**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 (clasificarla en una de las 20 categorías de tejidos del TCGA)?.

Variación de la pregunta:

* Lo mismo pero considerando un algoritmo random forest.
* Lo mismo pero considerando un algoritmo árbol de decisión. Este algoritmo tiene la ventaja de ser fácilmente interpretable, por lo que en el supuesto de que tuviera una capacidad predictiva aceptable podríamos buscar la coherencia de sus criterios de clasificación con los resultados de estudios previos de metilación diferencial entre grupos.
* Lo mismo pero en lugar de utilizar datos de metilación usando los datos (picos) de histonas modificadas.

Una primera propuesta de Título para el trabajo sería:

**Clasificador deep learning de muestras tumorales basado en perfiles de metilación a partir de los datos procesados de NCI Genomic Data Commons (GDC)**.

**Palabras clave: methylation, classifier, TCGA, NCI Genomic Data Commons (GDC), epigenomics, machine learning, deep leraning, cancer, tumours, Illumina methylation 450k.**

Se partiría de los datos depositados en TCGA (aproximadamente ¿? 11000 muestras existentes en TCGA) cuyas tablas de fenotipos contemplan unas ¿? 400 variables, entre las que se encuentran los tipos de tejido.

El planteamiento trata de ser algo diferente del que he observado en otros trabajos y que resumido consistiría en partir de dos grupos de muestras (dos grupos fenotípicamente distintos) y obtener las “features” (expresión de genes, locis metilados, picos de histonas) que estadísticamente poseen valores diferenciales.

**Descripción general:**

* Descarga del mayor volumen de ensayos de análisis de metilación posibles. Alternativas:
  + Usar la librería TCGAbiolinks. Acceso al portal GDC Genimic Data Commons.
  + Usar la librería curatedTCGAData. Acceso al repositorio BROAD Genome Data Analysis Center Firehose.
  + Usar la librería RTCGATollbox.
* Categorización de las muestras:
  + Opción 1: etiquetándolas como muestras de control (no patológicas) y muestras patológicas etiquetadas por patologías.
  + Opción 2: etiquetándolas como muestras con una patología concreta y resto de muestras.
* Desglose del conjunto de muestras entre conjunto de entrenamiento y conjunto test.
* A partir de las muestras de entrenamiento, determinación de las sondas con valores (proporciones de metilación, valores betas) diferenciales entre cada patología y el conjunto de las muestras de control o entre una patología y el resto de muestras.
* Selección de las sondas (solo muestras de entrenamiento). Alternativas:
  + Sondas con valores diferenciales entre cada grupo diseasecode y un grupo de control formado por todas las muestras de control de todos los estudios. Se seleccionan las sondas comunes a todos los contrastes.
  + Sondas con valores diferenciales entre cada diseasecode y el resto de muestras. Cada patología en concreto vs todas las demás muestras. Se seleccionan la unión de todos los subconjuntos de sondas obtenidos.
  + Seleccionar el subconjunto de sondas con mayor variabilidad (un número manejable de cómo máximo 30.000 sondas.
  + Seleccionar aleatoriamente las sondas en un número similar al del puntos anterior.
* Con el subconjunto de muestras, formulación de los clasificadores: red neuronal y random forest.
* Determinar los genes asociados (o sus promotores) a las sondas seleccionadas, analizando su ontología, al menos de forma preliminar.
* Evaluación de clasificador en las muestras test.

