

Biol 432 Group 1 Project

Chenyang Wu, Caroline Tang

2022/3/21

Project Info

Group name: Teambits

Date: 2022/3/21

GitHub Link: <https://github.com/carolinetang77/BIOL432-group1>

Load the packages we will need

```
library(dplyr)
library(ggplot2)
library(BiocManager)
library(genbankr)
library(rentrez)
library(muscle)
library(ape)
library(reshape2)
library(ggtree)
library(tidyverse)
library(Biostrings)
library(annotate)
```

Input the data

```
Table1 <- read.csv("../InputData/TableS1.csv")
Table9 <- read.csv("../InputData/TableS9.csv")

readLines("../InputData/TableS8.dat", n = 10) # Take a look at the .dat file

## [1] "BCoV1.1\t0.000147672\t1939" "BCoV1.main\t0.00016565\t1947"
## [3] "CHIKV.1\t0.000335409\t1953" "CHIKV.2\t8.95474e-05\t1829"
## [5] "CHIKV.grafted\tNaN\t1829" "CHIKV.main\t0.000117516\t1758"
## [7] "DENV.1\t0.000678122\t1980" "DENV.10\t0.000615014\t1962"
## [9] "DENV.11\t0.00077911\t1974" "DENV.12\t0.00152533\t2004"

Table8 <- read.table("../InputData/TableS8_edited.dat", header = F)
names(Table8) <- c("virus", "rate", "year")

Table11 <- read.csv("../InputData/TableS11.csv")
```

**Research question 1: Are mutation rates correlated with transmission method?
Do certain transmission methods have higher mutation rates?**

```
# Check how many types of virus in table S9
table(Table9$virus)
```

Coding by Chenyang Wu

```
##
##   BCoV1   CHIKV   DENV   Ebola   EVA   EVB   EVC   EVCr   EVD   H3N2
##   110     54    187     32    317   121    88    75    77     3
##   HCV    HCVr   HDV    HMPV   HRSV   HRV3   MERS   MMV    MRV  Norwalk
##   190    190    16     45    77    64    100    80    24   112
##   OHVA   PeVA   RVA    SARS2   SV    TBEV   WNV    YFV    ZIKV
##   86     79    33     25    71    76    66    55    55
```

```
# Check the corresponding transmission type to the abbreviation in table S1
list(Table1$Abbreviation)
```

```
## [[1]]
## [1] "BCoV1" "CHIKV" "DENV" "Ebola" "EVA" "EVB" "EVC"
## [8] "EVD" "H3N2" "HCV" "HDV" "HMPV" "HRSV" "HRV3"
## [15] "MERS" "MMV" "MRV" "Norwalk" "OHVA" "PeVA" "RVA"
## [22] "SARS2" "SV" "TBEV" "WNV" "YFV" "ZIKV"
```

```
list(Table1$Transmission)
```

```
## [[1]]
## [1] "aerosolic" "vector" "vector" "body fluids"
## [5] "fecal-oral" "fecal-oral" "fecal-oral" "fecal-oral"
## [9] "aerosolic" "blood/ sexual" "blood/ sexual" "aerosolic"
## [13] "aerosolic" "aerosolic" "aerosolic" "aerosolic"
## [17] "aerosolic" "fecal-oral" "fecal-oral" "aerosolic"
## [21] "aerosolic" "aerosolic" "fecal-oral" "vector"
## [25] "vector" "vector" "vector"
```

