20211212_Final_Project_Vargas

Benni Vargas

12/12/2021

INTRODUCTION SARS-CoV-2 (COVID-19) is a virus that has become an unprecedented worldwide problem that affects everyone's way of life–from survival, politics, economy, mental health, and more. COVID-19 has taken millions of lives. Bioinformatics methods are necessary to study the virus in order to expedite the development of therapeutics including vaccines, drugs, and antibodies against it. The GC dinucleotide content of viruses are common in molecular virology to predict the stability and infectivity of viruses. It has been shown in HIV that CG dinucleotides are suppressed in order to avoid host antiviral defense. (Takata et al., CG-dinucleotide suppression enables antiviral defense targeting non-self RNA. Nature. 2017 Oct 5; 550(7674): 124–127. doi: 10.1038/nature24039). SARS-CoV-2, also an RNA virus just like HIV-1, may possess unique GC nucleotide compositions that may be responsible for its success as a pandemic virus. The first goal is to assess if SARS-CoV-2 along with other epidemic/famous coronaviruses such as MERS and SARS have a low GC content of which may serve as an indicator of its success as a pandemic-causing virus. The second goal is described next. The Spike protein of coronaviruses is the target of interest for vaccine and drug development. It is highly mutable and the high mutation rate has been a concern due to the potential ability to escape vaccine antibody protection. Understanding the codon usage of the Spike gene of pandemic/famous coronaviruses will enable prediction of susceptibility to infection, understanding of the evolution of coronaviruses, and development of effective vaccines and other therapeutics against the Spike protein. Thus, the second goal is to compare the codon usage of SARS-CoV-2, omicron variant, MERS, SARS, and coronavirus B814 to show how codon optimization has evolved and how it correlates with the pandemic/epidemic nature of these coronaviruses