**Objetivos:**

1. **Obtener un clasificador que a partir del perfil de metilación, valores betas de un subconjunto seleccionado de sondas del array Illumina 450k, determinar si la muestra es patológica o no, y de serlo, a qué patología de la clasificación de TCGA corresponde.**
2. **Aproximación, a partir de la ubicación de las sondas con valores diferenciales entre muestras patológicas y controles, de procesos biológicos comunes al conjunto de tumores.**

**En cuanto a la búsqueda de bibliografía:**

Partiendo de Búsquedas en GOOGLE scholar:

* + - epigenetic methylation
    - promoter methylation
    - [histone methylation](https://scholar.google.es/scholar?hl=es&as_sdt=0,5&qsp=3&q=histone+methylation&qst=ib)
    - dna methylation

Ordenaría por fecha de publicación, puesto que las librerías R para manejar los repositorios de datos de TCGA son recientes no creo que sea preciso remontarse muchos años atrás para comprobar si existe publicado algún trabajo con este enfoque.

En cualquier caso necesitaría un criterio de selección de artículos dado que su número es abrumador.

**En concreto, los artículos que he consultado**:

Capper, D., Jones, D. T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., ... & Pfister, S. M. (2018). DNA methylation-based classification of central nervous system tumours. Nature, 555(7697), 469-474.

El clasificador es Randomforest. De este artículo veo que es factible un clasificador que a partir de 2801 muestras es capaz de clasificar los tumores cerebrales en más de 90 categoría, lo hace a partir de 32000 sondas seleccionadas por su variabilidad, (de este artículo saco la idea de usar la librería Rtsne en lugar del análisis de componentes principales.

Koelsche, C., Schrimpf, D., Stichel, D., Sill, M., Sahm, F., Reuss, D. E., ... & von Deimling, A. (2021). Sarcoma classification by DNA methylation profiling. Nature communications, 12(1), 1-10.

Clasificador de sarcomas. Utiliza la misma metodología que el estudio de Capper. Se entrena con 1077 muestras clasificadas en 62 categorías.

He consultado también los Trabajos Fin de Máster de Lorena Ponce y Ana Díez.

Librerías R:

* Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G and Noushmehr H. "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data." Nucleic acids research (2015): gkv1507.
* Silva TC, Colaprico A, Olsen C, Angelo FD, Bontempi G, Ceccarelli M, Noushmehr H (2022). *TCGAWorkflow: TCGA Workflow Analyze cancer genomics and epigenomics data using Bioconductor packages*. R package version 1.19.1, <https://f1000research.com/articles/5-1542/v2>.
* curatedTCGAData, TCGAutils, RTCGATollbox, IlluminaHumanMethylation450kanno.ilms12.hg19, FDb.,InfiniumMethylatyon.hg19...

**PROBLEMÁTICA ESPERADA**

* Obtener una Conclusión del trabajo de dos palabras: “NO FUNCIONA”. Si el objetivo es la comparación de dos grupos de muestras fenotípicamente distintas siempre vamos a encontrar algún grupo de “features” con valores significativamente diferentes y por lo tanto llegaremos a algo que nos permite redactar una “Conclusión”. En esta propuesta no esta garantizado que sea capaz de diseñar y entrenar un modelo con resultados predictivos aceptables.
* Que el volumen de muestras disponible no sea ni de lejos el necesario para entrenar una red neuronal y por lo tanto la matriz de confusión final sea decepcionante (estaría en la situación descrita en el punto anterior).
* Que las capacidades de cálculo no sean suficientes y no consiga entrenar el algoritmo.
* Las habituales de todo el mundo de limitación de tiempo y conocimientos a las que hay que añadir las mías propias.

Los borradores, documentos y scripts del trabajo Fin de Máster de máster UOC se almacenan en la ubicación:

C:\TFM UOC

<https://github.com/ajoven10/TFM_UOC.git>

En concreto, **LOS DATOS A EMPLEAR**:

Utilizando el repositorio Broad Genome Data Analysis (GDAC) Firehose descargo la tabla con los estudios disponibles asociados a tipo de experimento “Methyl\*”:

*library(curatedTCGAData)*

*curatedTCGAData("\*", "Methyl\*", dry.run = TRUE, version = "1.1.38")*



Se descargan los diferentes códigos de enfermedad, diseasecodes, hasta completar 32 códigos de 37 (no se descargan: STAD -dio error-, MISC, FFPP, CNTL y LCML –no hay estudio de metilación-):

