Análisis de biología computacional

Situacion problema

Juan Pablo Sebastián Escobar Juárez, Carol Jatziry Rendon Guerrero, Carlos Ito Miyasaki

Propuesta

Hacer una comparación entre tres variantes del Virus SARS-CoV-2: la variante original, una variante de hace uno o dos años, y una variante reciente.

En base a las comparaciones, analizar que porcentaje del virus ha cambiado e identificar las mutaciones de nucleótido único entre las secuencias de las mismas.

La finalidad de este análisis es ver que tanto ha cambiado el virus con el tiempo, basándonos en la cantidad de mutaciones.

El SARS-COV-2 sera tomado como la variante original, y se basaran los resultados en las graficas dadas por:

“https://ourworldindata.org/grapher/covid-variants-bar”

Se decidio utilizar la variante Delta (B.1.617.2, VBM, DELTA) ya que fue la variante predominante en 2021 y la variante Omicron (BA.5, VOC, OMICRON) que ha sido una de interes recientemente.

Hipotesis

Creemos que la variante más reciente del COVID-19, Omicron, tendra un mayor número de mutaciones únicas en comparación con la variante Delta, ya que Omicron ha surgido más recientemente y ha tenido menos tiempo para evolucionar. Entre mas tiempo pasa, esperamos ver una mayor cantidad de mutaciones relevantes esperamos ver en las variantes.

Análisis de la varaiante reciente (DELTA)

```
cat("\14")
```

```
trad =    c(UUU="F", UUC="F", UUA="L", UUG="L",
            UCU="S", UCC="S", UCA="S", UCG="S",
            UAU="Y", UAC="Y", UAA="STOP", UAG="STOP",
            UGU="C", UGC="C", UGA="STOP", UGG="W",
            CUU="L", CUC="L", CUA="L", CUG="L",
            CCU="P", CCC="P", CCA="P", CCG="P",
            CAU="H", CAC="H", CAA="Q", CAG="Q",
            CGU="R", CGC="R", CGA="R", CGG="R",
            AUU="I", AUC="I", AUA="I", AUG="M",
            ACU="T", ACC="T", ACA="T", ACG="T",
            AAU="N", AAC="N", AAA="K", AAG="K",
            AGU="S", AGC="S", AGA="R", AGG="R",
            GUU="V", GUC="V", GUA="V", GUG="V",
            GCU="A", GCC="A", GCA="A", GCG="A",
            GAU="D", GAC="D", GAA="E", GAG="E",
            GGU="G", GGC="G", GGA="G", GGG="G")

library(seqinr)
```

Importamos la secuencia de referencia, y 200 secuencias de la variante.

```
original = read.fasta("original.txt")
mexa = read.fasta("delta200.fasta")
```

Definimos el dataframe

```
df = data.frame(
  Mutation = character(),
  Nucleotide = numeric(),
  Codon = character(),
  Protein = character(),
  Gene = character(),
  Sequ = character(),
  LongSequ= numeric()
)
```

Encontramos las mutaciones, utilizando el open reading frame buscamos las diferencias.

```
for (g in seq(1,length(original))){
  if (g==2 ) next
  anotaciones = attr(original[[g]], "Annot")
  atributos = unlist(strsplit(anotaciones,"\\[|\\]|:|=|\\.|\\("));
  geneName = atributos[which(atributos=="gene")+1]
  if (length(which(atributos=="join"))>0) inicioGen = as.integer(atributos[which(atributos=="join")+1])
  else inicioGen = as.integer(atributos[which(atributos=="location")+1])
  cat ("------ gene:", geneName, "inicioGen:",inicioGen,"\n")
```

```
  arnOri = as.vector(original[[g]])
  arnOri[arnOri=="t"] = "u"
  arnOri = toupper(arnOri)

  for (k in seq(g,length(mexa),12)){
    a= names(mexa)[k]
    b= length(mexa[[k]])
    arnMexa = as.vector(mexa[[k]])
    arnMexa[arnMexa=="t"] = "u"
    arnMexa = toupper(arnMexa)
    if (length(arnOri) != length(arnMexa)) next
    dif = which(arnOri != arnMexa)
    for (x in dif){
      muta = paste(arnOri[x],"to",arnMexa[x], sep="")
      inicioCodon = x - (x-1)%%3
      posGlobal = inicioCodon + inicioGen
      numCodon = as.integer((x-1)/3+1)
      codonOri = paste(arnOri[inicioCodon], arnOri[inicioCodon+1], arnOri[inicioCodon+2],sep="")
      codonMex = paste(arnMexa[inicioCodon], arnMexa[inicioCodon+1], arnMexa[inicioCodon+2],sep="")
      codonChange = paste(codonOri,"to",codonMex, sep="")
      aminoChange = paste(trad[codonOri],numCodon,trad[codonMex], sep="")
      if (!is.na(trad[codonMex])){
        newRow = list(muta, posGlobal, codonChange, aminoChange, geneName, a, b)
        df[nrow(df)+1, ] = newRow
      }
    }
  }
}
```

```
## ------ gene: ORF1ab inicioGen: 266
## ------ gene: S inicioGen: 21563
## ------ gene: ORF3a inicioGen: 25393
## ------ gene: E inicioGen: 26245
## ------ gene: M inicioGen: 26523
## ------ gene: ORF6 inicioGen: 27202
## ------ gene: ORF7a inicioGen: 27394
## ------ gene: ORF7b inicioGen: 27756
## ------ gene: ORF8 inicioGen: 27894
## ------ gene: N inicioGen: 28274
## ------ gene: ORF10 inicioGen: 29558
```

```
nrow(df)
```

```
## [1] 1776
```

```
head(df)
```

```
##   Mutation Nucleotide    Codon Protein   Gene       Sequ LongSequ
## 1     CtoU       3036 UUCtoUUU   F924F ORF1ab WGP16278.1    21291
## 2     GtoU       4182 GCUtoUCU  A1306S ORF1ab WGP16278.1    21291
## 3     CtoU       6402 CCAtoCUA  P2046L ORF1ab WGP16278.1    21291
## 4     CtoU       7125 CCUtoUCU  P2287S ORF1ab WGP16278.1    21291
```

```
## 5     CtoU       8985 GACtoGAU  D2907D ORF1ab WGP16278.1    21291
## 6     GtoU       9054 GUAtoUUA  V2930L ORF1ab WGP16278.1    21291
```

```
nrow(df)
```

```
## [1] 1776
```

