Descripción clínica de pacientes de la base de datos TCA-SKCM

Avila Aylén

2023

Trabajo Integrador

DIPLOMATURA EN VISUALIZACIÓN DE DATOS

FCEyE - UNR - AÑO 2023

Introducción

Cancer

El término cáncer incluye un gran grupo de enfermedades que afectan a diferentes partes del organismo caracterizadas por fallas en mecanismos relacionados con la proliferación y diseminación celular. Durante el 2020 la mortalidad por cáncer ascendió a 10 millones de personas a nivel mundial (GLOBOCAN, 2020).

El desarrollo tumoral es un proceso complejo de múltiples pasos que puede culminar en una transformación de células normales a malignas siendocaracterísticas la evasión del sistema inmune, resistencia a apoptosis y cambios morfológicos, entre otros.
La metástasis consiste en la diseminación de las células cancerígenas desde su sitio de origen hacia otro órgano o tejido donde proliferan formando un nuevo tumor. Depende de propiedades tanto de la célula tumoral como de la persona huésped. El proceso metastásico es multifactorial y consiste en una compleja serie de pasos. La cascada metastásica es selectiva, parcialmente adaptativa, ineficiente y posee regulación genética y epigenética. El proceso de migración, desde el sitio tumoral de origen, puede ser por vía sanguínea, linfática o transcelómica y requiere que las células adquieran propiedades invasivas. Si bien no hay una única causa que conlleve a la desregulación del ciclo celular, proliferación celular y transición epitelio-mesénquima asociadas al desarrollo tumoral, hay ciertas familias de proteínas que controlan muchos procesos relacionados con estas funciones.

Melanoma

Particularmente, el melanoma es originado por la transformación maligna y posterior proliferación descontrolada de los melanocitos, células derivadas embriológicamente de células madres pluripotentes de la cresta neural. Durante el desarrollo fetal, no sólo migran y se diferencian predominantemente dentro de la epidermis, sino también en otros sitios que contienen pigmentos extra cutáneos como los ojos, las meninges, el esófago y las membranas mucosas. Por lo tanto, se pueden caracterizar tres subtipos de melanoma: el melanoma cutáneo (el más común) que surge de los melanocitos en la epidermis, el melanoma mucoso y el melanoma uveal (Yousaf & Larkin, 2013). El melanoma cutáneo es la forma más severa del cáncer de piel y constituye un gran problema sanitario a nivel nacional y mundial. La incidencia global de melanoma continúa en aumento y los principales factores que predisponen a su desarrollo parecen estar relacionados con un aumento de exposición crónica a radiación UV (ASCO, 2019).

Características clínico-patológicas de interés

Clínicas Se han seleccionado algunas características que han sido reportadas como interesantes para el análisis según la American Cancer Society (https://www.cancer.org/es/cancer/tipos/cancer-de-piel-tipo-melanoma/acerca/estadisticas-clave.html).

- Sexo manifestado
- País en el que viven al momento del diagnóstico
- Edad al momento del diagnóstico
Puede resultar de interés identificar la variable sexo en conjunto con la edad ya que, en personas mayores a 50 años, los diagnosticos en mujeres se han incrementado respecto de los de varones.

Patológicas El pronóstico y las opciones de tratamiento dependen de varios factores. A continuación se detallan alugnas características típicamente utilizadas para la estadificación tumoral:
- Tipo tumoral: El melanoma puede ser un tumor primario o un tumor metastásico.
- Nivel de Clark: Es una característica *cualitativa* de la profundidad de la diseminación del cáncer de piel.
- Clark I: solo se encuentra en la epidermis.
- Clark II: comenzó a diseminarse a la dermis papilar (capa superior de la dermis).
- Clark III: se diseminó a través de la dermis papilar a la unión entre la dermis papilar y reticular (capa inferior de la dermis).
- Clark IV: diseminó a la dermis reticular.
- Clark V: diseminó al tejido subcutáneo.
- Ulceración: Se denomina ulceración a la ruptura de la piel que se encuentra sobre el melanoma. Es una característica *cualitativa* dicotómica.
- Profundidad de Breslow: Es una medida de la profundidad *cuantitativa* del crecimiento del melanoma en la piel. Es utilizado, junto a otros indicadores, para determinar el estadio del cáncer. Se mide en milímetros.

Atlas genómico del cáncer (TCGA, The Cancer Genome Atlas)

En 2006 y con el objetivo de recopilar y analizar datos clínicos y moleculares en distintos tipos tumorales, el National Cancer Institute junto al National Human Genome Research Institute crearon el consorcio TCGA (https://portal.gdc.cancer.gov/) (TCGA, 2006). Este repositorio cuenta con datos clínicos, información histopatológica, datos genómicos, transcriptómicos, proteómicos y epigenéticos que se encuentran, en su gran mayoría, de libre acceso. Dentro de este repositorio se encuentran diferentes proyectos asociados a cada tipo de cáncer, para el caso del melanoma la denominación del proyecto es TCGA-SKCM (por sus siglas en inglés, Skin Cancer Cutaneous Melanoma).

R (CRAN: The Comprehensive R Archive Network)

R es un lenguaje y un entorno diseñado para la implementación de técnicas y análisis estadísticos. Se trata de un lenguaje abierto y se encuentra disponible de manera gratuita, con una licencia GPL (general public license). Todas las variables, funciones, datos, salidas, operaciones (lógicas, comparativas, etc.), etc. son guardadas como objetos, por lo que R es un lenguaje orientado a objetos. Fue creado en 1993 por Ross Ihaka y Robert Gentleman en el Departamento de Estadística de la Universidad de Auckland, Nueva Zelanda. A partir de 1997 se conformó el equipo llamado “R Core Team” quienes tienen acceso a las modificaciones del código fuente.

En este trabajo se propone acceder a la información clínica disponible en el proyecto TCGA-SKCM, utilizando el entorno R, para explorar y describir algunas características del set de datos.