Los 37 códigos:

"ACC" "BLCA" "BRCA" "CESC" "CHOL" "CNTL" "COAD" "DLBC" "ESCA" "FPPP" "GBM" "HNSC" "KICH" "KIRC" "KIRP" "LAML" "LCML" "LGG" "LIHC" "LUAD" "LUSC" "MESO" "MISC" "OV" "PAAD" "PCPG" "PRAD" "READ" "SARC" "SKCM" "STAD" "TGCT" "THCA" "THYM" "UCEC" "UCS" "UVM"

La función usada para la descarga fue:

*estudios <- curatedTCGAData(*

*diseaseCode = c("ACC", "BLCA"), assays = "Methyl\*", version = "1.1.38", dry.run = FALSE*

*)*

Así hasta descargar los 32 códigos.

Extraigo las matrices de metilación estudio a estudio, tomando tan solo los estudios que usaron el array 450k :

*experiments(estudios)*

*matriz1 <- assays(estudios)*

*matriz1 <- matriz1[["BLCA\_Methylation-20160128"]]*

*matriz1 <- as.matrix(matriz1)*

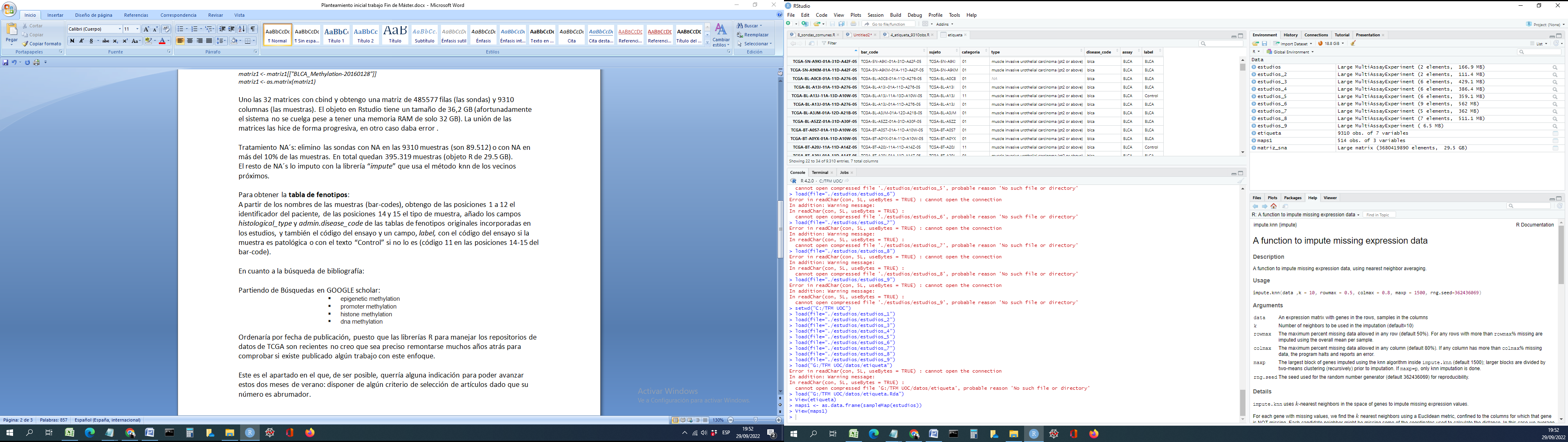
Uno las 32 matrices con cbind y obtengo una matriz de 485577 filas (las sondas) y 9310 columnas (las muestras). El objeto en Rstudio tiene un tamaño de 36,2 GB (afortunadamente el sistema no se cuelga pese a tener una memoria RAM de solo 32 GB). La unión de las matrices las hice de forma progresiva, en otro caso daba error .

