Clasificador diseasecodes TCGA. Red neuronal con sondas diferencialmente metiladas Tumor vs. Resto

Alberto Joven Álvarez

9 de noviembre, 2022

Índice

1 Lista de librerías empleadas **2**

2 Planteamiento general del trabajo **2**

2.1 Tabla de códigos de tumor del proyecto TCGA 3

2.2 Tabla con los códigos de las muestras 4

2.3 Detalle de las muestras descargadas para el estudio 5

3 Descarga de los datos **5**

3.1 Obtención de los beta_values 5

3.2 Tratamiento de los datos faltantes 6

3.3 Construcción de Data Frame con los fenotipos de las muestras 7

3.4 Granges con las anotaciones de las sondas 10

3.5 Construcción del objeto Summarized Experiment 10

4 Selección de las sondas elegidas por su metilación diferencial Tumor vs resto **10**

4.1 Exclusión de sondas problemáticas de acuerdo a (Price et al. 2013) y (Zhou, Laird, y Shen 2017) 11

4.2 Desglose de las muestras en los grupos train y test 12

4.3 Subsetting del objeto SummarizedExperiment 13

5 Análisis gráfico previo de los beta-values **15**

5.1 Gráfico previo Rtsne 15

5.2 Gráfico previo Rtsne para las muestras de control exclusivamente 16

5.3 Gráfico revisión normalidad de las sondas 17

5.4 Histograma de los valores beta 18

6 El clasificador RED neuronal **19**

6.1 Selección de muestras con código tumor primario y muestras de control 19

6.2 Desglose de las sondas de acuerdo al tipo de sonda 20

6.3 Formulación del modelo 20

6.4 Validación cruzada para determinar la capacidad predictiva del modelo 22

6.5 Predicción con lo valores test utilizando el modelo con epoch 70 seleccionado 23

7 Entrenamiento de la red usando las betas ajustadas con ComBat **28**

7.1 Ajuste de los valores betas del subset train con la función Combat 29

7.2 Entrenamiento del modelo con los nuevos valores ajustados 29

7.3 Evaluación del modelo 30

8 Prueba algoritmo randomforest **30**

8.1 Extracción de los valores betas y fenotipos correspondientes 30

8.2 Algoritmo randomforest 31

Bibliografía **34**

1 Lista de librerías empleadas

```
library(knitr)
library(dplyr)
library(readr)
library(curatedTCGAData)
library(TCGAutils)
library(IlluminaHumanMethylation450kanno.ilmn12.hg19)
library(FDb.InfiniumMethylation.hg19)
library(impute)
library(SummarizedExperiment)
library(GEOquery)
library(caret)
library(FactoMineR)
library(factoextra)
library(Rtsne)
library(ggplot2)
library(keras)
library(sva)
library(randomForest)
```

2 Planteamiento general del trabajo

Se plantea como pregunta inicial de interés biológico:

¿Es posible entrenar una red neuronal para que a partir de los datos de posiciones metiladas de una muestra sea capaz de identificar a qué categoría de tejido pertenece clasificándola en una de las 34 clases de tumores que establece el proyecto The Cancer Genome Atlas o bien considerarla como muestra no tumoral?

Este trabajo se inspira en el artículo (Capper et al. 2018) que propone un clasificador Machine Learning para la determinación del tipo de tumor del SNC a partir de los datos de metilación (betavalues) de las 10.000 sondas del array Illumina 450k que presentan una mayor variabilidad. También ha servido de guía el planteamiento descrito en (Maros et al. 2020) que expone una aplicación general de esa metodología.

En breve síntesis el trabajo ha consistido en:

1. Descarga de los datos: desde el repositorio del Broad Institute GDAC Firehouse se descargaron todos los análisis del proyecto TCGA con datos de metilación obtenidos con sondas Illumina 450k.
2. Se construye un objeto R *Summarized Experiment* a partir de los valores betas de todos los estudios, una vez suprimidas las sondas con más de un 10% de muestras con valor ausente.
3. Se eliminan las sondas previsiblemente problemáticas de acuerdo con los trabajos (Price et al. 2013) y (Zhou, Laird, y Shen 2017).
4. Se desglosan las muestras en los grupos train y test (75% y 25%) utilizando la librería *caret* (Kuhn 2017).
5. Se seleccionan 10.000 sondas elegidas aleatoriamente.
6. Suprimo de ambos grupos, *train* y *test* aquellas muestras no procedentes de tumores primarios.
7. Análisis gráfico previo: se muestran los gráficos de componentes principales y la técnica t-SNE (Stochastic Neighbor Embedding), así como los histogramas las sondas con mayor variabilidad, para ello empleo el grupo de sondas de mayor variabilidad y el grupo de muestras *train*.
8. Se diseña el algoritmo Red Neuronal y se evalúa mediante cross-validation entrenando tan solo con las muestras *train*.
9. Se selecciona el algoritmo de mejor resultado y se evalúa con las muestras *test*.
10. Se entrena un algoritmo random Forest y se evalúa su grado de acierto.

2.1 Tabla de códigos de tumor del proyecto TCGA

El detalle de los códigos de tumor del TCGA usados en este cuaderno es:

```
read.delim("C:/TFM UOC/R/codigos_tumores.txt",  sep="\t") %>% kable()
```

CODIGO	DESCRIPCION
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
CNTL	Controls
COAD	Colon adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma

CODIGO	DESCRIPCION
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma

2.2 Tabla con los códigos de las muestras

El detalle de los códigos de las muestras TCGA usados en este cuaderno es:

```
read.table("C:/TFM UOC/R/codigos_muestras.txt", head=TRUE, sep="\t") %>% kable()
```

Code	Definition	Short.Letter.Code
1	Primary Solid Tumor	TP
2	Recurrent Solid Tumor	TR
3	Primary Blood Derived Cancer - Peripheral Blood	TB
4	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
5	Additional - New Primary	TAP
6	Metastatic	TM
7	Additional Metastatic	TAM
8	Human Tumor Original Cells	THOC
9	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
15	sample type 15	15SH
16	sample type 16	16SH
20	Control Analyte	CELLC
40	Recurrent Blood Derived Cancer - Peripheral Blood	TRB
50	Cell Lines	CELL
60	Primary Xenograft Tissue	XP
61	Cell Line Derived Xenograft Tissue	XCL
99	sample type 99	99SH

En este trabajo usaré solo las muestras de códigos 01 Primary Solid Tumor (TP), 03 Primary Blood Derived

Cancer - Peripheral Blood (TB) y 11 Solid Tissus Normal (NT).

2.3 Detalle de las muestras descargadas para el estudio

```
read_delim("C:/TFM UOC/R/muestras.txt", delim = "\t",
    escape_double = FALSE, col_types = cols(TP = col_integer()),
    trim_ws = TRUE)  %>%  kable()
```

title assay	TP	TB	NT	Total
ACC_Methylation-20160128_assays	80	0	0	80
BLCA_Methylation-20160128_assays	412	0	21	433
BRCA_Methylation_methyl450-20160128_assays	783	0	97	880
CESC_Methylation-20160128_assays	307	0	3	310
CHOL_Methylation-20160128_assays	36	0	9	45
COAD_Methylation_methyl450-20160128_assays	293	0	38	331
DLBC_Methylation-20160128_assays	48	0	0	48
ESCA_Methylation-20160128_assays	185	0	16	201
GBM_Methylation_methyl450-20160128_assays	140	0	1	141
HNSC_Methylation-20160128_assays	528	0	50	578
KICH_Methylation-20160128_assays	66	0	0	66
KIRC_Methylation_methyl450-20160128_assays	319	0	160	479
KIRP_Methylation_methyl450-20160128_assays	275	0	45	320
LAML_Methylation_methyl450-20160128_assays	0	194	0	194
LGG_Methylation-20160128_assays	516	0	0	516
LIHC_Methylation-20160128_assays	377	0	50	427
LUAD_Methylation_methyl450-20160128_assays	458	0	32	490
LUSC_Methylation_methyl450-20160128_assays	370	0	42	412
MESO_Methylation-20160128_assays	87	0	0	87
OV_Methylation_methyl450-20160128_assays	10	0	0	10
PAAD_Methylation-20160128_assays	184	0	10	194
PCPG_Methylation-20160128_assays	179	0	3	182
PRAD_Methylation-20160128_assays	498	0	50	548
READ_Methylation_methyl450-20160128_assays	98	0	7	105
SARC_Methylation-20160128_assays	261	0	4	265
SKCM_Methylation-20160128_assays	105	0	2	107
STAD_Methylation_methyl450-20160128	395	0	2	397
TGCT_Methylation-20160128_assays	134	0	0	134
THCA_Methylation-20160128_assays	503	0	56	559
THYM_Methylation-20160128_assays	124	0	2	126
UCEC_Methylation_methyl450-20160128_assays	431	0	34	465
UCS_Methylation-20160128_assays	57	0	0	57
UVM_Methylation-20160128_assays	80	0	0	80
Total	8339	194	734	9267

3 Descarga de los datos

3.1 Obtención de los beta_values

El código empleado para la descarga de los datos fue:

```
estudios <- curatedTCGAData(
  diseaseCode = c("ACC", "BLCA"), assays = "Methyl*", version = "1.1.38", dry.run = FALSE
)
```

Sucesivamente se descargaron los 33 códigos de tumor que se han detallado anteriormente (en el ejemplo de código solo estan los códigos **ACC** y **BLCA**).

Con el código siguiente se extrajo uno a uno la información (objetos) contenida en las descargas anteriores. Este código se ejecuta para cada uno de los 33 assays expuestos en la primera columna de la tabla anterior:

```
matriz1 <- assays(estudios)
matriz1 <- matriz1[["BLCA_Methylation-20160128"]]
matriz1 <- as.matrix(matriz1)

matriz2 <- assays(estudios)
matriz2 <- matriz2[["ACC_Methylation-20160128"]]
matriz2 <- as.matrix(matriz2)
```

El paso siguiente fue unir sucesivamente una a una (en otro caso daba error) las matrices con los valores betas con *cbind*:

```
matriz <- cbind(matriz1, matriz2)
rm(matriz1, matriz2)

matriz <- cbind(matriz, matriz3)
rm(matriz3)
```

3.2 Tratamiento de los datos faltantes

Tratamiento de los valores faltantes **NAs**: se eliminaron las sondas con más del 10% de las observaciones NAs, quedan 395319:

```
matriz2 <- matriz[nas < 9707 * 0.1, ]
```

El resto de valors faltantes se imputaron con la función *impute* de la librería del mismo nombre, que utiliza el nearest neighbor averaging para su cálculo. Se divide la matriz de datos en 9 tramos para posibilitar el cálculo y, una vez calculada la estimación de lo valores faltantes, se vuelven a unir. El código utilizado fue:

```
library(impute)

matriz3 <- matriz2[ , 1:1000]
matriz4 <- matriz2[ , 1001:2000]
matriz5 <- matriz2[ , 2001:3000]
matriz6 <- matriz2[ , 3001:4000]
matriz7 <- matriz2[ , 4001:5000]
matriz8 <- matriz2[ , 5001:6000]
matriz9 <- matriz2[ , 6001:7000]
matriz10 <- matriz2[ , 7001:8000]
matriz11 <- matriz2[ , 8001:9000]
matriz12 <- matriz2[ , 9001:9707]

rm(matriz2)
```

```
matriz33 <- impute.knn(matriz3)
matriz44 <- impute.knn(matriz4)
matriz55 <- impute.knn(matriz5)
matriz66 <- impute.knn(matriz6)
matriz77 <- impute.knn(matriz7)
matriz88 <- impute.knn(matriz8)
matriz99 <- impute.knn(matriz9)
matriz1010 <- impute.knn(matriz10)
matriz1111 <- impute.knn(matriz11)
matriz1212 <- impute.knn(matriz12)


matriz_sna <- cbind(matriz33$data, matriz44$data, matriz55$data, matriz66$data,
                    matriz77$data, matriz88$data,
                    matriz99$data, matriz1010$data, matriz1111$data, matriz1212$data)
```

