Actividad 5

El objetivo de esta actividad es conocer las bases de datos de mutaciones. Para ello consultarás dos bases de datos, COSMIC (https://cancer.sanger.ac.uk/cosmic), que contiene mutaciones relacionadas a cáncer, y gnomAD (https://gnomad.broadinstitute.org/), que contiene mutaciones en personas sanas y enfermedades. Se espera que adquieras la habilidad para manejar la información que ofrecen estas bases de datos.

Leer los datos de COSMIC

```r
#load file
library(readxl)
setwd("C:\\Users\\Choy\\Documents\\Semestre 2\\Análisis de biología computacional")
cosmic <- read_excel("Gene_samples.xlsx")
head(cosmic)
```

```
## # A tibble: 6 x 19
##   Gene_Name Transcript Census_Tier_1 Sample_Name Sample_ID AA_Mutation
##   <chr>     <chr>      <chr>         <chr>           <dbl> <chr>
## 1 KRAS      ENST00000~ Yes           T189255       2658275 p.?
## 2 KRAS      ENST00000~ Yes           1319563       1319563 p.?
## 3 KRAS      ENST00000~ Yes           TCGA-AA-35~   1650974 p.?
## 4 KRAS      ENST00000~ Yes           T3235         2658250 p.?
## 5 KRAS      ENST00000~ Yes           TCGA-DM-A2~   1651287 p.?
## 6 KRAS      ENST00000~ Yes           CC1813        2640225 p.?
## # ... with 13 more variables: CDS_Mutation <chr>, Primary_Tissue <chr>,
## #   Tissue_Subtype_1 <chr>, Tissue_Subtype_2 <chr>, Histology <chr>,
## #   Histology_Subtype_1 <chr>, Histology_Subtype_2 <chr>, Pubmed_Id <chr>,
## #   CGP_Study <chr>, Somatic_Status <chr>, Sample_Type <chr>, Zygosity <chr>,
## #   Genomic_Coordinates <chr>
```

```r
dim(cosmic)
```

```
## [1] 25012    19
```

```r
names(cosmic)
```

```
##  [1] "Gene_Name"           "Transcript"          "Census_Tier_1"
##  [4] "Sample_Name"         "Sample_ID"           "AA_Mutation"
##  [7] "CDS_Mutation"        "Primary_Tissue"      "Tissue_Subtype_1"
## [10] "Tissue_Subtype_2"    "Histology"           "Histology_Subtype_1"
## [13] "Histology_Subtype_2" "Pubmed_Id"           "CGP_Study"
## [16] "Somatic_Status"      "Sample_Type"         "Zygosity"
## [19] "Genomic_Coordinates"
```

Filtrar mutaciones por ciertos parámetros: a) Tipo de muestra "Sample.Type", conservar "Tumour Sample"

```
table(cosmic$Sample_Type)
```

```
## 
##      Cultured Tumour Sample       Unknown 
##            285          23550          1177
```

```
cosmic <- cosmic[which(cosmic$Sample_Type == "Tumour Sample"),]
```

```
dim(cosmic)
```

```
## [1] 23550    19
```

b) Estatus somático, quitar las variantes de origen desconocido

```
table(cosmic$Somatic_Status)
```

```
## 
##         Confirmed Somatic       Previously Reported Variant of unknown origin 
##                      4009                      19473                        68
```

```
cosmic <- cosmic[-which(cosmic$Somatic_Status == "Variant of unknown origin"),]
dim(cosmic)
```

```
## [1] 23482    19
```