Tratamiento NA´s: elimino las sondas con NA en las 9310 muestras (son 89.512) o con NA en más del 10% de las muestras. En total quedan 395.319 muestras (objeto R de 29.5 GB).

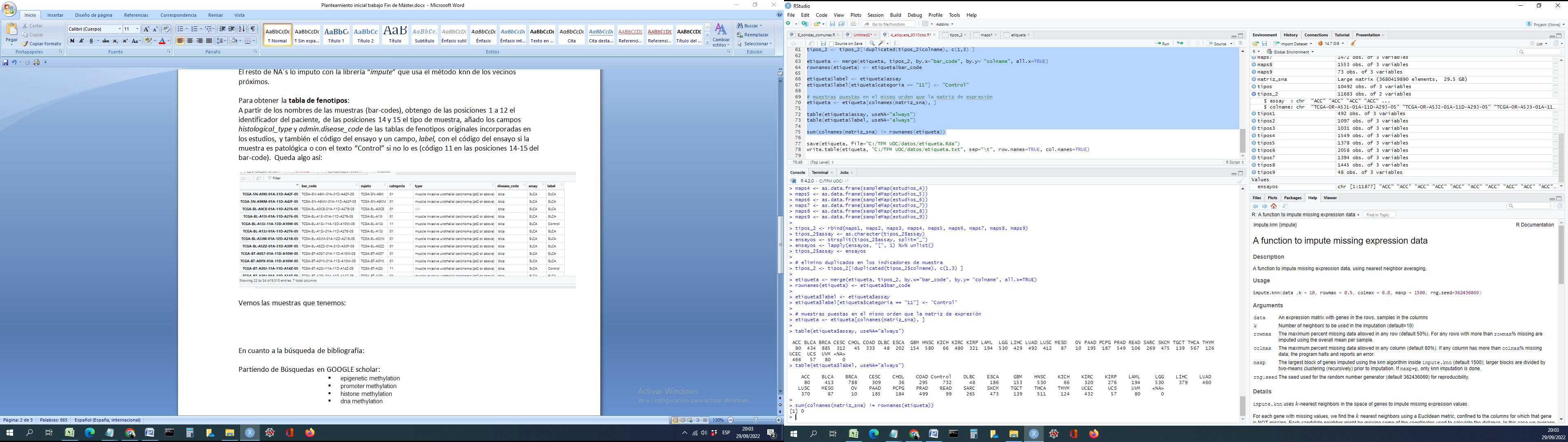
El resto de NA´s lo imputo con la librería “*impute*” que usa el método knn de los vecinos próximos.

Para obtener la **tabla de fenotipos**:

A partir de los nombres de las muestras (bar-codes), obtengo de las posiciones 1 a 12 el identificador del paciente, de las posiciones 14 y 15 el tipo de muestra, añado los campos *histological\_type* y *admin.disease\_code* de las tablas de fenotipos originales incorporadas en los estudios, y también el código del ensayo y un campo, *label,* con el código del ensayo si la muestra es patológica o con el texto “Control” si no lo es (código 11 en las posiciones 14-15 del bar-code). Queda algo así:



Vemos las muestras que tenemos:



De las 9310 muestras, 732 (Control) son no patológicas y la variabilidad en el número de muestras cancerosas es muy alta. A priori esto va a ser un hándicap muy notable para construir el clasificador.

**Los RANGOS:** información de las sondas obtenida de:

*library(FDb.InfiniumMethylation.hg19)*

*sondas <- get450k()*

*sondas <- sondas[rownames(matriz\_sna)]*

**y con la matriz de betas, tabla de fenotipos e información de rangos de las sondas, se construye el objeto SummarizedExperiment:**