Materiales y métodos

Librerías utilizadas

- RTCGA.clinical

```
library("RTCGA.clinical")
```

```
## Loading required package: RTCGA
```

```
## Welcome to the RTCGA (version: 1.26.0).
```

- Tidyverse

```
library("tidyverse")
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

La paleta de colores utilizada en los gráficos de este informe fueron seleccionados siguiendo las pautas propuestas por las simulaciones de KHROMA (https://cran.r-project.org/web/packages/khroma/readme/README.html) para que sean amigables a los diferentes tipos de daltonismo.

Acceso y limpieza del set de datos

- Descarga de datos

```
clinica<-as.data.frame(SKCM.clinical)
```

- Exploración de los datos disponibles

```
dim(clinica)
```

```
## [1]  470 1875
```

```
head(colnames(clinica))
```

```
## [1] "admin.bcr"                          "admin.day_of_dcc_upload"
## [3] "admin.disease_code"                 "admin.file_uuid"
## [5] "admin.month_of_dcc_upload"          "admin.patient_withdrawal.withdrawn"
```

- Selección, recodificación y reasignaciónde datos interesantes para este estudio

```
datos<-clinica %>%
  select(c(patient.age_at_initial_pathologic_diagnosis,
           patient.clinical_cqcf.country,
           patient.gender,
           patient.clinical_cqcf.tumor_type,
           patient.melanoma_clark_level_value,
           patient.breslow_depth_value,
           patient.melanoma_ulceration_indicator))%>%
  mutate(patient.gender = recode(patient.gender,
                                 female= "Mujer",
                                 male= "Varón")) %>%
  mutate(patient.clinical_cqcf.tumor_type = recode(patient.clinical_cqcf.tumor_type,
                                       primary= "Primario",
                                       metastatic= "Metastasico")) %>%
  mutate(patient.melanoma_ulceration_indicator = recode(patient.melanoma_ulceration_indicato
r,
                                                        yes = "Con",
                                                        no = "Sin"))%>%
  rename(Ulceracion=patient.melanoma_ulceration_indicator)%>%
  rename(Pais=patient.clinical_cqcf.country)%>%
  rename(Tipo_tumoral=patient.clinical_cqcf.tumor_type)%>%
  rename(Sexo=patient.gender)%>%
  rename(Edad=patient.age_at_initial_pathologic_diagnosis)%>%
  rename(Nivel_Clark=patient.melanoma_clark_level_value)%>%
  rename(Profundidad_Breslow=patient.breslow_depth_value)%>%
  mutate(Pais = recode(Pais,
                       australia = "Australia",
                       canada = "Canada",
                       germany = "Alemania",
                       italy = "Italia",
                       poland = "Polonia",
                       "puerto rico" = "Puerto Rico",
                       russia = "Rusia",
                       ukraine = "Ucrania",
                       "united kingdom" = "Reino Unido",
                       "united states" = "Estados Unidos",
                       "vietnam" = "Vietnam"))


datos[,c(1,6)]<-sapply(datos[,c(1,6)],as.numeric)
```

- Verificación de los datos seleccionados

```
summary(datos)
```

```
##       Edad           Pais               Sexo           Tipo_tumoral
##  Min.   :15.00   Length:470         Length:470         Length:470
##  1st Qu.:48.00   Class :character   Class :character   Class :character
##  Median :58.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :58.22
##  3rd Qu.:71.00
##  Max.   :90.00
##  NA's   :8
##  Nivel_Clark        Profundidad_Breslow  Ulceracion
##  Length:470         Min.   : 0.000      Length:470
##  Class :character   1st Qu.: 1.300      Class :character
##  Mode  :character   Median : 3.000      Mode  :character
##                     Mean   : 5.585
##                     3rd Qu.: 6.950
##                     Max.   :75.000
##                     NA's   :111
```

- Guardado de datos

```
write.table(datos,"datos.txt")
```

Resultados

La tabla con los datos seleccionados se encuentra disponible en el siguiente link (https://github.com/ImoPupato/DiploVisDatos/blob/main/datos.txt).

Características de las personas incluidas en la base de datos

- Edad al momento del diagnóstico

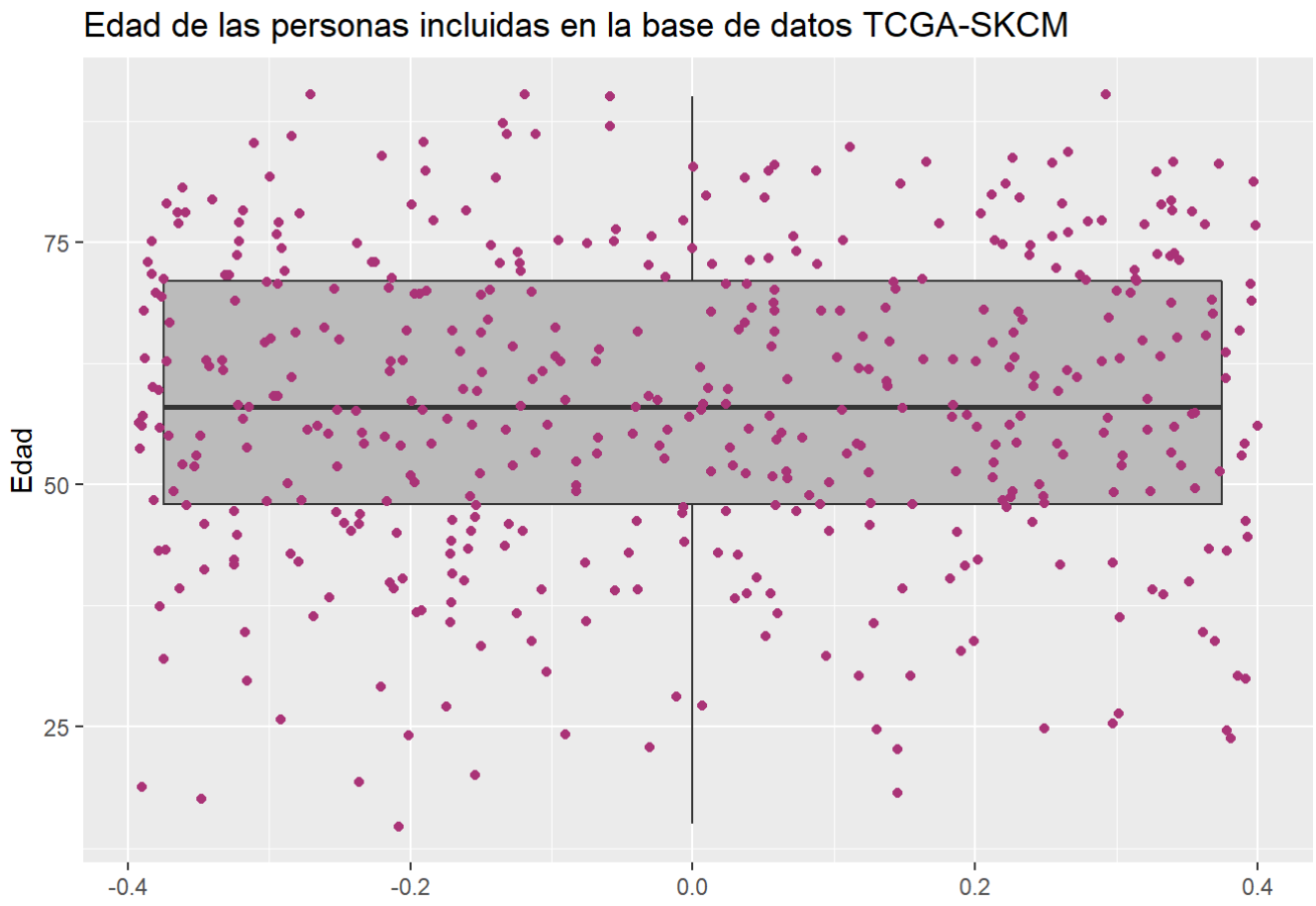
```
summary(datos$Edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.00   48.00   58.00   58.22   71.00   90.00       8
```