Leer los datos de gnomAD

```
setwd("C:\\Users\\Choy\\Documents\\Semestre 2\\Análisis de biología computacional")
gnomAD <- read_excel("gnomAD_v2.1.1_ENSG00000133703_2020_03_18_08_52_58.xlsx")
```

```
dim(gnomAD)
```

```
## [1] 265  50
```

```
names(gnomAD)
```

```
##  [1] "Chromosome"
##  [2] "Position"
##  [3] "rsID"
##  [4] "Reference"
##  [5] "Alternate"
##  [6] "Source"
##  [7] "Filters - exomes"
##  [8] "Filters - genomes"
##  [9] "Consequence"
## [10] "Protein Consequence"
## [11] "Transcript Consequence"
## [12] "Annotation"
## [13] "Flags"
## [14] "Allele Count"
## [15] "Allele Number"
## [16] "Allele Frequency"
## [17] "Homozygote Count"
## [18] "Hemizygote Count"
## [19] "Allele Count African"
## [20] "Allele Number African"
## [21] "Homozygote Count African"
## [22] "Hemizygote Count African"
## [23] "Allele Count Latino"
## [24] "Allele Number Latino"
## [25] "Homozygote Count Latino"
## [26] "Hemizygote Count Latino"
## [27] "Allele Count Ashkenazi Jewish"
## [28] "Allele Number Ashkenazi Jewish"
## [29] "Homozygote Count Ashkenazi Jewish"
## [30] "Hemizygote Count Ashkenazi Jewish"
## [31] "Allele Count East Asian"
## [32] "Allele Number East Asian"
## [33] "Homozygote Count East Asian"
## [34] "Hemizygote Count East Asian"
## [35] "Allele Count European (Finnish)"
## [36] "Allele Number European (Finnish)"
## [37] "Homozygote Count European (Finnish)"
## [38] "Hemizygote Count European (Finnish)"
## [39] "Allele Count European (non-Finnish)"
## [40] "Allele Number European (non-Finnish)"
## [41] "Homozygote Count European (non-Finnish)"
## [42] "Hemizygote Count European (non-Finnish)"
## [43] "Allele Count Other"
## [44] "Allele Number Other"
## [45] "Homozygote Count Other"
## [46] "Hemizygote Count Other"
## [47] "Allele Count South Asian"
## [48] "Allele Number South Asian"
## [49] "Homozygote Count South Asian"
## [50] "Hemizygote Count South Asian"
```

Recordar que gnomAD tiene datos de población latina, lo cual nos puede servir para hacer comparaciones

```
table(gnomAD$Annotation)
```

```
##
##     3_prime_UTR_variant     5_prime_UTR_variant      frameshift_variant
##                      20                       3                       6
##        inframe_deletion          intron_variant        missense_variant
##                       3                     109                      49
## splice_acceptor_variant    splice_donor_variant   splice_region_variant
##                       1                       1                      22
##             stop_gained      synonymous_variant
##                       5                      46
```

Para poder comparar la locación de las mutaciones en cosmic con las de gnomAD, tenemos que agregar variables a la tabla de cosmic.

```
aux_loc <- unlist(strsplit(x=cosmic$Genomic_Coordinates,split=":"))[seq(from=2, to=nrow(cosmic)*2, by=2
```

```
aux_loc[1:5]
```

```
## [1] "25362805..25362805" "25368440..25368440" "25368462..25368462"
## [4] "25368462..25368462" "25368462..25368462"
```

```
length(aux_loc)
```

```
## [1] 23482
```

```
aux_loc2 <- unlist(strsplit(x=aux_loc, split="\\.\\."))
```

```
aux_loc2[1:5]
```

```
## [1] "25362805" "25362805" "25368440" "25368440" "25368462"
```

```
length(aux_loc2)
```

```
## [1] 46964
```

Los elementos impares son el inicio de la locación de la mutación y los elementos pares son el final

```
cosmic$start <- aux_loc2[seq(from=1, to=nrow(cosmic)*2, by=2)]
cosmic$end <- aux_loc2[seq(from=2, to=nrow(cosmic)*2, by=2)]
```

Podemos comparar la cantidad de locaciones de variantes diferentes que tienen cada base de datos.

```
cosmic_pos <- sort(as.numeric(unique(c(cosmic$start, cosmic$end))))
gnomAD_pos <- sort(unique(gnomAD$Position))
```