3.3 Construcción de Data Frame con los fenotipos de las muestras

El Data.Frame con los fenotipos de las muestras lo construyo manualmente. Los campos a incorporar son:

1. Nombre de las muestras: el bar-code que identifica cada muestra en la matriz con los beta-values.
2. Bar-code: El bar-code de identificación de las muestras.
3. El identificador de individuo (los 12 primeros caracteres del bar-code).
4. Categoría: es el tipo de muestra: 01 Tumor primario, 11 Tejido normal etc...
5. type: es la descripción del tumor extraída de los datos de fenotipos de cada ensayo: Histological Type
6. disease_code: tipo de tumor extraído de lkos datos de fenotipos de cada ensayo.
7. assay: identifica el ensayo.
8. label: si la muestra es tumoral, figura el identificador del ensayo. Si la muestra no es tumoral, código de muestra "11" figura la categoría "Control".

```
# Se carga la matriz con todos los datos betas de todos los ensayos descargados:
# sus columnas están rotuladas con los  bar-codes.

load("G:/TFM UOC/datos/matriz.Rda")

# Se cargan los objetos MultiassayExperiment descargados anteriormente

load(file="G:/TFM UOC/estudios/estudios_1")
load(file="G:/TFM UOC/estudios/estudios_2")
load(file="G:/TFM UOC/estudios/estudios_3")
load(file="G:/TFM UOC/estudios/estudios_4")
load(file="G:/TFM UOC/estudios/estudios_5")
load(file="G:/TFM UOC/estudios/estudios_6")
load(file="G:/TFM UOC/estudios/estudios_7")
load(file="G:/TFM UOC/estudios/estudios_8")
load(file="G:/TFM UOC/estudios/estudios_9")
load(file="G:/TFM UOC/estudios/estudios_10")

# Los bar-codes y los 12 primeros caracteres de los mismos
# forman los dos primeros campos.

etiqueta <- data.frame(bar_code = colnames(matriz),
```

```
                        sujeto=substring(colnames(matriz), 1, 12))
etiqueta$categoria <- substring(colnames(matriz), 14,15) %>% as.factor()

# para añadir el histological_type y el disease_code

fenotipos1 <- colData(estudios)
fenotipos2 <- colData(estudios_2)
fenotipos3 <- colData(estudios_3)
fenotipos4 <- colData(estudios_4)
fenotipos5 <- colData(estudios_5)
fenotipos6 <- colData(estudios_6)
fenotipos7 <- colData(estudios_7)
fenotipos8 <- colData(estudios_8)
fenotipos9 <- colData(estudios_9)
fenotipos10 <- colData(estudios_10)
tipos1 <- data.frame(sujeto = rownames(fenotipos1),
                     type = fenotipos1$histological_type,
                     disease_code=fenotipos1$admin.disease_code)
tipos2 <- data.frame(sujeto = rownames(fenotipos2),
                     type = fenotipos2$histological_type,
                     disease_code=fenotipos2$admin.disease_code)
tipos3 <- data.frame(sujeto = rownames(fenotipos3),
                     type = fenotipos3$histological_type,
                     disease_code=fenotipos3$admin.disease_code)
tipos4 <- data.frame(sujeto = rownames(fenotipos4),
                     type = fenotipos4$histological_type,
                     disease_code=fenotipos4$admin.disease_code)
tipos5 <- data.frame(sujeto = rownames(fenotipos5),
                     type = fenotipos5$histological_type,
                     disease_code=fenotipos5$admin.disease_code)
tipos6 <- data.frame(sujeto = rownames(fenotipos6),
                     type = fenotipos6$histological_type,
                     disease_code=fenotipos6$admin.disease_code)
tipos7 <- data.frame(sujeto = rownames(fenotipos7),
                     type = fenotipos7$histological_type,
                     disease_code=fenotipos7$admin.disease_code)
tipos8 <- data.frame(sujeto = rownames(fenotipos8),
                     type = fenotipos8$histological_type,
                     disease_code=fenotipos8$admin.disease_code)
tipos9 <- data.frame(sujeto = rownames(fenotipos9),
                     type = fenotipos9$histological_type,
                     disease_code=fenotipos9$admin.disease_code)
tipos10 <- data.frame(sujeto = rownames(fenotipos10),
                      type = fenotipos10$histological_type,
                      disease_code=fenotipos10$admin.disease_code)

tipos <- rbind(tipos1, tipos2, tipos3, tipos4, tipos5, tipos6,
               tipos7, tipos8, tipos9, tipos10)

etiqueta <- merge(etiqueta, tipos, by="sujeto", all.x=TRUE)

# Para añadir el estudio assay
```

```
maps1 <- as.data.frame(sampleMap(estudios))
maps2 <- as.data.frame(sampleMap(estudios_2))
maps3 <- as.data.frame(sampleMap(estudios_3))
maps4 <- as.data.frame(sampleMap(estudios_4))
maps5 <- as.data.frame(sampleMap(estudios_5))
maps6 <- as.data.frame(sampleMap(estudios_6))
maps7 <- as.data.frame(sampleMap(estudios_7))
maps8 <- as.data.frame(sampleMap(estudios_8))
maps9 <- as.data.frame(sampleMap(estudios_9))
maps10 <- as.data.frame(sampleMap(estudios_10))

tipos_2 <- rbind(maps1, maps2, maps3, maps4, maps5, maps6,
                 maps7, maps8, maps9, maps10)
tipos_2$assay <- as.character(tipos_2$assay)
ensayos <- strsplit(tipos_2$assay, split="_")
ensayos <- lapply(ensayos, "[", 1) %>% unlist()
tipos_2$assay <- ensayos

# elimino duplicados en los indicadores de muestra
tipos_2 <- tipos_2[!duplicated(tipos_2$colname), c(1,3) ]

etiqueta <- merge(etiqueta, tipos_2, by.x="bar_code", by.y= "colname", all.x=TRUE)
rownames(etiqueta) <- etiqueta$bar_code

etiqueta$label <- etiqueta$assay
etiqueta$label[etiqueta$categoria == "11"] <- "Control"

# muestras puestas en el mismo orden que la matriz de expresión
etiqueta <- etiqueta[colnames(matriz), ]

table(etiqueta$assay, useNA="always")
table(etiqueta$label, useNA="always")

sum(colnames(matriz) != rownames(etiqueta))

sujetos <- unique(etiqueta$sujeto)
table(table(etiqueta$sujeto))
```

El data.frame etiqueta, al final tiene el siguiente aspecto (se traspone para mejor visualización:

```
etiqueta <- read.table("C:/TFM UOC/R/etiqueta.txt")

kable(t(etiqueta[1:2, ] ))
```

	TCGA-2F-A9KO-01A-11D-A38H-05	TCGA-2F-A9KP-01A-11D-A38H-05
bar_code	TCGA-2F-A9KO-01A-11D-A38H-05	TCGA-2F-A9KP-01A-11D-A38H-05
sujeto	TCGA-2F-A9KO	TCGA-2F-A9KP
categoria	1	1
type	muscle invasive urothelial carcinoma (pt2 or above)	muscle invasive urothelial carcinoma (pt2 or above)
disease_code	blca	blca
assay	BLCA	BLCA

	TCGA-2F-A9KO-01A-11D-A38H-05	TCGA-2F-A9KP-01A-11D-A38H-05
label	BLCA	BLCA

3.4 Granges con las anotaciones de las sondas

Se obtienen de las anotaciones de Illumina.

```
sondas <- get450k()
sondas <- sondas[rownames(matriz_sna)]
```

3.5 Construcción del objeto Summarized Experiment

```
data <- SummarizedExperiment::SummarizedExperiment(
  assays=S4Vectors::SimpleList(counts=matriz_sna),
  rowRanges = sondas,
  colData = etiqueta
)
```

4 Selección de las sondas elegidas por su metilación diferencial Tumor vs resto

```
############################################################
# Carga de los ficheros con los datos de medias betas
############################################################

data("diseaseCodes", package="TCGAutils")
codigos <- diseaseCodes$Study.Abbreviation
codigos <- codigos[-c(6,10,17,23)]
codigos <- c(codigos, "Control")

comparativas <- vector("list", length=length(codigos))

ficheros <- c()
for (i in 1:length(codigos)){
  ficheros[i] <- paste0("G:/TFM UOC/datos/Clasificador_tumor_vs_resto/", codigos[i], "_vs_Resto.Rda")
                      }

for (i in 1:length(codigos)) {
  load(file.path(ficheros[i]))
  comparativas[[i]] <- save
  print(codigos[i])
}
```

```
## [1] "ACC"
## [1] "BLCA"
## [1] "BRCA"
```

```
## [1] "CESC"
## [1] "CHOL"
## [1] "COAD"
## [1] "DLBC"
## [1] "ESCA"
## [1] "GBM"
## [1] "HNSC"
## [1] "KICH"
## [1] "KIRC"
## [1] "KIRP"
## [1] "LAML"
## [1] "LGG"
## [1] "LIHC"
## [1] "LUAD"
## [1] "LUSC"
## [1] "MESO"
## [1] "OV"
## [1] "PAAD"
## [1] "PCPG"
## [1] "PRAD"
## [1] "READ"
## [1] "SARC"
## [1] "SKCM"
## [1] "STAD"
## [1] "TGCT"
## [1] "THCA"
## [1] "THYM"
## [1] "UCEC"
## [1] "UCS"
## [1] "UVM"
## [1] "Control"
```

```
names(comparativas) <- codigos
```

4.1 Exclusión de sondas problemáticas de acuerdo a (Price et al. 2013) y (Zhou, Laird, y Shen 2017)

4.1.1 Descarga de las anotaciones adicionales de las sondas

Se descargan las anotaciones adicionales de las sondas desde la documentación adicional de (Price et al. 2013)

```
elist <- getGEO("GSE42409")
GSE42409 <- elist[[1]] %>% featureData()
```

4.1.2 Supresión de sondas problemáticas

Se excluirán las sondas identificadas con:

1. Aquellas cuya denominación comienza con rs o ch: identificadas como SNP por las anotaciones de Illumina o que no están mapeadas al genoma.