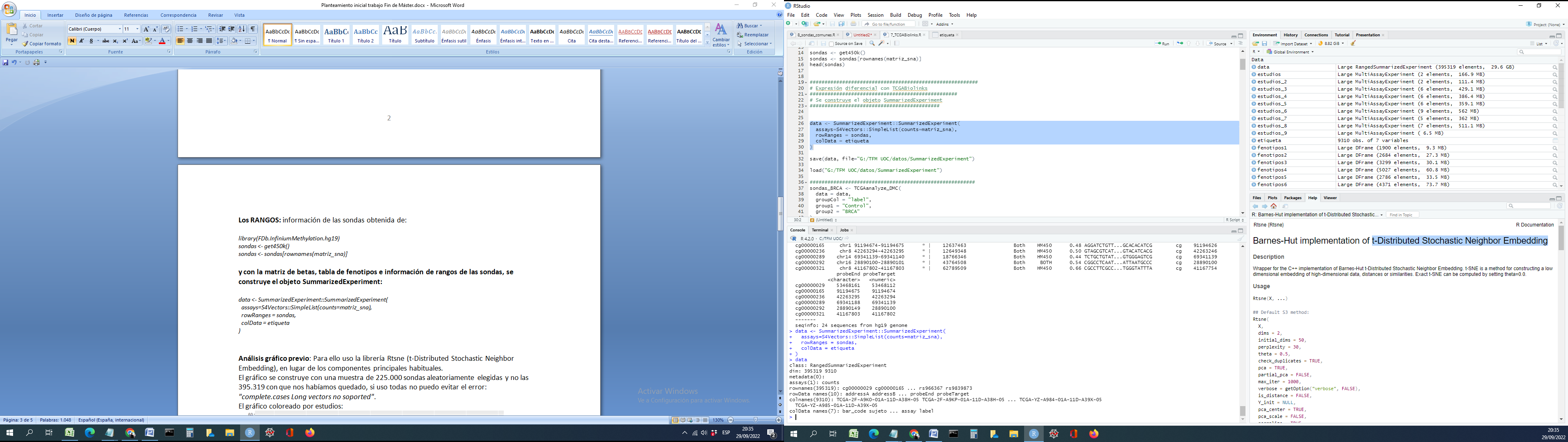
*data <- SummarizedExperiment::SummarizedExperiment(*