Combine the relevant columns into 1 dataframe

```
# Label the transmission type to the dataset
dNdSData <- Table9 %>%
  mutate(Transmission_Type = recode(virus,
                                     "BCoV1" = "aerosolic",
                                     "CHIKV" = "vector",
                                     "DENV" = "vector",
                                     "Ebola" = "body_fluids",
                                     "EVA" = "fecal-oral",
                                     "EVB" = "fecal-oral",
                                     "EVC" = "fecal-oral",
                                     "EVCr" = "fecal-oral",
                                     "EVD" = "fecal-oral",
                                     "H3N2" = "aerosolic",
                                     "HCV" = "blood/sexual",
                                     "HCVr" = "blood/sexual",
```

```

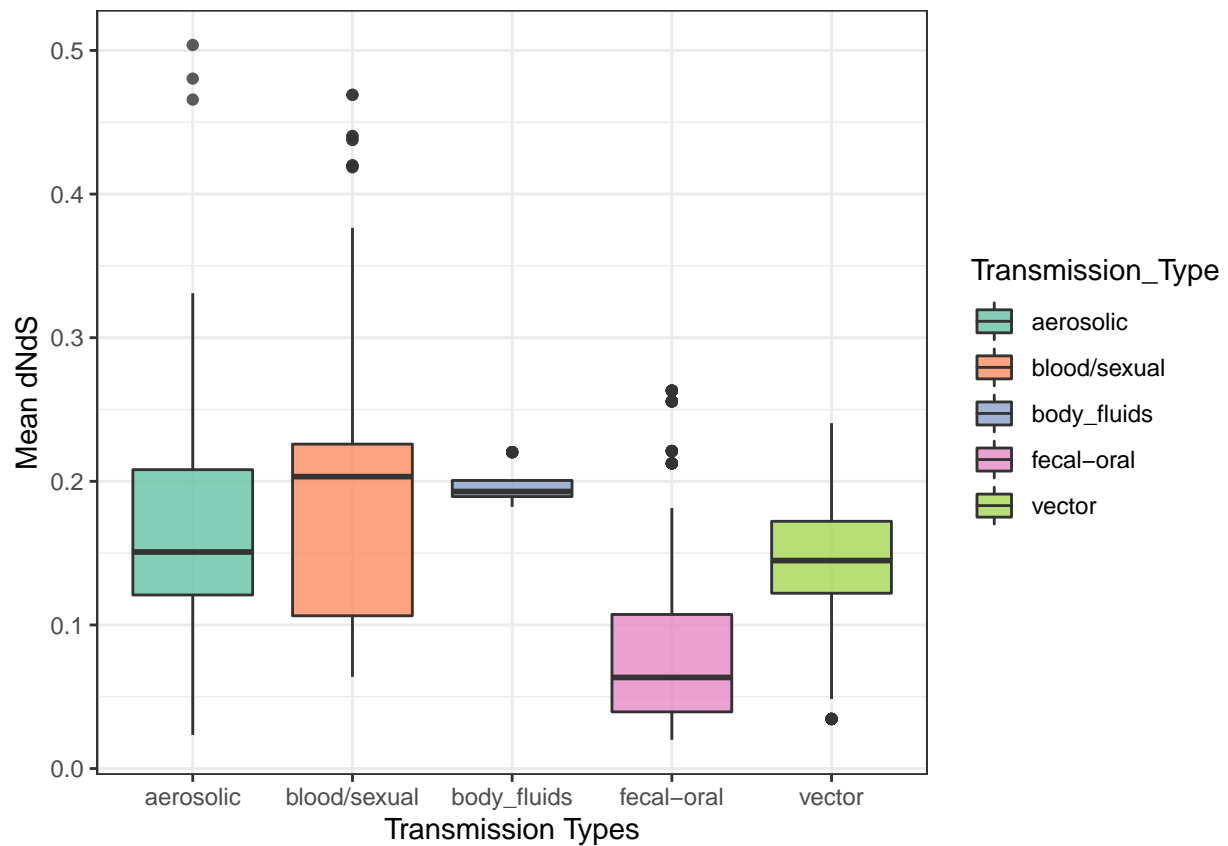
"HDV" = "blood/sexual",
"HMPV" = "aerosolic",
"HRSV" = "aerosolic",
"HRV3" = "aerosolic",
"MERS" = "aerosolic",
"MMV" = "aerosolic",
"MRV" = "aerosolic",
"Norwalk" = "fecal-oral",
"OHVA" = "fecal-oral",
"PeVA" = "aerosolic",
"RVA" = "aerosolic",
"SARS2" = "aerosolic",
"SV" = "fecal-oral",
"TBEV" = "vector",
"WNV" = "vector",
"YFV" = "vector",
"ZIKV" = "vector"))

