Group Project

10/04/2020

Biological Questions

1. Do the belly button microbiome differ between the two populations from which samples were taken from?
2. What is the evolutionary relationship between the different species found in one sample (one belly button)?

Libraries

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
library(ape)
```

```
## Warning: package 'ape' was built under R version 3.6.2
```

```
library(ggtree)
```

```
## Warning: package 'ggtree' was built under R version 3.6.1
```

```
## Registered S3 method overwritten by 'treeio':
##   method     from
##   root.phylo ape
```

```
## ggtree v2.0.1  For help: https://yulab-smu.github.io/treedata-book/
##
## If you use ggtree in published research, please cite the most appropriate paper(s):
##
## □[36m-□[39m Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods for mappin
g and visualizing associated data on phylogeny using ggtree. Molecular Biology and Evolution 201
8, 35(12):3041-3043. doi: 10.1093/molbev/msy194
## □[36m-□[39m Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: a
n R package for visualization and annotation of phylogenetic trees with their covariates and oth
er associated data. Methods in Ecology and Evolution 2017, 8(1):28-36, doi:10.1111/2041-210X.126
28
```

```
##
## Attaching package: 'ggtree'
```

```
## The following object is masked from 'package:ape':
##
##     rotate
```

```
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 3.6.2
```

```
## Loading required package: permute
```

```
## Warning: package 'permute' was built under R version 3.6.2
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-6
```

```
library(seqinr)
```

```
## Warning: package 'seqinr' was built under R version 3.6.2
```

```
##
## Attaching package: 'seqinr'
```

```
## The following object is masked from 'package:permute':
##
##     getType
```

```
## The following objects are masked from 'package:ape':
##
##     as.alignment, consensus
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.6.2
```

Question 1

Setup Sample Information

Import the sample information.

```
Samples <- read.csv("data/sample_info.csv")
```

Setup OTU table

Import OTU table.

```
OTU_table <- read.delim("data/OTU_file.txt", header = T, row.names = "X.OTU.ID")
```

Remove taxonomy column from OTU table.

```
OTU_data <- OTU_table[, -c(1, ncol(OTU_table))]
```

Calculate the total sequences in the dataset.

```
x <- rowSums(OTU_data)
sum(x)
```

```
## [1] 24000
```

With 24,000 reads there are most likely going to be some reads that are only contamination. Therefore, to remove these and to reduce the dataset slightly we will remove any OTUs that do not have more than one sequence in more than one sample.

```
drop <- rowSums(OTU_data) < 2
sum(drop)  # The number of sequences being removed.
```

```
## [1] 896
```

```
OTU_red <- OTU_data[!drop, ]
```

Transpose the table so that the species are across the top and the samples are along the side.

```
OTU_red[1:3, 1:3]  # Preview the layout of the original data.frame.
```

```
##      B1234 B1235 B1236
## 2151     0     0     0
## 347      0     0     0
## 1192     0     0     0
```

```
OTUs <- as.data.frame(t(OTU_red))
OTUs[1:3, 1:3]  # Ensure transpose worked.
```

```
##       2151 347 1192
## B1234    0   0    0
## B1235    0   0    0
## B1236    0   0    0
```

Analysis

Binary Method

Change OTUs data.frame into binary data by changing the read counts in each sequence to 1 or 0.

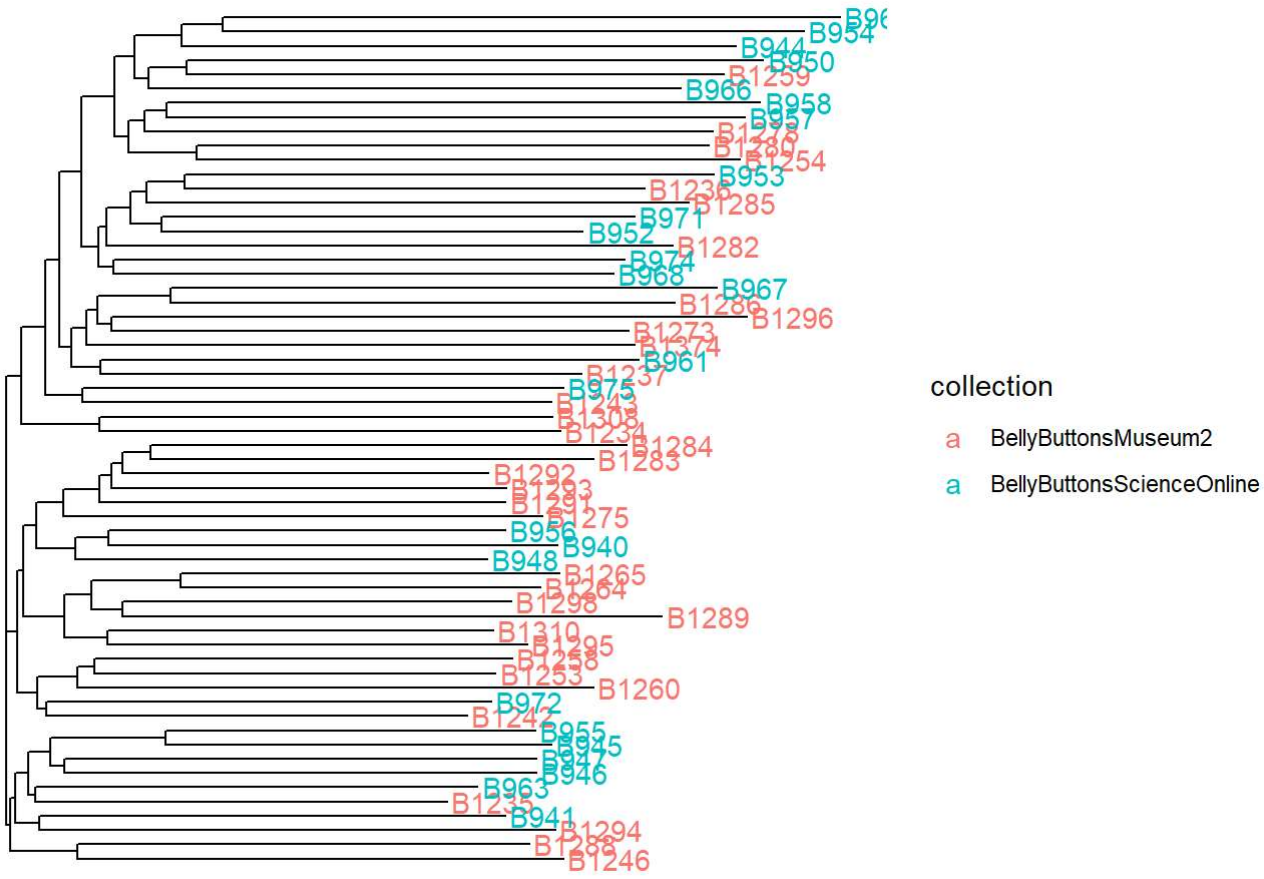
```
OTU_bin <- OTUs
OTU_bin[OTU_bin > 0] <- 1
```

Calculate the pairwise distance of the binary matrix.

```
OTU_bin_dist <- dist(OTU_bin, method = "binary")
```

Build the binary neighbour-joining tree, annotate it using the sample information, and output it as a pdf.

```
OTU_bin_tree <- nj(OTU_bin_dist)
ggtree(OTU_bin_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
```




```
pdf(width = 16, height = 20, "Binary_Tree.pdf")
ggtree(OTU_bin_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
dev.off()
```

```
## png
##   2
```

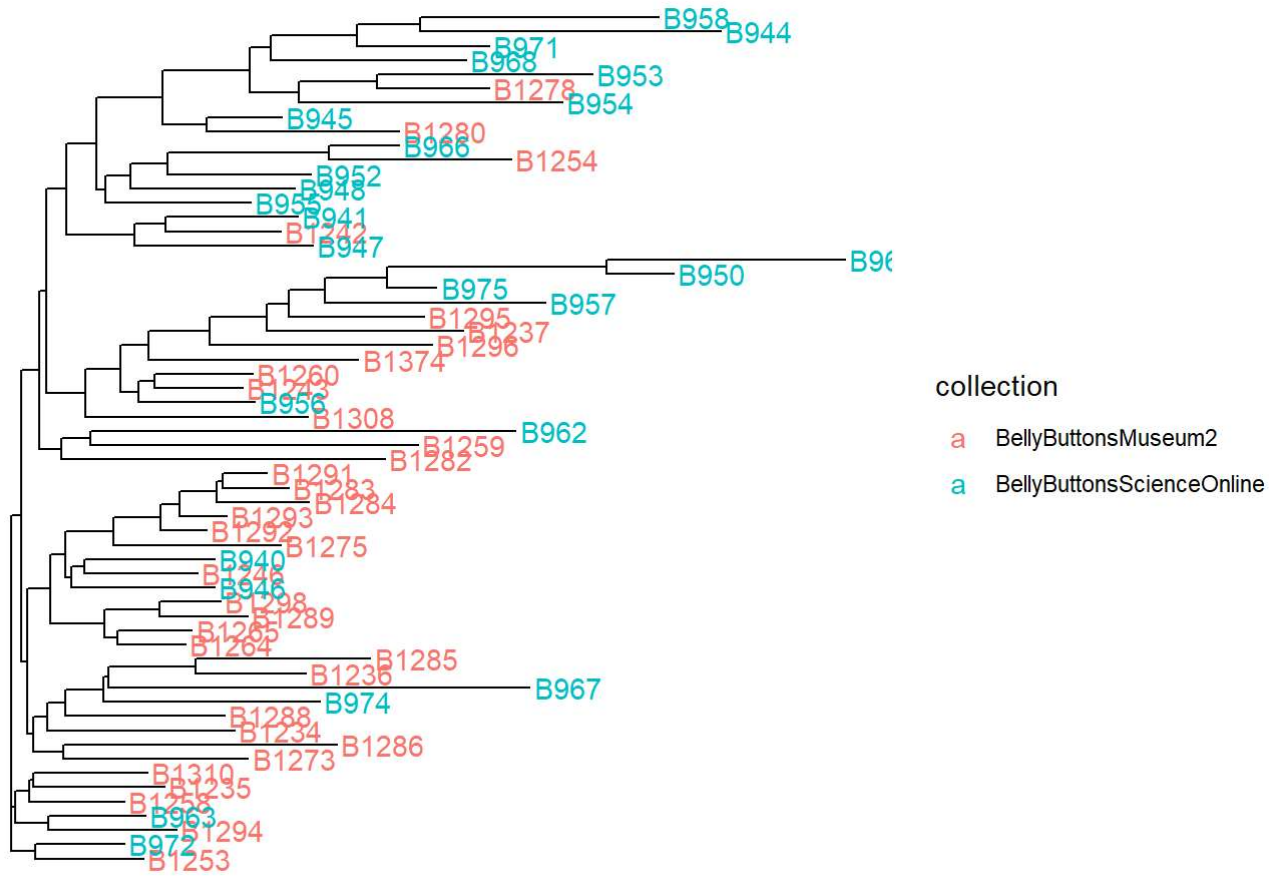
Euclidean Method

Calculate the euclidean distance.

```
OTU_euc_dist <- dist(OTUs, method = "euclidean")
```

Build the euclidean neighbour-joining tree, annotate it using the sample information, and output it as a pdf.

