# Amplicon 72 Analysis

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
color_ref <- "#1E88E5"
color_n <- "#B5B3AD"
color_mut <- "#D81B60"

residues <- read_delim("./residues/uk/residue21987.tsv", delim = " ", col_names = c("sequence_name", "lo
  separate(sequence_name, into = c("country", "coguk_id", "year"), sep = "/") %>%
  rename(res_21987 = value)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   sequence_name = col_character(),
##   location = col_double(),
##   value = col_character()
## )
```

```r
residues_21846 <- read_delim("./residues/uk/residue21846.tsv", delim = " ", col_names = c("sequence_name
  separate(sequence_name, into = c("country", "coguk_id", "year"), sep = "/") %>%
  select(coguk_id, value) %>%
  rename(residue21846 = value)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   sequence_name = col_character(),
##   location = col_double(),
##   value = col_character()
## )
```

```
metadata <- read_csv("./data/processed_metadata.csv.gz") %>% separate(sequence_name, into = c("country"
```

```
## Warning: Missing column names filled in: ’X1’ [1]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   sample_date = col_date(format = ""),
##   epi_week = col_double(),
##   lineage_conflict = col_double(),
##   lineage_ambiguity_score = col_double(),
##   scorpio_support = col_double(),
##   scorpio_conflict = col_double()
## )
## i Use ‘spec()‘ for the full column specifications.
```

```
subset_with_ct_data <- read_csv("./data/subset_with_ct_data_and_seqed_at_sanger.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   coguk_id = col_character()
## )
```

```
metadata <- metadata %>% inner_join(subset_with_ct_data)
```

```
## Joining, by = "coguk_id"
```

```
meta_residues <- inner_join(metadata, residues)
```

```
## Joining, by = c("country", "coguk_id", "year")
```

```
everything <- meta_residues


everything <- inner_join(everything, residues_21846)
```

```
## Joining, by = "coguk_id"
```

```
everything <- everything %>% mutate(week = lubridate::floor_date(sample_date, "weeks"))
```

```
delta <- everything %>%
  filter(sample_date < "2021-07-01", sample_date > "2021-03-01") %>%
  filter(scorpio_call == "Delta (B.1.617.2-like)") %>%
  mutate(has_g142d_call = grepl("G142D", mutations))
table(delta$has_g142d_call) / nrow(delta)
```

```
##
##     FALSE      TRUE
## 0.3510039 0.6489961
```

```
table(delta$value) / nrow(delta)
```

```
## Warning: Unknown or uninitialised column: ‘value‘.