```
tapply(datos$Edad,factor(datos$Sexo),summary,na.rm=TRUE)
```

```
## $Mujer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.00   47.00   59.00   58.53   72.50   90.00       5
##
## $Varón
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.00   48.00   58.00   58.03   70.00   90.00       3
```

```
ggplot(subset(datos,!is.na(Edad)),aes(y=Edad, x=0)) +
  geom_boxplot(fill=c("#BBBBBB")) +
  geom_jitter(colour = "#AA3377")+
  labs(title = "Edad de las personas incluidas en la base de datos TCGA-SKCM", x="")
```



El 50% de las personas tenía hasta 58 años de edad al momento de diagnóstica, siendo igual a la edad promedio. Tanto los cuartilos como el promedio de edad según el sexo, son bastante similares. - País de origen

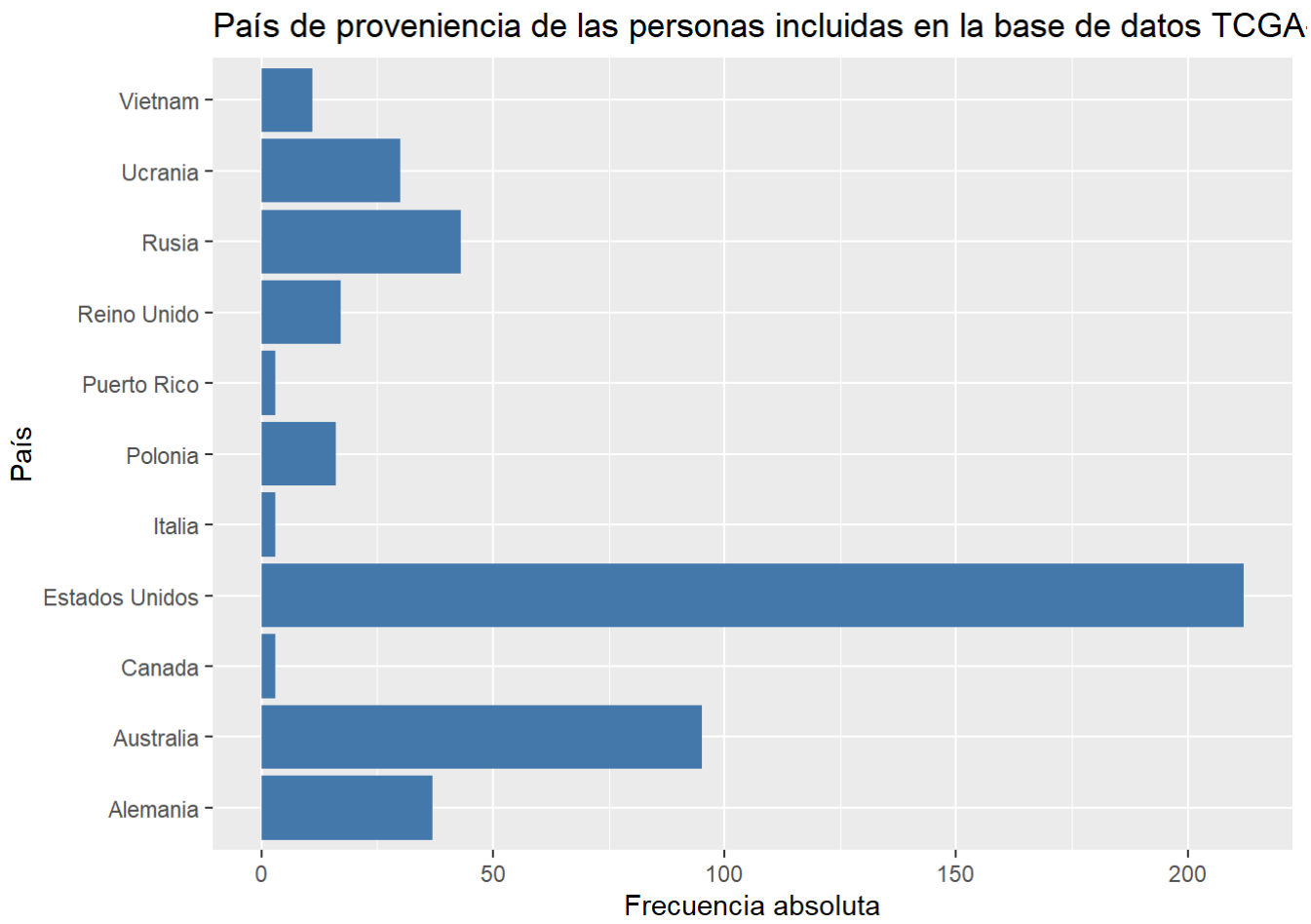
```
table(na.omit(datos$Pais))
```