2. Las que tienen en el fichero de anotación adicional de (Price et al. 2013) el campo `Target CpG SNP` no vacío.
3. Las que tiene más de una localización in silico de acuerdo al fichero de (Price et al. 2013).
4. Las que apuntan a ADN repetitivo según el fichero de (Price et al. 2013).

```
nombres_sondas <- row.names(comparativas[[1]])

sondas_en_GSE <- GSE42409[row.names(GSE42409) %in% nombres_sondas,  ]

sondas_excluir_1 <- nombres_sondas[substring(nombres_sondas, 1 , 2 ) %in% c("ch", "rs")]
sondas_excluir_2 <- nombres_sondas[sondas_en_GSE$`Target CpG SNP` != ""]
sondas_excluir_3 <- nombres_sondas[sondas_en_GSE$AlleleA_Hits != 1]
sondas_excluir_4 <- nombres_sondas[sondas_en_GSE$AlleleB_Hits != 0]
sondas_excluir_5 <- nombres_sondas[sondas_en_GSE$n_bp_repetitive != 0]

s1 <- union(sondas_excluir_1, sondas_excluir_2)
s1 <- union(sondas_excluir_3, s1)
s1 <- union(sondas_excluir_4, s1)
s1 <- union(sondas_excluir_5, s1)
```

4.1.3 Selección de sondas con mayores diferencias en las medias betas

```
sondas_s <- c()
for (i in 1:length(codigos)){
  estudio <- comparativas[[i]]
  estudio <- estudio[ !(row.names(estudio) %in% s1), ]
  orden <- order(estudio[ , 3])
  estudio_o_1 <- estudio[orden, ]
  orden <- order(estudio[ , 3], decreasing=TRUE)
  estudio_o_2 <- estudio[orden, ]
  sondas <- c(row.names(estudio_o_1)[1:100], row.names(estudio_o_2)[1:100])

  sondas_s <- c(sondas_s, sondas)
}

sondas <- unique(sondas_s)
```

4.2 Desglose de las muestras en los grupos train y test

Se desglosa el objeto summarized experiment que contiene todas las muestras, en dos, un grupo train con el 75 % de las muestras y un grupo test con el resto. Se usa la función `createDataPartition` de la librería `caret`.

```
load("G:/TFM UOC/datos/Clasificador_sondas_aleatorias/data_sondas_dep.Rda")

set.seed(321)
etiqueta <- data_sondas_dep$label

in_train <- createDataPartition(etiqueta, p=0.75, list=FALSE) %>%  as.vector()
```

```
train <- data_sondas_dep[ , in_train]
test <- data_sondas_dep[ , -in_train]
```

4.3 Subsetting del objeto SummarizedExperiment

```
#############################################################
# Subseting en el summarizedExperiment train
#############################################################

load("G:/TFM UOC/datos/Clasificador_variabilidad/data_sondas_dep_train.Rda")
train

train <- train[sondas, ]
train

load("G:/TFM UOC/datos/Clasificador_variabilidad/data_sondas_dep_test.Rda")

test <- test[sondas, ]
test

save(train, file= "G:/TFM UOC/datos/Clasificador_tumor_vs_resto/train.Rda")
save(test, file= "G:/TFM UOC/datos/Clasificador_tumor_vs_resto/test.Rda")
```

```
load("G:/TFM UOC/datos/Clasificador_tumor_vs_resto/train.Rda")
train
```

```
## class: RangedSummarizedExperiment
## dim: 6014 7292
## metadata(0):
## assays(1): counts
## rownames(6014): cg27007717 cg09234297 ... cg17138769 cg01395254
## rowData names(10): addressA addressB ... probeEnd probeTarget
## colnames(7292): TCGA-2F-A9KP-01A-11D-A38H-05
##   TCGA-2F-A9KQ-01A-11D-A38H-05 ... TCGA-ZA-A8F6-01A-23D-A365-05
##   TCGA-ZQ-A9CR-01A-11D-A398-05
## colData names(7): bar_code sujeto ... assay label
```

```
load("G:/TFM UOC/datos/Clasificador_tumor_vs_resto/test.Rda")
test
```

```
## class: RangedSummarizedExperiment
## dim: 6014 2415
## metadata(0):
## assays(1): counts
## rownames(6014): cg27007717 cg09234297 ... cg17138769 cg01395254
## rowData names(10): addressA addressB ... probeEnd probeTarget
## colnames(2415): TCGA-2F-A9KO-01A-11D-A38H-05
##   TCGA-2F-A9KW-01A-11D-A38H-05 ... TCGA-VQ-AA6J-01A-11D-A411-05
##   TCGA-VQ-AA6K-01A-11D-A411-05
## colData names(7): bar_code sujeto ... assay label
```

4.3.1 Tabla de distribución de muestras entre train y test

Se detalla la distribución entre los grupos train y test de las muestras.

```
t1 <- train$label %>% table() %>% as.matrix()
t2 <- test$label %>% table() %>% as.matrix()

df <- data.frame(Train= table(train$label),
                 Test = table(test$label),
                 Total = t1+t2)

df[, c(2,4,5) ] %>% kable()
```

	Train.Freq	Test.Freq	Total
ACC	60	20	80
BLCA	310	103	413
BRCA	591	197	788
CESC	232	77	309
CHOL	27	9	36
COAD	222	73	295
Control	551	183	734
DLBC	36	12	48
ESCA	140	46	186
GBM	115	38	153
HNSC	398	132	530
KICH	50	16	66
KIRC	240	80	320
KIRP	207	69	276
LAML	146	48	194
LGG	398	132	530
LIHC	285	94	379
LUAD	345	115	460
LUSC	278	92	370
MESO	66	21	87
OV	8	2	10
PAAD	139	46	185
PCPG	138	46	184
PRAD	375	124	499
READ	75	24	99
SARC	199	66	265
SKCM	355	118	473
STAD	297	98	395
TGCT	105	34	139
THCA	384	127	511
THYM	93	31	124
UCEC	324	108	432
UCS	43	14	57
UVM	60	20	80

4.3.2 Tabla de ubicaciones de las sondas seleccionadas

Se detalla los cromosomas a los que se mapean las 10.000 sondas seleccionadas.

```
rowRanges(train) %>% seqnames() %>% table()
```

```
## .
##  chr1  chr2  chr3  chr4  chr5  chr6  chr7  chr8  chr9 chr10 chr11 chr12 chr13
##   591   445   319   247   312   434   394   281   103   299   313   339   161
## chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22  chrX  chrY
##   173   149   290   377    81   354   148    51    91    61     1
```

4.3.3 Tabla con la tipología de las sondas seleccionadas

El desglose de las sondas seleccionadas entre Tipo I (Green y Red) y Tipo II (Both)

```
rowRanges(train)$channel %>% table() %>% kable()
```

.	Freq
Both	2609
Grn	1069
Red	2336

4.3.4 Tabla con tipos de targets HIL en las sondas seleccionadas

```
GSE42409$HIL_CpG_class[GSE42409$ID %in% names(train)]  %>%  table() %>%  kable()
```

.	Freq
HC	3000
IC	1280
ICshore	351
LC	1138

5 Análisis gráfico previo de los beta-values

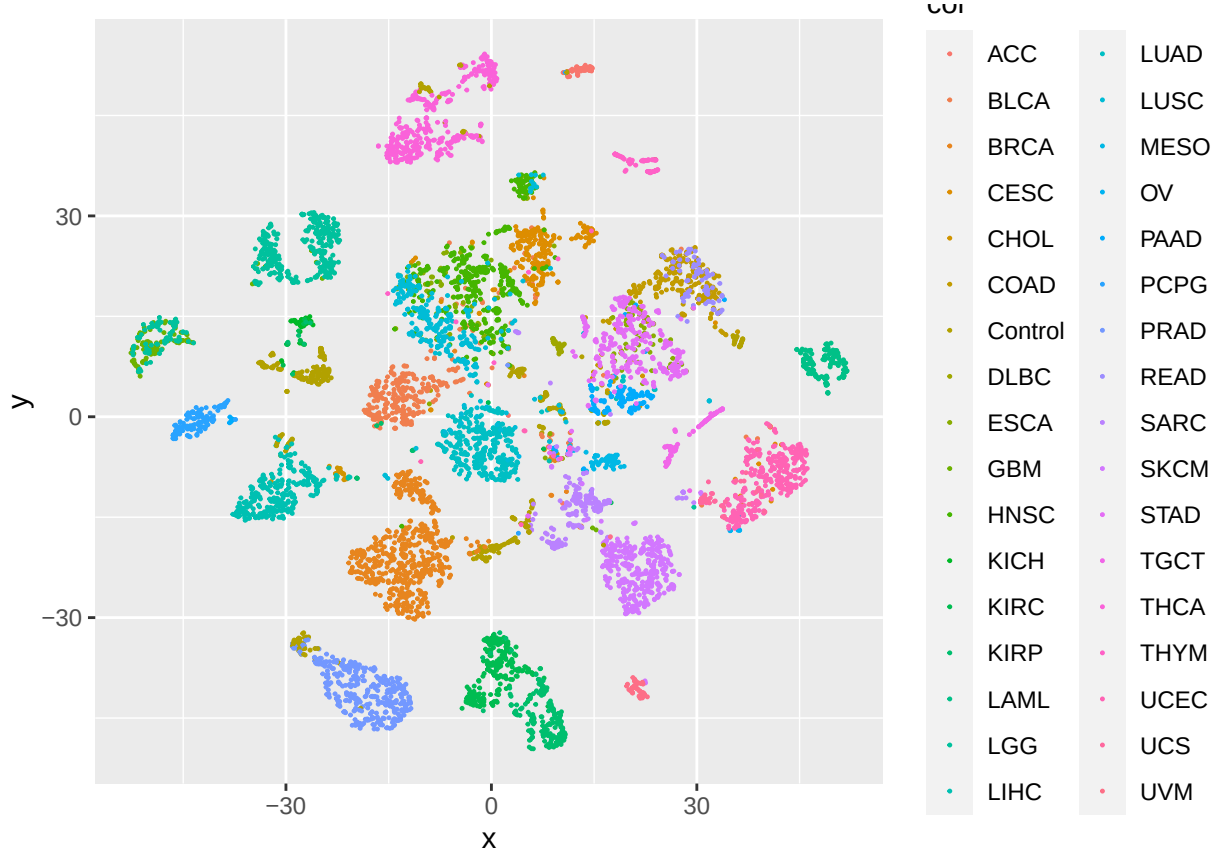
Se utilizan para construir los gráficos tan solo las muestras del grupo train.

```
train_data <- assay(train, "counts") %>%  t()
label <- train$label %>% factor()
```

5.1 Gráfico previo Rtsne

Tan solo utilizaremos los datos train:

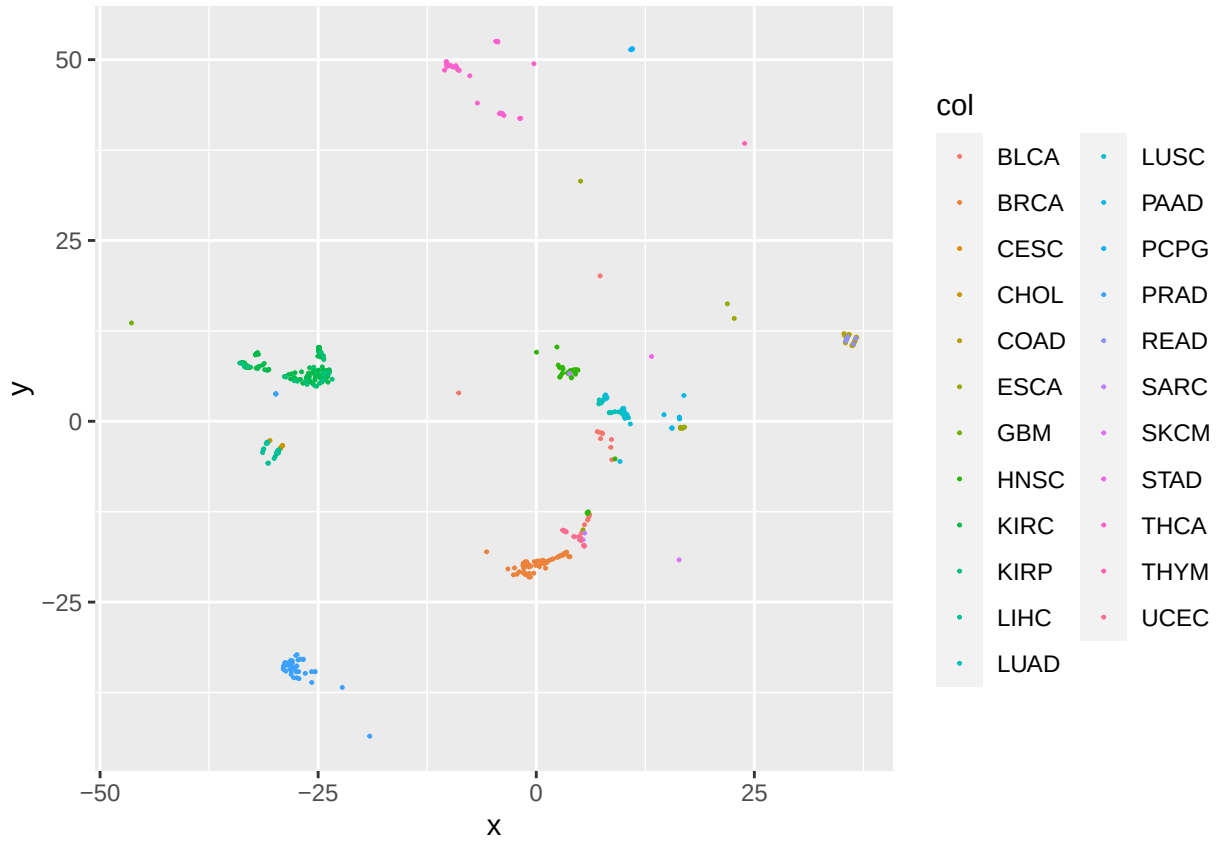
```
sed.seed=123
tsne <- Rtsne(train_data, partial_pca=TRUE, dims=2, perplexity=30, verbose =FALSE, max_iter=1000 )

# Gráfico por patologías
tsne_plot <- data.frame(x = tsne$Y[,1], y = tsne$Y[,2], col = label)
ggplot(tsne_plot) + geom_point(aes(x=x, y=y, color=col), size=0.2)
```