```

```

# Draw a box plot for mean dNdS data with different transmission types
ggplot(dNdSData, aes(x = Transmission_Type, y = meandNdS,
                     na.rm = TRUE, fill = Transmission_Type)) +
  geom_boxplot(alpha = 0.8) +
  theme_bw() +
  scale_fill_brewer(palette = "Set2") +
  labs(x = "Transmission Types", y = "Mean dNdS")

```



```

MutaRate <- Table8 %>%
  mutate(Transmission_Type = recode(virus,
                                     "BCoV1" = "aerosolic",
                                     "CHIKV" = "vector",
                                     "DENV" = "vector",
                                     "Ebola" = "body_fluids",
                                     "EVA" = "fecal-oral",
                                     "EVB" = "fecal-oral",
                                     "EVC" = "fecal-oral",
                                     "EVCr" = "fecal-oral",
                                     "EVD" = "fecal-oral",
                                     "H3N2" = "aerosolic",
                                     "HCV" = "blood/sexual",
                                     "HCVr" = "blood/sexual",
                                     "HDV" = "blood/sexual",
                                     "HMPV" = "aerosolic",
                                     "HRSV" = "aerosolic",
                                     "HRV3" = "aerosolic",
                                     "MERS" = "aerosolic",
                                     "MMV" = "aerosolic",
                                     "MRV" = "aerosolic",
                                     "Norwalk" = "fecal-oral",
                                     "OHVA" = "fecal-oral",
                                     "PeVA" = "aerosolic",
                                     "RVA" = "aerosolic",
                                     "SARS2" = "aerosolic",
                                     "SV" = "fecal-oral",
                                     "TBEV" = "vector",
                                     "WNV" = "vector",
                                     "YFV" = "vector",
                                     "ZIKV" = "vector"))

# Draw a box plot for the mutation rate under different transmission types
ggplot(MutaRate, aes(x = Transmission_Type,
                     y = rate, na.rm = TRUE,
                     fill = Transmission_Type)) +
  geom_boxplot(alpha = 0.8) +
  theme_bw() +
  scale_fill_brewer(palette = "Set2") +
  labs(x = "Transmission Types", y = "Mutation Rate")

```

Fig.1 Boxplot of mean dNdS values among different transmission types.

Warning: Removed 21 rows containing non-finite values (stat_boxplot).