```
##
##       Alemania      Australia         Canada Estados Unidos         Italia
##             37             95              3            212              3
##        Polonia    Puerto Rico    Reino Unido          Rusia        Ucrania
##             16              3             17             43             30
##        Vietnam
##             11
```

```
prop.table(table(na.omit(datos$Pais)))*100
```

```
##
##       Alemania      Australia         Canada Estados Unidos         Italia
##      7.8723404     20.2127660      0.6382979     45.1063830      0.6382979
##        Polonia    Puerto Rico    Reino Unido          Rusia        Ucrania
##      3.4042553      0.6382979      3.6170213      9.1489362      6.3829787
##        Vietnam
##      2.3404255
```

```
auxiliar<-as.data.frame(datos$Pais)
ggplot(auxiliar, aes(x = `datos$Pais`)) +
  geom_bar(fill="#4477AA") +
  coord_flip() +
  labs(title = "País de proveniencia de las personas incluidas en la base de datos TCGA-SKC
M", y = "Frecuencia absoluta",  x= "País")
```



El 45% de las personas se encontraban viviendo en Estados Unidos al momento del diagnóstico y el 20% en Australia. - Sexo manifestado

```
table(datos$Sexo)
```

```
##
## Mujer Varón
##   180   290
```

```
prop.table(table(na.omit(datos$Sexo)))*100
```

```
##
##    Mujer    Varón
## 38.29787 61.70213
```

Casi el 62% de los diagnósticos corresponde a varones. #### Características de los tumores - Tipo tumoral

```
table(na.omit(datos$Tipo_tumoral))
```

```
##
## Metastasico    Primario
##         366         104
```

```
prop.table(table(na.omit(datos$Tipo_tumoral)))*100
```

```
##
## Metastasico    Primario
##    77.87234    22.12766
```

Casi el 78% de las muestras corresponde a tumores metastásicos. - Nivel Clark

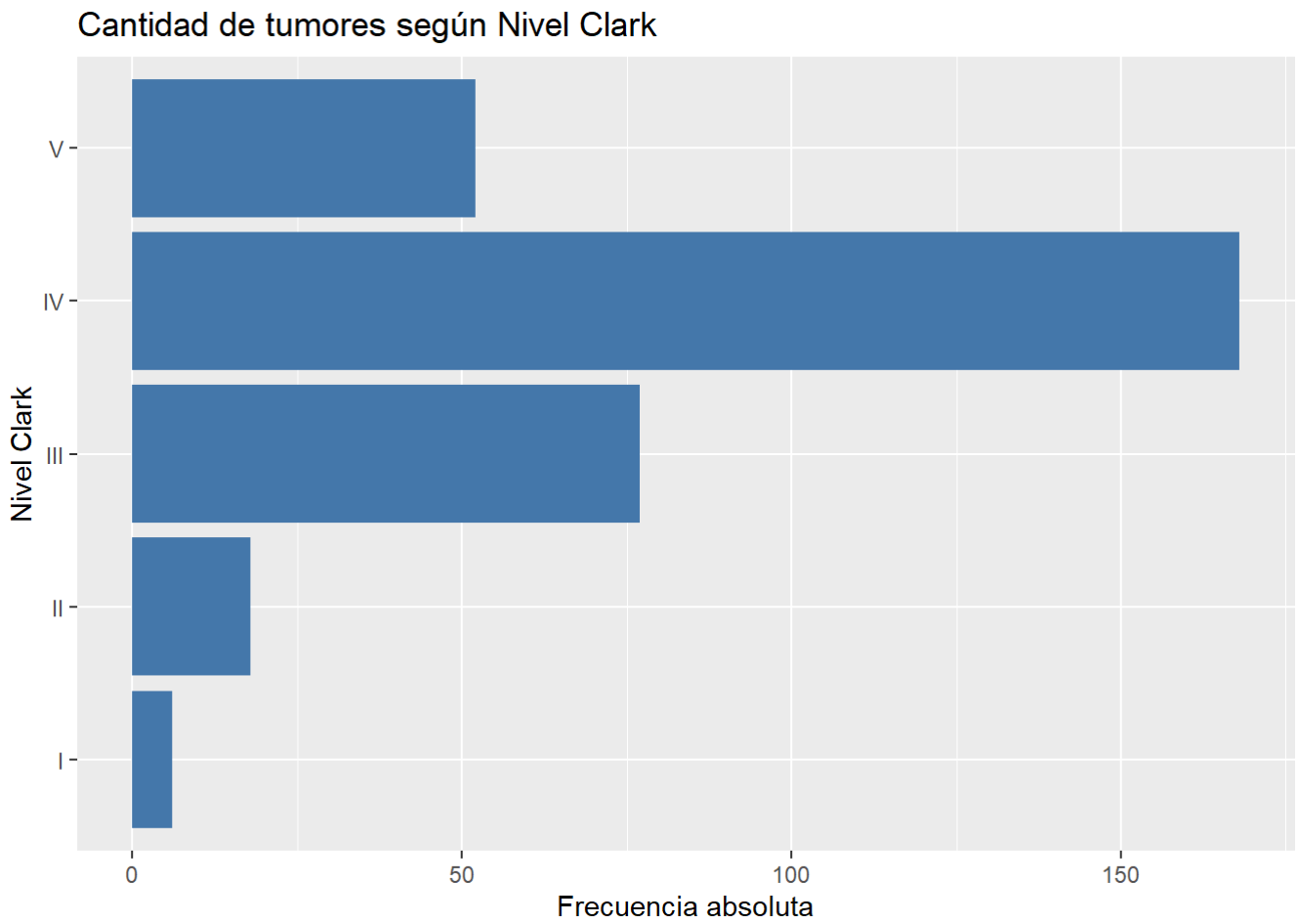
```
table(na.omit(datos$Nivel_Clark))
```

```
##
##   i  ii iii  iv   v
##   6  18  77 168  52
```

```
prop.table(table(na.omit(datos$Nivel_Clark)))*100
```

```
##
##         i        ii       iii        iv         v
##  1.869159  5.607477 23.987539 52.336449 16.199377
```