## numeric(0)
```

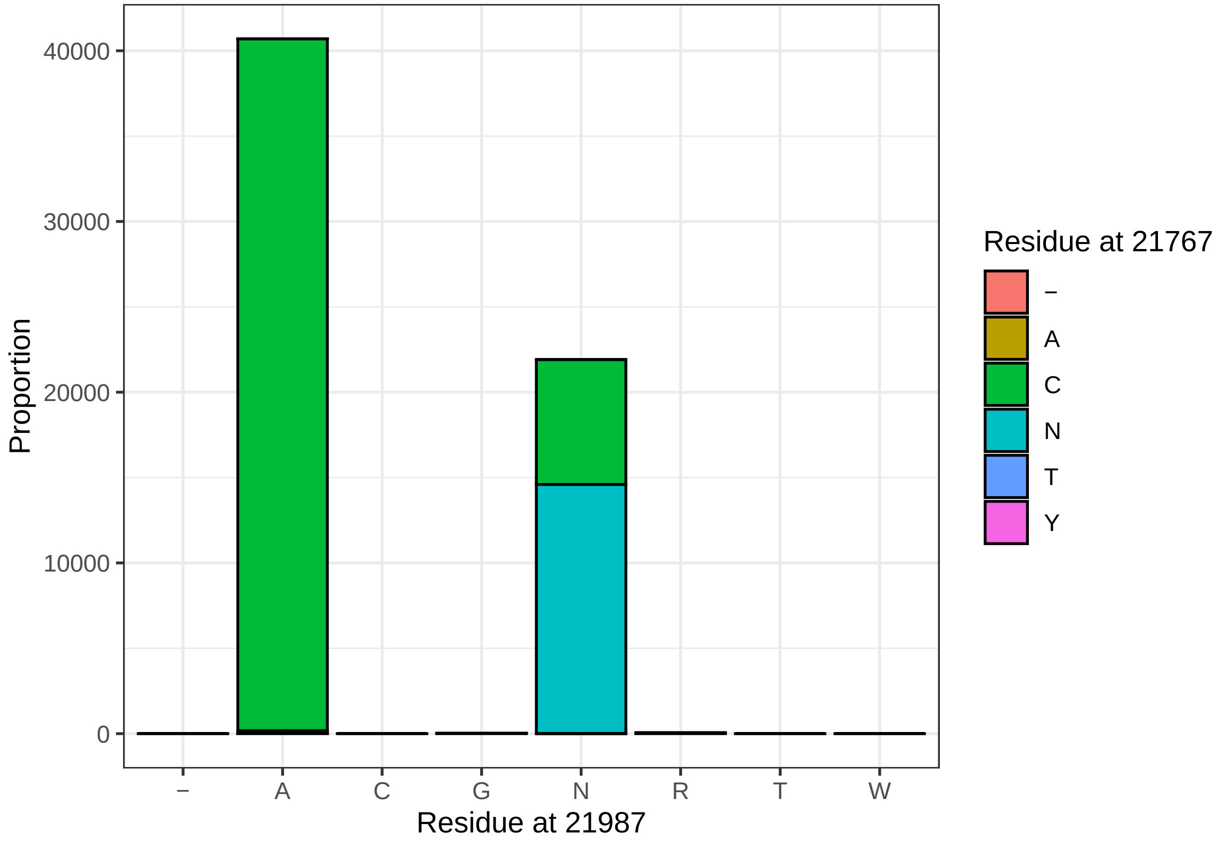
```
apparent_revertants <- delta %>%
  filter(scorpio_call == "Delta (B.1.617.2-like)") %>%
  filter(res_21987 == "G")
res21767 <- read_delim("./residues/uk/residue21767.tsv", delim = " ", col_names = c("sequence_name", "lo
  separate(sequence_name, into = c("country", "coguk_id", "year"), sep = "/") %>%
  mutate(is_revertant = coguk_id %in% apparent_revertants$coguk_id) %>%
  select(coguk_id, value) %>%
  rename(res_21767 = value)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   sequence_name = col_character(),
##   location = col_double(),
##   value = col_character()
## )
```