Filtramos los datos.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:seqinr':
##
##     count

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dfgraph = filter(
  summarise(
    select(
      group_by(df, Protein),
      Mutation:Gene
    ),
    Mutation = first(Mutation),
    Codon = first(Codon),
    Gene = first(Gene),
    Cuenta = n()
  ),
  Cuenta>20
)

df2graph = filter(
  summarise(
    select(
      group_by(df, Sequ),
      Mutation:LongSequ
    ),
    LongSequ = first(LongSequ),
    Nmuta = n()
  ),
  Nmuta>15
```

```
)
df2graph <- cbind(df2graph, Ncodones=c((df2graph$LongSequ-df2graph$LongSequ%%3)/3  +1))
df2graph <- cbind(df2graph, Porcentaje=c(100 - df2graph$Nmuta*100/df2graph$Ncodones))

head(dfgraph)
```

```
## # A tibble: 6 x 5
##   Protein Mutation Codon    Gene   Cuenta
##   <chr>   <chr>    <chr>    <chr>   <int>
## 1 A1306S  GtoU     GCUtoUCU ORF1ab     52
## 2 A6319V  CtoU     GCUtoGUU ORF1ab     52
## 3 D2907D  CtoU     GACtoGAU ORF1ab     52
## 4 D377Y   GtoU     GAUtoUAU N          64
## 5 D63G    AtoG     GACtoGGC N          64
## 6 F924F   CtoU     UUCtoUUU ORF1ab     60
```

```
nrow(dfgraph)
```

```
## [1] 22
```

```
str(dfgraph)
```

```
## tibble [22 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Protein : chr [1:22] "A1306S" "A6319V" "D2907D" "D377Y" ...
##  $ Mutation: chr [1:22] "GtoU" "CtoU" "CtoU" "GtoU" ...
##  $ Codon   : chr [1:22] "GCUtoUCU" "GCUtoGUU" "GACtoGAU" "GAUtoUAU" ...
##  $ Gene    : chr [1:22] "ORF1ab" "ORF1ab" "ORF1ab" "N" ...
##  $ Cuenta  : int [1:22] 52 52 52 64 64 60 54 57 64 52 ...
```

