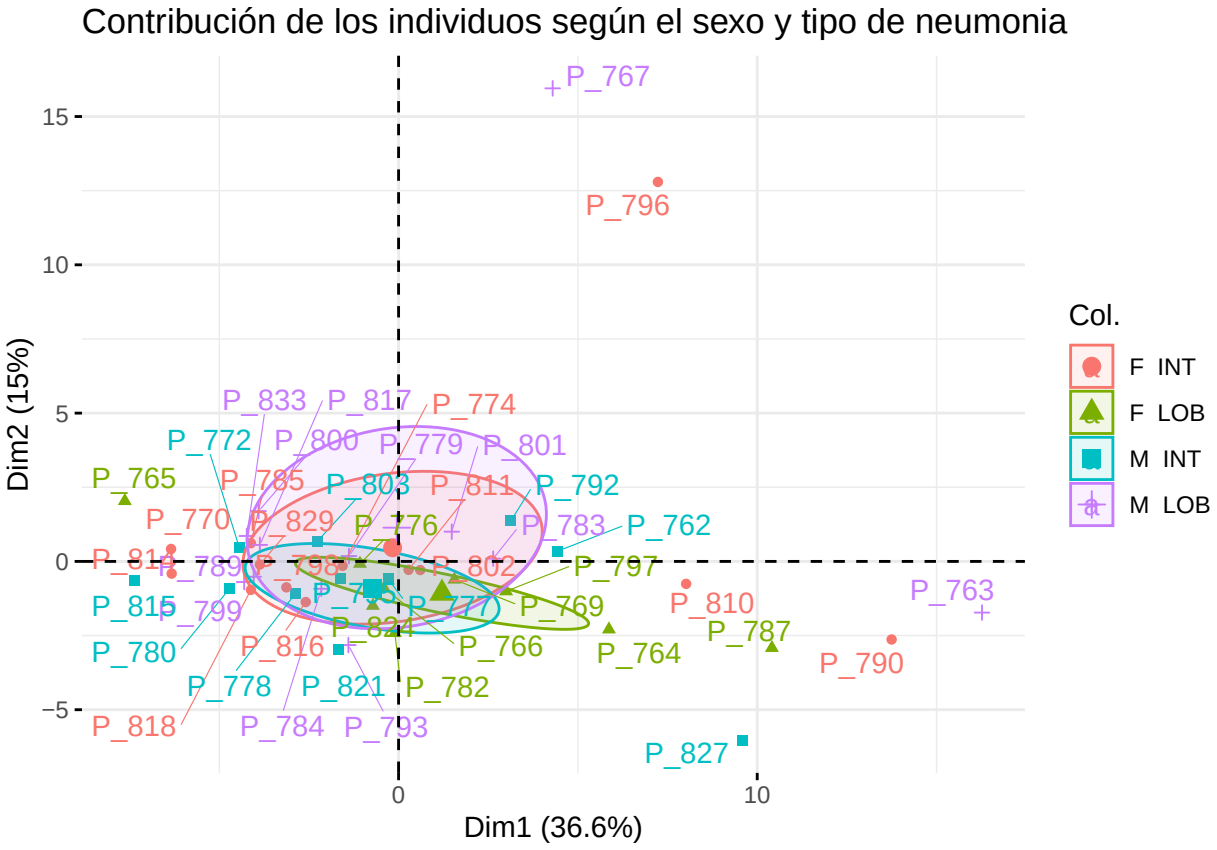


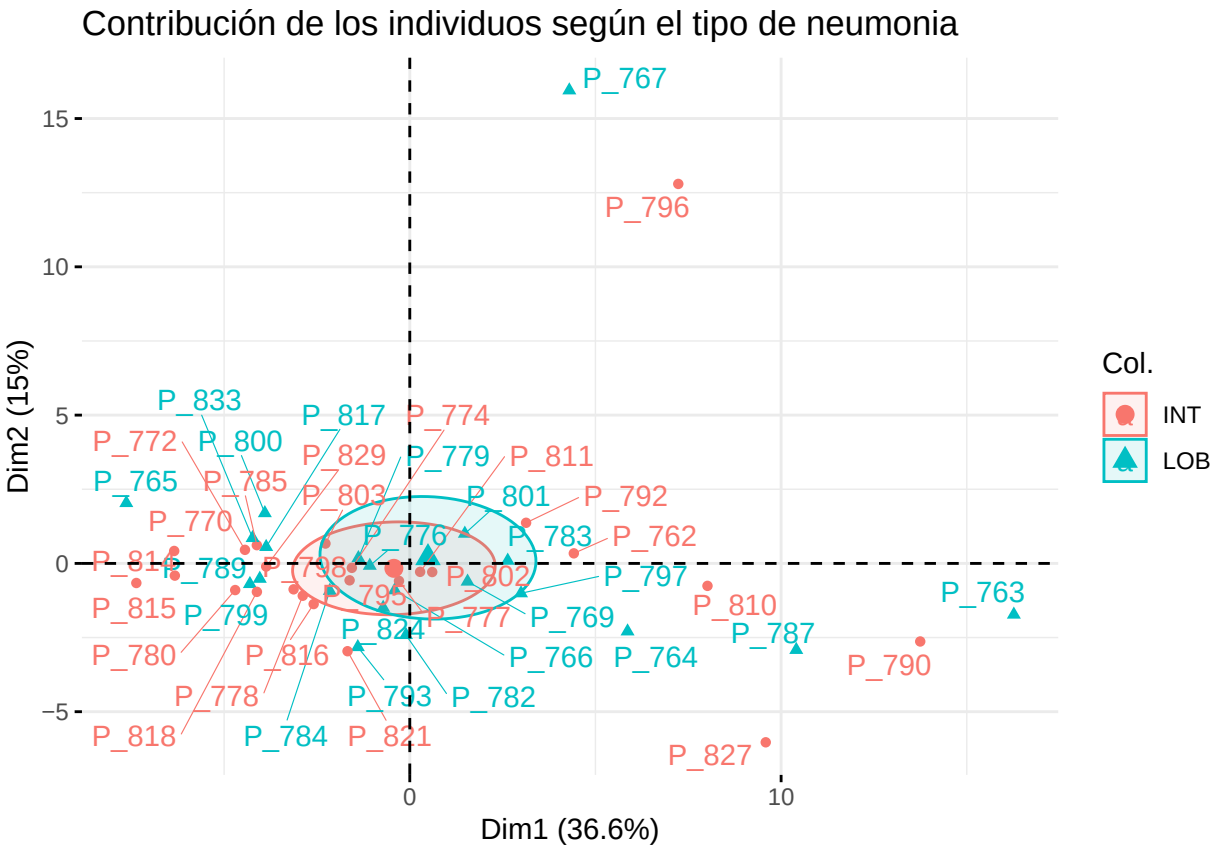
El scree plot nos muestra que entre las primeras 3 componentes se explica la mayor parte de la variabilidad, especialmente en la primera componente, por lo que se representarán las 2 primeras componentes en los siguientes gráficos, para facilitar la visualización.

En el siguiente gráfico se visualizan las 20 variables que más han contribuido a las 2 primeras componentes.



Y, a continuación, se visualizan las contribuciones de los individuos del experimento a los dos primeros componentes, por un lado separados por sexo y tipo de neumonia, y por el otro solamente por el tipo de neumonia.


Contribución de los individuos según el sexo y tipo de neumonia
Dim2 (15%)
Dim1 (36.6%)
Col.
F INT
F LOB
M INT
M LOB
P_767
P_796
P_833
P_817
P_774
P_772
P_800
P_779
P_801
P_765
P_785
P_803
P_811
P_792
P_770
P_829
P_776
P_783
P_762
P_797
P_763
P_815
P_799
P_769
P_810
P_787
P_824
P_766
P_764
P_790
P_780
P_816
P_778
P_821
P_782
P_818
P_784
P_793
P_827


Contribución de los individuos según el tipo de neumonia
Dim2 (15%)
Dim1 (36.6%)
15
10
5
0
−5
0
10
P_767
P_796
P_833
P_817
P_774
P_772
P_800
P_829
P_779
P_811
P_765
P_785
P_803
P_801
P_792
P_770
P_776
P_783
P_762
P_814
P_789
P_798
P_802
P_797
P_763
P_815
P_799
P_795
P_777
P_769
P_810
P_787
P_824
P_780
P_816
P_766
P_764
P_790
P_778
P_793
P_782
P_818
P_784
P_821
P_827
Col.
INT
LOB