Muchos tumores aparecen claramente diferenciados en el gráfico, sin embargo, otros se presume que va a ser difícil de discriminar con la información contenida en los beta-values de las 10.000 sondas seleccionadas.

5.2 Gráfico previo Rtsne para las muestras de control exclusivamente

```
tsne_plot <- data.frame(x = tsne$Y[train$label=="Control", 1],
                        y = tsne$Y[train$label=="Control", 2], col = train$assay[train$label=="Control"]
ggplot(tsne_plot) + geom_point(aes(x=x, y=y, color=col), size=0.2)
```



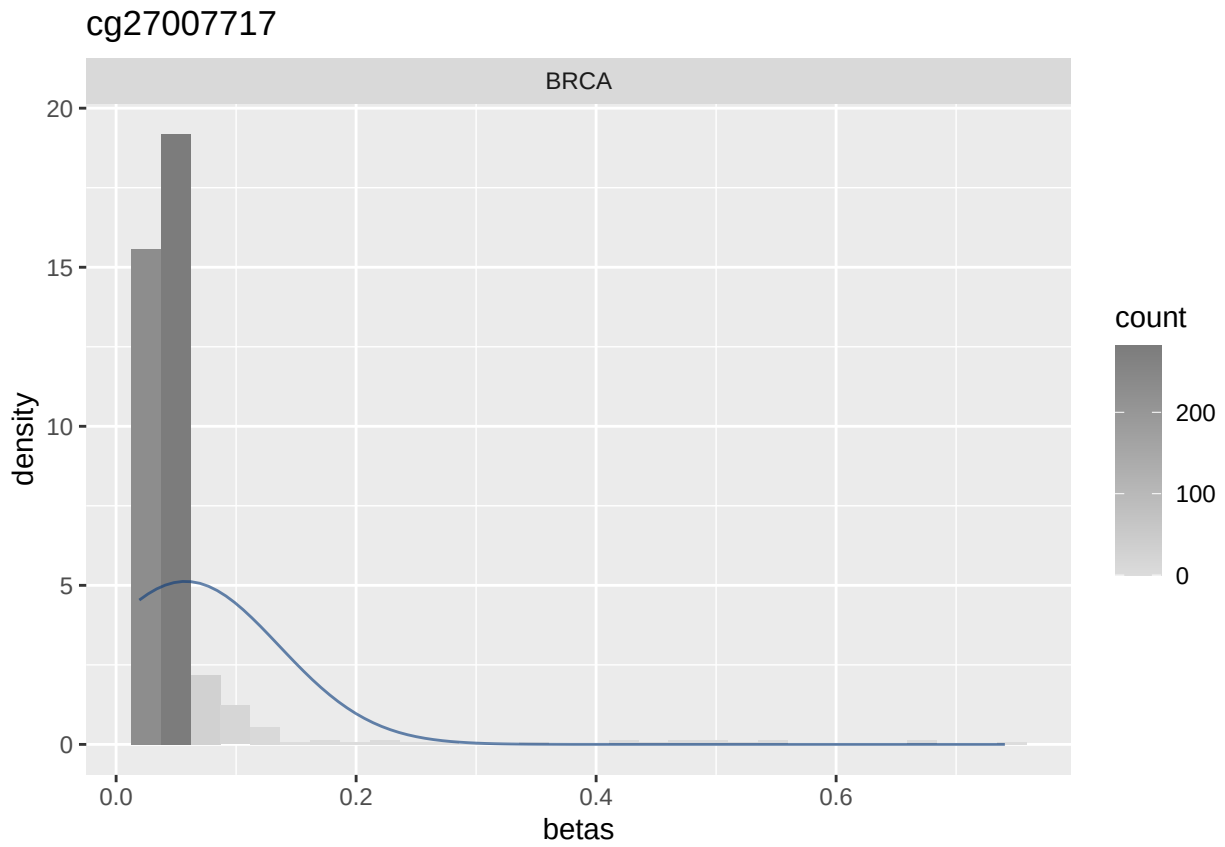
Las muestras de control, al proceder de tejidos dispares presentan un alto grado de dispersión en este análisis de similiritud.

5.3 Gráfico revisión normalidad de las sondas

Se muestra el histograma de los valores betas de una sonda elegida aleatoriamente para las muestras del estudio BRCA

```
label_g <- as.character(label)
l <- label_g[label == "BRCA" ]

df <- data.frame(betas = train_data[ label =="BRCA" , sample(1:dim(train_data)[2], 1)], label = l )

ggplot(data = df , aes(x=betas)) +
  geom_histogram(aes(y=..density.., fill = ..count..)) +
  scale_fill_gradient(low="#DCDCDC", high = "#7C7C7C") +
  stat_function(fun=dnorm, colour="#0C3D7D9F", args=list(mean=mean(df$betas), sd = sd(df$betas))) +
  ggtitle(colnames(train_data)[1]) + facet_grid( . ~ df$label)
```

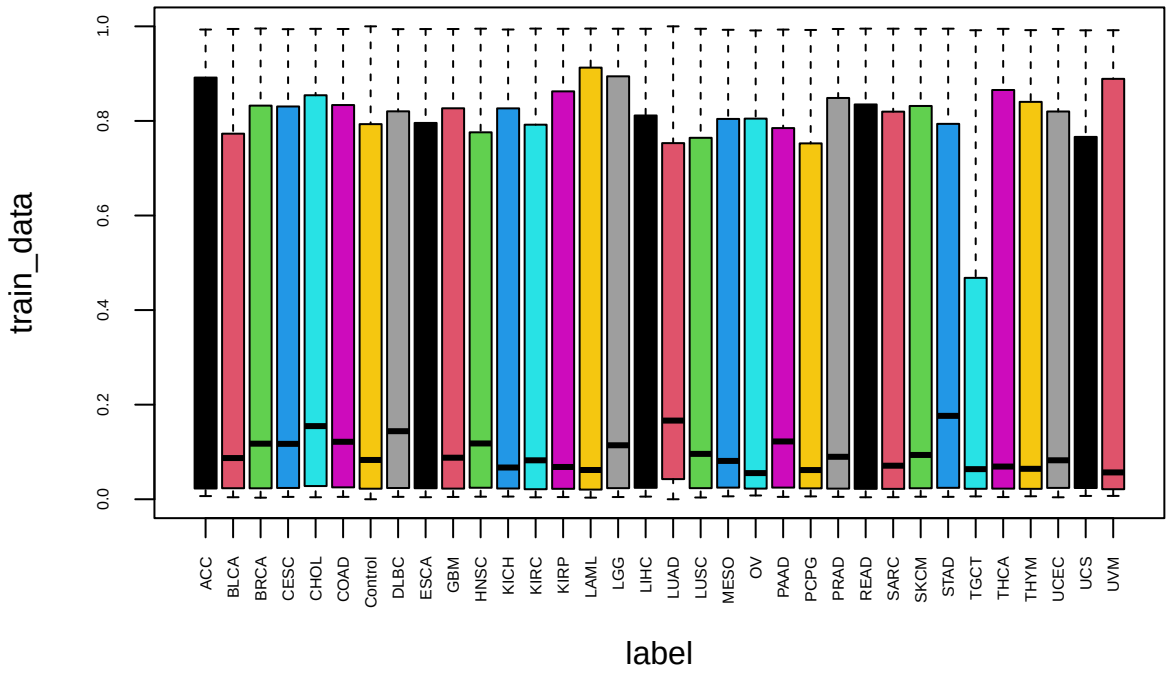
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



No podemos presumir que los valores betas se distribuyan normalmente ni en el conjunto de todas las muestras ni dentro de cada tumor en particular. Esto sería un impedimento en una clasificador que utilizara la regresión logística, sin embargo no aparece como restricción ni en los algoritmos red neuronal ni random forest que son los que aplicaremos seguidamente.

5.4 Histograma de los valores beta

```
boxplot(train_data ~ label, cex.axis=0.5, las=3, col=palette("Polychrome 36"))
```



El histograma de valores betas por tumor (solo consideradas las 10.000 sondas seleccionadas) muestra diferencias significativas entre los diferentes tumores, sin embrago el grupo de control representado en el séptimo lugar no se identifica o discrimina claramente con respecto al resto.