```
dfgraph = as.data.frame(dfgraph)
df2graph = as.data.frame(df2graph)
str(df2graph)
```

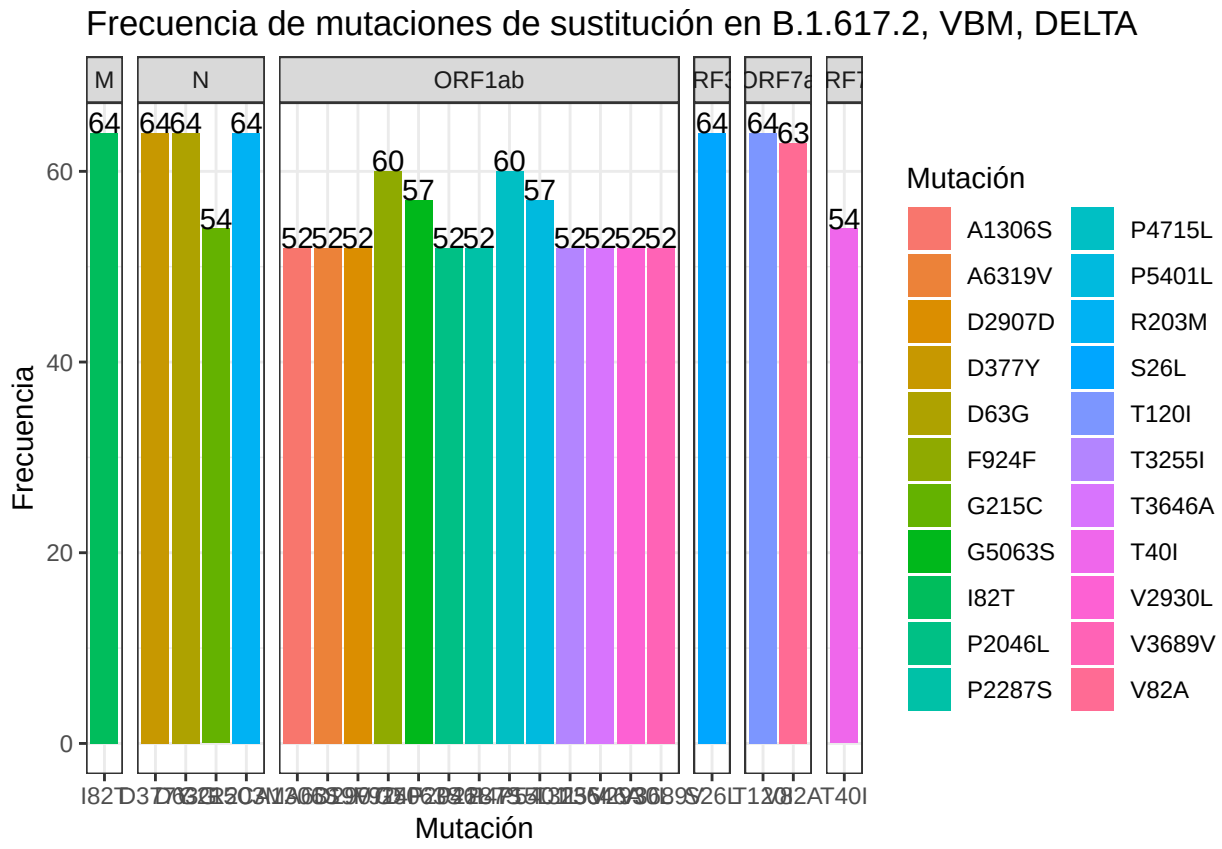
```
## 'data.frame':    51 obs. of  5 variables:
##  $ Sequ      : chr  "WGP16446.1" "WGP17281.1" "WGP62604.1" "WGP70239.1" ...
##  $ LongSequ  : num  21291 21291 21291 21291 21291 ...
##  $ Nmuta     : int  22 17 19 16 17 17 19 17 20 20 ...
##  $ Ncodones  : num  7098 7098 7098 7098 7098 ...
##  $ Porcentaje: num  99.7 99.8 99.7 99.8 99.8 ...
```

Resultados

Grafica 1

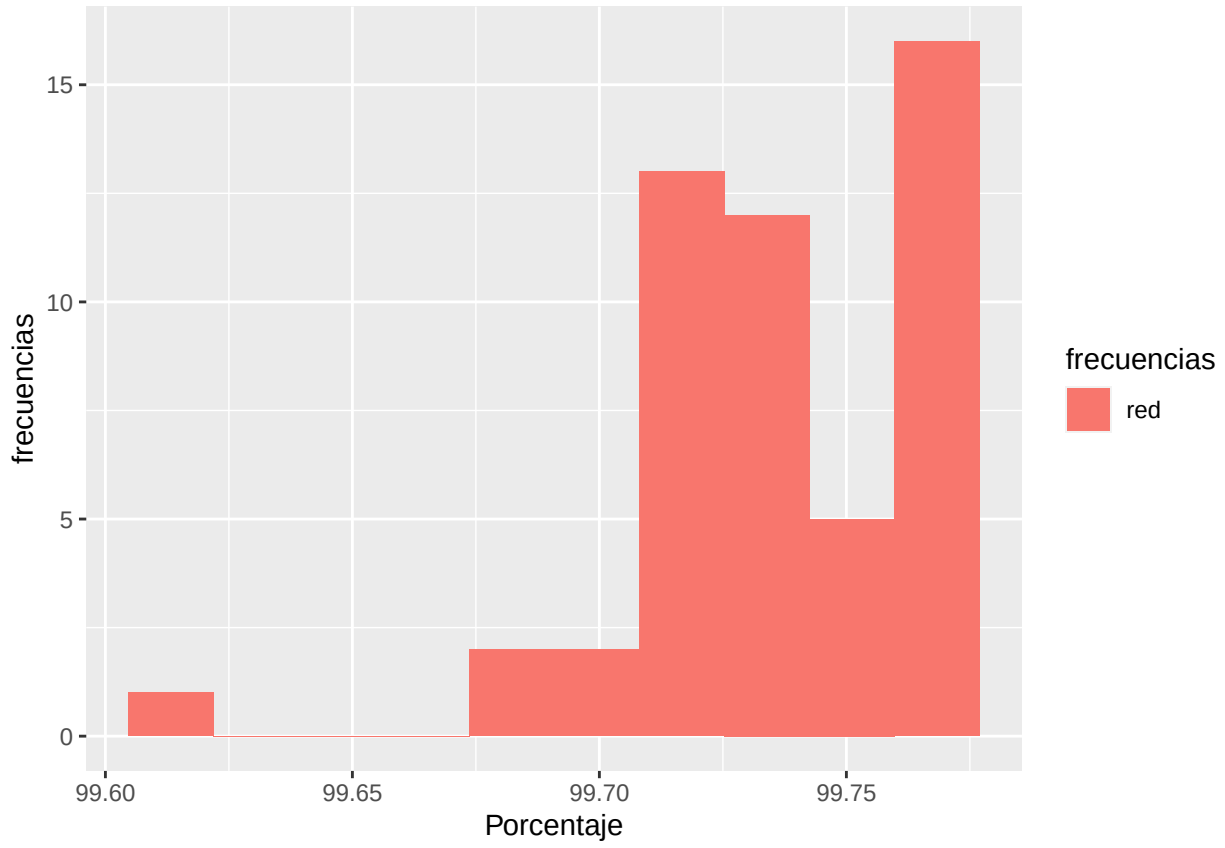
```
library(ggplot2)
p = ggplot(dfgraph)
p = p + aes(x=Protein, y=Cuenta, fill=Protein, label=Cuenta)
```