Los rangos de regiones del gen son similares entre las 2, aunque un poco más grande en gnomAD.

```
range(cosmic_pos)
```

```
## [1] 25362805 25398407
```

```
range(gnomAD_pos)
```

```
## [1] 25362664 25398392
```

```
diff(range(cosmic_pos))
```

```
## [1] 35602
```

```r
diff(range(gnomAD_pos))
```

```
## [1] 35728
```

¿Cuáles son las más frecuentes en gnomad?

```r
unique(gnomAD$"Allele Count")
```

```
##  [1]      1      3      5      2    166  53595      8   8462     19     11
## [11]     75      6     51      9     78      4     21    139     12 282512
## [21]     24     17    488     62     30      7    353     10     71     14
```

En gnomAD, 3 mutaciones tienen una frecuencia de alelos mayor a 1000

```r
gnomAD[which(gnomAD$'Allele Count' > 1000),c(1:5,9,10,12,14:16)]
```

```
## # A tibble: 3 x 11
##   Chromosome Position rsID  Reference Alternate Consequence `Protein Conseq~
##        <dbl>    <dbl> <chr> <chr>     <chr>     <chr>       <chr>
## 1         12 25362777 rs11~ A         G         p.Asp173Asp c.519T>C(p.=)
## 2         12 25362854 rs12~ C         T         c.*5-9G>A   <NA>
## 3         12 25368462 rs43~ C         T         p.Arg161Arg c.483G>A(p.=)
## # ... with 4 more variables: Annotation <chr>, `Allele Count` <dbl>, `Allele
## #   Number` <dbl>, `Allele Frequency` <dbl>
```

En cosmic, 6 mutaciones están presentes en más de 1000 pacientes

```r
data.frame(sort(table(cosmic$AA_Mutation),decreasing = T))
```

```
##                 Var1 Freq
## 1             p.G12D 8131
## 2             p.G12V 5293
## 3             p.G13D 4338
## 4             p.G12C 1887
## 5             p.G12A 1344
## 6             p.G12S 1322
## 7             p.G12R  289
## 8            p.A146T  132
## 9             p.Q61H  129
## 10            p.G13C  113
## 11            p.Q61L   53
## 12            p.G13R   43
## 13           p.A146V   32
## 14            p.G13V   29
## 15            p.Q61R   29
## 16            p.G13S   28
## 17           p.R161=   22
## 18            p.G12F   21
## 19            p.Q61K   20
## 20            p.G13A   19
## 21           p.K117N   19
## 22            p.L19F   14
## 23            p.V14I   13
## 24               p.?   11
## 25            p.A59T   11
## 26            p.G13=   10
## 27           p.A146P    7
## 28            p.Q22K    6
## 29            p.G12I    5
## 30      p.A11_G12dup    4
## 31            p.A18D    4
## 32            p.D57N    4
## 33          p.G10dup    4
## 34            p.A59E    3
## 35            p.D33E    3
## 36            p.E31K    3
## 37            p.G12=    3
## 38           p.G138=    3
## 39          p.G13dup    3
## 40            p.G60D    3
## 41           p.K117E    3
## 42            p.R68S    3
## 43           p.A134V    2
## 44            p.A59G    2
## 45            p.C51R    2
## 46           p.D108N    2
## 47            p.E49K    2
## 48            p.E63K    2
## 49            p.E98*    2
## 50            p.G10E    2
## 51            p.G10V    2
## 52            p.G13E    2
## 53           p.Q150*    2
## 54            p.Q22R    2
## 55            p.T20M    2
## 56            p.Y64H    2
## 57           p.A134T    1
## 58            p.C51=    1
## 59            p.D92Y    1
## 60           p.E107K    1
## 61           p.E143K    1
## 62            p.E49*    1
## 63      p.E62_A66dup    1
## 64            p.G10R    1
## 65           p.G115E    1
## 66            p.G12L    1
## 67           p.G138E    1
## 68            p.G60=    1
## 69            p.G60V    1
## 70      p.I171Nfs*14    1
## 71            p.I84M    1
## 72           p.K147=    1
## 73           p.K147T    1
## 74             p.K5E    1
## 75       p.K88Nfs*26    1
## 76            p.L23I    1
## 77            p.L23R    1
## 78             p.L6H    1
## 79           p.M111V    1
## 80           p.N116H    1
## 81            p.P34L    1
## 82            p.Q22H    1
## 83            p.Q61E    1
## 84            p.Q61P    1
## 85            p.Q70P    1
## 86       p.R102Sfs*2    1
## 87           p.R135K    1
## 88           p.R149G    1
## 89           p.S136N    1
## 90           p.S145L    1
## 91           p.T144P    1
## 92            p.T20=    1
## 93            p.T58I    1
## 94           p.V125I    1
## 95            p.V44E    1
## 96             p.V8I    1
## 97 p.Y71_M72delinsSV    1
```