```
auxiliar<-as.data.frame(datos$Nivel_Clark)
ggplot(subset(auxiliar,!is.na(auxiliar$`datos$Nivel_Clark`)), aes(x = `datos$Nivel_Clark`)) +
  geom_bar(fill="#4477AA") +
  coord_flip() +
  scale_x_discrete(label = c("I","II","III","IV","V"))+
  labs(title = "Cantidad de tumores según Nivel Clark", y = "Frecuencia absoluta",  x= "Nivel
Clark")
```

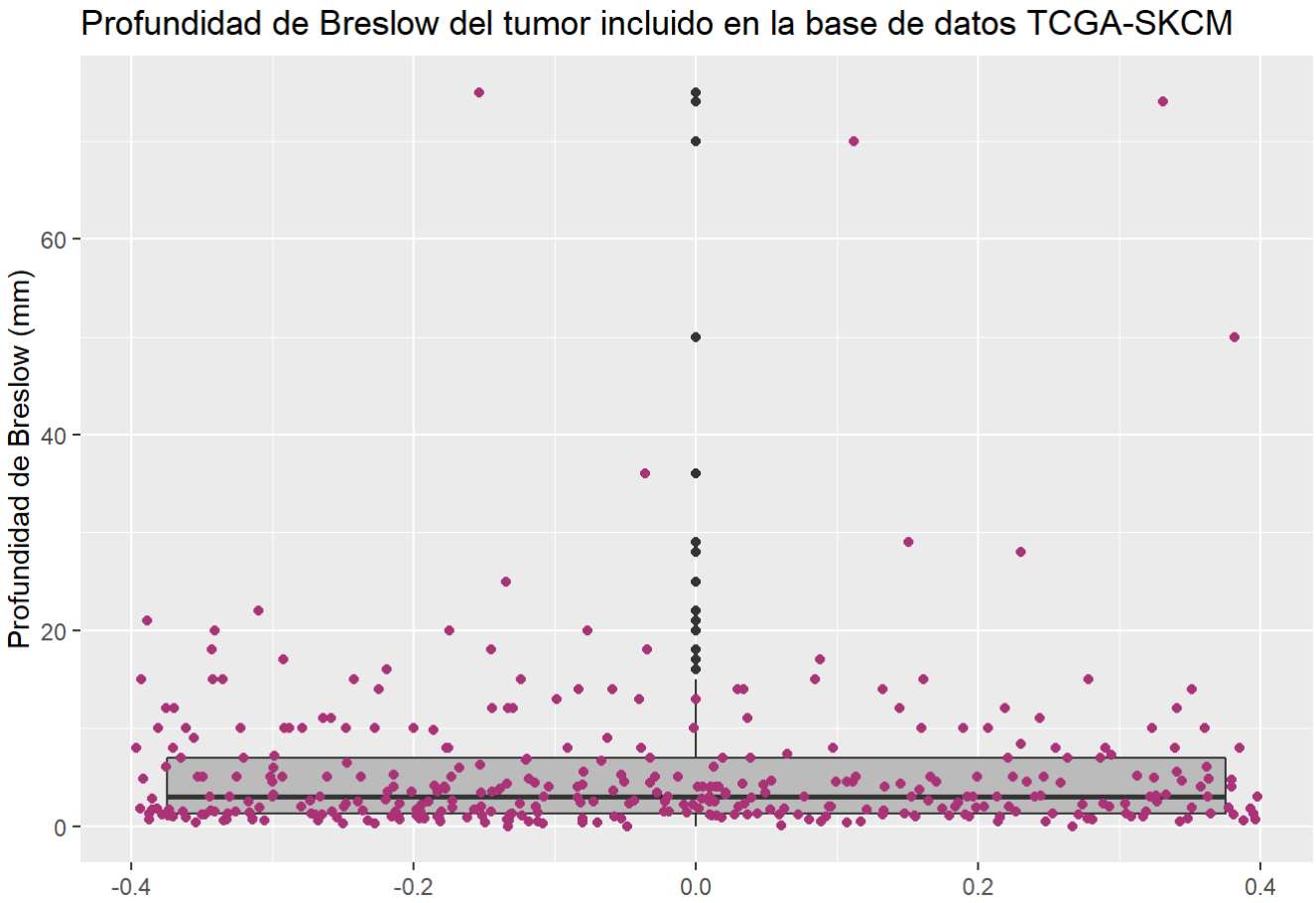


El 93% de los tumores presenta un Nivel de Clark de III o más. - Profundidad de Breslow

```
summary(na.omit(datos$Profundidad_Breslow))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.300   3.000   5.585   6.950  75.000
```

```
ggplot(subset(datos,!is.na(Profundidad_Breslow)),aes(y=Profundidad_Breslow, x=0)) +
  geom_boxplot(fill=c("#BBBBBB")) +
  geom_jitter(colour=c("#AA3377"))+
  labs(title = "Profundidad de Breslow del tumor incluido en la base de datos TCGA-SKCM", x
="", y="Profundidad de Breslow (mm)")
```



El 50% de los tumores presenta una Profundidad de Breslow de 3mm o menos - Ulceración