*assays=S4Vectors::SimpleList(counts=matriz\_sna),*

*rowRanges = sondas,*

*colData = etiqueta*

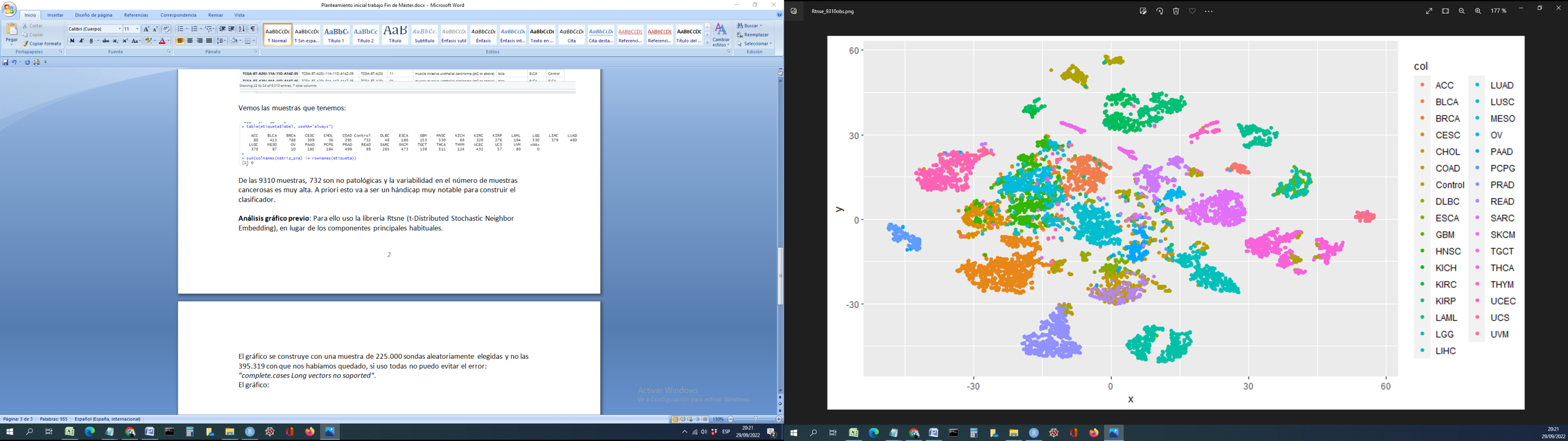
*)*

****

**Análisis gráfico previo**: Para ello uso la librería Rtsne (t-Distributed Stochastic Neighbor Embedding), en lugar de los componentes principales habituales.

El gráfico se construye con una muestra de 225.000 sondas aleatoriamente elegidas y no las 395.319 con que nos habíamos quedado, si uso todas no puedo evitar el error: *"complete.cases Long vectors no soported"*.

El gráfico coloreado por estudios:

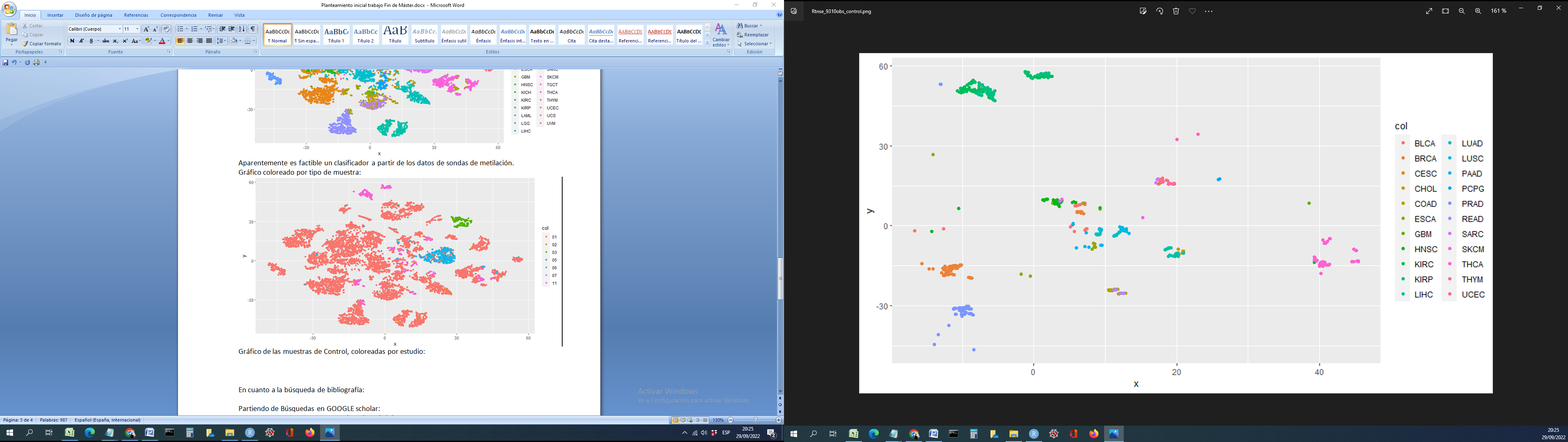


Aparentemente es factible un clasificador a partir de los datos de sondas de metilación.

Gráfico coloreado por tipo de muestra:



Gráfico de las muestras de Control, coloreadas por estudio:



El patrón de metilación varía mucho según el tejido no patológico seleccionado.