```
OTU_euc_tree <- nj(OTU_euc_dist)
ggtree(OTU_euc_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
```




```
pdf(width = 16, height = 20, "Euclidean_Tree.pdf")
ggtree(OTU_euc_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
dev.off()
```

```
## png
##   2
```

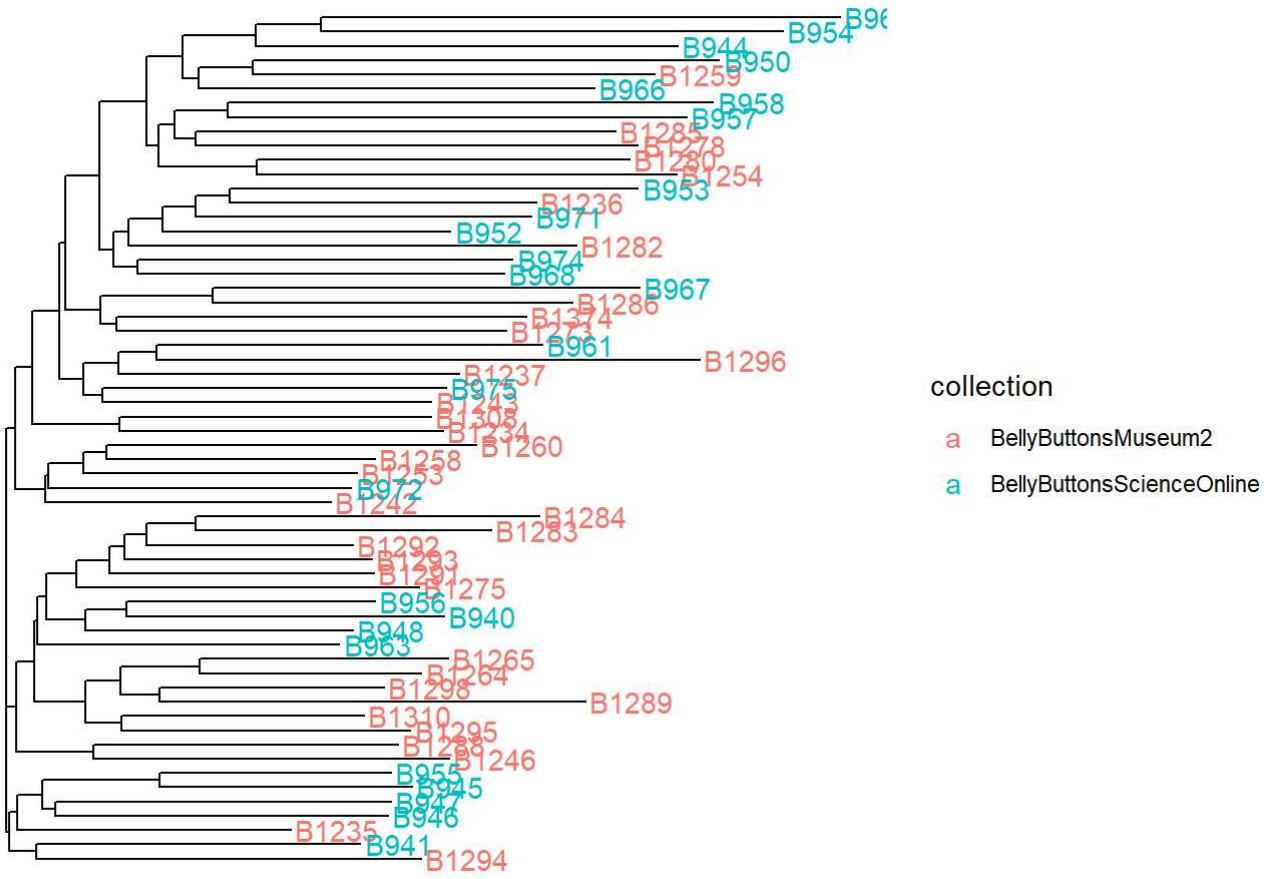
Bray-Curtis Dissimilarity Method

Calculate the Bray-Curtis dissimilarity.

```
OTU_bc_dist <- vegdist(OTUs, method = "bray", binary = T)
```

Build the Bray-Curtis dissimilarity neighbour-joining tree, annotate it using the sample information, and output it as a pdf.

```
OTU_bc_tree <- nj(OTU_bc_dist)
ggtree(OTU_bc_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
```




```
pdf(width = 16, height = 20, "Bray_Curtis_Tree.pdf")
ggtree(OTU_bc_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
dev.off()
```

```
## png
##   2
```

Non-Metric Multidimensional Scaling (NMDS)

Perform the model.

```
set.seed(13)
NMDSdat <- metaMDS(OTU_bc_dist, k = 2, trymax = 100)
```

```
## Run 0 stress 0.2034762
## Run 1 stress 0.2023931
## ... New best solution
## ... Procrustes: rmse 0.06109644  max resid 0.3518649
## Run 2 stress 0.2018038
## ... New best solution
## ... Procrustes: rmse 0.07063142  max resid 0.3473908
## Run 3 stress 0.2104821
## Run 4 stress 0.2099949
## Run 5 stress 0.2018383
## ... Procrustes: rmse 0.01063148  max resid 0.06581382
## Run 6 stress 0.2018036
## ... New best solution
## ... Procrustes: rmse 0.0004402051  max resid 0.002531702
## ... Similar to previous best
## Run 7 stress 0.2129357
## Run 8 stress 0.2040082
## Run 9 stress 0.220175
## Run 10 stress 0.2021465
## ... Procrustes: rmse 0.09039165  max resid 0.3156751
## Run 11 stress 0.2039927
## Run 12 stress 0.2219921
## Run 13 stress 0.2038741
## Run 14 stress 0.2123407
## Run 15 stress 0.2021928
## ... Procrustes: rmse 0.08552759  max resid 0.3238095
## Run 16 stress 0.2091613
## Run 17 stress 0.2088129
## Run 18 stress 0.2038626
## Run 19 stress 0.2118643
## Run 20 stress 0.2141294
## *** Solution reached
```

Create data for plotting.

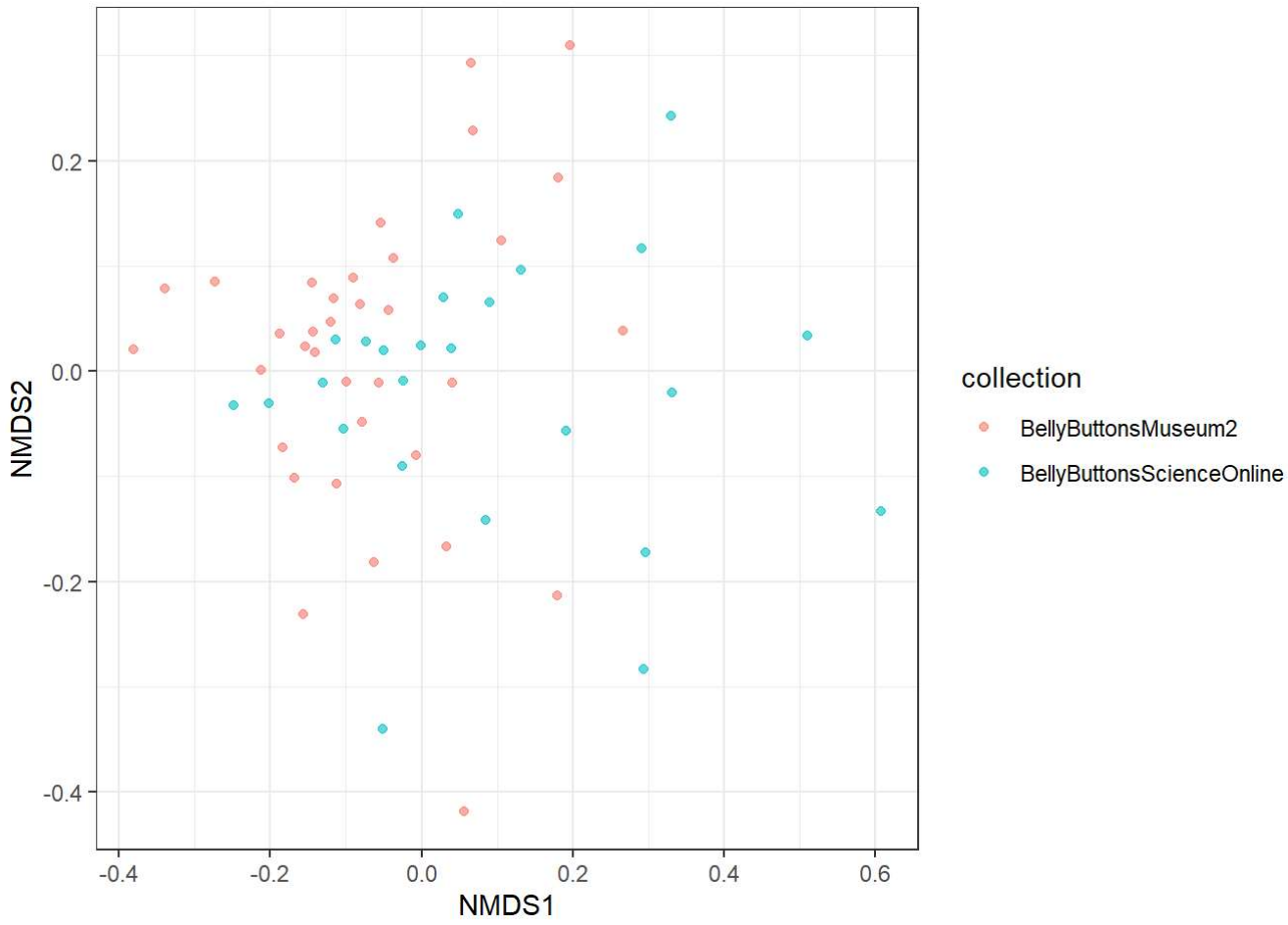
```
PDat <- data.frame(NMDS1 = NMDSdat$points[ , 1],
                   NMDS2 = NMDSdat$points[ , 2],
                   sample = row.names(OTUs))
```

Add species labels.

```
PDat <- merge(PDat, Samples, by = "sample", all.x = T, all.y = F)
```

Plot the NMDS.

```
qplot(x = NMDS1, NMDS2, colour = collection, alpha = I(0.6), data = PDat) +
  theme_bw()
```



Question 2

Data setup

Importing the Fasta file and selecting a single sample, B1285, in order to determine the evolutionary diversity among the bacteria and archaea within that single sample. A for loop was utilized to gather all the sequence reads which pertained to the individual sample collected (B1285)

```
myFasta <- read.fasta(file = "data/raw_seqs_BB.fna", seqtype = "AA", as.string = TRUE, set.attri
butes = FALSE)
nam <- names(myFasta)

indexes <- grep("B1285", nam)

sub <- rep(NA, 1170)

for (i in 1:length(indexes)){
  sub[i] <- myFasta[indexes[i]]
}
```

Creating a dataframe using index numbers as IDs and the sequence data from the myFasta file pasted in the seq column. A new object ‘dna’ was created using the sapply function to seperate each base pair into seperate columns. Following, the names function was utilized to re-name all row names to their corresponding indices from the orignial file read into the myFasta file.

```
df <- data.frame(ID = as.factor(indexes), Seq = paste(sub), stringsAsFactors = FALSE)
dna <- sapply(df$Seq, strsplit, split = "")
names(dna) <- paste(1:nrow(df),df$ID,sep="_")
dna_bin<-as.DNAbin(dna)
```

Alignment

The ‘dna’ file was converted to a DNAbin object which can be acted upon the muscle to align the sequence data.

```
dna_align <- muscle(dna_bin,quiet=F)
```

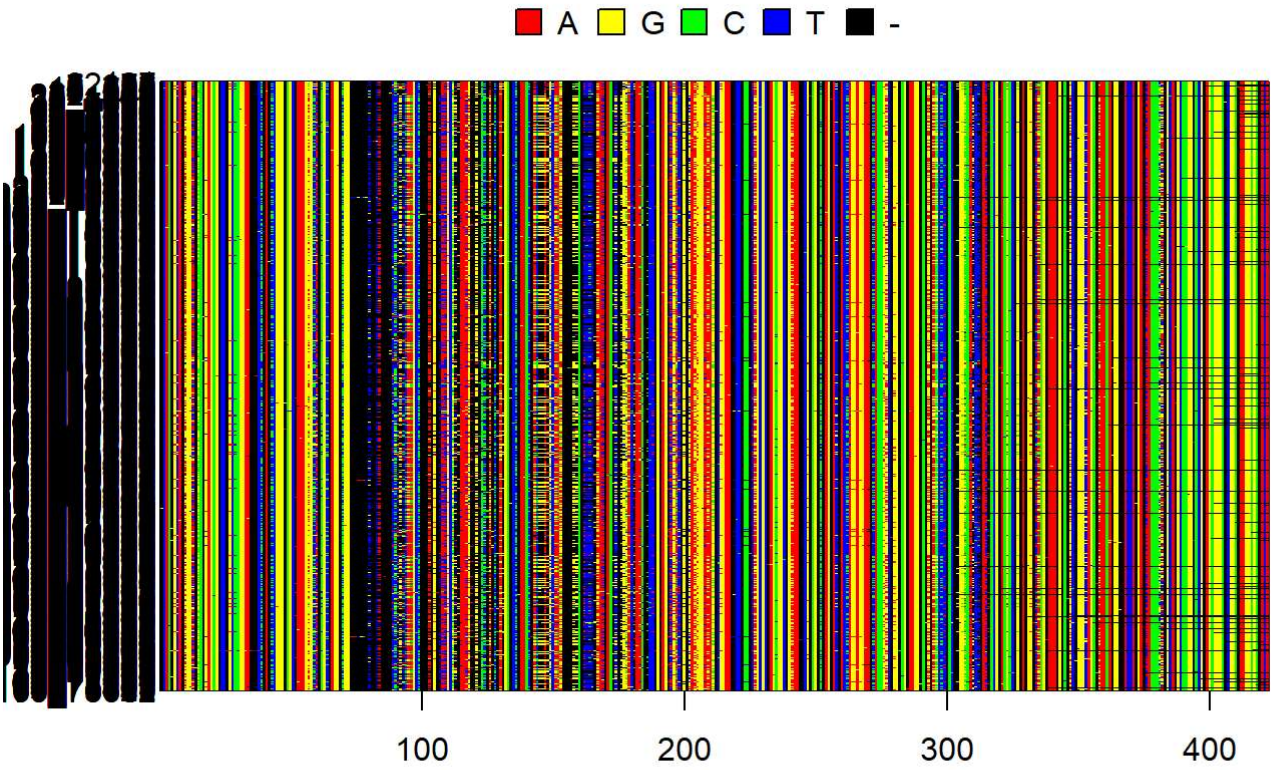
Inspect the alignment

Visual of the alignment of all sequences at all sites. There is too much going on to determine anything so we will zoom in on the beginning, middle, and end of the alignment for the first half of the sequences and then the second half.