```
together <- inner_join(delta, res21767)
```

```
## Joining, by = "coguk_id"
```

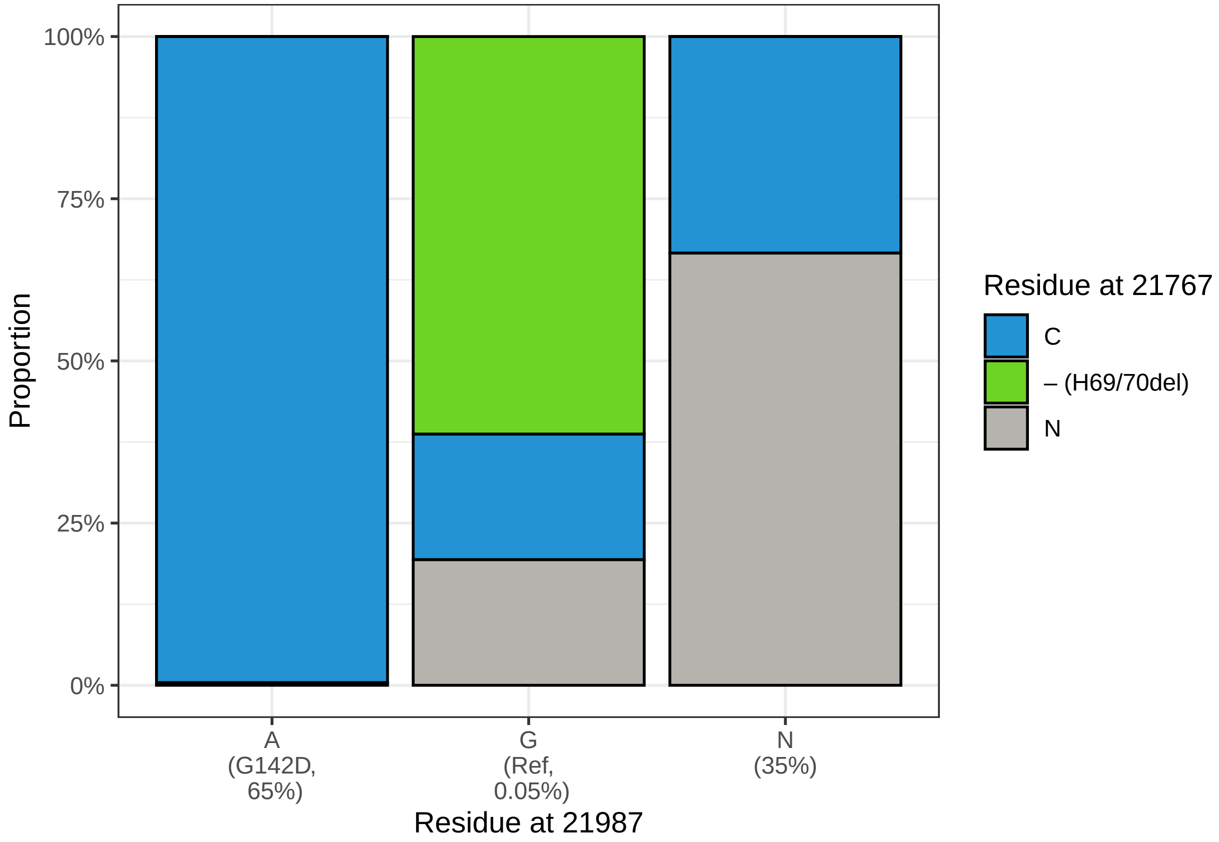
```
ggplot(together, aes(x = res_21987, fill = res_21767)) +
  geom_bar(color = "black", position = "stack") +
  labs(x = "Residue at 21987", fill = "Residue at 21767", y = "Proportion") +
  theme_bw()
```



```
subset <- together %>%
  filter(res_21767 %in% c("-", "N", "C"), res_21987 %in%c("A","G","N") )  %>%
  mutate(res_21767 = case_when(res_21767 == "-" ~ "- (H69/70del)", TRUE ~ res_21767)) %>%
  mutate(res_21987 = case_when(res_21987 == "A" ~ "A\n(G142D,\n 65%)", res_21987 == "G" ~ "G\n(Ref,\n 0