Obtención de las sondas con **betas diferenciales** entre grupos patológicos y grupo de Control (que agrupa las 732 muestras de control de todos los estudios):

De la asignatura de Regresión: dado que las betas son proporciones (varían entre 0 y 1) no es posible aplicar el test t-Student, y las posibles estrategias serían (copio de un material de la asignatura):

* Aplicar un test de Welch a los datos transformados con la función de normalización .
* Hacer un test no paramétrico como el de Mann-Whitney-Wilcoxon.
* Aplicar un test de permutaciones para comparar las dos muestras (aquí tenemos 33).
* Realizar una regresión logística binomial.
* Calcular una *regresión beta* que es especialmente apropiada para proporciones.

En términos de librerías R:

* Test de permutaciones: paquete coin o paquete lmPerm.
* Regresión beta: <https://rcompanion.org/handbook/J_02.html> librería *betareg*
* Para el test no paramétrico como el de Mann-Whitney-Wilcoxon puedo usar la librería *TCGABiolinks.*

Uso el test Wilcoxon: Lo aplico para cada estudio. Para los 32 estudios se calcula la prueba Wilcoxon entre los valores de las sondas de tejidos patológicos de ese estudio y el conjunto de las muestras de control:

*library(TCGAbiolinks)*

*sondas\_BRCA <- TCGAanalyze\_DMC(*

*data = data,*

*groupCol = "label",*

*group1 = "Control",*

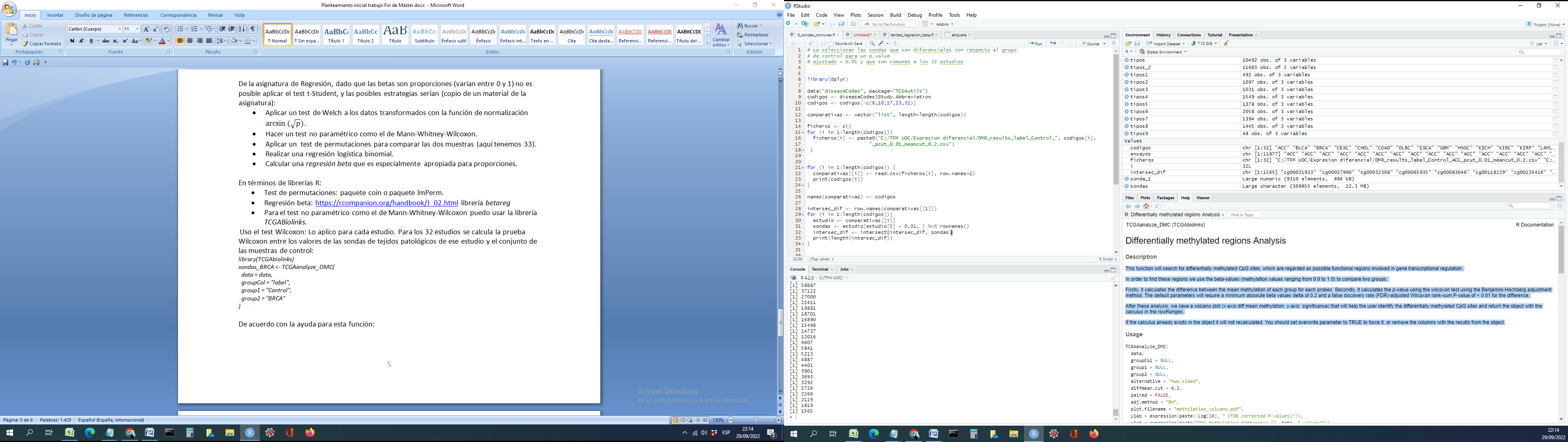
*group2 = "BRCA"*

*)*

De acuerdo con la ayuda para esta función:

*Se calcula la diferencia entre la metilación media de cada grupo para cada sonda. En segundo lugar, calcula el valor p mediante la prueba de Wilcoxon utilizando el método de ajuste de Benjamini-Hochberg*

Una vez obtenidos los 32 subconjuntos de muestras, obtengo el subconjunto de muestras comunes a todos ellos:



El resultado son 1565 sondas que, en principio, han de ser los datos de entrada del clasificador para esta primera opción.