```
p = p + ggtitle("Frecuencia de mutaciones de sustitución en B.1.617.2, VBM, DELTA")
p = p + labs(x="Mutación", y="Frecuencia", fill="Mutación")
p = p + geom_bar(stat = "identity")
p = p + geom_text(stat = "identity", vjust=0)
p = p + theme_bw()
p = p + facet_grid(~Gene,scales="free", space="free_x")
print(p)
```



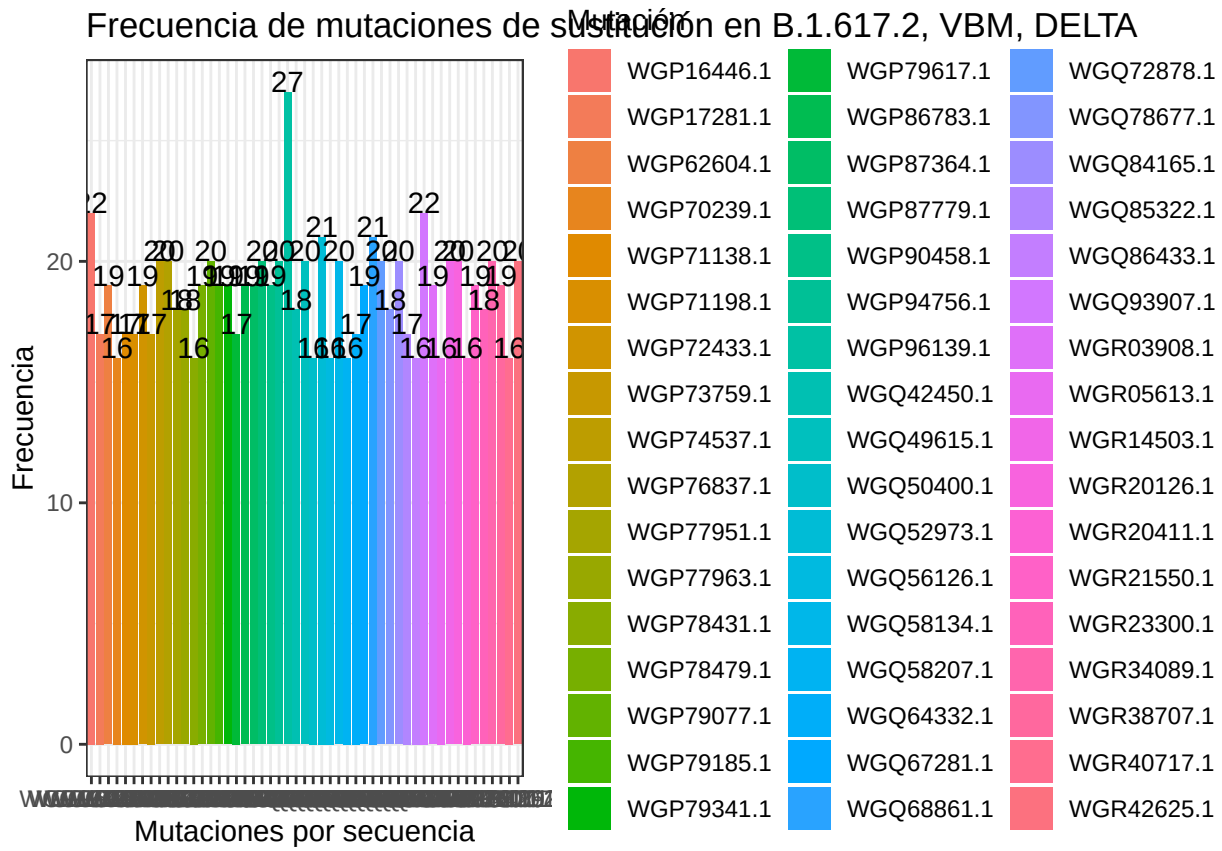
Grafica 2

```
m= ggplot(data= df2graph,
       mapping= aes(x= Porcentaje, fill= "red")) +
         geom_histogram(bins=10, alpha=1) +
        labs(tittle= 'Frecuencias de porcentajes de mutaciones por secuencia',
             fill='frecuencias',
             y='frecuencias')
print(m)
```



Grafica 3