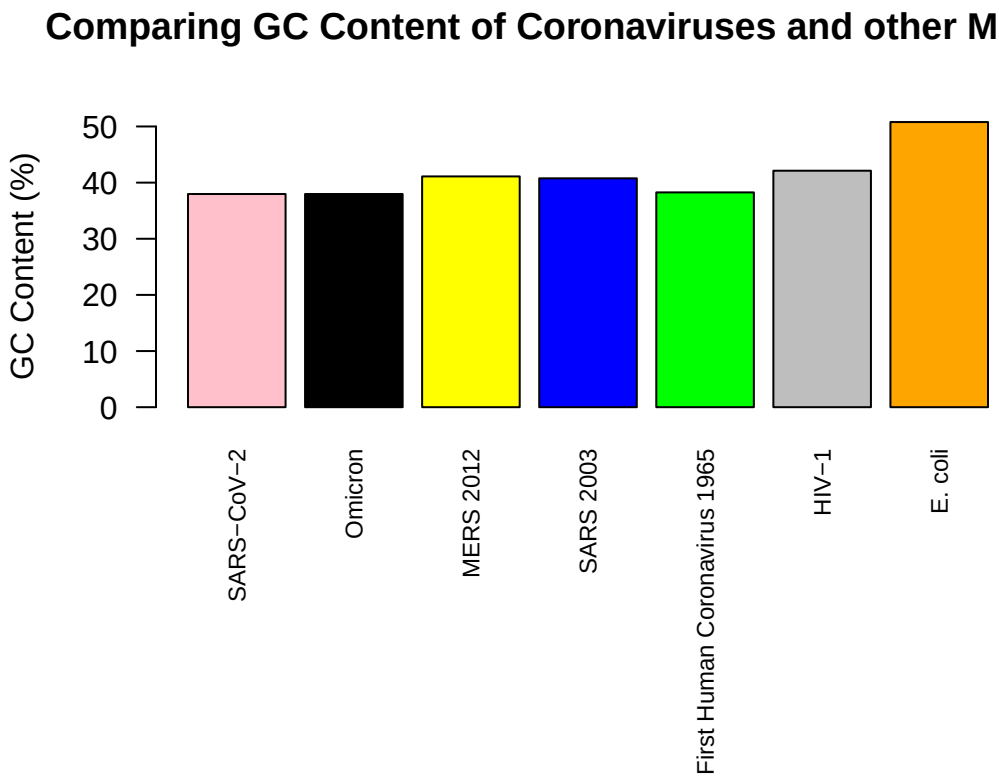
METHODS Coronavirus FASTA sequences were downloaded from NCBI. The only library that was used was the Biostrings package. Each of the code blocks contains error checking (e.g. stopifnot). For GC content analysis, I used Biostrings to compute the alphabet frequency of GC and stored the results for each virus in a list. For controls, I included HIV-1 and E. coli as comparison. Next, I used Biostrings to compute the windows frequency (100bp windwos) of the GC content of COVID-19 versus the B814 coronavirus strain (first isolated human coronavirus). For codon usage analysis, I initialized a matrix of codons and reserved each row for each virus. The following coronaviruses were analyzed: SARS-CoV-2, Omicron, MERS 2012, SARS 2003, and B814 1965. I used for loops for each virus to store codon usage in the matrix. I then plotted the relative codon frequency for each virus in a barplot.

```
library(Biostrings)
```

```
#In this block, I am comparing the GC Content of coronaviruses with other microbes

#Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
#https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta
covid <- readDNAStringSet("H:\\ID\\Shared_Mellors_Cremer\\Benni\\BIOINF 2018\\Final Project\\sequence.fa
covidFreq <- alphabetFrequency(covid, as.prob=T)

#prepare GC_list for barplot
GC_list <- numeric()
```

```
GC_list[1] <- sum(covidFreq[2:3]) * 100

#B.1.1.529 (Omicron)
#https://www.ncbi.nlm.nih.gov/nuccore/2156809762
omicron <- readDNAStringSet("H:\\ID\\Shared_Mellors_Cremer\\Benni\\BIOINF 2018\\Final Project\\omicron.
omicronFreq <- alphabetFrequency(omicron, as.prob=T)

GC_list[2] <- sum(omicronFreq[2:3]) * 100

#MERS 2012
#https://www.ncbi.nlm.nih.gov/nuccore/MH454272.1?report=fasta
MERS <- readDNAStringSet("H:\\ID\\Shared_Mellors_Cremer\\Benni\\BIOINF 2018\\Final Project\\MERS.fasta"
MERSFreq <- alphabetFrequency(MERS, as.prob=T)
#another striking observation - an endemic virus has higher GC than SARS-CoV-2

GC_list[3] <- sum(MERSFreq[2:3]) * 100

#https://www.science.org/doi/10.1126/science.1085953?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rf
#https://www.ncbi.nlm.nih.gov/nuccore/AY274119.3?report=fasta

SARS2003 <- readDNAStringSet("H:\\ID\\Shared_Mellors_Cremer\\Benni\\BIOINF 2018\\Final Project\\SARS200
SARS2003Freq <- alphabetFrequency(SARS2003, as.prob=T)

GC_list[4] <- sum(SARS2003Freq[2:3]) * 100


#https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7098031/
# ---> Article says, "The first HCoV-229E strain, B814, was isolated from the nasal discharge of patien
#https://www.ncbi.nlm.nih.gov/nuccore/NC_002645.1

originalCoronavirus <- readDNAStringSet("H:\\ID\\Shared_Mellors_Cremer\\Benni\\BIOINF 2018\\Final Proje
originalCoronavirusFreq <- alphabetFrequency(originalCoronavirus, as.prob=T)

GC_list[5] <- sum(originalCoronavirusFreq[2:3]) * 100

#HIV-1
#https://www.ncbi.nlm.nih.gov/nuccore/AF033819.3?report=fasta

HIV1 <- readDNAStringSet("H:\\ID\\Shared_Mellors_Cremer\\Benni\\BIOINF 2018\\Final Project\\HIV-1.fasta
HIV1Freq <- alphabetFrequency(HIV1, as.prob=T)

GC_list[6] <- sum(HIV1Freq[2:3]) * 100

#E. coli
Ecoli <- readDNAStringSet("H:\\ID\\Shared_Mellors_Cremer\\Benni\\BIOINF 2018\\Ecoli_Genome.fas.gz")
EcoliFreq <- alphabetFrequency(Ecoli, as.prob=T)

GC_list[7] <- sum(EcoliFreq[2:3]) * 100
names(GC_list) <- c("SARS-CoV-2", "Omicron", "MERS 2012", "SARS 2003", "First Human Coronavirus 1965",

stopifnot(length(GC_list) == 7)

par(mar=c(10,5,5,5))
```