Mutaciones en cosmic presentes en más de 1000 muestras

```
cosmic[c(5174,5177,615,19968,5175,19970),c(1,2,4,6,7,9,16,17,19:21)]
```

```
## # A tibble: 6 x 11
##   Gene_Name Transcript Sample_Name AA_Mutation CDS_Mutation Tissue_Subtype_1
##   <chr>     <chr>      <chr>       <chr>       <chr>        <chr>
## 1 KRAS      ENST00000~ 2           p.G12D      c.35G>A      Colon
## 2 KRAS      ENST00000~ P-0012269-~ p.G12V      c.35G>T      Appendix
## 3 KRAS      ENST00000~ 3           p.G13D      c.38G>A      Colon
## 4 KRAS      ENST00000~ 2           p.G12C      c.34G>T      NS
## 5 KRAS      ENST00000~ AC-P15-Tum~ p.G12A      c.35G>C      Anus
## 6 KRAS      ENST00000~ P-0012100-~ p.G12S      c.34G>A      Rectum
## # ... with 5 more variables: Somatic_Status <chr>, Sample_Type <chr>,
## #   Genomic_Coordinates <chr>, start <chr>, end <chr>
```

De estas variantes, la única presente en gnomAD, está en la locación 25398284, no es tan frecuente y es diferente a las de cosmic.

```
gnomAD[which(gnomAD$Position == 25398284),c(1:5,9,10,12,14:16)]
```

```
## # A tibble: 1 x 11
##   Chromosome Position rsID  Reference Alternate Consequence `Protein Conseq~
##        <dbl>    <dbl> <chr> <chr>     <chr>     <chr>       <chr>
## 1         12 25398284 rs12~ C         T         p.Gly12Asp  p.Gly12Asp
## # ... with 4 more variables: Annotation <chr>, `Allele Count` <dbl>, `Allele
## #   Number` <dbl>, `Allele Frequency` <dbl>
```

Por otro lado, de las 3 más frecuentes en gnomAD, sólo una también está en cosmic, pero también es diferente, cambio de G a A

```
cosmic[which(cosmic$start %in% c(253627777,25362854,25368462))[1],]
```

```
## # A tibble: 1 x 21
##   Gene_Name Transcript Census_Tier_1 Sample_Name Sample_ID AA_Mutation
##   <chr>     <chr>      <chr>         <chr>           <dbl> <chr>
## 1 KRAS      ENST00000~ Yes           CC1757        2628444 p.R161=
## # ... with 15 more variables: CDS_Mutation <chr>, Primary_Tissue <chr>,
## #   Tissue_Subtype_1 <chr>, Tissue_Subtype_2 <chr>, Histology <chr>,
## #   Histology_Subtype_1 <chr>, Histology_Subtype_2 <chr>, Pubmed_Id <chr>,
## #   CGP_Study <chr>, Somatic_Status <chr>, Sample_Type <chr>, Zygosity <chr>,
## #   Genomic_Coordinates <chr>, start <chr>, end <chr>
```