```
checkAlignment(dna_align, what = 1)
```

```
##
## Number of sequences: 1170
## Number of sites: 422
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 11 13 14 16 17 20 22 25 26 28 29 31 34
35 37 40 41 49 53 61 67 74 77 85 86 95 97 98 131
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 83845  9968  2702  1966   435  1057   562   102     6    11     1    33     9
##    14    15    16    17    18    20    21    22    24    25    26    27    28
##     1    13   620     2     1     4    13     3     2     8     5     3     3
##    29    31    33    34    35    36    37    40    41    42    45    48    49
##     1     1     3     1     1     4     1     1     1     2     2     6     3
##    51    53    57    60    61    66    67    69    74    77    81    85    86
##     1     1     4     2     5     2     1     1     2     3     1     4     1
##    95    97    98   114   120   129   131
##     1     1     1     2     1     1     2
##    => length of gaps on the left border of the alignment: 1 0
##    => length of gaps on the right border of the alignment: 131 131 129 120 114 114 97 95 86 8
5 85 85 85 81 77 77 77 74 74 69 67 66 66 61 61 60 60 57 57 57 57 51 49 49 49 48 48 48 48 48 48 4
2 41 40 36 36 36 36 33 33 33 31 27 27 26 26 26 26 26 25 25 25 25 25 25 25 22 22 21 21 21 21 21 2
1 21 21 21 21 21 15 15 15 15 15 15 15 15 15 13 13 13 13 13 12 12 12 12 12 12 12 12 12 12 12 12 1
2 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 9 9 9 7 7 6 5 5 5 5 5 5 4 4 4 4 4 4 2
##
## Number of unique contiguous base segments defined by gaps: 357
## Number of segment lengths not multiple of 3: 250
##     => on the left border of the alignement: 1
##     => on the right border                : 3
##     => positions of these segments inside the alignment: 10..13 15..16 18..22 24..25 27..34 3
8..39 41..41 43..47 53..60 80..80 83..84 89..90 103..103 105..105 108..112 115..118 121..121 12
3..124 126..126 128..128 130..130 132..133 135..136 138..141 143..144 147..147 149..153 160..160
162..165 171..172 174..174 177..180 187..188 190..190 192..193 195..198 200..200 202..211 213..2
17 220..221 223..224 240..243 246..247 249..250 252..252 254..255 257..258 260..270 272..278 28
0..281 283..284 286..289 291..291 293..293 295..299 301..302 305..309 311..312 314..315 321..325
327..327 329..329 331..332 334..337 343..343 345..345 347..348 350..356 358..362 364..365 368..3
71 373..374 376..382 399..405 407..410 412..418 407..417 89..99 108..118 147..153 177..184 192..
198 102..103 107..113 143..153 176..180 305..311 115..121 158..158 121..124 130..133 176..179 37
6..386 88..88 37..38 53..56 93..99 305..308 143..147 246..250 123..126 254..258 89..92 321..327
270..270 202..203 205..211 95..99 157..160 171..174 202..217 168..169 412..413 260..269 53..68 1
08..121 167..174 257..270 77..80 104..105 249..255 295..302 80..84 283..289 373..373 187..190
6..13 126..133 373..382 54..60 114..118 178..184 388..389 135..141 138..144 395..395 70..80 83..
83 324..325 337..337 43..46 53..59 10..16 15..22 95..101 103..118 80..81 91..91 166..169 227..23
1 280..284 359..362 38..41 233..243 345..348 350..362 376..380 382..382 293..299 260..278 388..3
91 41..47 132..136 407..407 24..34 158..165 399..400 384..385 412..415 162..169 221..221 295..29
6 298..299 305..312 41..51 146..147 397..397 90..90 314..314 149..149 384..393 138..153 62..68 3
31..337 358..365 347..356 89..98 152..153 156..160 385..386 306..309 66..72 395..396 365..365 8
4..90 368..374 388..388 390..393 347..362 159..160 396..397 358..361 122..126 190..193 272..281
395..405 84..84 388..397 144..144 226..238 332..332 76..80 339..343 393..393 240..240 420..420 2
02..206 415..418 2..8 18..25 329..332 89..93 364..371 334..341 90..99 86..86 389..393 143..156 1
74..175 311..315 105..112 304..307 304..311 153..153 162..166 186..190 213..220 220..224 240..24
7 412..416 31..38 350..353 249..252 252..255 334..334 336..337
##
```

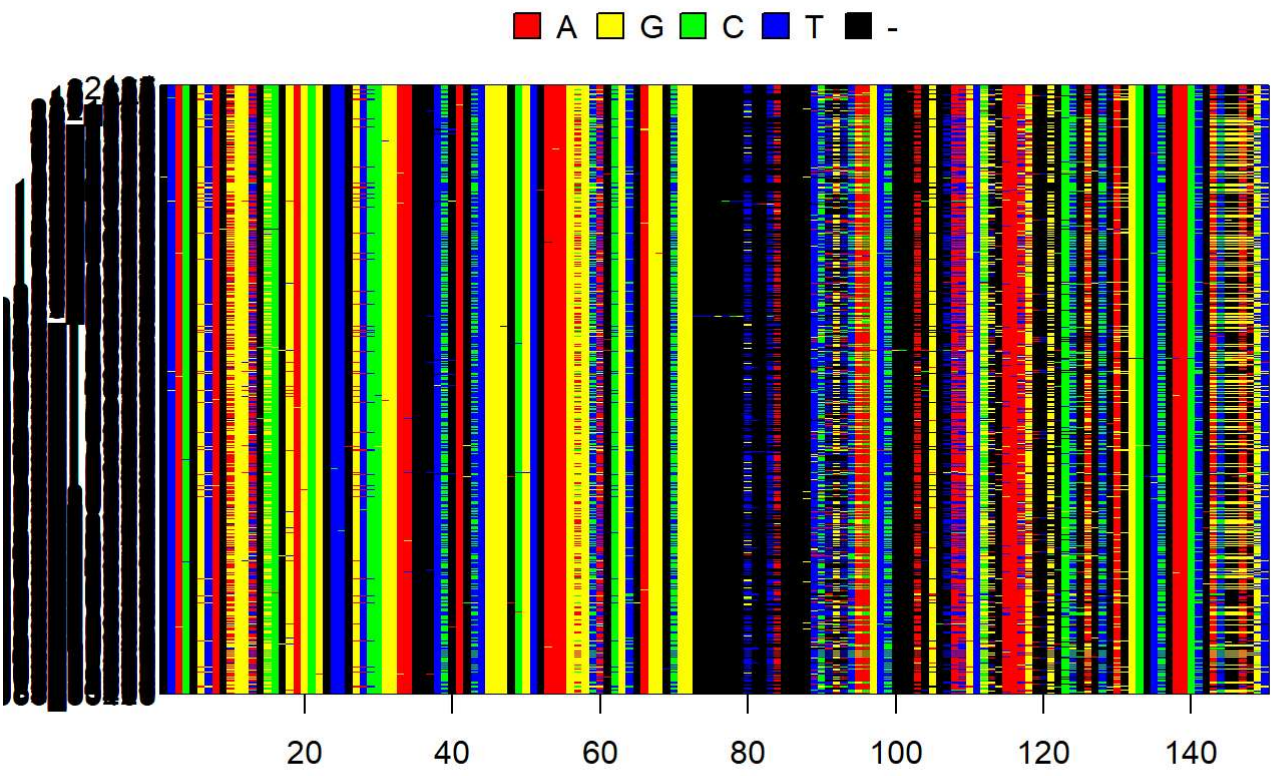
```
## Number of segregating sites (including gaps): 408
## Number of sites with at least one substitution: 322
## Number of sites with 1, 2, 3 or 4 observed bases:
##   1   2   3   4
##  14 141 102  79
```



The beginning of the alignment for the first half of the sequences. Note the large gaps right before and after site 80.

```
checkAlignment(dna_align[1:585, 1:150], what = 1)
```

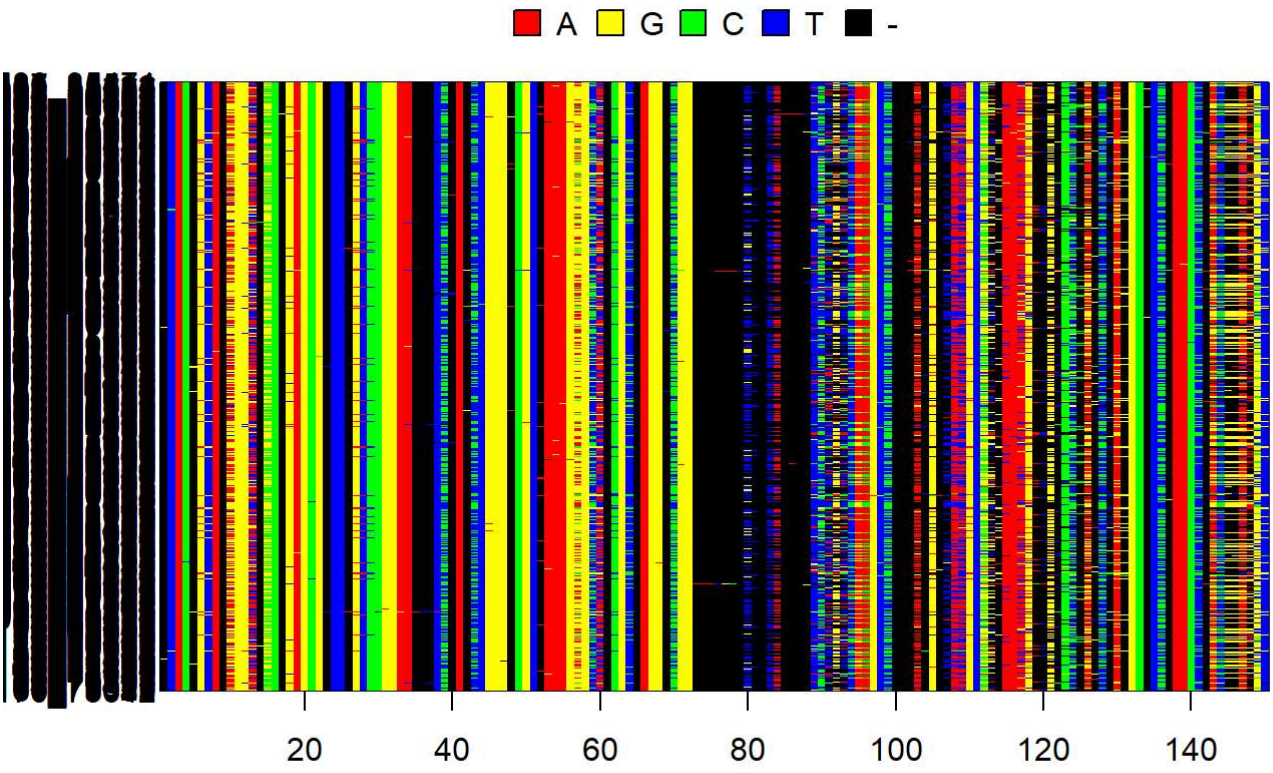
```
##
## Number of sequences: 585
## Number of sites: 150
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 13 16 22
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8    10    13    16    22
## 13245  1650  1264   564   215    40   283    45     3     1   301     1
##    => length of gaps on the left border of the alignment: 1 0
##    => length of gaps on the right border of the alignment: 6 1
##
## Number of unique contiguous base segments defined by gaps: 106
## Number of segment lengths not multiple of 3: 78
##     => on the left border of the alignement: 1
##     => on the right border                 : 4
##     => positions of these segments inside the alignment: 10..13 15..16 18..22 24..25 27..34 3
8..39 41..41 43..47 53..60 80..80 83..84 89..90 103..103 105..105 108..112 115..118 121..121 12
3..124 126..126 128..128 130..130 132..133 135..136 138..141 143..144 147..147 89..99 108..118 1
02..103 107..113 115..121 121..124 130..133 88..88 37..38 53..56 93..99 143..147 123..126 89..92
95..99 53..68 108..121 77..80 104..105 80..84 6..13 126..133 54..60 114..118 135..141 138..144 7
0..80 83..83 43..46 53..59 10..16 15..22 95..101 103..118 80..81 91..91 38..41 41..47 132..136 2
4..34 41..51 146..147 90..90 149..149 62..68 89..98 66..72
##
## Number of segregating sites (including gaps): 128
## Number of sites with at least one substitution: 105
## Number of sites with 1, 2, 3 or 4 observed bases:
##  1  2  3  4
## 22 44 34 27
```



The beginning of the alignment for the second half of the sequences. Note the large gaps right before and after site 80.