```
table(datos$Ulceracion)
```

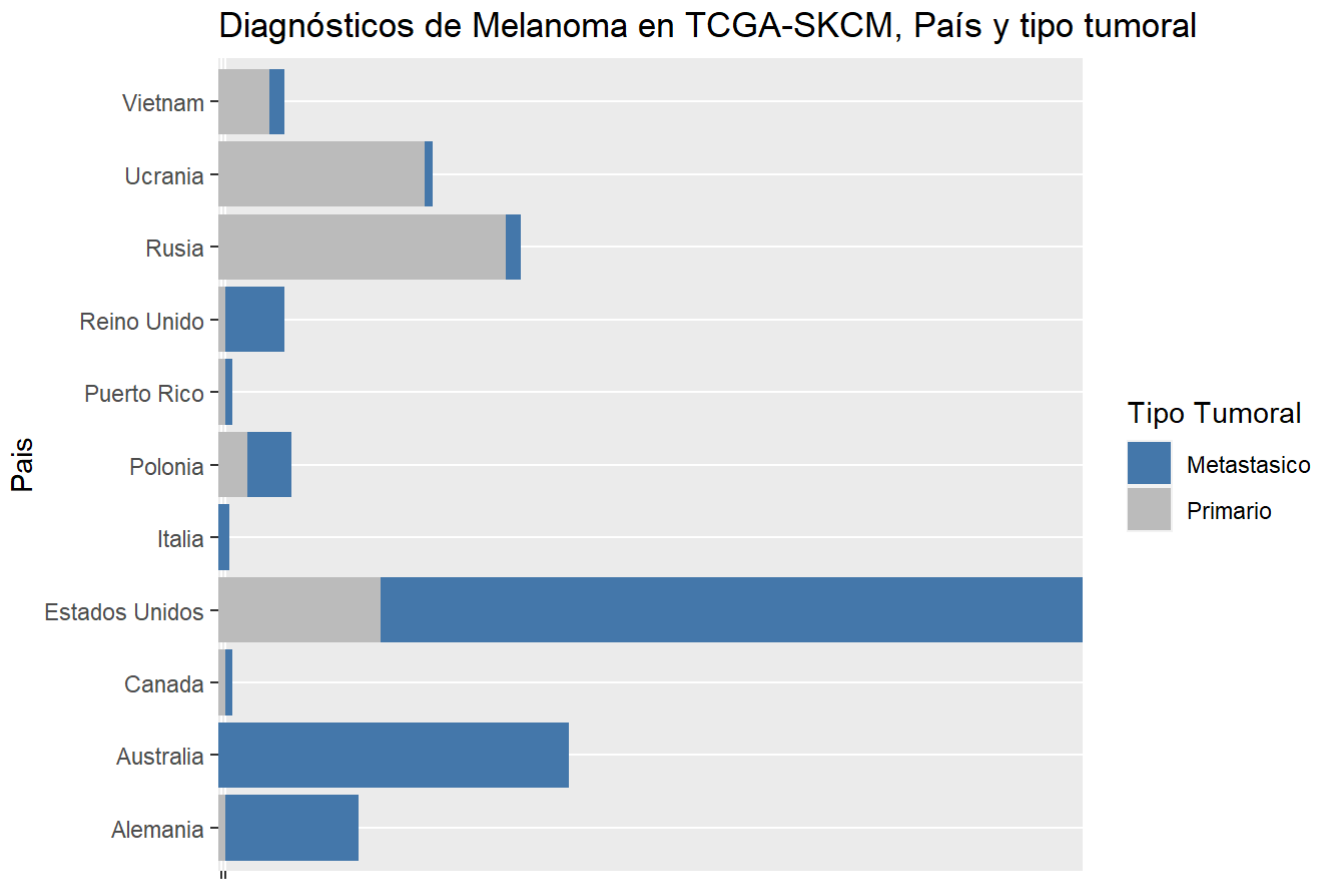
```
##
## Con Sin
## 167 146
```