6 El clasificador RED neuronal

6.1 Selección de muestras con código tumor primario y muestras de control

Se seleccionan en los grupos train y test las muestras cuyos códigos son 01 Primary Solid Tumor (TP), 03 Primary Blood Derived Cancer - Peripheral Blood (TB) y 11 Solid Tissus Normal (NT).

```
codigos <- c("01", "03", "11")
train <- train[ , train$categoria %in% codigos]
train
```

```
## class: RangedSummarizedExperiment
## dim: 6014 6969
## metadata(0):
## assays(1): counts
## rownames(6014): cg27007717 cg09234297 ... cg17138769 cg01395254
## rowData names(10): addressA addressB ... probeEnd probeTarget
## colnames(6969): TCGA-2F-A9KP-01A-11D-A38H-05
##   TCGA-2F-A9KQ-01A-11D-A38H-05 ... TCGA-ZA-A8F6-01A-23D-A365-05
##   TCGA-ZQ-A9CR-01A-11D-A398-05
## colData names(7): bar_code sujeto ... assay label
```

```
test <- test[ , test$categoria %in% codigos]
test
```

```
## class: RangedSummarizedExperiment
## dim: 6014 2298
## metadata(0):
## assays(1): counts
## rownames(6014): cg27007717 cg09234297 ... cg17138769 cg01395254
## rowData names(10): addressA addressB ... probeEnd probeTarget
## colnames(2298): TCGA-2F-A9KO-01A-11D-A38H-05
##   TCGA-2F-A9KW-01A-11D-A38H-05 ... TCGA-VQ-AA6J-01A-11D-A411-05
##   TCGA-VQ-AA6K-01A-11D-A411-05
## colData names(7): bar_code sujeto ... assay label
```

6.2 Desglose de las sondas de acuerdo al tipo de sonda

Se desglosan las sondas en los grupos "Green" y "Red", sondas tipo I y "Both" sondas tipo II de acuerdo a las anotaciones de Illumina 450k. También se preparan las etiquetas de las muestras (el tumor al que pertenecen) para poder incorporarse a la red neuronal: se convierte un factor (label) en un array con el procedimiento One-hot encoding.

```
train_data <- assay(train, "counts") %>%  t()
label <- train$label %>% factor()
label_c <- to_categorical(as.integer(label))
```

```
## Loaded Tensorflow version 2.9.2
```

```
tipos_sondas <- rowData(train)$channel %>%  as.factor()
train_green_data <- train_data[ , tipos_sondas=="Grn"]
train_red_data <- train_data[ , tipos_sondas == "Red"]
train_both_data <- train_data[ , tipos_sondas =="Both"]
```

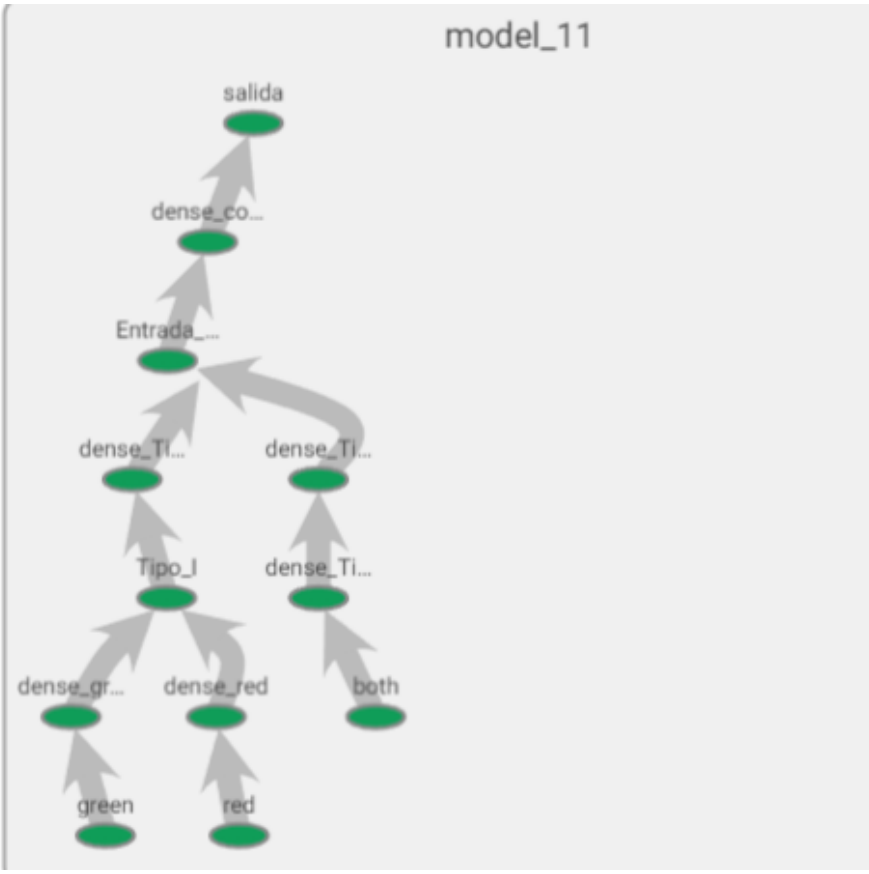
6.3 Formulación del modelo

```
build_model <- function() {
  green <- layer_input(shape= dim(train_green_data)[[2]], name="green")
  red <- layer_input(shape= dim(train_red_data)[[2]], name="red")
  both <- layer_input(shape= dim(train_both_data)[[2]], name="both")

  salida_green <- green %>%
    layer_dense(units=1000, activation="relu", name ="dense_green")

  salida_red <- red %>%
    layer_dense(units=1000, activation="relu", name="dense_red")

  salida_tipo_I <- layer_concatenate(list(salida_green, salida_red),
                                     name = "Tipo_I")

  tipo_I <- salida_tipo_I %>%
```

```
    layer_dense(units = 32, activation = "relu", name="dense_Tipo_I")

  tipo_II <- both %>%
    layer_dense(units=1000, activation="relu", name="dense_Tipo_II_1") %>%
    layer_dense(units=32, activation="relu", name = "dense_Tipo_II_2")

  concatenated <- layer_concatenate(list(tipo_I, tipo_II), name="Entrada_Tipo_I_TipoII")

  salida <- concatenated %>%
    layer_dense(units=1000, activation = "relu", name="dense_conjunta") %>%
    layer_dense(units=35, activation = "softmax", name="salida")


  model <- keras_model(list(green, red, both), salida)

  model %>%  compile(
    optimizer = "rmsprop",
    loss = "categorical_crossentropy",
    metrics = c("accuracy")
  )
}

build_model() %>% summary()
```

```
## Loaded Tensorflow version 2.9.2

## Model: "model"
## __________________________________________________________________________________
##  Layer (type)              Output Shape      Param #  Connected to
## ==================================================================================
##  green (InputLayer)        [(None, 1069)]    0        []
##  red (InputLayer)          [(None, 2336)]    0        []
##  dense_green (Dense)       (None, 1000)      1070000  ['green[0][0]']
##  dense_red (Dense)         (None, 1000)      2337000  ['red[0][0]']
##  both (InputLayer)         [(None, 2609)]    0        []
##  Tipo_I (Concatenate)      (None, 2000)      0        ['dense_green[0][0]',
##                                                       'dense_red[0][0]']
##  dense_Tipo_II_1 (Dense)   (None, 1000)      2610000  ['both[0][0]']
##  dense_Tipo_I (Dense)      (None, 32)        64032    ['Tipo_I[0][0]']
##  dense_Tipo_II_2 (Dense)   (None, 32)        32032    ['dense_Tipo_II_1[0][0]']
##  Entrada_Tipo_I_TipoII (C  (None, 64)        0        ['dense_Tipo_I[0][0]',
##  oncatenate)                                          'dense_Tipo_II_2[0][0]']
##  dense_conjunta (Dense)    (None, 1000)      65000    ['Entrada_Tipo_I_TipoII[0][
##                                                       0]']
##  salida (Dense)            (None, 35)        35035    ['dense_conjunta[0][0]']
## ==================================================================================
## Total params: 6,213,099
## Trainable params: 6,213,099
## Non-trainable params: 0
## __________________________________________________________________________________
```

El grafo de modelo extraído de tensorboard:



6.4 Validación cruzada para determinar la capacidad predictiva del modelo

Se realiza la validación cruzada con 4 particiones, pliegues. El grupo train se divide en 4 bloques, se entrena el modelo con tres de ellos, validándose con el restante. Una vez entrenado y validado el modelo las cuatro veces, el `accuracy` propuesto es la media de los 4 valores obtenidos.

```
k=4
folds <- createFolds(label, k)

num_epoch = 70
all_scores = c()

for (i in 1:k) {
  cat("procesing fold #", i, "\n")
  partial_green_data <- train_green_data[ -folds[[i]] , ]
  partial_red_data <- train_red_data[ -folds[[i]] , ]
  partial_both_data <- train_both_data[ -folds[[i]] , ]
  partial_train_label <- label_c[-folds[[i]] , ]

  val_green_data <- train_green_data[folds[[i]] , ]
  val_red_data <- train_red_data[folds[[i]] , ]
  val_both_data <- train_both_data[folds[[i]] , ]
  val_label <- label_c[folds[[i]],]

  model <- build_model()
```

```
  history <- model %>% fit(list(partial_green_data, partial_red_data, partial_both_data ),
                partial_train_label,
                epoch = num_epoch
              )

  results <- model %>%  evaluate(list(val_green_data, val_red_data, val_both_data), val_label)
  all_scores <- c(all_scores, results[2])
  }
```

```
## procesing fold # 1
## procesing fold # 2
## procesing fold # 3
## procesing fold # 4
```

```
all_scores
```

```
##  accuracy  accuracy  accuracy  accuracy
## 0.8363950 0.9173838 0.9076305 0.9264790
```

```
mean(all_scores)
```

```
## [1] 0.8969721
```

6.5 Predicción con lo valores test utilizando el modelo con epoch 70 seleccionado

Una vez ajustado el parámetro `epoch` al repetir la validación cruzada con varios valores, el modelo elegido final, que se entrena con todos los valores del subset train, se evalúa con los datos de test.

6.5.1 Preparación de los datos del subset test

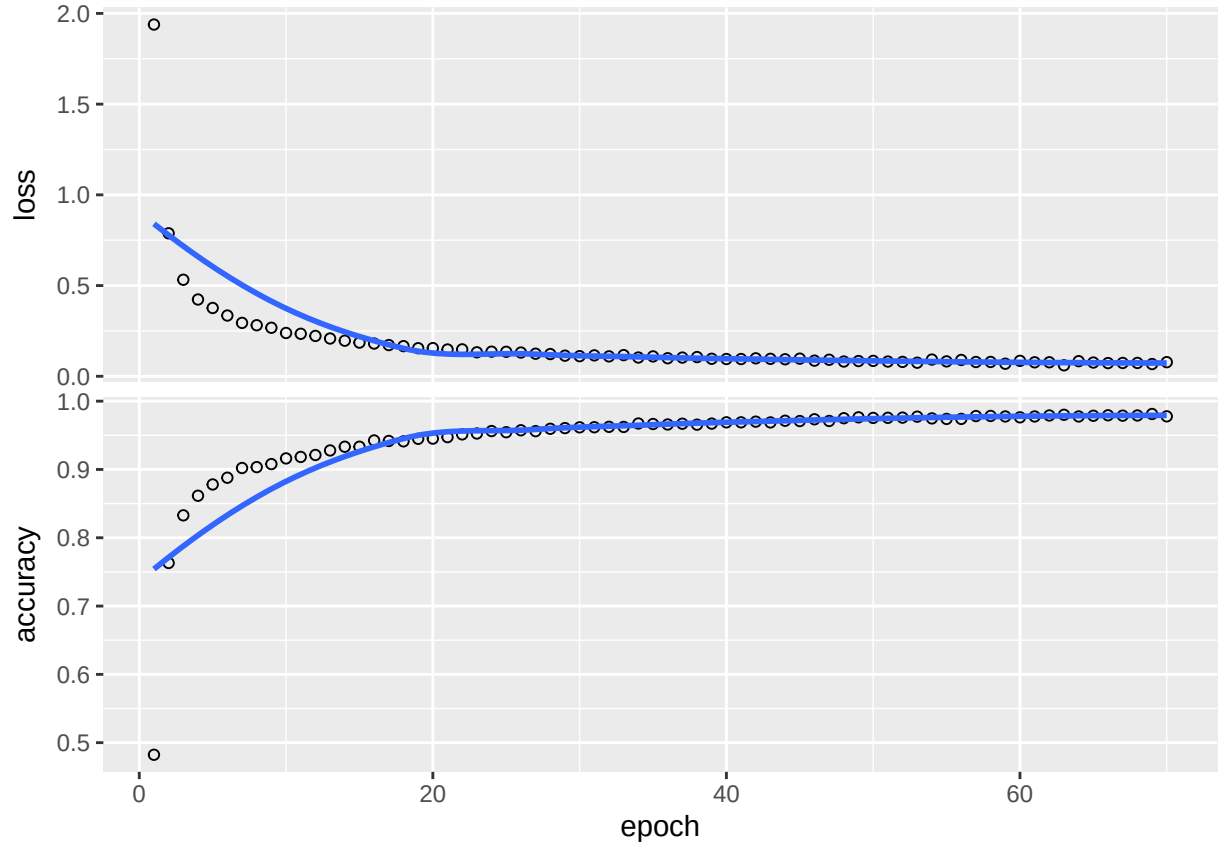
```
betas_test <- assay(test, "counts") %>% t()

test_green_data <- betas_test[ , tipos_sondas=="Grn"]
test_red_data <- betas_test[ , tipos_sondas == "Red"]
test_both_data <- betas_test[ , tipos_sondas =="Both"]

fenotipos_test <- colData(test)$label %>% factor(ordered=TRUE)
test_labels <- to_categorical(as.integer(fenotipos_test))
```

