Comparison of DENV and ZIKV genomes with SeqinR package

Ruth Gómez Graciani, Mercè Alemany Chavarria

26 de octubre de 2017

Introduction

Zika virus (ZIKV) and Dengue virus (DENV) are two single-stranded RNA viruses affecting humans in the genus *Flavivirus*. They are very similar in their transmission -via different species of *Aedes* mosquitoes- although ZIKV can also be transmitted through sex and blood transfusions. Early symptoms are also similar -mostly fever and also several kinds of pain-, but differ in their medium-to-long term consequences; infection by ZIKV lasts usually few days but can derive in birth defects when affecting pregnant women while DENV infection can last several weeks and derive in worse symptoms as haemorrhagic fevers that can be fatal. Even though there are no effective vaccines or antiviral drugs against these viruses, in both cases early diagnosis is determinant to give the best treatment and prevent their spreading.

The analysis of their sequences has been useful to better know their mechanisms of infection and the evolution of the different serotypes, as well as to develop diagnostic tests. Here, we explore the different options that the R package `seqinr` offers to distinguish them.

Analysis

Construction and application of the functions:

```
library (seqinr)
library(reshape2)
library(ggplot2)
library(zoo)

FAS<-read.fasta('sequence.fasta')

seqGC<-function(seq){
  allGC<-rollapply(seq,width=500,by=10,FUN=GC)
  counter<-c(1:length(seq))
  positions<-rollapply(counter, width=500,by=10,FUN=function(x){
    median(x)
  })
   return(data.frame(allGC, positions))

  }


General_study<-function(sequence,name){
  len<-length(sequence) #1
  nuc_count<-count(sequence, 1) #2
  GC_cont<-GC(sequence) #3
  Dimer_z<-as.data.frame(zscore(sequence, modele = 'codon')) #4
  colnames(Dimer_z)<-c('Dimers', name)
  Codon_us_ratio<-as.data.frame(uco(sequence)/sum(uco(sequence))) #5
  colnames(Codon_us_ratio)<-c('Codons', name)
```

```r
  Windows<-seqGC(sequence) #6
  return(list(len,nuc_count, GC_cont,Dimer_z,Codon_us_ratio, Windows))

  }

Dengue<-General_study(FAS[[1]],"Dengue")
Zika<-General_study(FAS[[2]], "Zika")
```

General description of the two sequences:

Dengue general descripton:

- Length of the sequence

```
## [1] 10735
```

- Nucleotide count of the sequence

```
##
##    a    c    g    t
## 3426 2240 2770 2299
```

- GC content of the sequence

```
## [1] 0.4666977
```

Zika general descripton:

- Length of the sequence

```
## [1] 10794
```

- Nucleotide count of the sequence

```
##
##    a    c    g    t
## 2991 2359 3139 2305
```