```
prop.table(table(na.omit(datos$Ulceracion)))*100
```

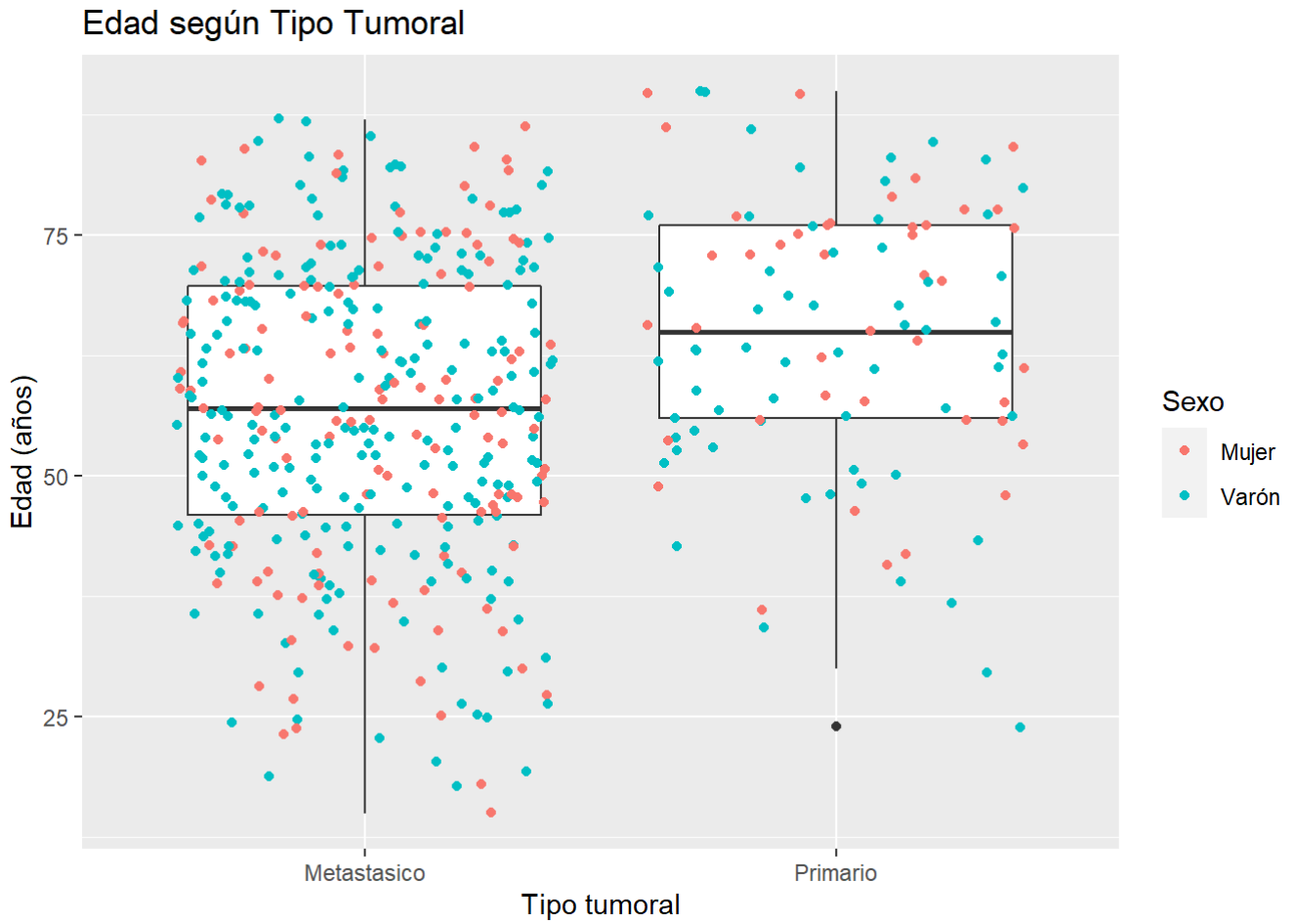
```
##
##      Con      Sin
## 53.35463 46.64537
```

Gráficos de variables conjuntas

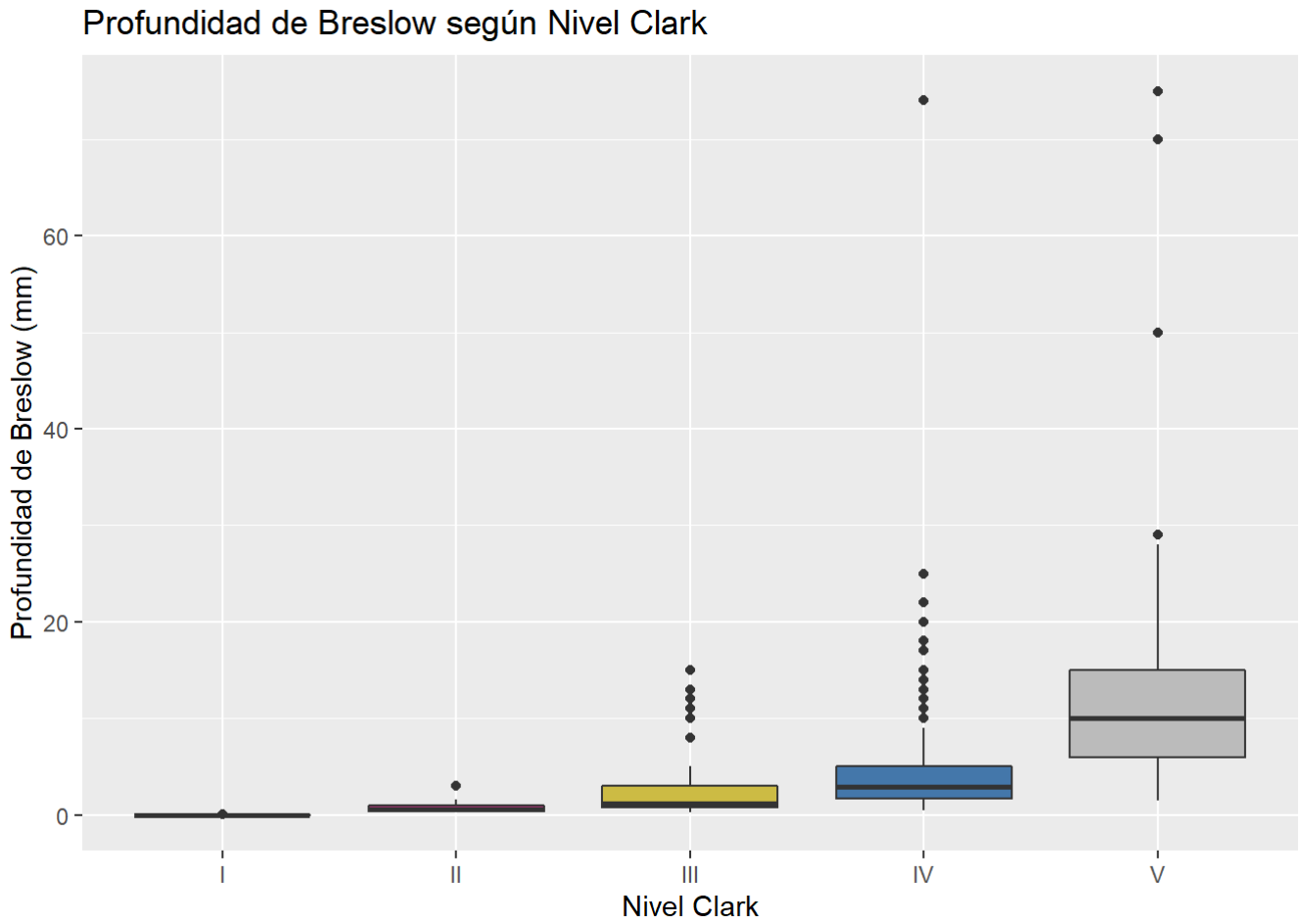
```
ggplot(datos, aes(x = Pais, y = Tipo_tumoral, fill = Tipo_tumoral)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_manual(values = c("#4477AA" ,"#BBBBBB"))+
  labs(title = "Diagnósticos de Melanoma en TCGA-SKCM, País y tipo tumoral", y = " ",  x= "Pa
is", fill = "Tipo Tumoral")+
  theme(axis.text.x = element_text(color = "white"))
```




```
ggplot(datos, aes(x = Tipo_tumoral , y = Edad)) +
  geom_boxplot() +
  geom_jitter(aes(colour=Sexo))+
  labs(title = "Edad según Tipo Tumoral", y = "Edad (años)", x= "Tipo tumoral")
```