6.5.2 Entrenamiento del modelo con todos los datos train

```
modelo <- build_model()

tensorboard("C:/TFM UOC/R/tf_log_dir")
```

```
## Started TensorBoard at http://127.0.0.1:7502
```

```
callbacks =list(
  callback_tensorboard(
    log_dir="C:/TFM UOC/R/tf_log_dir",
    histogram_freq=1,
  )
)
history <- modelo %>% fit(list(train_green_data, train_red_data, train_both_data ),
                          label_c,
                          epoch = num_epoch,
                          callbacks = callbacks)

plot(history)
```




```
metrics <- modelo %>% evaluate(list(test_green_data, test_red_data, test_both_data), test_labels)
metrics
```

```
##      loss  accuracy
## 1.0394118 0.8816362
```

6.5.3 Matriz de contingencia de los resultados

```r
prediccion <- modelo %>% predict(list(test_green_data, test_red_data, test_both_data)) %>%
  k_argmax() %>%
  as.array() %>% as.integer()

l <- as.list(1:34)
names(l) <- levels(fenotipos_test)
f <- names(l)[prediccion]
lev <- levels(fenotipos_test)
prediccion <- factor(f, levels=lev)

c3 <- confusionMatrix(fenotipos_test, prediccion)
c3
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction ACC BLCA BRCA CESC CHOL COAD Control DLBC ESCA GBM HNSC KICH KIRC
##    ACC      17    0    0    0    0    0       0    0    0   0    0    0    0
##    BLCA      0  101    0    0    0    0       1    0    0   0    0    0    0
##    BRCA      0    0  193    0    0    0       0    0    0   0    0    0    0
##    CESC      0   13    2   31    0    0       6    0    0   0   15    0    0
##    CHOL      0    0    0    0    5    0       0    0    0   0    0    0    0
##    COAD      0    2    0    0    0   34      13    0    0   0    0    0    0
##    Control   0    0    6    0    0    0     154    0    1   0    1    0    0
##    DLBC      0    1    0    0    0    0       0    9    0   0    0    0    0
##    ESCA      0    0    0    0    0    0       0    0   11   0    8    0    0
##    GBM       0    0    0    0    0    0       0    0    0  12    0    0    0
##    HNSC      0    1    0    0    0    0       0    0    0   0  129    0    0
##    KICH      0    0    0    0    0    0       0    0    0   0    0    3    0
##    KIRC      0    0    0    0    0    0       0    0    0   0    0    0   59
##    KIRP      0    1    0    0    0    0       0    0    0   0    0    0    0
##    LAML      0    0    0    0    0    0       0    0    0   0    1    0    0
##    LGG       0    0    0    0    0    0       0    0    0   0    0    0    0
##    LIHC      0    0    0    0    0    0       1    0    0   0    0    0    0
##    LUAD      0    0    0    0    0    0       0    0    0   0    0    0    0
##    LUSC      0    1    0    0    0    0       0    0    0   0    2    0    0
##    MESO      0    0    0    0    0    0       0    0    0   0    0    0    0
##    OV        0    0    0    0    0    0       0    0    0   0    0    0    0
##    PAAD      0    0    0    0    0    0       0    0    0   0    0    0    0
##    PCPG      1    0    0    0    0    0       0    0    0   0    0    0    0
##    PRAD      0    0    0    0    0    0       1    0    0   0    0    0    0
##    READ      0    1    0    0    0    2       3    0    0   0    0    0    0
##    SARC      0    6    0    0    0    0       0    0    0   0    1    0    0
##    SKCM      0    0    1    0    0    0       0    0    0   0    0    0    0
##    STAD      0    0    0    0    0    0       1    1    0   0    0    0    0
##    TGCT      0    0    0    0    0    0       0    0    0   0    0    0    0
##    THCA      0    0    0    0    0    0       1    0    0   0    0    0    0
##    THYM      0    0    0    0    0    0       0    0    1   0    0    0    0
##    UCEC      0    0    2    0    0    0       0    0    0   0    0    0    0
##    UCS       0    0    1    0    0    0       0    0    0   0    0    0    0
##    UVM       0    0    0    0    0    0       0    0    0   0    0    0    0
##            Reference
## Prediction KIRP LAML LGG LIHC LUAD LUSC MESO  OV PAAD PCPG PRAD READ SARC SKCM
```

```
##     ACC        1    0   0    1    0    1    0   0    0    0    0    0    0    0
##     BLCA       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     BRCA       0    0   0    0    0    0    0   0    0    0    0    0    2    0
##     CESC       0    0   0    0    0    2    0   0    1    0    0    0    0    0
##     CHOL       0    0   0    4    0    0    0   0    0    0    0    0    0    0
##     COAD       0    0   0    4    0    0    0   0    0    0    0    3    0    0
##     Control    3    0   0    4    4    1    0   0    2    0    6    0    0    0
##     DLBC       0    0   0    0    1    0    0   0    0    0    0    0    0    0
##     ESCA       0    0   0    1    0    3    0   0    0    0    0    0    0    0
##     GBM        0    0  22    0    0    0    0   0    0    0    0    0    1    0
##     HNSC       0    0   0    0    0    2    0   0    0    0    0    0    0    0
##     KICH      13    0   0    0    0    0    0   0    0    0    0    0    0    0
##     KIRC      19    0   0    0    0    0    0   0    0    0    0    0    2    0
##     KIRP      68    0   0    0    0    0    0   0    0    0    0    0    0    0
##     LAML       0   46   0    1    0    0    0   0    0    0    0    0    0    0
##     LGG        0    0 124    0    0    0    0   0    0    0    0    0    0    0
##     LIHC       0    0   0   92    0    0    0   0    0    0    0    0    0    0
##     LUAD       0    0   0    0  112    1    0   0    0    0    0    0    0    0
##     LUSC       1    0   0    0    1   87    0   0    0    0    0    0    0    0
##     MESO       0    0   0    0    0    1   20   0    0    0    0    0    0    0
##     OV         0    0   0    0    0    0    0   1    0    0    0    0    0    0
##     PAAD       0    0   0    0    0    0    0   0   45    0    0    0    0    0
##     PCPG       0    0   0    0    0    0    0   0    1   42    0    0    0    0
##     PRAD       0    0   0    0    0    0    0   0    0    0  122    0    0    0
##     READ       0    0   0    0    0    0    0   0    0    0    0   12    0    0
##     SARC       1    0   0    0    0    0    0   0    1    0    0    0   57    0
##     SKCM       0    0   0    0    0    0    0   0    0    0    0    0    0   24
##     STAD       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     TGCT       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     THCA       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     THYM       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     UCEC       0    0   0    1    0    0    0   0    0    0    0    0    0    0
##     UCS        0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     UVM        0    0   0    0    0    0    0   0    0    0    0    0    0    0
##              Reference
## Prediction STAD TGCT THCA THYM UCEC UCS UVM
##     ACC       0    0    0    0    0   0   0
##     BLCA      0    0    0    0    0   0   0
##     BRCA      0    0    0    0    0   0   0
##     CESC      1    0    0    0    4   0   0
##     CHOL      0    0    0    0    0   0   0
##     COAD     17    0    0    0    0   0   0
##     Control   1    0    0    0    0   0   0
##     DLBC      1    0    0    0    0   0   0
##     ESCA     23    0    0    0    0   0   0
##     GBM       0    0    0    0    0   0   0
##     HNSC      0    0    0    0    0   0   0
##     KICH      0    0    0    0    0   0   0
##     KIRC      0    0    0    0    0   0   0
##     KIRP      0    0    0    0    0   0   0
##     LAML      0    0    0    0    0   0   0
##     LGG       1    0    0    0    0   0   1
##     LIHC      0    0    0    0    0   0   0
##     LUAD      0    0    0    0    0   0   0
```

```
##    LUSC       0    0    0    0    0   0   0
##    MESO       0    0    0    0    0   0   0
##    OV         0    0    0    0    1   0   0
##    PAAD       1    0    0    0    0   0   0
##    PCPG       0    0    0    0    0   0   0
##    PRAD       0    0    0    0    0   0   0
##    READ       6    0    0    0    0   0   0
##    SARC       0    0    0    0    0   0   0
##    SKCM       0    0    0    0    0   0   0
##    STAD      96    0    0    0    0   0   0
##    TGCT       0   33    0    0    0   0   0
##    THCA       0    0  124    0    0   0   0
##    THYM       0    0    0   30    0   0   0
##    UCEC       0    0    0    0  104   0   0
##    UCS        0    0    0    0    4   9   0
##    UVM        0    0    0    0    0   0  20
##
## Overall Statistics
##
##                Accuracy : 0.8816
##                  95% CI : (0.8677, 0.8946)
##     No Information Rate : 0.0892
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8758
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL
## Sensitivity            0.944444     0.79528     0.94146     1.00000    1.000000
## Specificity            0.998684     0.99954     0.99904     0.98059    0.998256
## Pos Pred Value         0.850000     0.99020     0.98974     0.41333    0.555556
## Neg Pred Value         0.999561     0.98816     0.99429     1.00000    1.000000
## Prevalence             0.007833     0.05527     0.08921     0.01349    0.002176
## Detection Rate         0.007398     0.04395     0.08399     0.01349    0.002176
## Detection Prevalence   0.008703     0.04439     0.08486     0.03264    0.003916
## Balanced Accuracy      0.971564     0.89741     0.97025     0.99030    0.999128
##                      Class: COAD Class: Control Class: DLBC Class: ESCA
## Sensitivity              0.94444         0.85083    0.900000    0.846154
## Specificity              0.98276         0.98630    0.998689    0.984683
## Pos Pred Value           0.46575         0.84153    0.750000    0.239130
## Neg Pred Value           0.99910         0.98723    0.999563    0.999112
## Prevalence               0.01567         0.07876    0.004352    0.005657
## Detection Rate           0.01480         0.06701    0.003916    0.004787
## Detection Prevalence     0.03177         0.07963    0.005222    0.020017
## Balanced Accuracy        0.96360         0.91857    0.949344    0.915418
##                      Class: GBM Class: HNSC Class: KICH Class: KIRC Class: KIRP
## Sensitivity            1.000000     0.82166    1.000000     1.00000     0.64151
## Specificity            0.989939     0.99860    0.994336     0.99062     0.99954
## Pos Pred Value         0.342857     0.97727    0.187500     0.73750     0.98551
## Neg Pred Value         1.000000     0.98707    1.000000     1.00000     0.98295
## Prevalence             0.005222     0.06832    0.001305     0.02567     0.04613
```

```
## Detection Rate         0.005222     0.05614    0.001305     0.02567     0.02959
## Detection Prevalence   0.015231     0.05744    0.006963     0.03481     0.03003
## Balanced Accuracy      0.994969     0.91013    0.997168     0.99531     0.82053
##                      Class: LAML Class: LGG Class: LIHC Class: LUAD Class: LUSC
## Sensitivity              1.00000    0.84932     0.85185     0.94915     0.88776
## Specificity              0.99911    0.99907     0.99954     0.99954     0.99773
## Pos Pred Value           0.95833    0.98413     0.98925     0.99115     0.94565
## Neg Pred Value           1.00000    0.98987     0.99274     0.99725     0.99501
## Prevalence               0.02002    0.06353     0.04700     0.05135     0.04265
## Detection Rate           0.02002    0.05396     0.04003     0.04874     0.03786
## Detection Prevalence     0.02089    0.05483     0.04047     0.04917     0.04003
## Balanced Accuracy        0.99956    0.92419     0.92570     0.97435     0.94274
##                      Class: MESO Class: OV Class: PAAD Class: PCPG Class: PRAD
## Sensitivity             1.000000 1.0000000     0.90000     1.00000     0.95312
## Specificity             0.999561 0.9995646     0.99956     0.99911     0.99954
## Pos Pred Value          0.952381 0.5000000     0.97826     0.95455     0.99187
## Neg Pred Value          1.000000 1.0000000     0.99778     1.00000     0.99724
## Prevalence              0.008703 0.0004352     0.02176     0.01828     0.05570
## Detection Rate          0.008703 0.0004352     0.01958     0.01828     0.05309
## Detection Prevalence    0.009138 0.0008703     0.02002     0.01915     0.05352
## Balanced Accuracy       0.999781 0.9997823     0.94978     0.99956     0.97633
##                      Class: READ Class: SARC Class: SKCM Class: STAD
## Sensitivity              0.800000     0.91935     1.00000     0.65306
## Specificity              0.994744     0.99597     0.99956     0.99907
## Pos Pred Value           0.500000     0.86364     0.96000     0.97959
## Neg Pred Value           0.998681     0.99776     1.00000     0.97682
## Prevalence               0.006527     0.02698     0.01044     0.06397
## Detection Rate           0.005222     0.02480     0.01044     0.04178
## Detection Prevalence     0.010444     0.02872     0.01088     0.04265
## Balanced Accuracy        0.897372     0.95766     0.99978     0.82607
##                      Class: TGCT Class: THCA Class: THYM Class: UCEC Class: UCS
## Sensitivity              1.00000     1.00000     1.00000     0.92035   1.000000
## Specificity              1.00000     0.99954     0.99956     0.99863   0.997816
## Pos Pred Value           1.00000     0.99200     0.96774     0.97196   0.642857
## Neg Pred Value           1.00000     1.00000     1.00000     0.99589   1.000000
## Prevalence               0.01436     0.05396     0.01305     0.04917   0.003916
## Detection Rate           0.01436     0.05396     0.01305     0.04526   0.003916
## Detection Prevalence     0.01436     0.05440     0.01349     0.04656   0.006092
## Balanced Accuracy        1.00000     0.99977     0.99978     0.95949   0.998908
##                      Class: UVM
## Sensitivity            0.952381
## Specificity            1.000000
## Pos Pred Value         1.000000
## Neg Pred Value         0.999561
## Prevalence             0.009138
## Detection Rate         0.008703
## Detection Prevalence   0.008703
## Balanced Accuracy      0.976190
```