```
checkAlignment(dna_align[586:1170, 1:150], what = 1)
```

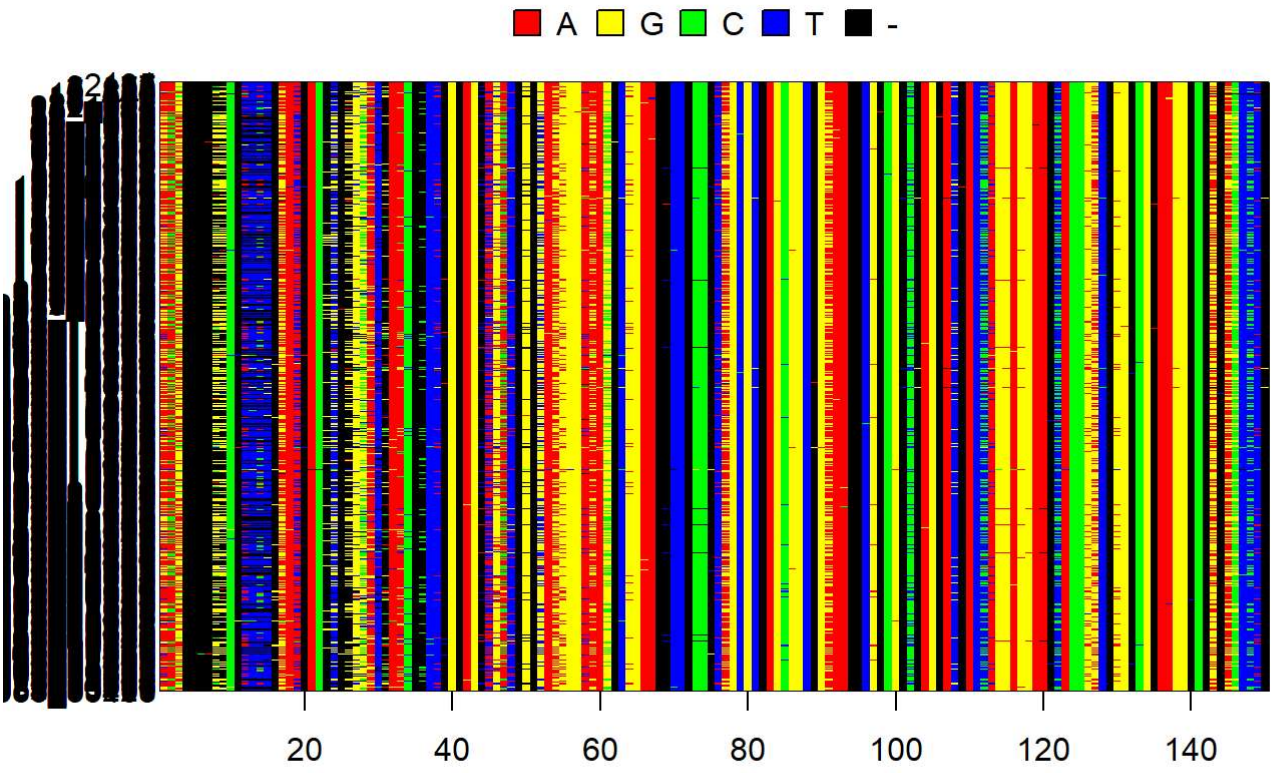
```
##
## Number of sequences: 585
## Number of sites: 150
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 11 13 16 17
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8    11    13    15    16    17
## 13075  1588  1267   568   207    64   271    51     1     2     1   314     1
##    => length of gaps on the left border of the alignment: 1 0
##    => length of gaps on the right border of the alignment: 1
##
## Number of unique contiguous base segments defined by gaps: 110
## Number of segment lengths not multiple of 3: 78
##     => on the left border of the alignement: 1
##     => on the right border                 : 4
##     => positions of these segments inside the alignment: 10..13 15..16 18..22 24..25 27..34 3
8..39 41..41 43..47 53..60 89..99 105..105 107..113 115..118 123..124 126..126 128..128 130..130
132..133 135..136 138..141 103..103 108..112 121..121 143..144 147..147 80..80 83..84 89..90 11
5..121 123..126 132..136 80..81 91..91 121..124 104..105 108..118 114..118 15..22 84..90 95..99
54..60 130..133 138..144 102..103 122..126 143..147 146..147 135..141 38..41 84..84 144..144 1
0..16 24..34 41..47 76..80 41..51 62..68 66..72 93..99 2..8 18..25 89..93 126..133 37..38 90..99
86..86 108..121 105..112 88..88 90..90 31..38 43..46 149..149
##
## Number of segregating sites (including gaps): 126
## Number of sites with at least one substitution: 103
## Number of sites with 1, 2, 3 or 4 observed bases:
##  1  2  3  4
## 24 48 32 23
```



The middle of the alignment for the first half of the sequences. Note the large gap before site 10.

```
checkAlignment(dna_align[1:585, 151:300], what = 1)
```

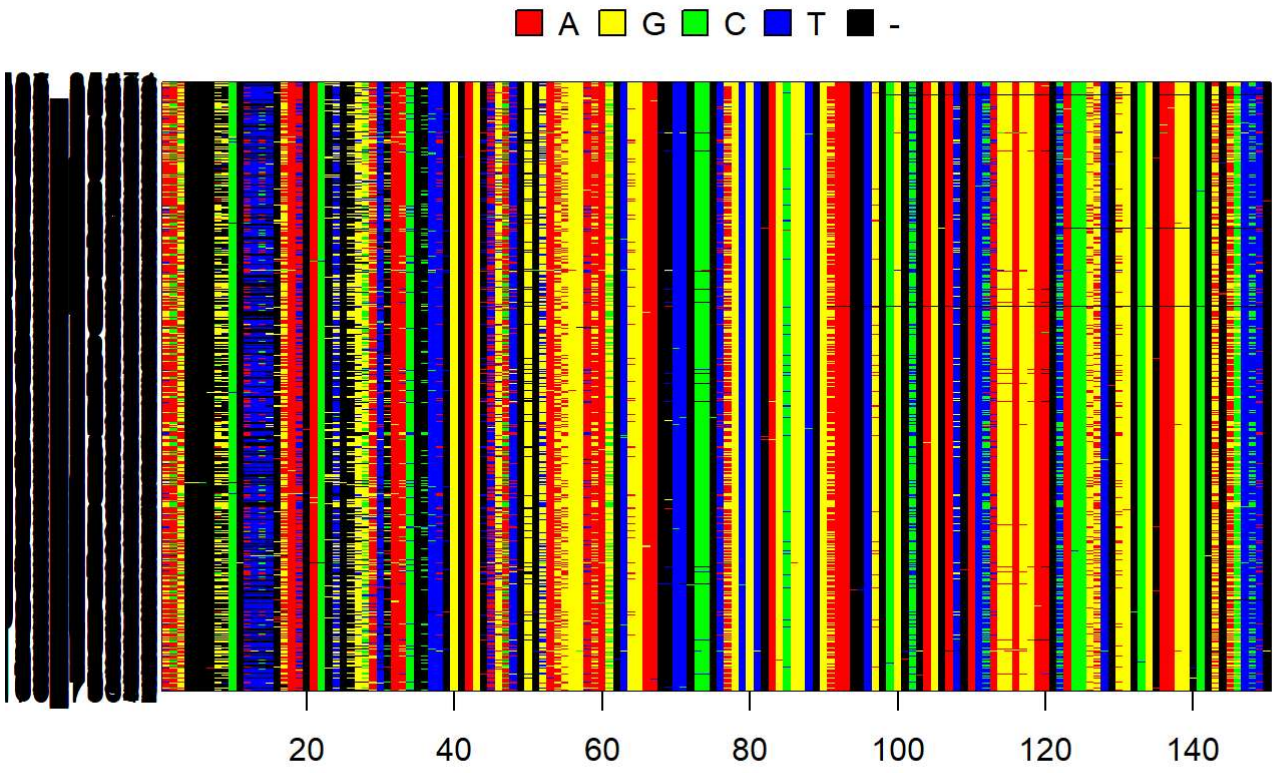
```
##
## Number of sequences: 585
## Number of sites: 150
##
## Some gap lengths are not multiple of 3: 1 2 4 7 8
##
## Frequencies of gap lengths:
##     1     2     3     4     6     7     8     9
## 15837  2088    67   349   495     1     1     1
##    => length of gaps on the left border of the alignment: 1
##    => length of gaps on the right border of the alignment: 9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Number of unique contiguous base segments defined by gaps: 84
## Number of segment lengths not multiple of 3: 62
##     => on the left border of the alignement: 0
##     => on the right border                 : 0
##     => positions of these segments inside the alignment: 10..10 12..15 21..22 24..24 27..30 3
7..38 40..40 42..43 45..48 50..50 52..61 63..67 70..71 73..74 90..93 96..97 99..100 102..102 10
4..105 107..108 110..120 122..128 130..131 133..134 136..139 141..141 143..143 145..149 27..34 4
2..48 26..30 8..8 26..29 96..100 104..108 120..120 52..53 55..61 7..10 21..24 52..67 18..19 11
0..119 17..24 107..120 99..105 133..139 37..40 28..34 16..19 77..81 130..134 83..93 143..149 11
0..128 8..15 12..19 71..71 145..146 148..149 2..3 6..10
##
## Number of segregating sites (including gaps): 119
## Number of sites with at least one substitution: 99
## Number of sites with 1, 2, 3 or 4 observed bases:
##  1  2  3  4
## 31 50 30 19
```



The middle of the alignment for the second half of the sequences. Note the large gap before site 10.

```
checkAlignment(dna_align[586:1170, 151:300], what = 1)
```

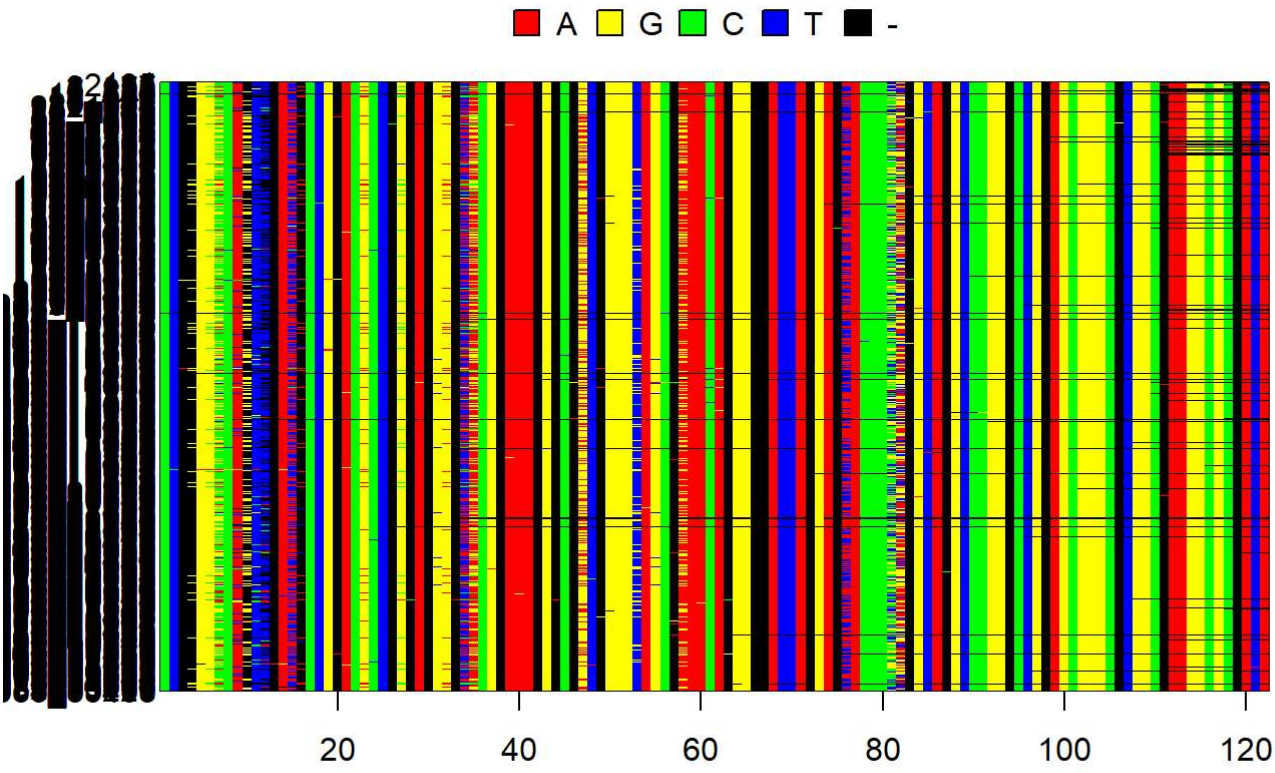
```
##
## Number of sequences: 585
## Number of sites: 150
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 35
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8     9    10    12    35    60
## 15776  2049    90   388     6   458     1     1     2     1     1     1     1
##    => no gap on the left border of the alignment
##    => length of gaps on the right border of the alignment: 60 9 9 7 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Number of unique contiguous base segments defined by gaps: 96
## Number of segment lengths not multiple of 3: 66
##     => on the left border of the alignement: 1
##     => on the right border                 : 0
##     => positions of these segments inside the alignment: 21..24 27..30 37..38 40..40 42..43 4
5..48 50..50 52..61 63..67 70..71 73..74 90..93 96..97 99..100 102..102 104..105 107..108 110..1
20 122..128 130..131 133..134 136..139 141..141 143..143 145..149 10..10 12..15 21..22 24..24 2
6..30 8..8 16..19 28..34 27..34 42..48 26..29 37..40 9..10 110..119 40..43 133..139 143..149 12
2..131 12..19 76..88 104..108 90..90 52..56 7..10 17..24 24..25 3..3 12..16 18..19 8..15 36..40
107..120 63..70 70..74 90..97 130..134 99..105 52..67 99..102 102..105
##
## Number of segregating sites (including gaps): 139
## Number of sites with at least one substitution: 110
## Number of sites with 1, 2, 3 or 4 observed bases:
##  1  2  3  4
## 11 57 31 22
```



The end of the alignment for the first half of the sequences. Note that there does not seem to be any large gaps as in the previous alignments but there are still some smaller gaps that are visible.