```
ggplot(subset(datos,!is.na(Nivel_Clark)),aes(x = Nivel_Clark, y = Profundidad_Breslow)) +
  geom_boxplot(fill=c("#228833", "#AA3377","#CCBB44", "#4477AA" ,"#BBBBBB")) +
  labs(title = "Profundidad de Breslow según Nivel Clark", y = "Profundidad de Breslow (mm)",
x= "Nivel Clark")+
  scale_x_discrete(label = c("I","II","III","IV","V"))
```



Como era de esperarse, a mayor Nivel de Clark, mayor profundidad de Breslow.

Conclusiones

El proyecto TCGA-SKCM brinda una enorme cantidad de datos respecto de las característica clinico-patológicas y moleculares (no descriptas aquí) de muestras tumorales. El uso de estas fuentes de datos requiere de una mirada criteriosa y cautelosa ya que no se cuenta, por ejemplo, con información epidemiológica respecto de las posibles fuentes de exposición de las personas incluidas en el proyecto. La caracterización epidemiológica enriquece el análisis, ya que, a modo de ejeplo, ha sido reportado que la exposición frecuente a luz UV tiene efectos facilitadores en el desarrollo del melanoma. Poder incorporar esta dimensión para el análisis nos permitiría construir grupos de contraste mejor caracterizados.

Es imporante aclarar también que sólo el 22% de las muestras corresponden a tumores primarios, entendiendo al cáncer como un proceso multifactorial y que en la metástasis se desregulan muchas vías biológicas, no considerar este aspecto para el análisis de la información molecular representa, sin dudas, un sesgo.

Referencias bibliográficas

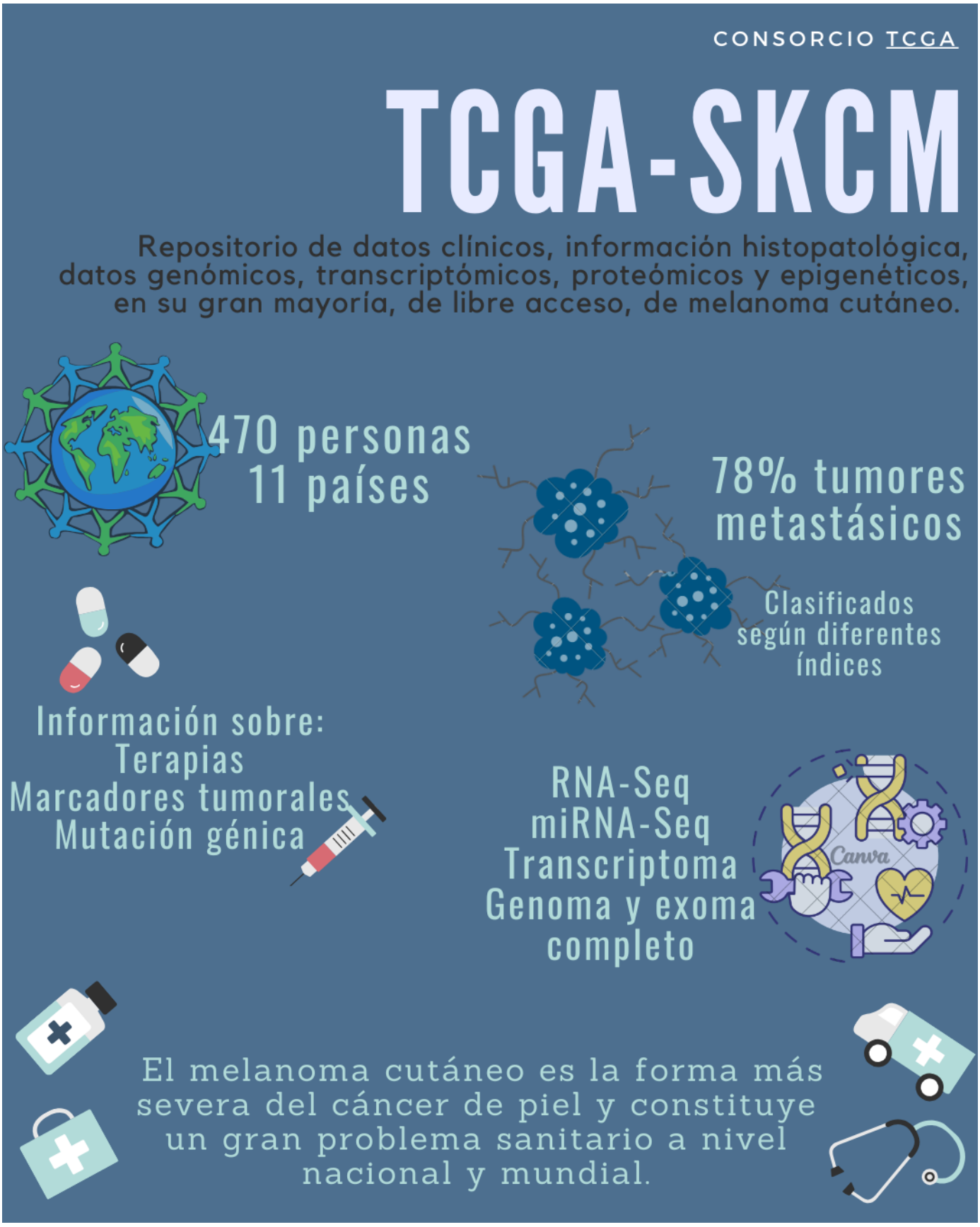
- Ali Z., Yousaf N., Larkin J. (2013). Melanoma epidemiology, biology and prognosis. EJC Supp. 11(2):81-91.
- American Cancer Society (https://www.cancer.org/es/cancer/tipos/cancer-de-piel-tipo-melanoma/acerca/estadisticas-clave.html).
- ASCO (2020). Onco Americas (https://www.oncoamericas.com/es/home-espanol/).
- Colaprico A., Silva T.C., Olsen C., Garofano L., Cava C., Garolini D., Sabedot T.S., Malta T.M., Pagnotta S.M., Castiglioni I., Ceccarelli M., Bontempi G., Noushmehr H. 2015. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 44(8):e71. doi: 10.1093/nar/gkv1507.

- Frerebeau N (2023). *khroma: Colour Schemes for Scientific Data Visualization*. Université Bordeaux Montaigne, Pessac, France. doi:10.5281/zenodo.1472077 (https://doi.org/10.5281/zenodo.1472077), R package version 1.11.0 (https://packages.tesselle.org/khroma/).
- Geiger T.R. & Peeper D.S. 2009. Metastasis mechanisms. Biochimica et Biophysica Acta (BBA). Reviews on Cancer, 1796(2), 293–308. doi:10.1016/j.bbcan.2009.07.006 (doi:10.1016/j.bbcan.2009.07.006).
- Mounir M., Lucchetta M., Silva C.T., Olsen C., Bontempi G., Chen X., Noushmehr H., Colaprico A., Papaleo E. 2019. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS computational biology, 15(3), e1006701.
- Silva C.T., Colaprico A., Olsen C., D'Angelo F., Bontempi G., Ceccarelli M., Noushmehr H. 2016. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Research, 5.

Anexos

Infografía

Utilizando Canva se realizó la siguiente infografía que tiene por objetivo presentar el portal TCGA, el proyecto TCGA-SKCM y algunos números de interés.



Infografia

Github

Se generó un repositorio (https://github.com/ImoPupato/DiploVisDatos) en donde se encuentran todos los archivos generados, este informe y el script utilizado.