7 Entrenamiento de la red usando las betas ajustadas con ComBat

En este apartado ajustamos los valores betas por el posible efecto **batch** de la variable `plate_id`. El ajuste solo se realiza a los valores del subset train, dado que si aspiramos a emplear el algoritmo en muestras ajenas

al proyecto TCGA, éstas no van a estar ajustadas por esa variable, propia delproyecto.

7.1 Ajuste de los valores betas del subset train con la función Combat

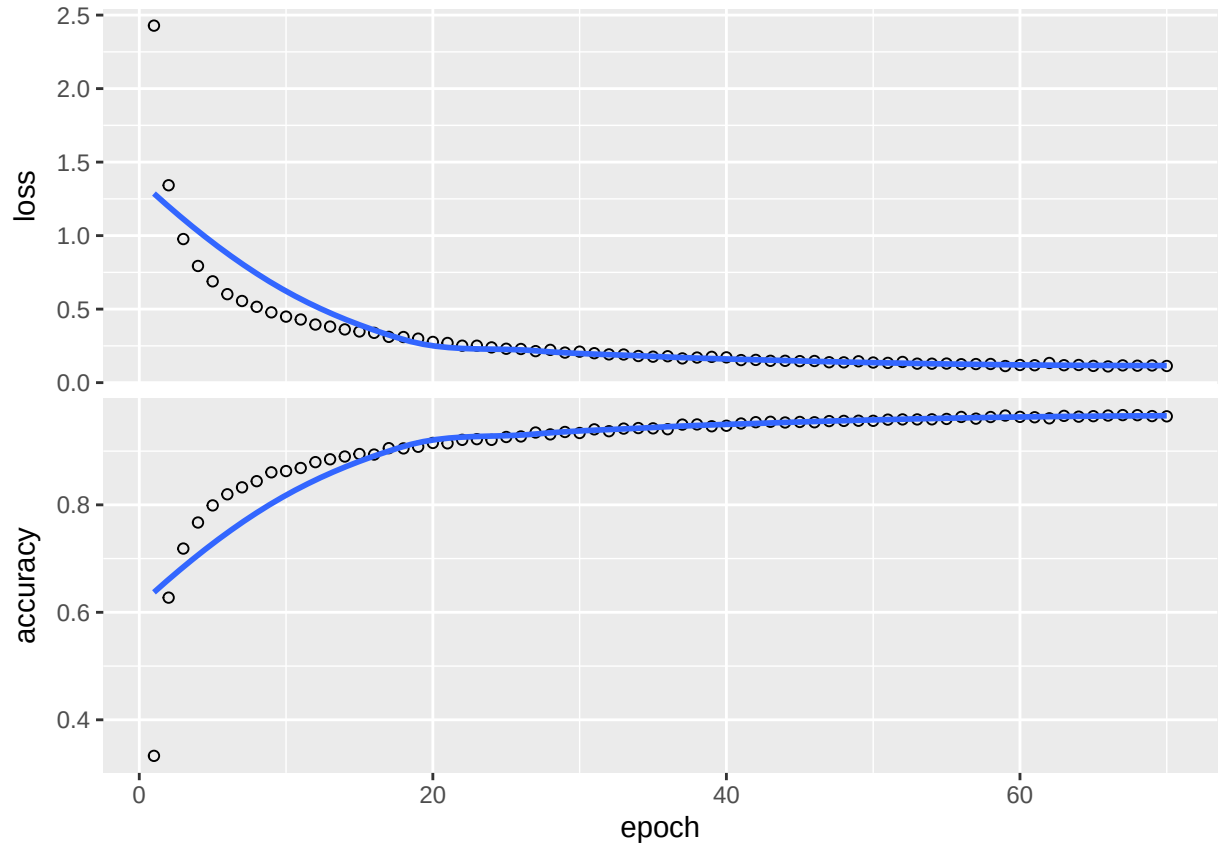
```
edata <- assay(train, "counts")
pdata <- colData(train)

nombres <- assay(train, "counts") %>% colnames()
plate_id <- TCGAbiospec(nombres)
plate_id <- plate_id$plate

train$plate_id <- plate_id
```

```
mod0 <- model.matrix( ~ 1, data = pdata)
mod <- model.matrix( ~ as.factor(label), data = pdata)

combat_edata <- ComBat(dat = edata, batch = plate_id,
                       mod = mod0, mean.only=TRUE, par.prior=TRUE)

assay(train, "counts") <- combat_edata
```

```
train_data <- assay(train, "counts") %>%  t()
label <- train$label %>% factor()
label_c <- to_categorical(as.integer(label))

tipos_sondas <- rowData(train)$channel %>%  as.factor()
train_green_data <- train_data[ , tipos_sondas=="Grn"]
train_red_data <- train_data[ , tipos_sondas == "Red"]
train_both_data <- train_data[ , tipos_sondas =="Both"]
```

7.2 Entrenamiento del modelo con los nuevos valores ajustados

```
modelo <- build_model()

history <- modelo %>% fit(list(train_green_data, train_red_data, train_both_data ),
                          label_c,
                          epoch = num_epoch
                          )

plot(history)
```



7.3 Evaluación del modelo

```
metrics <- modelo %>% evaluate(list(test_green_data, test_red_data, test_both_data), test_labels)
metrics
```

```
##      loss  accuracy
## 0.6606339 0.9129678
```

8 Prueba algoritmo randomforest

8.1 Extracción de los valores betas y fenotipos correspondientes

```
betas_train <- assay(train, "counts") %>% t()
fenotipos_train <- colData(train)$label %>% factor(ordered=TRUE)

betas_test <- assay(test, "counts") %>% t()
fenotipos_test <- colData(test)$label %>% factor(ordered=TRUE)
```