```
checkAlignment(dna_align[1:585, 301:422], what = 1)
```

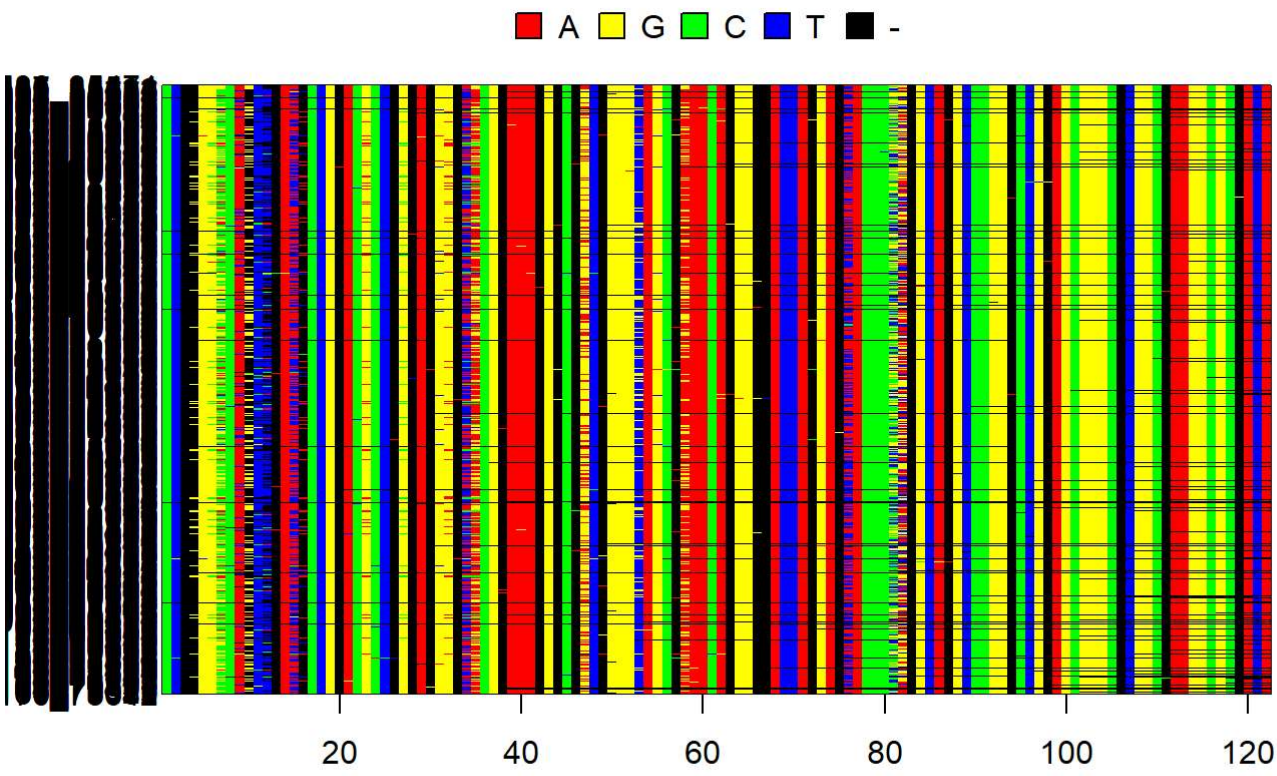
```
##
## Number of sequences: 585
## Number of sites: 122
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 13 16 20 22 23 25 26 31 40 41 49 97 12
2
##
## Frequencies of gap lengths:
##     1     2     3     4     5     7     8     9    10    12    13    15    16
## 13160  1312     3    47     5     4     1     2     2    23     3     2     3
##    20    21    22    23    24    25    26    27    31    33    36    40    41
##     1     4     1     1     1     6     1     3     1     1     2     1     1
##    42    48    49    60    97   114   120   122
##     1     2     3     1     1     1     1     1
##    => length of gaps on the left border of the alignment: 122 23
##    => length of gaps on the right border of the alignment: 122 120 114 97 60 49 49 49 48 48 4
1 40 36 36 33 31 27 27 26 25 25 25 25 25 22 21 21 15 15 13 13 13 12 12 12 12 12 12 12 12 12 12 1
2 12 12 12 12 12 12 12 12 12 12 12 9 7 5 5 5 5 4
##
## Number of unique contiguous base segments defined by gaps: 81
## Number of segment lengths not multiple of 3: 54
##     => on the left border of the alignement: 1
##     => on the right border                 : 0
##     => positions of these segments inside the alignment: 5..9 11..12 14..15 21..25 27..27 2
9..29 31..32 34..37 43..43 45..45 47..48 50..56 58..62 64..65 68..71 73..74 76..82 99..105 107..
110 112..118 107..117 5..11 76..86 5..8 21..27 112..113 73..73 73..82 88..89 95..95 24..25 37..3
7 59..62 45..48 50..62 76..80 82..82 88..91 107..107 99..100 84..85 112..115 5..12 97..97 14..14
84..93 31..37 58..65 47..56 85..86 6..9 95..96 65..65
##
## Number of segregating sites (including gaps): 113
## Number of sites with at least one substitution: 54
## Number of sites with 1, 2, 3 or 4 observed bases:
##  1  2  3  4
##  9 32 12 10
```



The end of the alignment for the second half of the sequences. Note that there does not seem to be any large gaps as in the previous alignments but there are still some smaller gaps that are visible.

```
checkAlignment(dna_align[586:1170, 301:422], what = 1)
```

```
##
## Number of sequences: 585
## Number of sites: 122
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 13 14 16 17 20 22 25 26 28 29 34 37 38
44 61 67 74 77 85 86 95 122
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8     9    10    12    13    14
## 12756  1281    11    50     2     1     2     3     4     5     9     3     1
##    15    16    17    18    20    21    22    25    26    28    29    33    34
##    10     2     1     1     3     9     1     2     4     3     1     2     1
##    36    37    38    42    44    48    51    57    60    61    66    67    69
##     2     1     1     1     3     4     1     4     1     5     2     1     1
##    74    77    81    85    86    95   114   122
##     2     3     1     4     1     1     1     2
##    => length of gaps on the left border of the alignment: 122 44 38
##    => length of gaps on the right border of the alignment: 122 122 114 95 86 85 85 85 85 81 7
7 77 77 74 74 69 67 66 66 61 61 60 57 57 57 57 51 48 48 48 48 42 36 36 33 33 26 26 26 26 25 25 2
2 21 21 21 21 21 21 21 21 21 15 15 15 15 15 15 15 13 13 12 12 12 12 12 12 12 12 12 9 9 7 6 5 5 4
4 4 4 4 2
##
## Number of unique contiguous base segments defined by gaps: 101
## Number of segment lengths not multiple of 3: 73
##     => on the left border of the alignement: 1
##     => on the right border                : 3
##     => positions of these segments inside the alignment: 5..11 14..15 21..25 27..27 29..29 3
1..32 34..37 43..43 45..45 47..48 50..56 58..62 64..65 68..71 73..74 76..82 99..105 107..110 11
2..118 5..9 11..12 59..62 95..96 50..62 68..74 84..93 88..88 90..93 47..62 5..8 96..97 5..12 1
4..14 58..61 45..48 95..105 88..97 58..65 32..32 112..113 39..43 88..91 93..93 85..86 120..120 1
15..118 29..32 112..115 99..100 107..107 64..71 31..37 47..56 88..89 34..41 73..82 21..27 89..93
11..15 4..7 95..95 97..97 4..11 76..80 112..116 50..53 34..34 36..37 107..117
##
## Number of segregating sites (including gaps): 119
## Number of sites with at least one substitution: 63
## Number of sites with 1, 2, 3 or 4 observed bases:
##  1  2  3  4
##  3 37 19  7
```

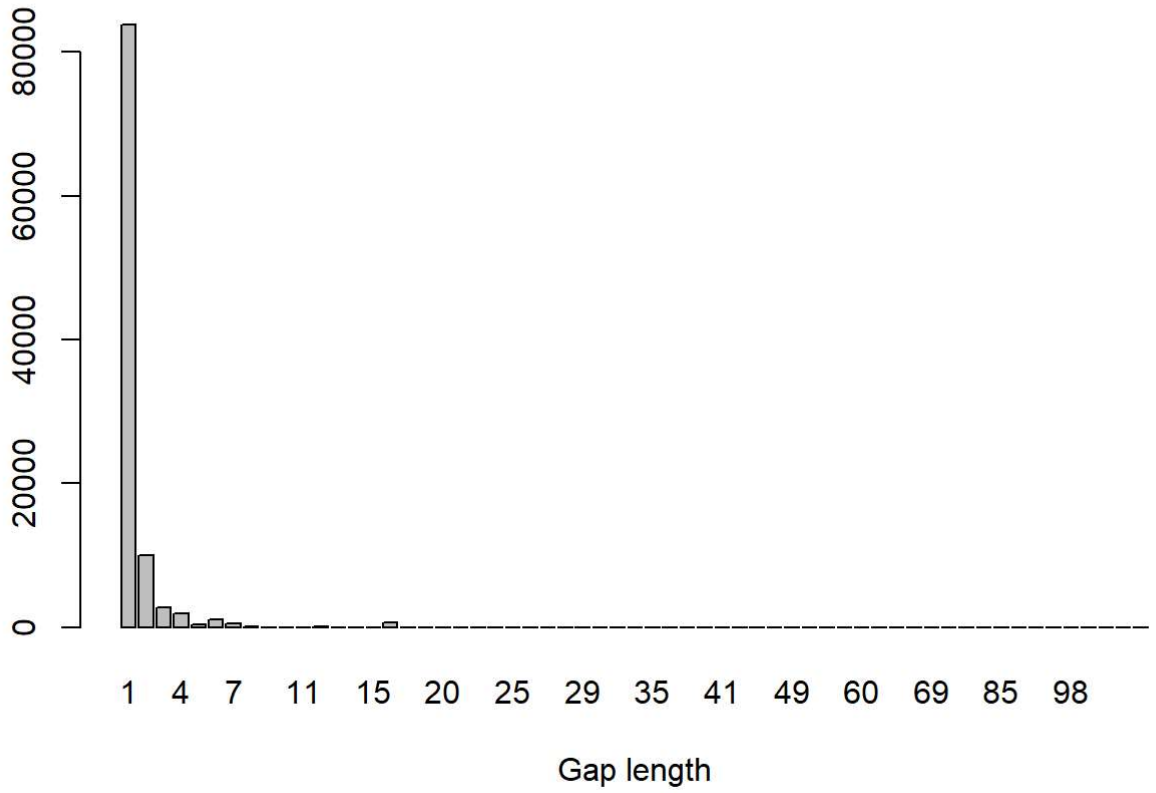


The distribution of gap lengths. Note that there are over 8000 gaps of length 1.