```
q = ggplot(df2graph)
q = q + aes(x=Sequ, y=Nmuta, fill=Sequ, label=Nmuta)
q = q + ggtitle("Frecuencia de mutaciones de sustitución en B.1.617.2, VBM, DELTA")
q = q + labs(x="Mutaciones por secuencia", y="Frecuencia", fill="Mutación")
q = q + geom_bar(stat = "identity")
q = q + geom_text(stat = "identity", vjust=0)
q = q + theme_bw()
print(q)
```



Analisis de la varaiante de hace años (OMICRON)

```
cat("\14")
```

```
trad =    c(UUU="F", UUC="F", UUA="L", UUG="L",
            UCU="S", UCC="S", UCA="S", UCG="S",
            UAU="Y", UAC="Y", UAA="STOP", UAG="STOP",
            UGU="C", UGC="C", UGA="STOP", UGG="W",
            CUU="L", CUC="L", CUA="L", CUG="L",
            CCU="P", CCC="P", CCA="P", CCG="P",
            CAU="H", CAC="H", CAA="Q", CAG="Q",
            CGU="R", CGC="R", CGA="R", CGG="R",
            AUU="I", AUC="I", AUA="I", AUG="M",
            ACU="T", ACC="T", ACA="T", ACG="T",
            AAU="N", AAC="N", AAA="K", AAG="K",
            AGU="S", AGC="S", AGA="R", AGG="R",
            GUU="V", GUC="V", GUA="V", GUG="V",
            GCU="A", GCC="A", GCA="A", GCG="A",
            GAU="D", GAC="D", GAA="E", GAG="E",
            GGU="G", GGC="G", GGA="G", GGG="G")

library(seqinr)
```

Importamos la secuencia de referencia, y 200 secuencias de la variante.

```
original = read.fasta("original.txt")
mexa = read.fasta("omicron200.fasta")
```

Definimos el dataframe

```
df = data.frame(
  Mutation = character(),
  Nucleotide = numeric(),
  Codon = character(),
  Protein = character(),
  Gene = character(),
  Sequ = character(),
  LongSequ= numeric()
)
```

Encontramos las mutaciones, utilizando el open reading frame buscamos las diferencias.

```
for (g in seq(1,length(original))){
  if (g==2 ) next
  anotaciones = attr(original[[g]], "Annot")
  atributos = unlist(strsplit(anotaciones,"\\[|\\]|:|=|\\.|\\("));
  geneName = atributos[which(atributos=="gene")+1]
  if (length(which(atributos=="join"))>0) inicioGen = as.integer(atributos[which(atributos=="join")+1])
  else inicioGen = as.integer(atributos[which(atributos=="location")+1])
  cat ("------ gene:", geneName, "inicioGen:",inicioGen,"\n")
```

```
  arnOri = as.vector(original[[g]])
  arnOri[arnOri=="t"] = "u"
  arnOri = toupper(arnOri)

  for (k in seq(g,length(mexa),12)){
    a= names(mexa)[k]
    b= length(mexa[[k]])
    arnMexa = as.vector(mexa[[k]])
    arnMexa[arnMexa=="t"] = "u"
    arnMexa = toupper(arnMexa)
    if (length(arnOri) != length(arnMexa)) next
    dif = which(arnOri != arnMexa)
    for (x in dif){
      muta = paste(arnOri[x],"to",arnMexa[x], sep="")
      inicioCodon = x - (x-1)%%3
      posGlobal = inicioCodon + inicioGen
      numCodon = as.integer((x-1)/3+1)
      codonOri = paste(arnOri[inicioCodon], arnOri[inicioCodon+1], arnOri[inicioCodon+2],sep="")
      codonMex = paste(arnMexa[inicioCodon], arnMexa[inicioCodon+1], arnMexa[inicioCodon+2],sep="")
      codonChange = paste(codonOri,"to",codonMex, sep="")
      aminoChange = paste(trad[codonOri],numCodon,trad[codonMex], sep="")
      if (!is.na(trad[codonMex])){
        newRow = list(muta, posGlobal, codonChange, aminoChange, geneName, a, b)
        df[nrow(df)+1, ] = newRow
      }
    }
  }
}
```

```
## ------ gene: ORF1ab inicioGen: 266
## ------ gene: S inicioGen: 21563
## ------ gene: ORF3a inicioGen: 25393
## ------ gene: E inicioGen: 26245
## ------ gene: M inicioGen: 26523
## ------ gene: ORF6 inicioGen: 27202
## ------ gene: ORF7a inicioGen: 27394
## ------ gene: ORF7b inicioGen: 27756
## ------ gene: ORF8 inicioGen: 27894
## ------ gene: N inicioGen: 28274
## ------ gene: ORF10 inicioGen: 29558
```

```
nrow(df)
```

```
## [1] 472
```

```
head(df)
```

```
##   Mutation Nucleotide    Codon Protein  Gene       Sequ LongSequ
## 1     CtoU      25583 ACCtoACU    T64T ORF3a WGP80321.1      828
## 2     CtoU      26060 ACUtoAUU   T223I ORF3a WGP80321.1      828
## 3     CtoU      25583 ACCtoACU    T64T ORF3a WGP89349.1      828
## 4     CtoU      26060 ACUtoAUU   T223I ORF3a WGP89349.1      828
```

```
## 5     CtoU      25583 ACCtoACU    T64T ORF3a WGP93866.1      828
## 6     CtoU      26060 ACUtoAUU   T223I ORF3a WGP93866.1      828
```

```
nrow(df)
```

```
## [1] 472
```

Filtramos los datos.