- GC content of the sequence

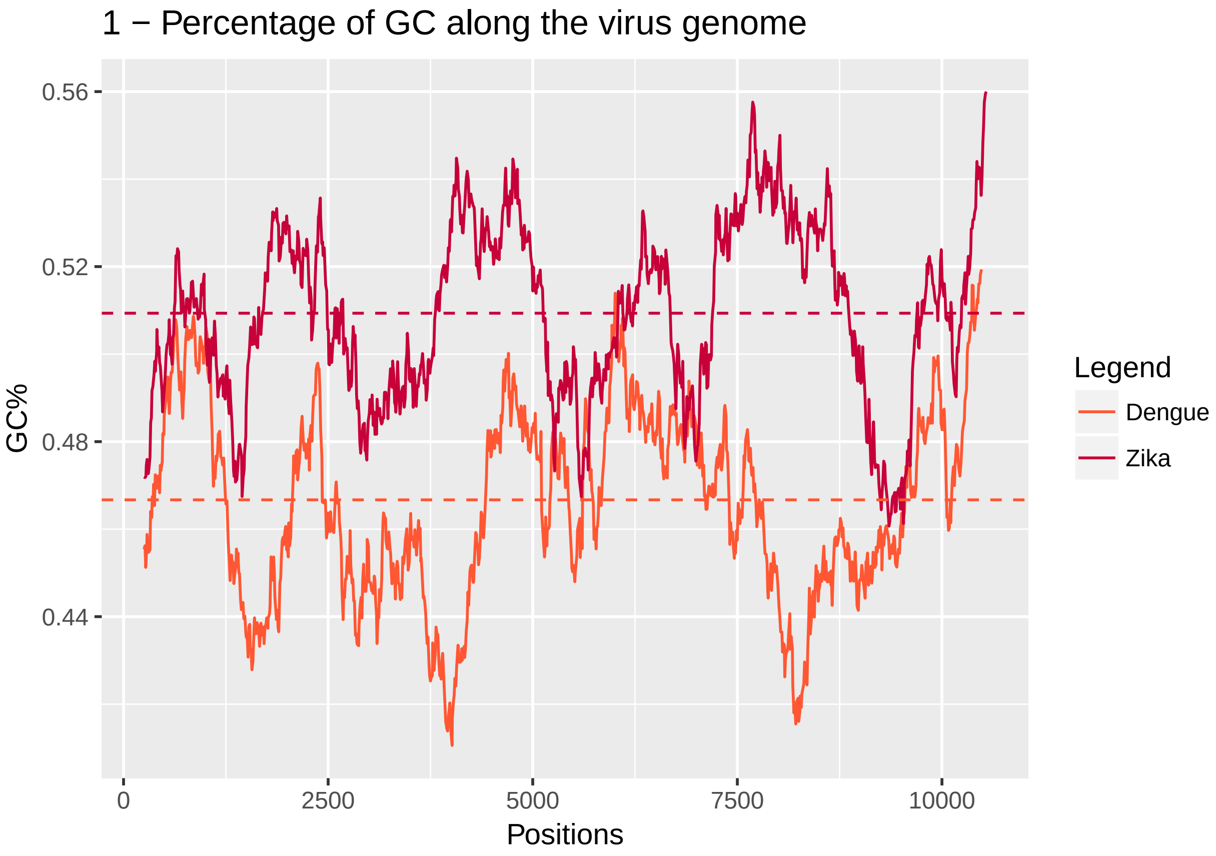
```
## [1] 0.5093571
```

Generation of the plot tables:

```r
# Table Dimers

Dimer<-merge(Dengue[[4]], Zika[[4]], by = "Dimers")
Dimer$`|Dengue - Zika|` <- abs(Dimer$Dengue-Dimer$Zika)
Dimer<-melt(Dimer)
colnames(Dimer)<-c('Dimers', 'Legend' , 'Dimer bias' )
Dimer$Group<-c(rep('Comparison',32), rep('Distance',16))


# Table codon

Codon<-merge(Dengue[[5]],Zika[[5]],by="Codons")
Codon$`|Dengue - Zika|`<-abs(Codon$Dengue-Codon$Zika)
Codon<-melt(Codon)
colnames(Codon)<-c('Codons', 'Legend' , 'Codon usage' )
```