8.2 Algoritmo randomforest

```
be <- as.data.frame(betas_train)
be$label <- fenotipos_train

modelo_rf <- randomForest(label ~., data=be, ntree=100,
                          importance=TRUE)

resultado <- predict(modelo_rf, newdata=betas_test, type="class")

c4 <- confusionMatrix(fenotipos_test, resultado)
c4
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction ACC BLCA BRCA CESC CHOL COAD Control DLBC ESCA GBM HNSC KICH KIRC
##    ACC       8    0    1    0    0    0       2    0    0   0    0    0    0
##    BLCA      0   95    1    0    0    0       2    1    0   0    1    0    0
##    BRCA      0    0  191    0    0    0       0    0    0   0    2    0    0
##    CESC      0    1    0   70    0    0       1    0    0   0    1    0    0
##    CHOL      0    0    0    0    5    0       1    0    0   0    0    0    0
##    COAD      0    0    0    0    0   69       0    0    0   0    0    0    0
##    Control   0    0    3    0    0    1     161    0    0   0    2    0    0
##    DLBC      0    0    0    0    0    0       0   11    0   0    0    0    0
##    ESCA      0    0    1    0    0    0       0    0    4   0   19    0    0
##    GBM       0    0    0    0    0    0       0    0    0   0    0    0    0
##    HNSC      0    1    0    2    0    0       1    0    1   0  121    0    0
##    KICH      0    0    0    0    0    0       5    0    0   0    0    5    1
##    KIRC      0    0    1    0    0    0       1    0    0   0    0    0   71
##    KIRP      0    1    0    0    0    0       2    0    0   0    0    0    1
##    LAML      0    0    7    0    0    0       0    2    0   0    0    0    0
##    LGG       0    0    0    0    0    0       1    0    0   0    0    0    0
##    LIHC      0    0    0    0    0    0       1    0    0   0    0    0    0
##    LUAD      0    0    0    0    0    0       0    0    0   0    1    0    0
##    LUSC      0    1    2    0    0    0       0    0    0   0    9    0    0
##    MESO      0    0    0    0    0    0       0    0    0   0    0    0    0
##    OV        0    0    0    0    0    0       0    0    0   0    0    0    0
##    PAAD      0    0    0    0    0    1       3    0    0   0    0    0    0
##    PCPG      0    0    0    0    0    0       8    0    0   0    0    0    0
##    PRAD      0    0    0    0    0    0       1    0    0   0    0    0    0
##    READ      0    0    0    0    0   24       0    0    0   0    0    0    0
##    SARC      0    0    0    0    0    0       0    0    0   0    0    0    0
##    SKCM      0    0    1    0    0    0       0    0    0   0    0    0    0
##    STAD      0    0    0    0    0    1       1    1    6   0    3    0    0
##    TGCT      0    0    0    0    0    0      10    0    0   0    0    0    0
##    THCA      0    0    0    0    0    0       1    0    0   0    0    0    0
##    THYM      0    0    0    0    0    0       0    0    0   0    3    0    0
##    UCEC      0    0    0    0    0    0       0    0    0   0    0    0    0
##    UCS       0    0    0    0    0    0       0    0    0   0    0    0    0
##    UVM       0    0    0    0    0    0       0    0    0   0    0    0    0
##           Reference
## Prediction KIRP LAML LGG LIHC LUAD LUSC MESO  OV PAAD PCPG PRAD READ SARC SKCM
```

```
##     ACC        0    0   0    0    0    1    0    0    0    0    0    0    8    0
##     BLCA       0    0   0    0    0    1    0    0    0    0    0    0    1    0
##     BRCA       0    0   0    0    0    0    0    0    0    0    0    0    2    0
##     CESC       0    0   0    0    0    0    0    0    0    0    0    0    0    0
##     CHOL       0    0   0    2    0    0    0    0    0    0    0    0    0    0
##     COAD       0    0   0    0    0    0    0    0    0    0    0    0    0    0
##     Control    0    0   0    0    1    0    0    0    1    0    6    0    1    0
##     DLBC       0    0   0    0    0    0    0    0    1    0    0    0    0    0
##     ESCA       0    0   0    0    1    2    0    0    0    0    0    0    0    0
##     GBM        0    0  35    0    0    0    0    0    0    0    0    0    0    0
##     HNSC       0    0   0    0    0    6    0    0    0    0    0    0    0    0
##     KICH       5    0   0    0    0    0    0    0    0    0    0    0    0    0
##     KIRC       5    0   0    0    0    0    0    0    0    0    0    0    2    0
##     KIRP      65    0   0    0    0    0    0    0    0    0    0    0    0    0
##     LAML       0   13   0    0    0    0    0    0    0    0    0    0    4    0
##     LGG        0    0 125    0    0    0    0    0    0    0    0    0    0    0
##     LIHC       0    0   0   91    0    0    0    0    0    0    0    0    1    0
##     LUAD       0    0   0    0  112    0    0    0    0    0    0    0    0    0
##     LUSC       0    0   0    0    0   78    1    0    1    0    0    0    0    0
##     MESO       0    0   0    0    0    0   20    0    0    0    0    0    1    0
##     OV         0    0   0    0    0    0    0    0    0    0    0    0    0    0
##     PAAD       0    0   0    0    0    0    0    0   40    0    0    0    1    0
##     PCPG       0    0   4    0    0    0    0    0    0   28    0    0    4    0
##     PRAD       0    0   0    0    0    0    0    0    0    0  122    0    0    0
##     READ       0    0   0    0    0    0    0    0    0    0    0    0    0    0
##     SARC       0    0   0    0    0    0    0    0    0    0    0    0   66    0
##     SKCM       0    0   0    0    0    0    0    0    0    0    0    0    0   24
##     STAD       0    0   0    0    0    1    0    0    3    0    0    0    0    0
##     TGCT       0    0   1    0    3    1    0    0    0    0    0    0    0    0
##     THCA       0    0   0    0    0    0    0    0    0    0    0    0    0    0
##     THYM       0    0   0    0    0    1    0    0    0    0    0    0    0    0
##     UCEC       0    0   0    0    0    0    0    0    0    0    0    0    0    0
##     UCS        0    0   0    0    0    0    0    0    0    0    0    0    0    0
##     UVM        0    0   0    0    0    0    0    0    0    0    0    0    0   20
##              Reference
## Prediction STAD TGCT THCA THYM UCEC UCS UVM
##     ACC        0    0    0    0    0   0   0
##     BLCA       0    0    0    0    0   0   0
##     BRCA       0    0    0    0    0   0   0
##     CESC       0    0    0    0    2   0   0
##     CHOL       1    0    0    0    0   0   0
##     COAD       4    0    0    0    0   0   0
##     Control    0    0    6    1    0   0   0
##     DLBC       0    0    0    0    0   0   0
##     ESCA      19    0    0    0    0   0   0
##     GBM        0    0    0    0    0   0   0
##     HNSC       0    0    0    0    0   0   0
##     KICH       0    0    0    0    0   0   0
##     KIRC       0    0    0    0    0   0   0
##     KIRP       0    0    0    0    0   0   0
##     LAML       0    0    0   22    0   0   0
##     LGG        0    0    0    0    0   0   0
##     LIHC       0    0    0    0    0   0   0
##     LUAD       0    0    0    0    0   0   0
```

```
##    LUSC       0    0    0    0    0   0   0
##    MESO       0    0    0    0    0   0   0
##    OV         0    0    0    0    2   0   0
##    PAAD       1    0    0    0    0   0   0
##    PCPG       0    0    0    0    0   0   0
##    PRAD       0    0    0    0    0   0   0
##    READ       0    0    0    0    0   0   0
##    SARC       0    0    0    0    0   0   0
##    SKCM       0    0    0    0    0   0   0
##    STAD      82    0    0    0    0   0   0
##    TGCT       0   18    0    0    0   0   0
##    THCA       0    0  124    0    0   0   0
##    THYM       0    0    0   27    0   0   0
##    UCEC       0    0    0    0  107   0   0
##    UCS        0    0    0    0   10   4   0
##    UVM        0    0    0    0    0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.8516
##                  95% CI : (0.8364, 0.8659)
##     No Information Rate : 0.0905
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8441
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL
## Sensitivity            1.000000     0.95960     0.91827     0.97222    1.000000
## Specificity            0.994760     0.99682     0.99809     0.99775    0.998256
## Pos Pred Value         0.400000     0.93137     0.97949     0.93333    0.555556
## Neg Pred Value         1.000000     0.99818     0.99192     0.99910    1.000000
## Prevalence             0.003481     0.04308     0.09051     0.03133    0.002176
## Detection Rate         0.003481     0.04134     0.08312     0.03046    0.002176
## Detection Prevalence   0.008703     0.04439     0.08486     0.03264    0.003916
## Balanced Accuracy      0.997380     0.97821     0.95818     0.98499    0.999128
##                      Class: COAD Class: Control Class: DLBC Class: ESCA
## Sensitivity              0.71875         0.79703    0.733333    0.363636
## Specificity              0.99818         0.98950    0.999562    0.981635
## Pos Pred Value           0.94521         0.87978    0.916667    0.086957
## Neg Pred Value           0.98787         0.98061    0.998250    0.996892
## Prevalence               0.04178         0.08790    0.006527    0.004787
## Detection Rate           0.03003         0.07006    0.004787    0.001741
## Detection Prevalence     0.03177         0.07963    0.005222    0.020017
## Balanced Accuracy        0.85847         0.89327    0.866448    0.672636
##                      Class: GBM Class: HNSC Class: KICH Class: KIRC Class: KIRP
## Sensitivity                  NA     0.74691    1.000000     0.97260     0.86667
## Specificity              0.98477     0.99485    0.995203     0.99596     0.99820
## Pos Pred Value               NA     0.91667    0.312500     0.88750     0.94203
## Neg Pred Value               NA     0.98107    1.000000     0.99910     0.99551
## Prevalence               0.00000     0.07050    0.002176     0.03177     0.03264
```

```
## Detection Rate          0.00000     0.05265    0.002176     0.03090     0.02829
## Detection Prevalence    0.01523     0.05744    0.006963     0.03481     0.03003
## Balanced Accuracy            NA     0.87088    0.997601     0.98428     0.93243
##                       Class: LAML Class: LGG Class: LIHC Class: LUAD Class: LUSC
## Sensitivity              1.000000    0.75758     0.97849     0.95726     0.85714
## Specificity              0.984683    0.99953     0.99909     0.99954     0.99366
## Pos Pred Value           0.270833    0.99206     0.97849     0.99115     0.84783
## Neg Pred Value           1.000000    0.98158     0.99909     0.99771     0.99411
## Prevalence               0.005657    0.07180     0.04047     0.05091     0.03960
## Detection Rate           0.005657    0.05440     0.03960     0.04874     0.03394
## Detection Prevalence     0.020888    0.05483     0.04047     0.04917     0.04003
## Balanced Accuracy        0.992341    0.87855     0.98879     0.97840     0.92540
##                       Class: MESO Class: OV Class: PAAD Class: PCPG Class: PRAD
## Sensitivity              0.952381        NA     0.86957     1.00000     0.95312
## Specificity              0.999561 0.9991297     0.99734     0.99295     0.99954
## Pos Pred Value           0.952381        NA     0.86957     0.63636     0.99187
## Neg Pred Value           0.999561        NA     0.99734     1.00000     0.99724
## Prevalence               0.009138 0.0000000     0.02002     0.01218     0.05570
## Detection Rate           0.008703 0.0000000     0.01741     0.01218     0.05309
## Detection Prevalence     0.009138 0.0008703     0.02002     0.01915     0.05352
## Balanced Accuracy        0.975971        NA     0.93345     0.99648     0.97633
##                       Class: READ Class: SARC Class: SKCM Class: STAD
## Sensitivity                   NA     0.72527     0.54545     0.76636
## Specificity              0.98956     1.00000     0.99956     0.99270
## Pos Pred Value                NA     1.00000     0.96000     0.83673
## Neg Pred Value                NA     0.98880     0.99120     0.98864
## Prevalence               0.00000     0.03960     0.01915     0.04656
## Detection Rate           0.00000     0.02872     0.01044     0.03568
## Detection Prevalence     0.01044     0.02872     0.01088     0.04265
## Balanced Accuracy             NA     0.86264     0.77251     0.87953
##                       Class: TGCT Class: THCA Class: THYM Class: UCEC Class: UCS
## Sensitivity              1.000000     0.95385     0.54000     0.88430   1.000000
## Specificity              0.993421     0.99954     0.99822     1.00000   0.995641
## Pos Pred Value           0.545455     0.99200     0.87097     1.00000   0.285714
## Neg Pred Value           1.000000     0.99724     0.98985     0.99361   1.000000
## Prevalence               0.007833     0.05657     0.02176     0.05265   0.001741
## Detection Rate           0.007833     0.05396     0.01175     0.04656   0.001741
## Detection Prevalence     0.014360     0.05440     0.01349     0.04656   0.006092
## Balanced Accuracy        0.996711     0.97669     0.76911     0.94215   0.997820
##                       Class: UVM
## Sensitivity                   NA
## Specificity             0.991297
## Pos Pred Value                NA
## Neg Pred Value                NA
## Prevalence              0.000000
## Detection Rate          0.000000
## Detection Prevalence    0.008703
## Balanced Accuracy             NA
```

Bibliografía

Capper, David, David TW Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, et al. 2018. «DNA methylation-based classification of central nervous system tumours». *Nature*

555 (7697): 469-74.
Kuhn, Max. 2017. *A Short Introduction to the caret Package*. https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf.
Lantz, Brett. 2015. *Machine learning with R*. Packt Publishing Ltd. http://www.packtpub.com/books/content/machine-learning-r.
Maros, Máté E, David Capper, David TW Jones, Volker Hovestadt, Andreas von Deimling, Stefan M Pfister, Axel Benner, Manuela Zucknick, y Martin Sill. 2020. «Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data». *Nature protocols* 15 (2): 479-512.
Price, E Magda, Allison M Cotton, Lucia L Lam, Pau Farré, Eldon Emberly, Carolyn J Brown, Wendy P Robinson, y Michael S Kobor. 2013. «Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array». *Epigenetics & chromatin* 6 (1): 1-15.
Zhou, Wanding, Peter W Laird, y Hui Shen. 2017. «Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes». *Nucleic acids research* 45 (4): e22-22.