```
library(dplyr)
dfgraph = filter(
  summarise(
    select(
      group_by(df, Protein),
      Mutation:Gene
    ),
    Mutation = first(Mutation),
    Codon = first(Codon),
    Gene = first(Gene),
    Cuenta = n()
  ),
  Cuenta>20
)

df2graph = filter(
  summarise(
    select(
      group_by(df, Sequ),
      Mutation:LongSequ
    ),
    LongSequ = first(LongSequ),
    Nmuta = n()
  ),
  Nmuta>15
)
df2graph <- cbind(df2graph, Ncodones=c((df2graph$LongSequ-df2graph$LongSequ%%3)/3  +1))
df2graph <- cbind(df2graph, Porcentaje=c(100 - df2graph$Nmuta*100/df2graph$Ncodones))

head(dfgraph)
```

```
## # A tibble: 6 x 5
##   Protein Mutation Codon    Gene  Cuenta
##   <chr>   <chr>    <chr>    <chr>  <int>
## 1 A63T    GtoA     GCUtoACU M         30
## 2 D3N     GtoA     GAUtoAAU M         24
## 3 L18L    CtoU     CUAtoUUA ORF7b     31
## 4 T223I   CtoU     ACUtoAUU ORF3a     31
## 5 T64T    CtoU     ACCtoACU ORF3a     31
## 6 T9I     CtoU     ACAtoAUA E         31
```

```
nrow(dfgraph)
```

```
## [1] 6
```

```
str(dfgraph)
```

```
## tibble [6 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Protein : chr [1:6] "A63T" "D3N" "L18L" "T223I" ...
##  $ Mutation: chr [1:6] "GtoA" "GtoA" "CtoU" "CtoU" ...
##  $ Codon   : chr [1:6] "GCUtoACU" "GAUtoAAU" "CUAtoUUA" "ACUtoAUU" ...
##  $ Gene    : chr [1:6] "M" "M" "ORF7b" "ORF3a" ...
##  $ Cuenta  : int [1:6] 30 24 31 31 31 31
```

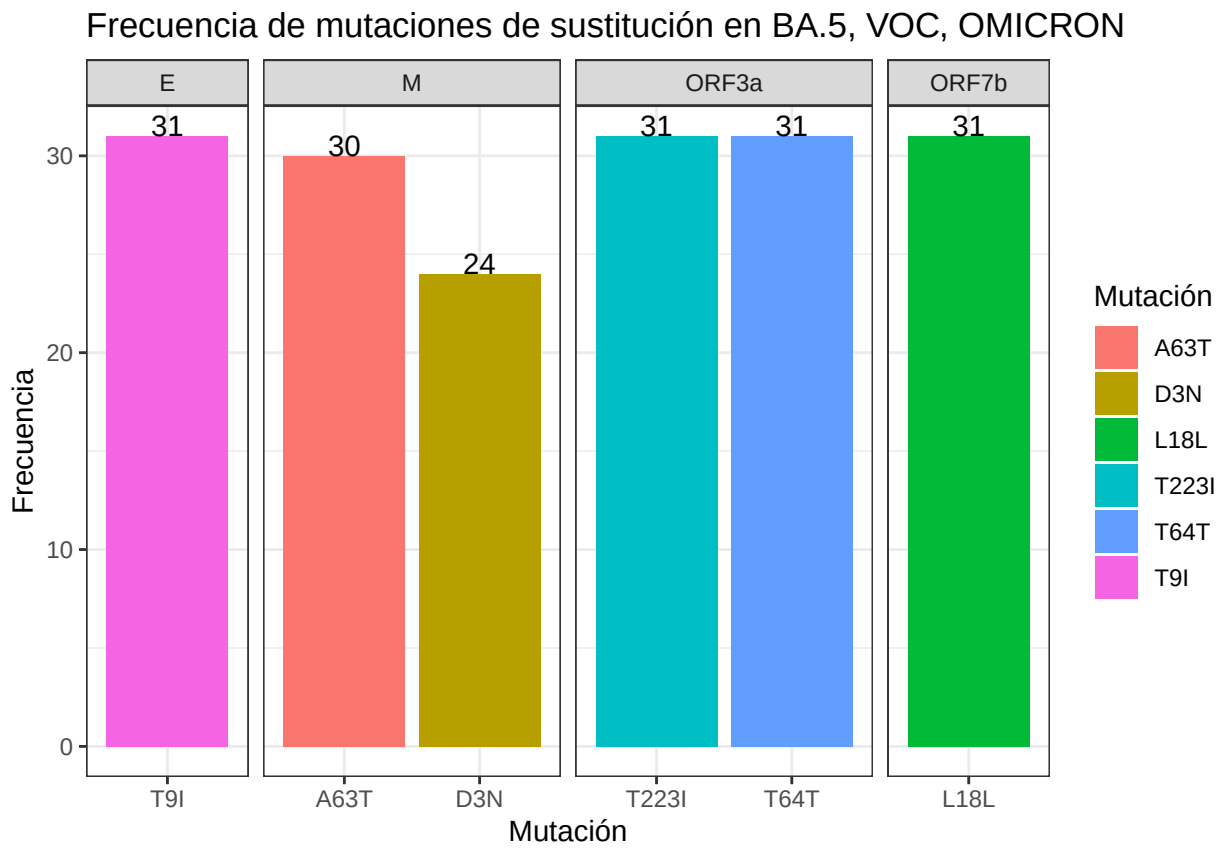
```
dfgraph = as.data.frame(dfgraph)
df2graph = as.data.frame(df2graph)
str(df2graph)
```