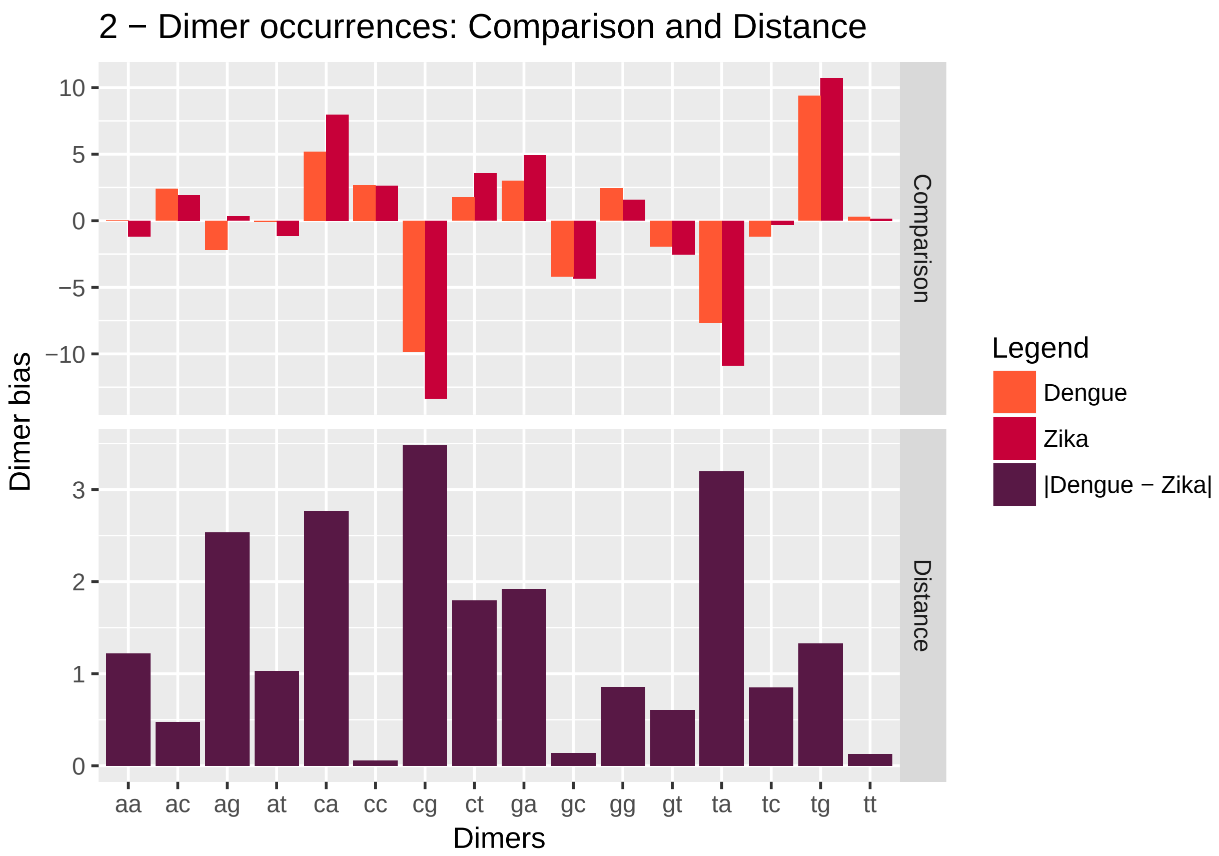
```
Codon$Group<-c(rep('Comparison',128), rep('Distance',64))


# %GC
GC_plot<-merge(Dengue[[6]], Zika[[6]], by = 'positions', all.y =  TRUE)
colnames(GC_plot)<-c('Positions', 'Dengue', 'Zika')
GC_plot<-melt(GC_plot, id = 'Positions')
colnames(GC_plot)<-c('Positions', 'Legend' , 'GC%' )
```