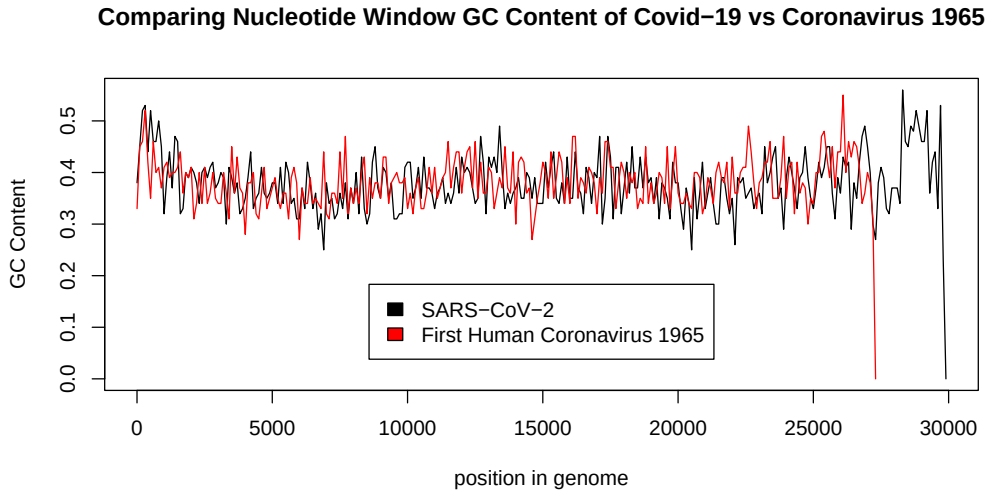
```
barplot(GC_list, cex.names= 0.75, las=2, ylab="GC Content (%)", col=c("pink", "black", "yellow", "blue"
```



RESULTS (part 1 of 3) The input data were FASTA sequences of coronaviruses, HIV-1, and E. coli downloaded from NCBI (E. coli was obtained through Dr. Erik Wright in BIOINF 2018 Fall 2021). The output is a barplot of the GC content (%) versus the associated microbe. Eukaryotic and prokaryotic organisms vary in terms of their GC content. RNA viruses, however, may have a distinct GC dinucleotide pattern. In the above figure, E. coli has a GC content of ~50%. HIV-1, where it was discovered in Nature 2017 to evade host antiviral defense by suppressing GC dinucleotide content, has a GC content of ~40%. It is striking to find that SARS-CoV-2, along with omicron and the original first human identified coronavirus, has a lower GC content of ~38%.

```
#In this block, I am looking at a sliding window (100bp) of GC content of Covid-19 versus first isolate

covidS <- seq(1, width(covid)[1], 100)
covidWindows <- extractAt(covid[[1]], IRanges(covidS, c(covidS[-1] - 1, width(covid))))
covidWindowsFreq <- alphabetFrequency(covidWindows, as.prob=T)

originalCoronavirusS <- seq(1, width(originalCoronavirus)[1], 100)
windows2 <- extractAt(originalCoronavirus[[1]], IRanges(originalCoronavirusS, c(originalCoronavirusS[-1
originalCoronavirusWindowsFreq <- alphabetFrequency(windows2, as.prob=T)

stopifnot(!is.na(covidWindowsFreq))
stopifnot(!is.na(originalCoronavirusFreq))
```

```
par(mar=c(5,8,5,7))
matplot(covidS, covidWindowsFreq[, 2] + covidWindowsFreq[, 3], type="l", xlab="position in genome", yla
matlines(originalCoronavirusS, originalCoronavirusWindowsFreq[,2] + originalCoronavirusWindowsFreq[,3],
legend("bottom",
       legend = c("SARS-CoV-2", "First Human Coronavirus 1965"),
       fill = c("black", "red"), inset=c(0, 0.1))
```