```
## 'data.frame':    1 obs. of  5 variables:
##  $ Sequ      : chr "WBD99210.1"
##  $ LongSequ  : num 366
##  $ Nmuta     : int 246
##  $ Ncodones  : num 123
##  $ Porcentaje: num -100
```

Resultados

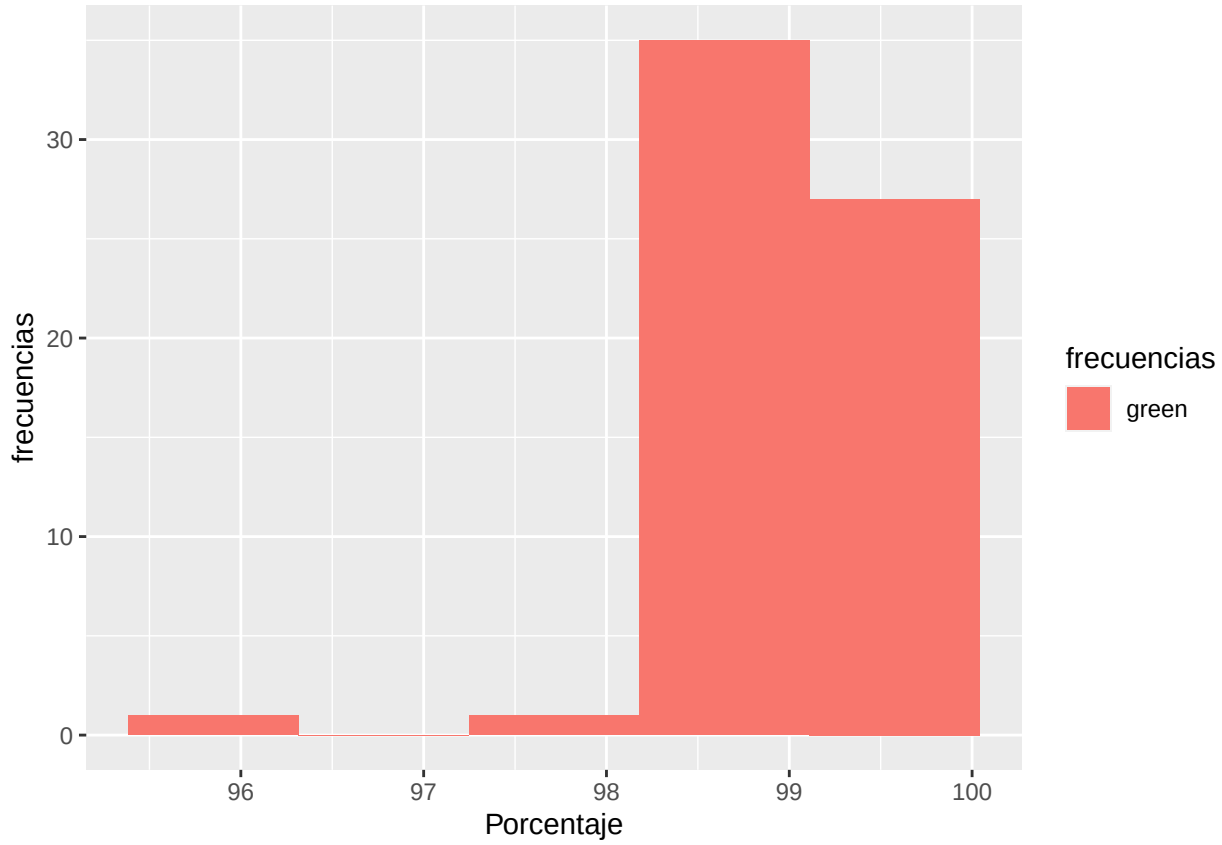
Grafica 1