```
checkAlignment(dna_align, what = 2)
```

```
##
## Number of sequences: 1170
## Number of sites: 422
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 11 13 14 16 17 20 22 25 26 28 29 31 34
35 37 40 41 49 53 61 67 74 77 85 86 95 97 98 131
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 83845  9968  2702  1966   435  1057   562   102     6    11     1    33     9
##    14    15    16    17    18    20    21    22    24    25    26    27    28
##     1    13   620     2     1     4    13     3     2     8     5     3     3
##    29    31    33    34    35    36    37    40    41    42    45    48    49
##     1     1     3     1     1     4     1     1     1     2     2     6     3
##    51    53    57    60    61    66    67    69    74    77    81    85    86
##     1     1     4     2     5     2     1     1     2     3     1     4     1
##    95    97    98   114   120   129   131
##     1     1     1     2     1     1     2
##    => length of gaps on the left border of the alignment: 1 0
##    => length of gaps on the right border of the alignment: 131 131 129 120 114 114 97 95 86 8
5 85 85 85 81 77 77 77 74 74 69 67 66 66 61 61 60 60 57 57 57 57 51 49 49 49 48 48 48 48 48 48 4
2 41 40 36 36 36 36 33 33 33 31 27 27 26 26 26 26 26 25 25 25 25 25 25 25 22 22 21 21 21 21 21 2
1 21 21 21 21 21 15 15 15 15 15 15 15 15 15 13 13 13 13 13 12 12 12 12 12 12 12 12 12 12 12 12 1
2 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 9 9 9 7 7 6 5 5 5 5 5 5 4 4 4 4 4 4 2
##
## Number of unique contiguous base segments defined by gaps: 357
## Number of segment lengths not multiple of 3: 250
##     => on the left border of the alignement: 1
##     => on the right border                : 3
##     => positions of these segments inside the alignment: 10..13 15..16 18..22 24..25 27..34 3
8..39 41..41 43..47 53..60 80..80 83..84 89..90 103..103 105..105 108..112 115..118 121..121 12
3..124 126..126 128..128 130..130 132..133 135..136 138..141 143..144 147..147 149..153 160..160
162..165 171..172 174..174 177..180 187..188 190..190 192..193 195..198 200..200 202..211 213..2
17 220..221 223..224 240..243 246..247 249..250 252..252 254..255 257..258 260..270 272..278 28
0..281 283..284 286..289 291..291 293..293 295..299 301..302 305..309 311..312 314..315 321..325
327..327 329..329 331..332 334..337 343..343 345..345 347..348 350..356 358..362 364..365 368..3
71 373..374 376..382 399..405 407..410 412..418 407..417 89..99 108..118 147..153 177..184 192..
198 102..103 107..113 143..153 176..180 305..311 115..121 158..158 121..124 130..133 176..179 37
6..386 88..88 37..38 53..56 93..99 305..308 143..147 246..250 123..126 254..258 89..92 321..327
270..270 202..203 205..211 95..99 157..160 171..174 202..217 168..169 412..413 260..269 53..68 1
08..121 167..174 257..270 77..80 104..105 249..255 295..302 80..84 283..289 373..373 187..190
6..13 126..133 373..382 54..60 114..118 178..184 388..389 135..141 138..144 395..395 70..80 83..
83 324..325 337..337 43..46 53..59 10..16 15..22 95..101 103..118 80..81 91..91 166..169 227..23
1 280..284 359..362 38..41 233..243 345..348 350..362 376..380 382..382 293..299 260..278 388..3
91 41..47 132..136 407..407 24..34 158..165 399..400 384..385 412..415 162..169 221..221 295..29
6 298..299 305..312 41..51 146..147 397..397 90..90 314..314 149..149 384..393 138..153 62..68 3
31..337 358..365 347..356 89..98 152..153 156..160 385..386 306..309 66..72 395..396 365..365 8
4..90 368..374 388..388 390..393 347..362 159..160 396..397 358..361 122..126 190..193 272..281
395..405 84..84 388..397 144..144 226..238 332..332 76..80 339..343 393..393 240..240 420..420 2
02..206 415..418 2..8 18..25 329..332 89..93 364..371 334..341 90..99 86..86 389..393 143..156 1
74..175 311..315 105..112 304..307 304..311 153..153 162..166 186..190 213..220 220..224 240..24
7 412..416 31..38 350..353 249..252 252..255 334..334 336..337
##
```

```
## Number of segregating sites (including gaps): 408
## Number of sites with at least one substitution: 322
## Number of sites with 1, 2, 3 or 4 observed bases:
##   1   2   3   4
##  14 141 102  79
```

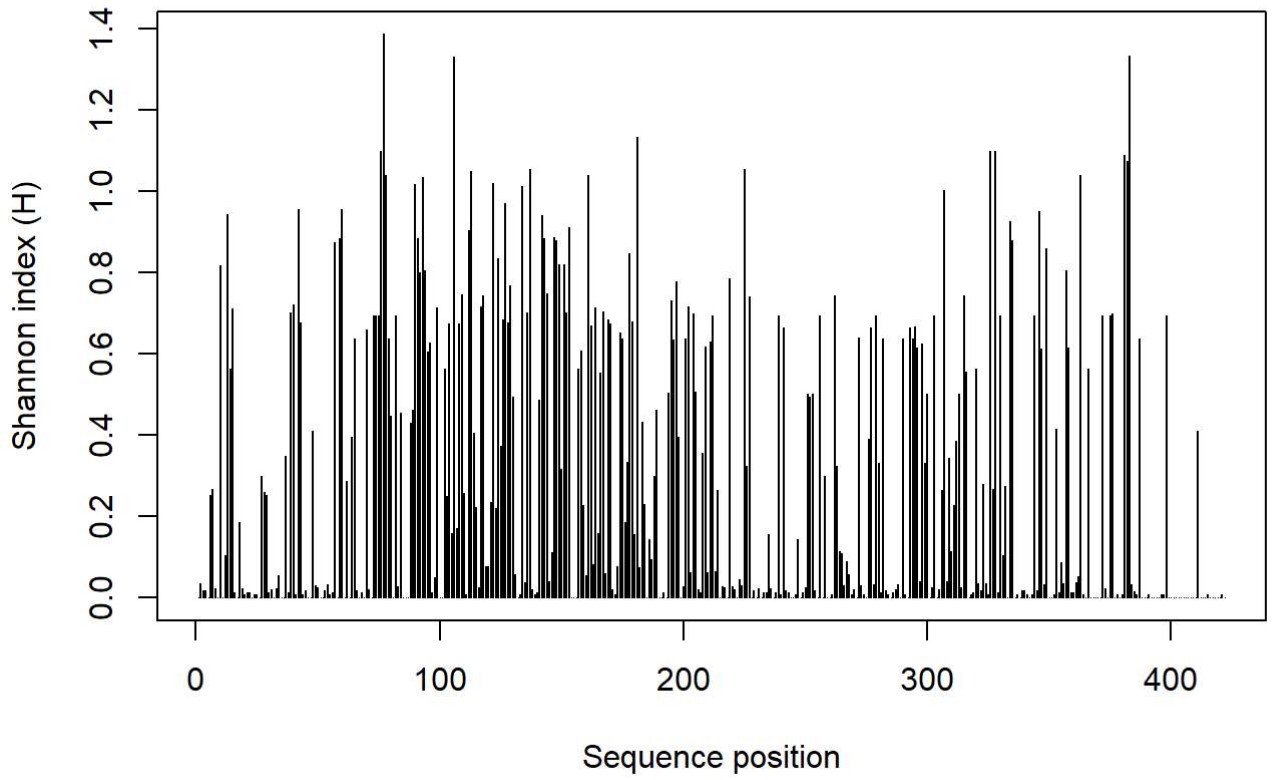


The shannon index (H) for each sequence position.

```
checkAlignment(dna_align, what = 3)
```

```
##
## Number of sequences: 1170
## Number of sites: 422
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 11 13 14 16 17 20 22 25 26 28 29 31 34
35 37 40 41 49 53 61 67 74 77 85 86 95 97 98 131
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 83845  9968  2702  1966   435  1057   562   102     6    11     1    33     9
##    14    15    16    17    18    20    21    22    24    25    26    27    28
##     1    13   620     2     1     4    13     3     2     8     5     3     3
##    29    31    33    34    35    36    37    40    41    42    45    48    49
##     1     1     3     1     1     4     1     1     1     2     2     6     3
##    51    53    57    60    61    66    67    69    74    77    81    85    86
##     1     1     4     2     5     2     1     1     2     3     1     4     1
##    95    97    98   114   120   129   131
##     1     1     1     2     1     1     2
##    => length of gaps on the left border of the alignment: 1 0
##    => length of gaps on the right border of the alignment: 131 131 129 120 114 114 97 95 86 8
5 85 85 85 81 77 77 77 74 74 69 67 66 66 61 61 60 60 57 57 57 57 51 49 49 49 48 48 48 48 48 48 4
2 41 40 36 36 36 36 33 33 33 31 27 27 26 26 26 26 26 25 25 25 25 25 25 25 22 22 21 21 21 21 21 2
1 21 21 21 21 21 15 15 15 15 15 15 15 15 15 13 13 13 13 13 12 12 12 12 12 12 12 12 12 12 12 12 1
2 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 9 9 9 7 7 6 5 5 5 5 5 5 4 4 4 4 4 4 2
##
## Number of unique contiguous base segments defined by gaps: 357
## Number of segment lengths not multiple of 3: 250
##     => on the left border of the alignement: 1
##     => on the right border                : 3
##     => positions of these segments inside the alignment: 10..13 15..16 18..22 24..25 27..34 3
8..39 41..41 43..47 53..60 80..80 83..84 89..90 103..103 105..105 108..112 115..118 121..121 12
3..124 126..126 128..128 130..130 132..133 135..136 138..141 143..144 147..147 149..153 160..160
162..165 171..172 174..174 177..180 187..188 190..190 192..193 195..198 200..200 202..211 213..2
17 220..221 223..224 240..243 246..247 249..250 252..252 254..255 257..258 260..270 272..278 28
0..281 283..284 286..289 291..291 293..293 295..299 301..302 305..309 311..312 314..315 321..325
327..327 329..329 331..332 334..337 343..343 345..345 347..348 350..356 358..362 364..365 368..3
71 373..374 376..382 399..405 407..410 412..418 407..417 89..99 108..118 147..153 177..184 192..
198 102..103 107..113 143..153 176..180 305..311 115..121 158..158 121..124 130..133 176..179 37
6..386 88..88 37..38 53..56 93..99 305..308 143..147 246..250 123..126 254..258 89..92 321..327
270..270 202..203 205..211 95..99 157..160 171..174 202..217 168..169 412..413 260..269 53..68 1
08..121 167..174 257..270 77..80 104..105 249..255 295..302 80..84 283..289 373..373 187..190
6..13 126..133 373..382 54..60 114..118 178..184 388..389 135..141 138..144 395..395 70..80 83..
83 324..325 337..337 43..46 53..59 10..16 15..22 95..101 103..118 80..81 91..91 166..169 227..23
1 280..284 359..362 38..41 233..243 345..348 350..362 376..380 382..382 293..299 260..278 388..3
91 41..47 132..136 407..407 24..34 158..165 399..400 384..385 412..415 162..169 221..221 295..29
6 298..299 305..312 41..51 146..147 397..397 90..90 314..314 149..149 384..393 138..153 62..68 3
31..337 358..365 347..356 89..98 152..153 156..160 385..386 306..309 66..72 395..396 365..365 8
4..90 368..374 388..388 390..393 347..362 159..160 396..397 358..361 122..126 190..193 272..281
395..405 84..84 388..397 144..144 226..238 332..332 76..80 339..343 393..393 240..240 420..420 2
02..206 415..418 2..8 18..25 329..332 89..93 364..371 334..341 90..99 86..86 389..393 143..156 1
74..175 311..315 105..112 304..307 304..311 153..153 162..166 186..190 213..220 220..224 240..24
7 412..416 31..38 350..353 249..252 252..255 334..334 336..337
##
```

```
## Number of segregating sites (including gaps): 408
## Number of sites with at least one substitution: 322
## Number of sites with 1, 2, 3 or 4 observed bases:
##   1   2   3   4
##  14 141 102  79
```

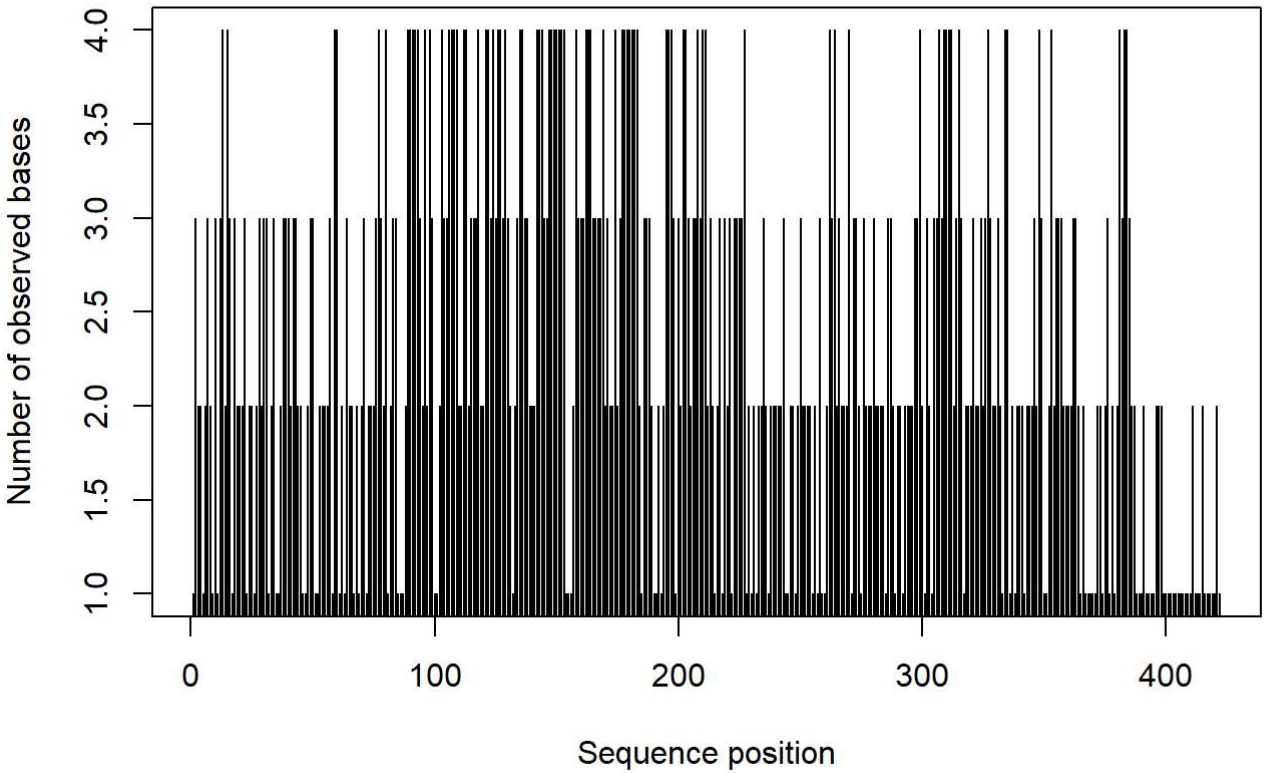


The number of observed bases for each sequence position. Note that all positions have 1.0 base and most positions have 2.0 bases.