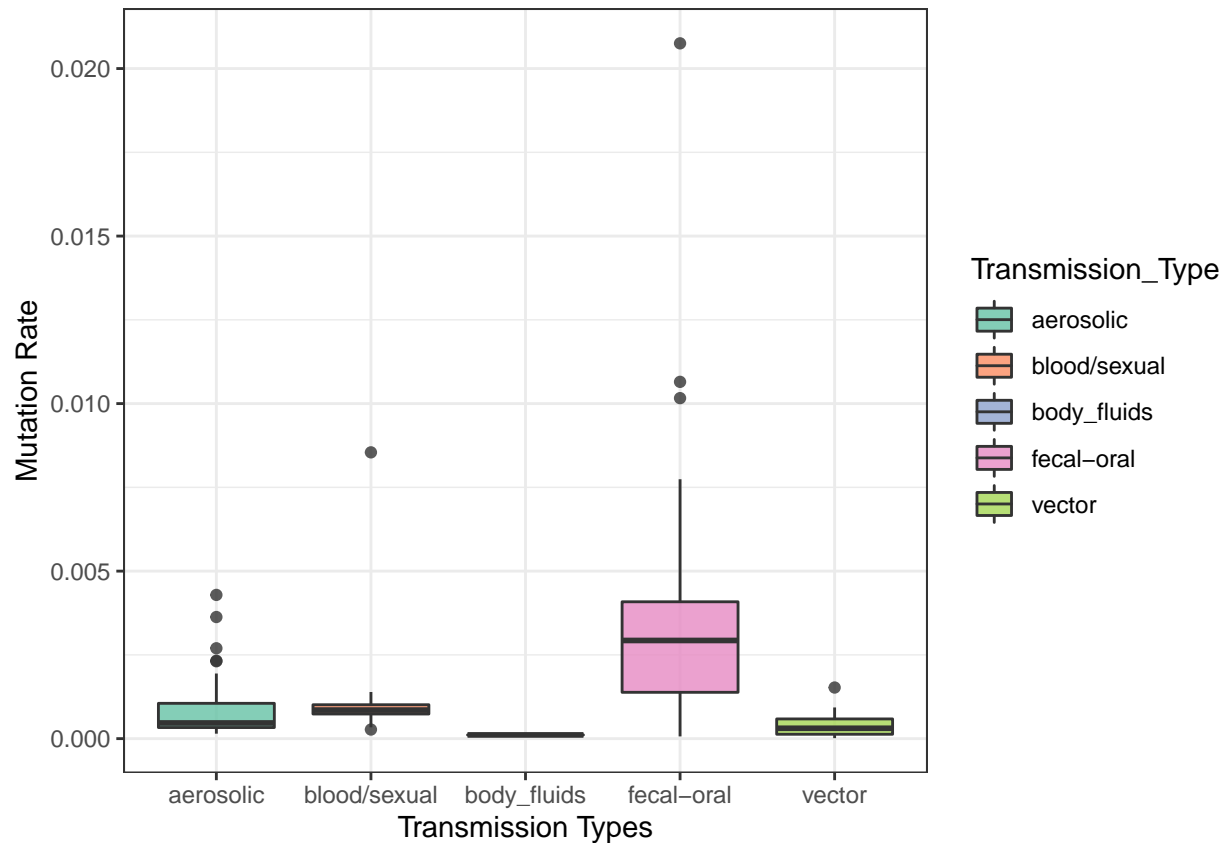


Fig.2 Boxplot of mutation rate among different Transmission Types

Adjust the sequence ID for SARS2

Covid sequences have different IDs from other viruses, not based on genbank accession number so we have to fix that

```
# Load the virus table and the COVID sequence acknowledgement table
Table3 <- read.delim("InputData/TableS3_GISAID_acknowledgements.dat")

# Merge the COVID rows in the virus table with the acknowledgement table
DataMerged <- merge(Table11[Table11$virus == "SARS2",], Table3, by.x = "ID", by.y = "internID")

# Filter for only rows with GenBank accession IDs
Datafilter <- DataMerged %>%
  filter(genbank_accession != "?")

# Replace the alldb ID with the accession ID
Datafilter$ID <- Datafilter$genbank_accession

# Select only the columns from the original virus table and rename them
Datafilter <- Datafilter %>%
  dplyr::select(1:12)
names(Datafilter) <- names(Table11)
```

```
# Replace the original covid rows with the ones we just made
finalIDs <- bind_rows(Table11[Table11$virus != "SARS2",], Datafilter)
```

```
# Check the number of IDs in table S11 for each virus
table(finalIDs$virus)
```