```
library(ggplot2)
p = ggplot(dfgraph)
p = p + aes(x=Protein, y=Cuenta, fill=Protein, label=Cuenta)
p = p + ggtitle("Frecuencia de mutaciones de sustitución en BA.5, VOC, OMICRON")
p = p + labs(x="Mutación", y="Frecuencia", fill="Mutación")
p = p + geom_bar(stat = "identity")
p = p + geom_text(stat = "identity", vjust=0)
p = p + theme_bw()
p = p + facet_grid(~Gene,scales="free", space="free_x")
print(p)
```



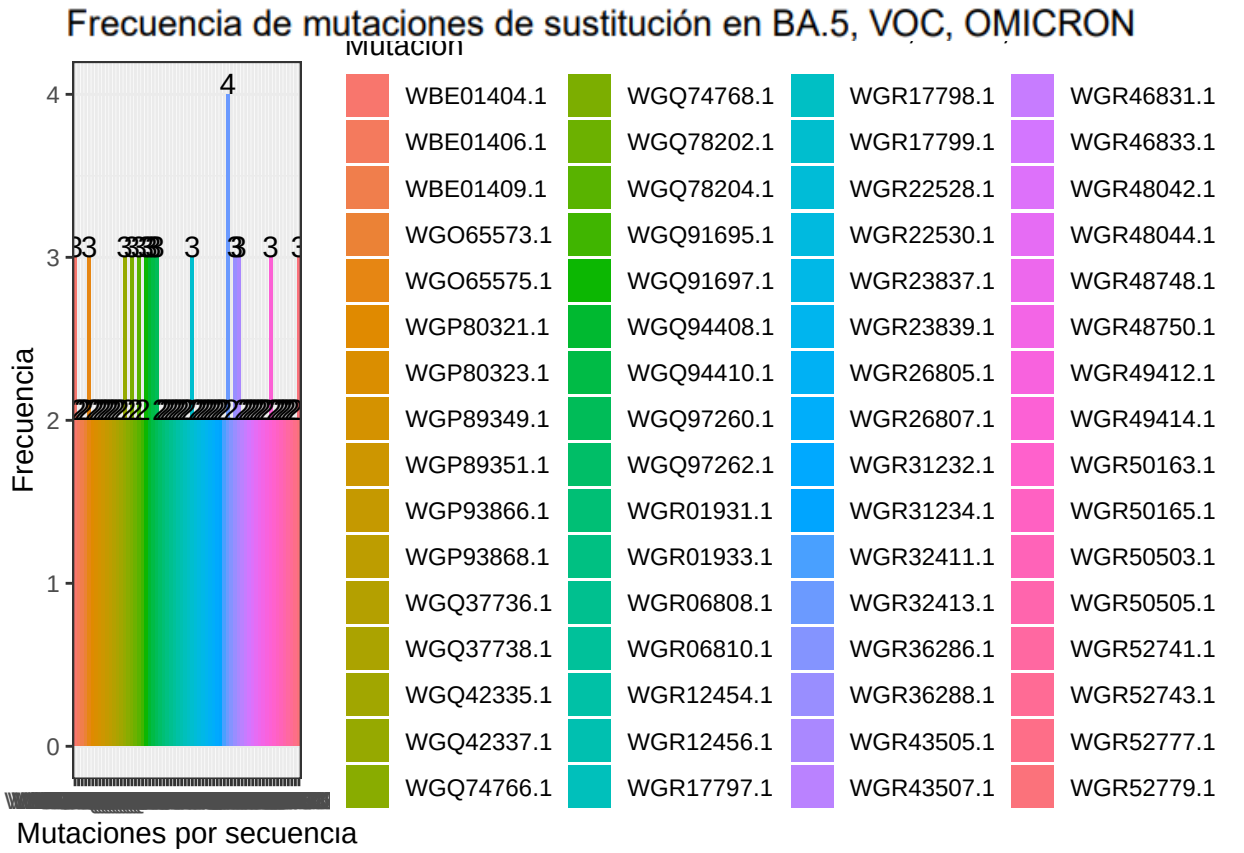
Grafica 2

```
m= ggplot(data= df2graph,
       mapping= aes(x= Porcentaje, fill= "red")) +
         geom_histogram(bins=10, alpha=1) +
        labs(tittle= 'Frecuencias de porcentajes de mutaciones por secuencia',
             fill='frecuencias',
             y='frecuencias')
print(m)
```




```
q = ggplot(df2graph)
q = q + aes(x=Sequ, y=Nmuta, fill=Sequ, label=Nmuta)
q = q + ggtitle("Frecuencia de mutaciones de sustitución en B.1.617.2, VBM, DELTA")
q = q + labs(x="Mutaciones por secuencia", y="Frecuencia", fill="Mutación")
q = q + geom_bar(stat = "identity")
q = q + geom_text(stat = "identity", vjust=0)
q = q + theme_bw()

print(q)
```




```
print("FIN")
```

```
## [1] "FIN"
```

Conclusion

Basandonos en los resultados obtenidos, podemos concluir que la cantidad de mutaciones encontradas en las variantes Delta y Omicron no difiere significativamente. La variante Omicron no parece tener muchas más mutaciones relevantes que la variante Delta, a pesar de ser más reciente. Creemos que esto puede ser relevante para la comprensión de la evolución del virus y su capacidad de propagación y transmisión.