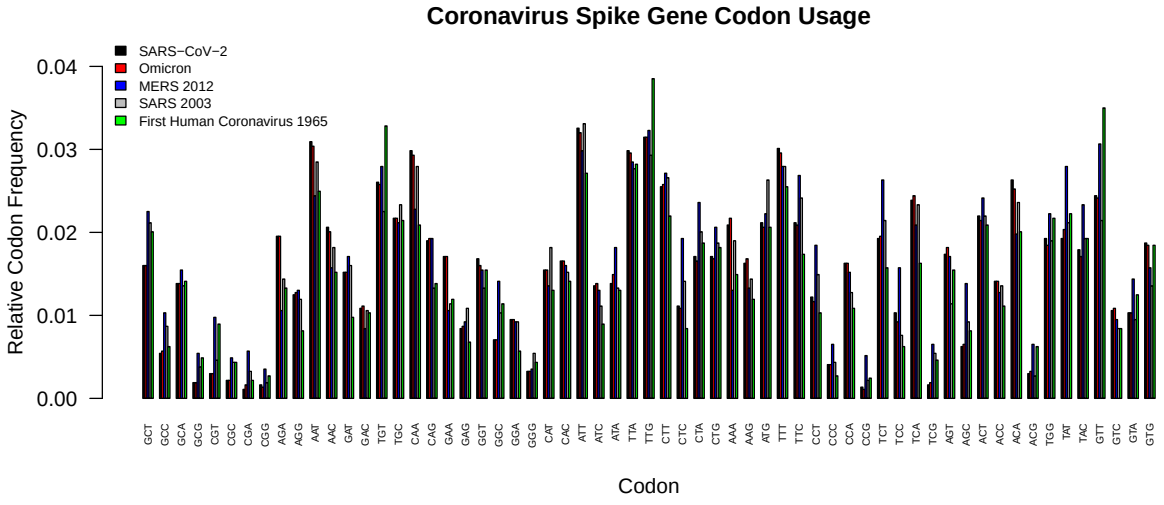


RESULTS (part 2 of 3) The input data were FASTA sequences of SARS-CoV-2 and B814 coronavirus (first human isolated coronavirus). The output is a matplot of the 100bp nucleotide window GC content versus the associated coronavirus. From the above figure, we see that SARS-CoV-2 has a larger genome approx. 3kb larger than coronavirus B814. There is also a trailing off at the 3'end of GC content, a trend that is a common among genomes. The 5' end has the highest GC content. The spread and variation of GC content from the matplot between the two viruses looks similar and are consistently lower than 0.5.

```
#In this block, I am comparing the codon usage of the Spike gene for various famous coronaviruses of in

#initialize matrix of codons
m1 <- matrix(0L, nrow = 5, ncol=64)
codonPattern <- c("GCT", "GCC", "GCA", "GCG", "CGT", "CGC", "CGA", "CGG", "AGA", "AGG", "AAT", "AAC", "
colnames(m1) <- codonPattern

r1 <- integer(length(codonPattern))

#store codon usage of Spike gene for each strain in codon matrix

#covidSpike gene 21563..25384
covidSpike <- subseq(covid, 21563, 25384)
for (i in seq_along(r1)) {
  stopifnot(!is.na(codonPattern))
  stopifnot(nchar(covidSpike) < 5000)
  if (grepl(codonPattern[i], covidSpike) == TRUE) {
    value <- length(unlist(gregexpr(codonPattern[i], covidSpike)))
    m1[1, i] <- m1[1, i] + value
  } else {
```

```
    next
  }
}


#omicronSpike gene 21497..25309
omicronSpike <- subseq(omicron, 21497, 25309)
omicronSpike

for (i in seq_along(r1)) {
  stopifnot(!is.na(codonPattern))
  stopifnot(nchar(omicronSpike) < 5000)
  if (grepl(codonPattern[i], omicronSpike) == TRUE) {
    value <- length(unlist(gregexpr(codonPattern[i], omicronSpike)))
    m1[2, i] <- m1[2, i] + value
  } else {
    next
  }
}

#MERSSpike gene 21456..25517
MERSSpike <- subseq(MERS, 21456, 25517)
MERSSpike

for (i in seq_along(r1)) {
  stopifnot(!is.na(codonPattern))
  stopifnot(nchar(MERSSpike) < 5000)
  if (grepl(codonPattern[i], MERSSpike) == TRUE) {
    value <- length(unlist(gregexpr(codonPattern[i], MERSSpike)))
    m1[3, i] <- m1[3, i] + value
  } else {
    next
  }
}

#SARS2003Spike gene 21492..25259
SARS2003Spike <- subseq(SARS2003, 21492, 25259)
SARS2003Spike

for (i in seq_along(r1)) {
  stopifnot(!is.na(codonPattern))
  stopifnot(nchar(SARS2003Spike) < 5000)
  if (grepl(codonPattern[i], SARS2003Spike) == TRUE) {
    value <- length(unlist(gregexpr(codonPattern[i], SARS2003Spike)))
    m1[4, i] <- m1[4, i] + value
  } else {
    next
  }
}

#originalCoronavirusSpike gene  20570..24091
originalCoronavirusSpike <- subseq(originalCoronavirus, 20570, 24091)
```

```r
for (i in seq_along(r1)) {
  stopifnot(!is.na(codonPattern))
  stopifnot(nchar(originalCoronavirusSpike) < 5000)
  if (grepl(codonPattern[i], originalCoronavirusSpike) == TRUE) {
    value <- length(unlist(gregexpr(codonPattern[i], originalCoronavirusSpike)))
    m1[5, i] <- m1[5, i] + value
  } else {
    next
  }
}

#normalize matrix by frequency of each codon per position
m2 <- matrix(0L, nrow=nrow(m1), ncol=ncol(m1))
colnames(m2) <- codonPattern

r2 <- numeric(nrow(m2))
for (i in seq_along(r2)) {
  m2[i,] <- m1[i,]/(sum(m1[1,]))
  stopifnot(m2[i,] < 2 | m2[i,] > 0)
}

#getting rid of stop codons
m3 <- m2[,1:61, drop=F]

colorMe <- c("black", "red", "blue", "gray", "green")

stopifnot(!is.na(m3))

barplot(m3, width=0.5, las=2, ylim=c(0, 0.04), beside=T, space=c(0, 3), cex.names=0.5, col=colorMe, yla
legend("topleft",    bty="n",                                    # Add legend to barplot
       legend = c("SARS-CoV-2", "Omicron", "MERS 2012", "SARS 2003", "First Human Coronavirus 1965"),
       fill = colorMe, xpd=T, cex=0.7, inset=c(0, -0.1))
```