This chunk was done with the help of Caroline Tang

```
##
##   BCoV1   CHIKV   DENV   Ebola   EVA   EVB   EVC   EVD   H3N2   HCV
##     321     1127   7263   2512   2866   959   1019   974   29706  4494
##     HDV     HMPV   HRSV   HRV3   MERS   MMV   MRV Norwalk   OHVA   PeVA
##     740     224   2751   477    629   329   608   2384   677    323
##     RVA     SARS2    SV    TBEV   WNV   YFV   ZIKV
##     337     1542   289    266   2731   435   1106
```

```
# Since there are so many IDs, We decided to pick 5 IDs from each virus for question 1
```

```
# Random pick 5 IDs for each virus
set.seed(1) # Set seed to make sure the output are constant
IDs <- finalIDs %>%
  group_by(virus) %>%
  sample_n(5)
```

```
# Double check if the code correctly pick 5 IDs from each virus
table(IDs$virus)
```

```
##
##   BCoV1   CHIKV   DENV   Ebola   EVA   EVB   EVC   EVD   H3N2   HCV
##     5      5      5      5      5      5      5      5      5      5
##     HDV     HMPV   HRSV   HRV3   MERS   MMV   MRV Norwalk   OHVA   PeVA
##     5      5      5      5      5      5      5      5      5      5
##     RVA     SARS2    SV    TBEV   WNV   YFV   ZIKV
##     5      5      5      5      5      5      5
```

```
# Create the id list
ncbi_ids <- IDs$ID
```

```
# Search the sequence info from NCBI
Q1Vir <- entrez_fetch(db = "nucleotide", id = ncbi_ids, rettype = "fasta")
Q1Seq <- strsplit(Q1Vir, split = "\n\n", fixed = T)
Q1Seq <- unlist(Q1Seq)
```

```
# Use regular expression to edit the search result
header <- gsub("(^>.*genome|*cds|*sequence|*SEQUENCES|*RNA)\\n([ATCG].*)", "\\1", Q1Seq)
```

```
seq <- gsub(">.*genome\\n([ATCG].*)", "\\1", Q1Seq)
seq <- gsub(">.*cds\\n([ATCG].*)", "\\1", seq)
seq <- gsub(">.*sequence\\n([ATCG].*)", "\\1", seq)
seq <- gsub(">.*SEQUENCES\\n([ATCG].*)", "\\1", seq)
seq <- gsub(">.*RNA\\n([ATCG].*)", "\\1", seq)
```

```
Q1SeqTable <- data.frame(Name = header, Sequence = seq)
Q1SeqTable$Sequence <- gsub("\n", "", Q1SeqTable$Sequence)
```

```
# There are several lines have different ending words, so we will output the data set and adjust it manually
```

```
write.csv(Q1SeqTable, "./InputData/Q1Seq.csv", row.names = F)
```

```
# Input the edited data
```

```
Q1Sequence <- read.csv("./InputData/Q1Seq_edited.csv")
```

Multiple Alignments

```
Q1DF <- data.frame(ID = IDs$ID,  
                  Virus = IDs$virus,  
                  Seq = Q1Sequence$Sequence,  
                  stringsAsFactors = FALSE)
```

```
Q1DF <- Q1DF %>%
```

```
  mutate(Transmission_Type = recode(Virus,  
                                     "BCoV1" = "aerosolic",  
                                     "CHIKV" = "vector",  
                                     "DENV" = "vector",  
                                     "Ebola" = "body_fluids",  
                                     "EVA" = "fecal-oral",  
                                     "EVB" = "fecal-oral",  
                                     "EVC" = "fecal-oral",  
                                     "EVCr" = "fecal-oral",  
                                     "EVD" = "fecal-oral",  
                                     "H3N2" = "aerosolic",  
                                     "HCV" = "blood/sexual",  
                                     "HCVr" = "blood/sexual",  
                                     "HDV" = "blood/sexual",  
                                     "HMPV" = "aerosolic",  
                                     "HRSV" = "aerosolic",  
                                     "HRV3" = "aerosolic",  
                                     "MERS" = "aerosolic",  
                                     "MMV" = "aerosolic",  
                                     "MRV" = "aerosolic",  
                                     "Norwalk" = "fecal-oral",  
                                     "OHVA" = "fecal-oral",  
                                     "PeVA" = "aerosolic",  
                                     "RVA" = "aerosolic",  
                                     "SARS2" = "aerosolic",  
                                     "SV" = "fecal-oral",  
                                     "TBEV" = "vector",  
                                     "WNV" = "vector",  
                                     "YFV" = "vector",  
                                     "ZIKV" = "vector"))
```

```
# Convert DNASbin to DNASringSet
```

```
VirString <- Q1DF$Seq %>%  
  as.character %>%  
  lapply(., paste0, collapse = "") %>%  
  unlist %>%  
  DNASringSet
```

```
# Give each sequence a unique names
```

```
names(VirString) <- paste(1:nrow(Q1DF), Q1DF$ID, sep = "_")
```

```

# Use MUSCLE to align the sequences
# This line will take about 1 hour for R to run it.
VirAlign <- muscle::muscle(stringset = VirString, diags = T, gapopen = -10)

# Convert our DNA Multiple Alignment object to a DNA Bin object
VirAlignBin <- as.DNABin(VirAlign)

SeqLen <- as.numeric(lapply(VirString, length))
# Show the distribution of sequence length
qplot(SeqLen) + theme_bw()