```
checkAlignment(dna_align, what = 4)
```

```
##
## Number of sequences: 1170
## Number of sites: 422
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 11 13 14 16 17 20 22 25 26 28 29 31 34
35 37 40 41 49 53 61 67 74 77 85 86 95 97 98 131
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 83845  9968  2702  1966   435  1057   562   102     6    11     1    33     9
##    14    15    16    17    18    20    21    22    24    25    26    27    28
##     1    13   620     2     1     4    13     3     2     8     5     3     3
##    29    31    33    34    35    36    37    40    41    42    45    48    49
##     1     1     3     1     1     4     1     1     1     2     2     6     3
##    51    53    57    60    61    66    67    69    74    77    81    85    86
##     1     1     4     2     5     2     1     1     2     3     1     4     1
##    95    97    98   114   120   129   131
##     1     1     1     2     1     1     2
##    => length of gaps on the left border of the alignment: 1 0
##    => length of gaps on the right border of the alignment: 131 131 129 120 114 114 97 95 86 8
5 85 85 85 81 77 77 77 74 74 69 67 66 66 61 61 60 60 57 57 57 57 51 49 49 49 48 48 48 48 48 48 4
2 41 40 36 36 36 36 33 33 33 31 27 27 26 26 26 26 26 25 25 25 25 25 25 25 22 22 21 21 21 21 21 2
1 21 21 21 21 21 15 15 15 15 15 15 15 15 15 13 13 13 13 13 12 12 12 12 12 12 12 12 12 12 12 12 1
2 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 9 9 9 7 7 6 5 5 5 5 5 5 4 4 4 4 4 4 2
##
## Number of unique contiguous base segments defined by gaps: 357
## Number of segment lengths not multiple of 3: 250
##     => on the left border of the alignement: 1
##     => on the right border                : 3
##     => positions of these segments inside the alignment: 10..13 15..16 18..22 24..25 27..34 3
8..39 41..41 43..47 53..60 80..80 83..84 89..90 103..103 105..105 108..112 115..118 121..121 12
3..124 126..126 128..128 130..130 132..133 135..136 138..141 143..144 147..147 149..153 160..160
162..165 171..172 174..174 177..180 187..188 190..190 192..193 195..198 200..200 202..211 213..2
17 220..221 223..224 240..243 246..247 249..250 252..252 254..255 257..258 260..270 272..278 28
0..281 283..284 286..289 291..291 293..293 295..299 301..302 305..309 311..312 314..315 321..325
327..327 329..329 331..332 334..337 343..343 345..345 347..348 350..356 358..362 364..365 368..3
71 373..374 376..382 399..405 407..410 412..418 407..417 89..99 108..118 147..153 177..184 192..
198 102..103 107..113 143..153 176..180 305..311 115..121 158..158 121..124 130..133 176..179 37
6..386 88..88 37..38 53..56 93..99 305..308 143..147 246..250 123..126 254..258 89..92 321..327
270..270 202..203 205..211 95..99 157..160 171..174 202..217 168..169 412..413 260..269 53..68 1
08..121 167..174 257..270 77..80 104..105 249..255 295..302 80..84 283..289 373..373 187..190
6..13 126..133 373..382 54..60 114..118 178..184 388..389 135..141 138..144 395..395 70..80 83..
83 324..325 337..337 43..46 53..59 10..16 15..22 95..101 103..118 80..81 91..91 166..169 227..23
1 280..284 359..362 38..41 233..243 345..348 350..362 376..380 382..382 293..299 260..278 388..3
91 41..47 132..136 407..407 24..34 158..165 399..400 384..385 412..415 162..169 221..221 295..29
6 298..299 305..312 41..51 146..147 397..397 90..90 314..314 149..149 384..393 138..153 62..68 3
31..337 358..365 347..356 89..98 152..153 156..160 385..386 306..309 66..72 395..396 365..365 8
4..90 368..374 388..388 390..393 347..362 159..160 396..397 358..361 122..126 190..193 272..281
395..405 84..84 388..397 144..144 226..238 332..332 76..80 339..343 393..393 240..240 420..420 2
02..206 415..418 2..8 18..25 329..332 89..93 364..371 334..341 90..99 86..86 389..393 143..156 1
74..175 311..315 105..112 304..307 304..311 153..153 162..166 186..190 213..220 220..224 240..24
7 412..416 31..38 350..353 249..252 252..255 334..334 336..337
##
```

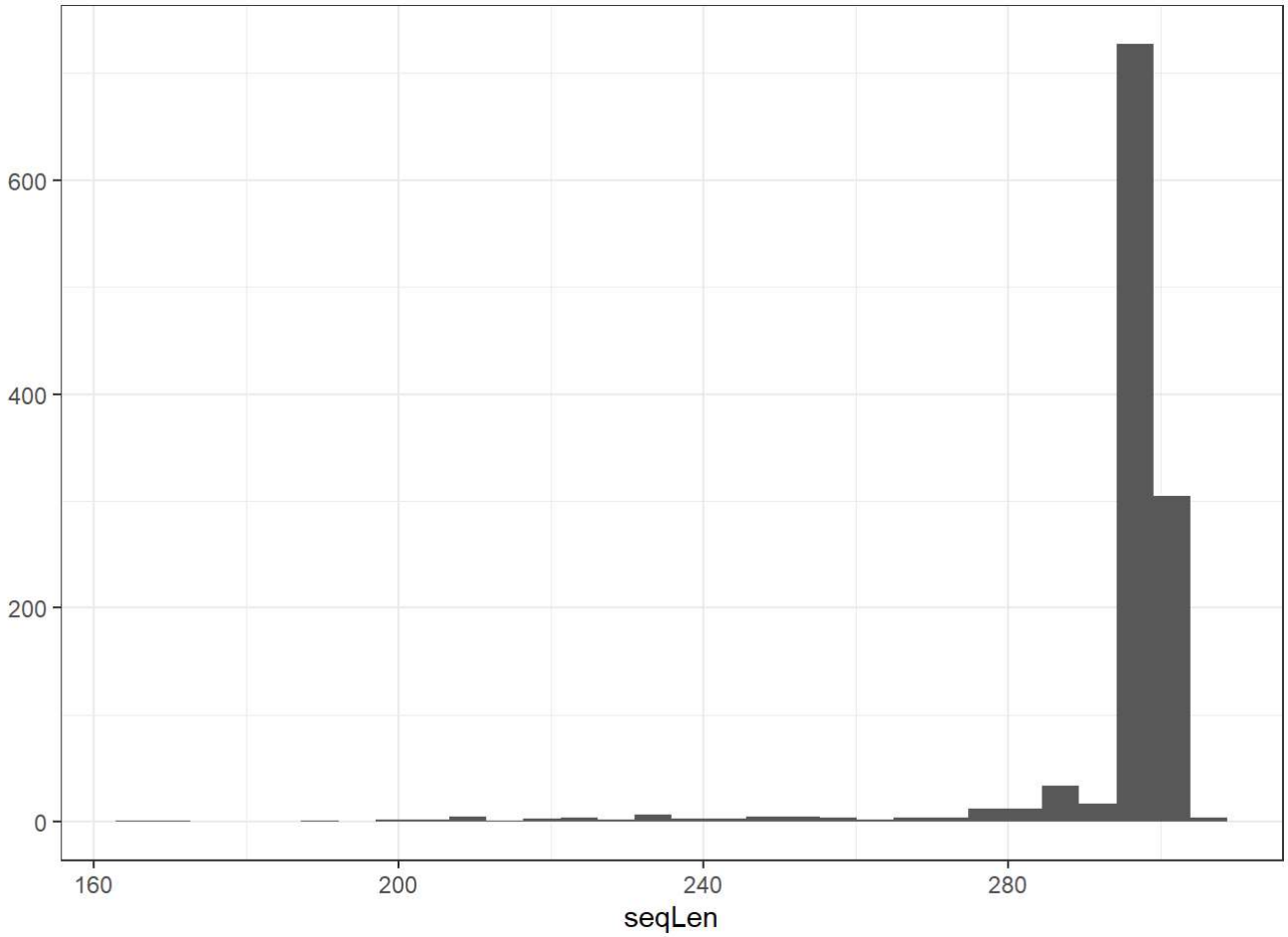
```
## Number of segregating sites (including gaps): 408
## Number of sites with at least one substitution: 322
## Number of sites with 1, 2, 3 or 4 observed bases:
##   1   2   3   4
##  14 141 102  79
```



We are going to remove the sequences with too many gaps. First we will visualize the gaps.

```
seqLen <- as.numeric(lapply(dna_bin, length))
qplot(seqLen) +
  theme_bw()
```

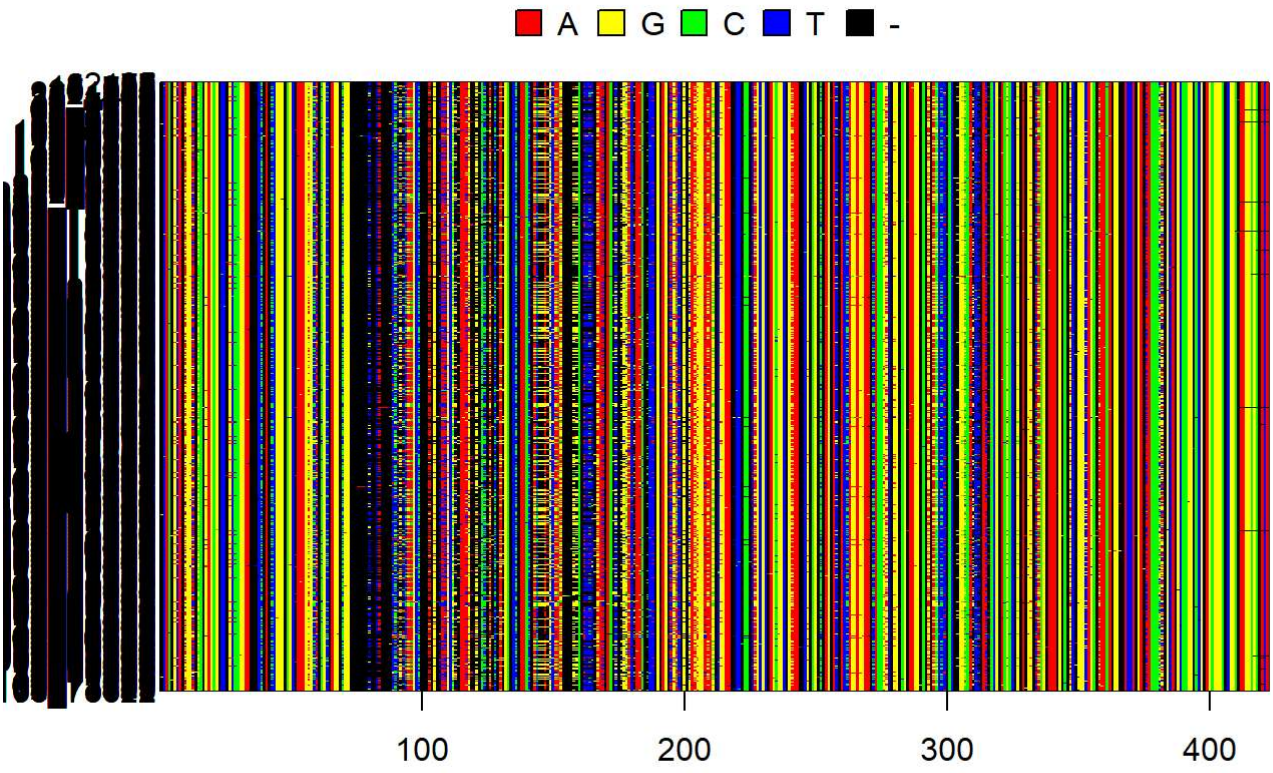
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the graph above, we will use 290 as a cutoff and then visualize the alignment without those gaps.