RESULTS (part 3 of 3) The input data were FASTA sequences of SARS-CoV-2, Omicron, MERS 2012, SARS 2003, and B814 coronavirus (first human isolated coronavirus). FASTA sequences were subject to subseq in order to extract the Spike gene at positions annotated in the code block. The output is a barplot of the codon usage of the Spike gene of the viruses omitting the stop codons. There is low usage for all viruses for the following codons: CCC, GGG, CGG, CGC, GCG, ACG. There is high usage for all viruses for the following codons: ATT, TTG, AAT, TGT, GTT, ATT, and TTG. It is interesting to note for B814 there is high usage in TTG and GTT versus the other more recent coronaviruses.

DISCUSSION/CONCLUSION The high codon usage for the Spike gene of TTG, AAT, TGT, GTT, ATT, and TTG correlates with what we have observed with the low GC content of the coronaviruses. It is also not surprising that due to the overall low GC content that the least used codons were CCC, GGG, CGG, CGC, GCG, and ACG. It is known that HIV-1 suppresses CG dinucleotide content to evade host antiviral defense (Takata et al., Nature, 2017). I have verified this in my analysis and observed HIV-1 has a GC content of ~40%. It was exciting to see that SARS-CoV-2 and Omicron both have GC contents of ~38%, which is less than the ~40% observed for the other epidemic coronaviruses–MERS 2012 and SARS 2003. This low GC content of SARS-CoV-2 and omicron as well as coronaviruses in general may be a key reason why they are successful as pandemic/epidemic viruses as this allows them to evade host antiviral defense. I learned many things working on this project. I learned how to work with the NCBI databases, exploring the literature, as well as working with the FASTA files and Biostrings package. I am traditionally a wet lab virologist. I am a PhD candidate studying host mechanisms regulating HIV-1 latency in the laboratory of Dr. Nicolas Sluis-Cremer. With the bioinformatics knowledge I have gained from working on this project, I can combine my expertise in wet lab virology with R programming and bioinformatics to solve complex problems to better understand bothersome viruses of interest, such as COVID-19. The knowledge obtained from this will allow better development of therapeutics against these threats.