```

Fig.3 Distribution of sequence length of the selected IDs.

Visualize the distance matrix

```

VirDM <- dist.dna(VirAlignBin, model = "K80")
VirDMmat <- as.matrix(VirDM)

# Plot the distance matrix
VirPDat <- melt(VirDMmat)
ggplot(data = VirPDat, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradientn(colours = c("white", "blue", "green", "red")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Fig.4 Figure of the distance matrix.

Create the phlogeny tree

```

VirTree <- nj(VirDM)

# Edit the tip label of the tree to better group the sequence by transmission types
VirTree$tip.label <- paste(rownames(Q1DF), Q1DF$Transmission_Type)
TransType <- split(VirTree$tip.label, Q1DF$Transmission_Type)
TransTree <- groupOTU(VirTree, TransType)

ggtree(TransTree, branch.length = 'none', layout = "circular", aes(colour = group)) +
  geom_tiplab(size = 2, aes(angle = angle)) +
  theme_bw()

```

Fig.5 Phylogenetic tree of the 135 selected sequences without consider the branch length.

Output the phylogeny tree

```

write.tree(TransTree, "./Output/Transmission_Type_Tree.tre")

```

Based on the boxplots, it appears that fecal-oral viruses have the highest mutation rates relative to other methods of transmission. However, they also have the lowest dN/dS rates, suggesting a trade-off between mutation rates and rates of non-synonymous mutations. Due to differences in viral genomes, when creating the tree, there were no common sequences found. As a result, the tree showed all sequences as equally distant from one another.

Research question 2: Is the mutation rate correlated with guanine-cytosine content?

Coding By Caroline Tang

Subset sequences (10 per virus)

```
set.seed(1)
virusSubset <- finalIDs %>%
  group_by(virus) %>%
  slice_sample(n = 10)
```

Get sequences from GenBank

```
virusID <- GBAccession(virusSubset$ID)
virusGBK <- read.GenBank(virusID, as.character = TRUE)
```

Calculate GC content of each sequence and merge with virus types

```
gc <- vector(length = length(virusGBK))
for (i in 1:length(virusGBK)) {
  gc[i] <- length(grep("[gc]", unlist(virusGBK[[i]]))) / length(virusGBK[[i]])
}
gcContent <- data.frame(ID = names(virusGBK), gc = gc)
virusSubset <- merge(virusSubset, gcContent, by = "ID")
```

Scatterplot of GC content vs. mutation rate and transmission method

```
# Calculate mean mutation rate per virus and merge with GC content
mutationMean <- Table8 %>%
  group_by(virus) %>%
  summarise(meanRate = mean(rate, na.rm = T))
virusSubset <- merge(virusSubset, mutationMean, by = "virus")

# Merge transmission data
virusSubset <- merge(virusSubset, Table1, by.x = "virus", by.y = "Abbreviation")

# Create scatter plot
ggplot(data = virusSubset, aes(x = gc, y = meanRate)) +
  geom_point(aes(colour = Transmission), alpha = 0.8) +
  theme_classic() +
  geom_smooth(method = "lm") +
  scale_fill_brewer(palette = "Set2") +
  labs(x = "GC content", y = "Mean mutation rate")