Plots



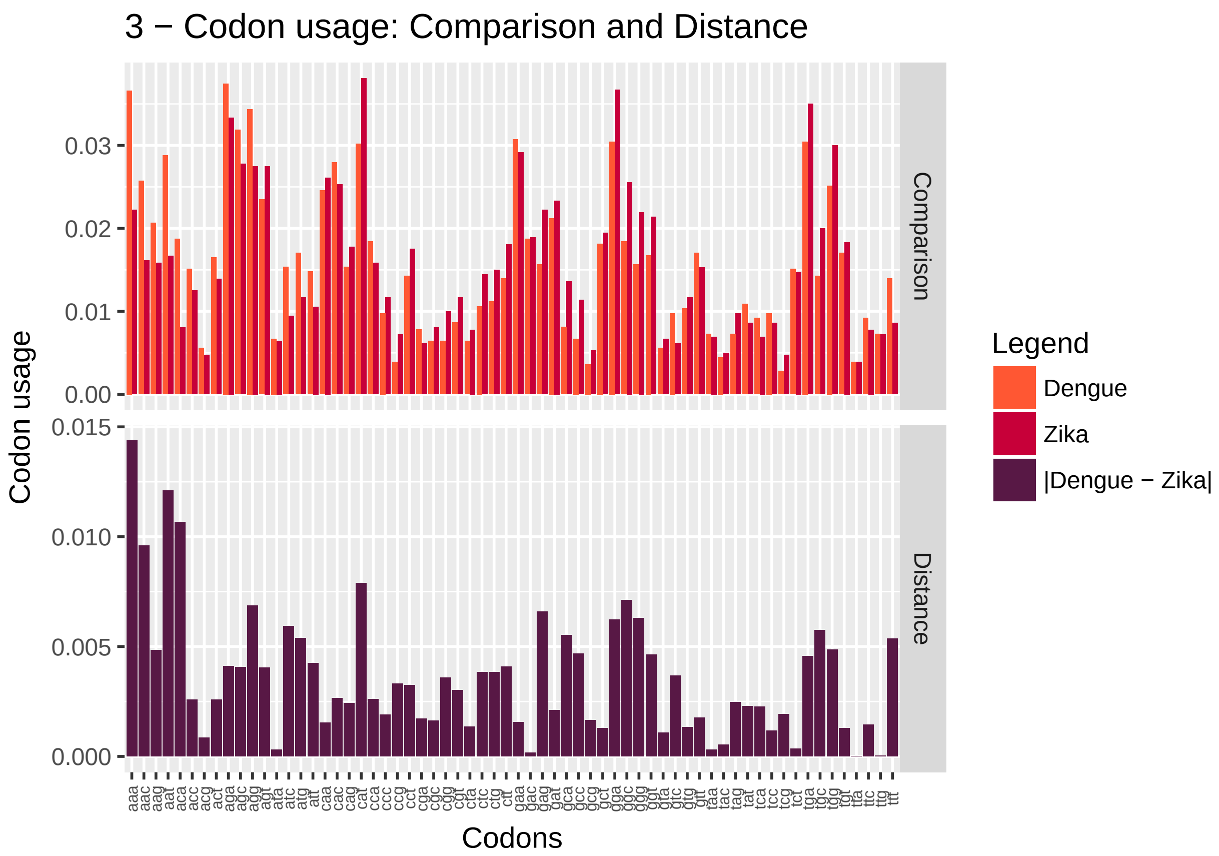
As calculated previously, GC percentage of DENV(showed in orange) genome is 46.6698%, and ZIKV's (showed in red) is 50.9357.

In plot 1 it can be seen that at both ends of DENV and ZIKV genomes the percentage of GC value is very similar and follows a similar pattern. In the central region of the genome, the percentage of GC in ZIKV is visibly higher than for DENV. Between nucleotides 7.500 and 8.750 the main difference can be found.



Zscore is represented in this plot. It is showed which dimer representation it is closer to the expected one by the random probability model (|zscore|<2), and which dimer representation is far of the expected.

There are some coincidences between both genomes; dimers CG, GC and TA are under-represented in DENV(orange) and also in ZIKV(red). TG, CA, GA and CC are over-represented because their value is bigger than +2. With that many dimers' zscore bigger than |2|, we can assume that the multinomial model does not apply in those sequences, implying that they were not build in a random way.



In this plot is showed a comparison of the frequency of appearence for each codon of the genetic code in genome of DENV (orange) and ZIKV (red).

The more frequent codons are DENV is AAA, AGG and GGA, and the least frequent are TCG, GCG ad TTA. In ZIKV, the more frequent are CAT, GGA and TGC, while the least frequent are TTA, TCG, and ACG. There can be identified some coincidence, being frequent to fiend GGA in both genomes. The least frequent codons for both are TTA and TCG. The codons with a bigger distance between genomes are AAA, AAT and ACA (predominant in DENV). The minimum distance is seen in codons TTA, ATA, and TAA.

Bibliography

[1] Rodriguez-Roche, Rosmari, and Ernest A. Gould. “Understanding the Dengue Viruses and Progress towards Their Control.” BioMed Research International 2013 (2013): 690835. PMC. Web. 29 Oct. 2017.

[2] Lorenzo Zammarchi, Giulia Stella, et al. “Zika virus infections imported to Italy: Clinical, immunological and virological findings, and public health implications.” Journal of Clinical Virology, Volume 63, 2015, Pages 32-35, ISSN 1386-6532

[3] Guzman MG, Halstead SB, Artsob H, et al. “Dengue: a continuing global threat.” Nature reviews Microbiology. 2010;8(12 0):S7-16.