```
keepSeq <- seqLen > 290
dna_subset <- dna_align[keepSeq, ]
checkAlignment(dna_subset, what = 1)
```

```
##
## Number of sequences: 1048
## Number of sites: 422
##
## Some gap lengths are not multiple of 3: 1 2 4 5 7 8 10 11 13 16 17 22
##
## Frequencies of gap lengths:
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 76122  8980  2428  1777   392   945   509    86     2     3     1     5     2
##    15    16    17    22
##     1   544     1     1
##    => length of gaps on the left border of the alignment: 1 0
##    => length of gaps on the right border of the alignment: 13 12 12 12 12 12 9 7 7 6 5 5 5 5
5 5 4 4 4 4 4 2
##
## Number of unique contiguous base segments defined by gaps: 309
## Number of segment lengths not multiple of 3: 223
##     => on the left border of the alignement: 1
##     => on the right border                 : 2
##     => positions of these segments inside the alignment: 10..13 15..16 18..22 24..25 27..34 3
8..39 41..41 43..47 53..60 80..80 83..84 89..90 103..103 105..105 108..112 115..118 121..121 12
3..124 126..126 128..128 130..130 132..133 135..136 138..141 143..144 147..147 149..153 160..160
162..165 171..172 174..174 177..180 187..188 190..190 192..193 195..198 200..200 202..211 213..2
17 220..221 223..224 240..243 246..247 249..250 252..252 254..255 257..258 260..270 272..278 28
0..281 283..284 286..289 291..291 293..293 295..299 301..302 305..309 311..312 314..315 321..325
327..327 329..329 331..332 334..337 343..343 345..345 347..348 350..356 358..362 364..365 368..3
71 373..374 376..382 399..405 407..410 412..418 407..417 89..99 108..118 147..153 177..184 192..
198 115..121 158..158 107..113 143..153 176..180 305..311 176..179 376..386 88..88 37..38 53..56
93..99 305..308 123..126 254..258 89..92 321..327 270..270 202..203 205..211 412..413 143..147 2
60..269 53..68 108..121 171..174 168..169 167..174 257..270 77..80 95..99 104..105 249..255 29
5..302 80..84 187..190 6..13 126..133 130..133 54..60 114..118 121..124 178..184 135..141 138..1
44 43..46 53..59 283..289 10..16 15..22 95..101 103..118 80..81 91..91 166..169 227..231 38..41
233..243 345..348 350..362 376..380 382..382 260..278 24..34 158..165 384..385 412..415 162..169
221..221 295..296 298..299 305..312 90..90 314..314 149..149 384..393 138..153 41..51 62..68 33
1..337 358..365 347..356 89..98 152..153 280..284 306..309 66..72 84..90 368..374 388..388 390..
393 347..362 159..160 396..397 102..103 122..126 293..299 395..405 84..84 388..397 144..144 76..
80 339..343 388..391 393..393 41..47 420..420 146..147 202..206 385..386 2..8 18..25 329..332 27
2..281 89..93 157..160 364..371 334..341 90..99 373..382 226..238 389..393 143..156 174..175 10
5..112 304..307 311..315 395..395 397..397 213..220 220..224 240..247 412..416 202..217 249..252
252..255 334..334 336..337
##
## Number of segregating sites (including gaps): 365
## Number of sites with at least one substitution: 293
## Number of sites with 1, 2, 3 or 4 observed bases:
##   1   2   3   4
##  57 139  84  70
```



Now we will redo the alignment without the sequences with the large gaps.

```
dna_sub_align <- muscle(dna_subset, quiet = F)
```

Analyze

We used the dist.dna function to compute a matrix of pairwise distances from DNA sequences using the k80 model of DNA evolution.

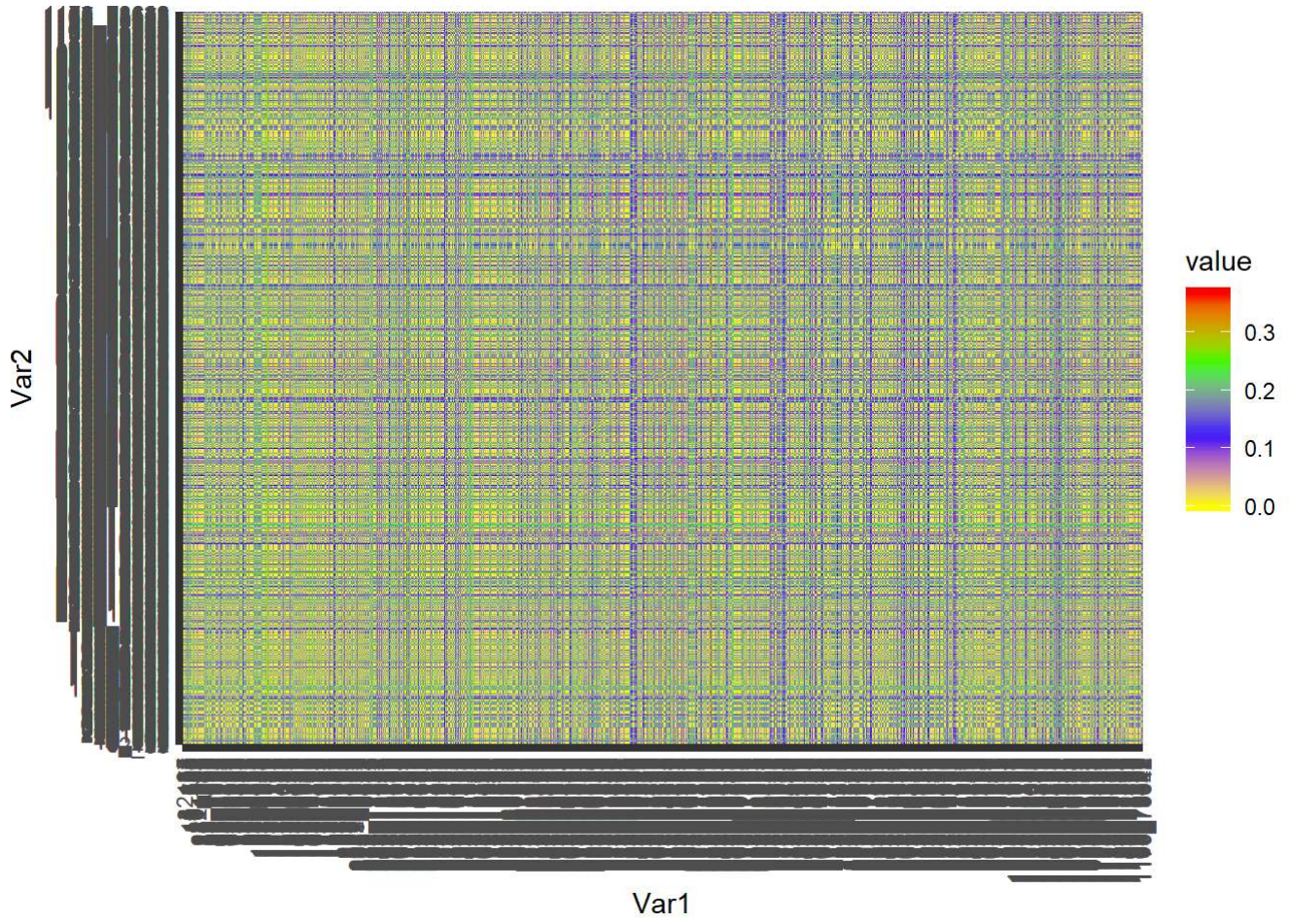
```
dnaDM<-dist.dna(dna_sub_align, model="K80")
dnaDMmat<-as.matrix(dnaDM)
```

Next we generated a heat-map comparing the sequence similarities, however, to do this we had to utilize the melt function to turn our 'DM' file into a linear matrix which can then be plotted as a heat map. We then used ggplot to construct the heatmap according to certain specifications.

```
PDat<-melt(dnaDMmat)
dim(PDat)
```

```
## [1] 1098304       3
```

```
ggplot(data = PDat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+scale_fill_gradientn(colours=c("yellow","blue","green","red")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```




```
pdf(width=20,height=10, "HeatMap_B1285.pdf")
ggplot(data = PDat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+scale_fill_gradientn(colours=c("yellow","blue","green","red")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
dev.off()
```

```
## png
##   2
```

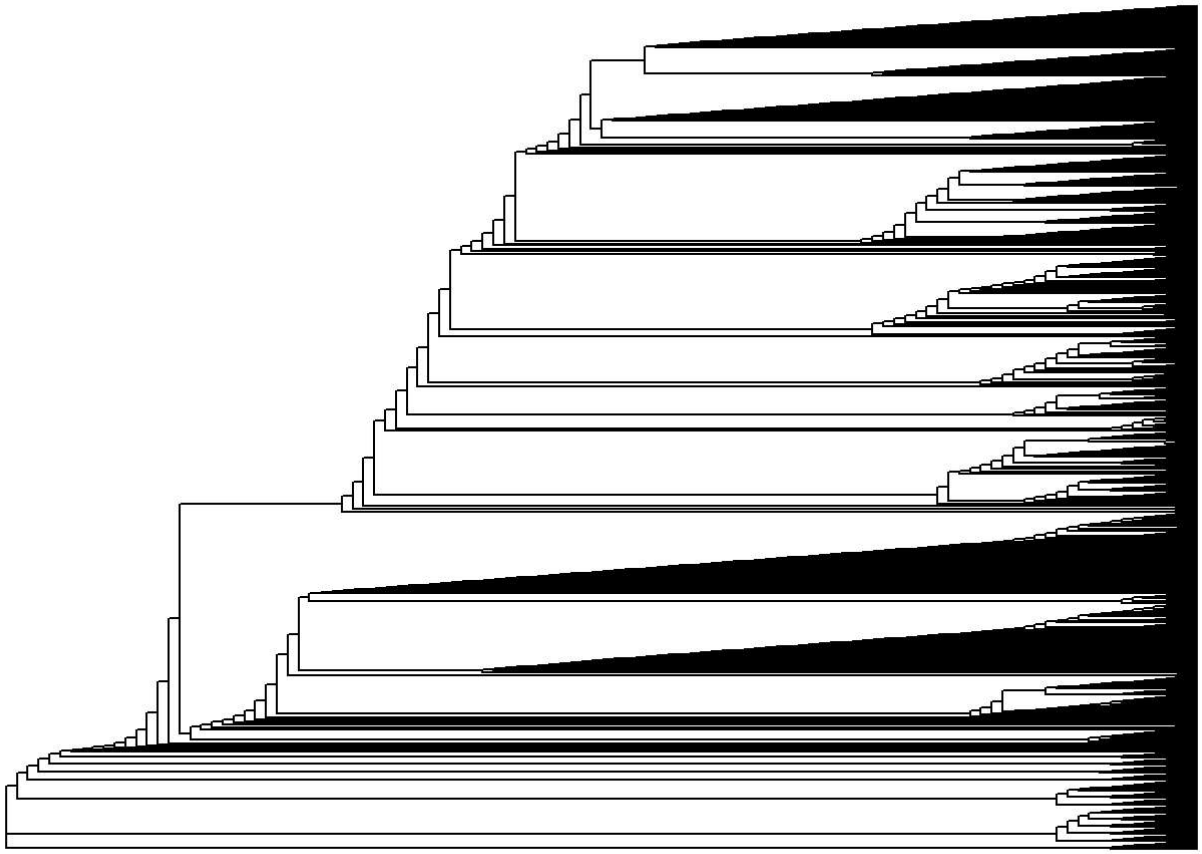
Phylogenetic Tree Building

To understand the evolutionary releationship among all 1170 samples, a phylogentic tree using the neighbour joining method was created. The distance matrix created earlier was inputted as the data. The neighbour joining method was used over the minimium evolution method becasuse each sequence read was relatively small (<300) while the number of taxa compared was high. The branch length was specified to ‘none’ to better focus on the relationship among the taxa rather than the evolutionary distance among them. We decided not to include tip labels because we are interested in the overall evolutionary relationship not the relationships between individual sequences. In addtion, not including the tip labels makes the tree more readable.

```
dnaTree<-nj(dnaDM)
str(dnaTree)
```

```
## List of 4
##  $ edge       : int [1:2093, 1:2] 1049 1064 1078 1092 1106 1120 1134 1148 1162 1162 ...
##  $ edge.length: num [1:2093] 3.39e-10 7.44e-10 1.59e-09 4.33e-09 1.31e-08 ...
##  $ tip.label  : chr [1:1048] "2_35" "3_127" "4_172" "5_194" ...
##  $ Nnode      : int 1046
##  - attr(*, "class")= chr "phylo"
##  - attr(*, "order")= chr "cladewise"
```

```
ggtree(dnaTree, branch.length='none')
```



```
pdf(width=20,height=10, "PhylogeneticTree_B1285")
ggtree(dnaTree, branch.length='none')
dev.off()
```

```
## png
##   2
```