## `geom_smooth()` using formula 'y ~ x'
```

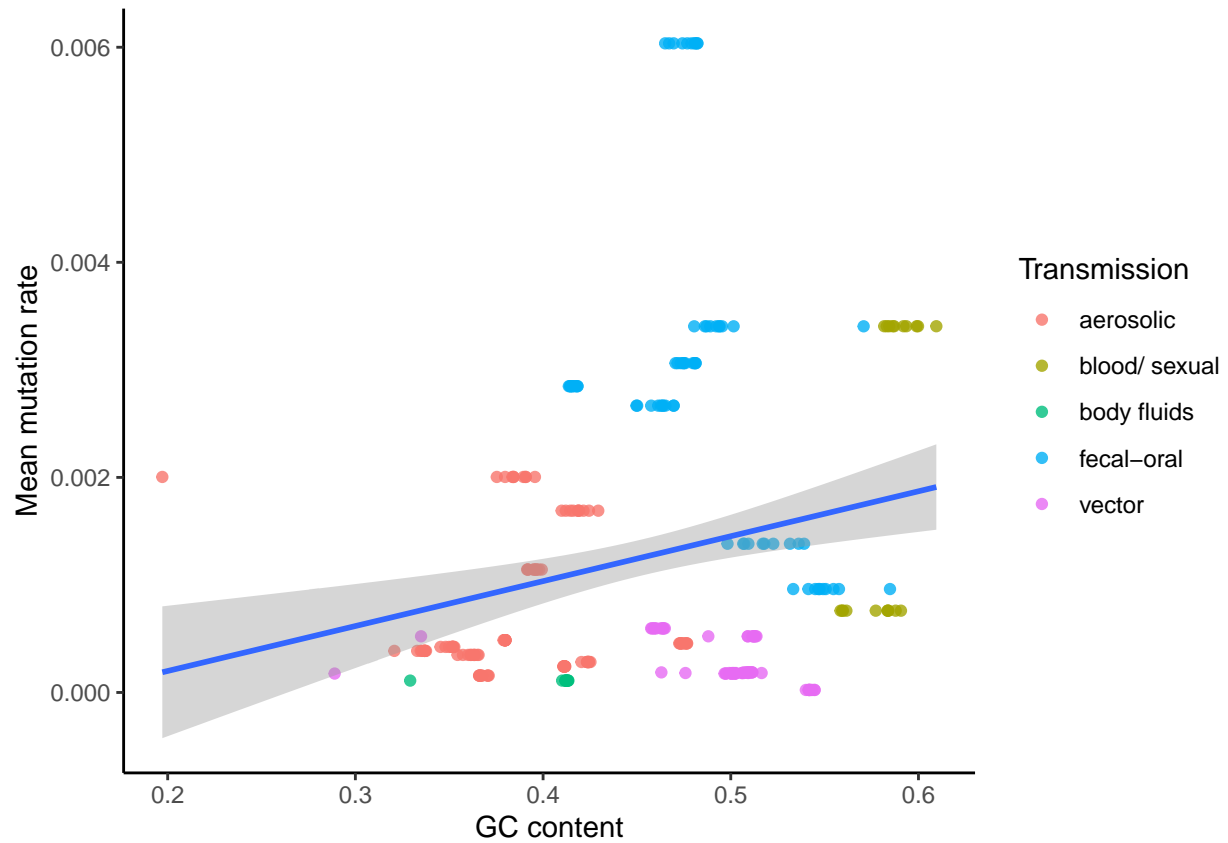


Fig.6 Scatter plot of the 130 selected sequences' GC content and mutation rates, colour coded by transmission type.

Based on the scatter plot, there is a slightly positive correlation between GC content and mutation rate, which supports the hypothesis that mutation rate increases with GC content. However, this trend varied among transmission methods, and the overall trend be skewed by the fecal-oral viruses with high mutation rates.