subset %>%
  group_by(res_21987) %>%
  summarise(n = n()) %>%
  mutate(p = (100 * n / sum(n)))
```

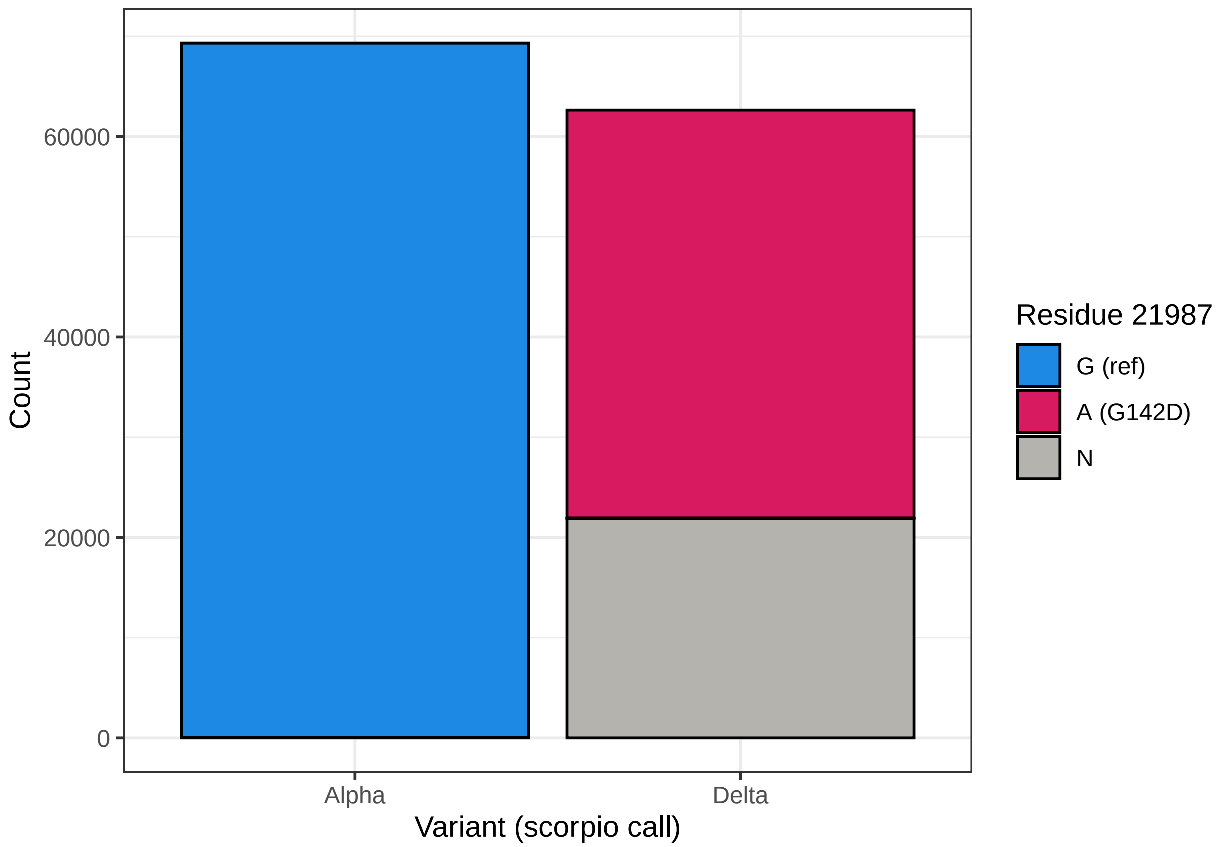
```
## # A tibble: 3 x 3
##   res_21987               n       p
##   <chr>               <int>   <dbl>
## 1 "A\n(G142D,\n 65%)" 40690 65.0
## 2 "G\n(Ref,\n 0.05%)"    31  0.0495
## 3 "N\n(35%)"          21906 35.0
```

```
ggplot(subset, aes(x = res_21987, fill = res_21767)) +
  geom_bar(color = "black", position = "fill") +
  labs(x = "Residue at 21987", fill = "Residue at 21767", y = "Proportion") +
  scale_y_continuous(label = scales::percent) +
  theme_bw() +
  scale_fill_manual(values = c("C" = "#2393d4", "- (H69/70del)" = "#6dd423", "N" = color_n))
```



```
caption <- "Relationship of the residue at 21987 in Delta lineage samples (representing Spike 142) with

cat(caption, file = "./Figures/h69.caption", sep = "\n")

ggsave("./Figures/h69.pdf", width = 3.5, height = 3)
```

```
ggplot(everything %>% filter(sample_date < "2021-07-01", sample_date > "2021-03-01", scorpio_call %in% c
  geom_bar(color = "black") +
  theme_bw() +
  labs(fill = "Residue 21987", x = "Variant (scorpio call)", y = "Count") +
  scale_fill_manual(values = c("G (ref)" = color_ref, "A (G142D)" = color_mut, "N" = color_n))
```



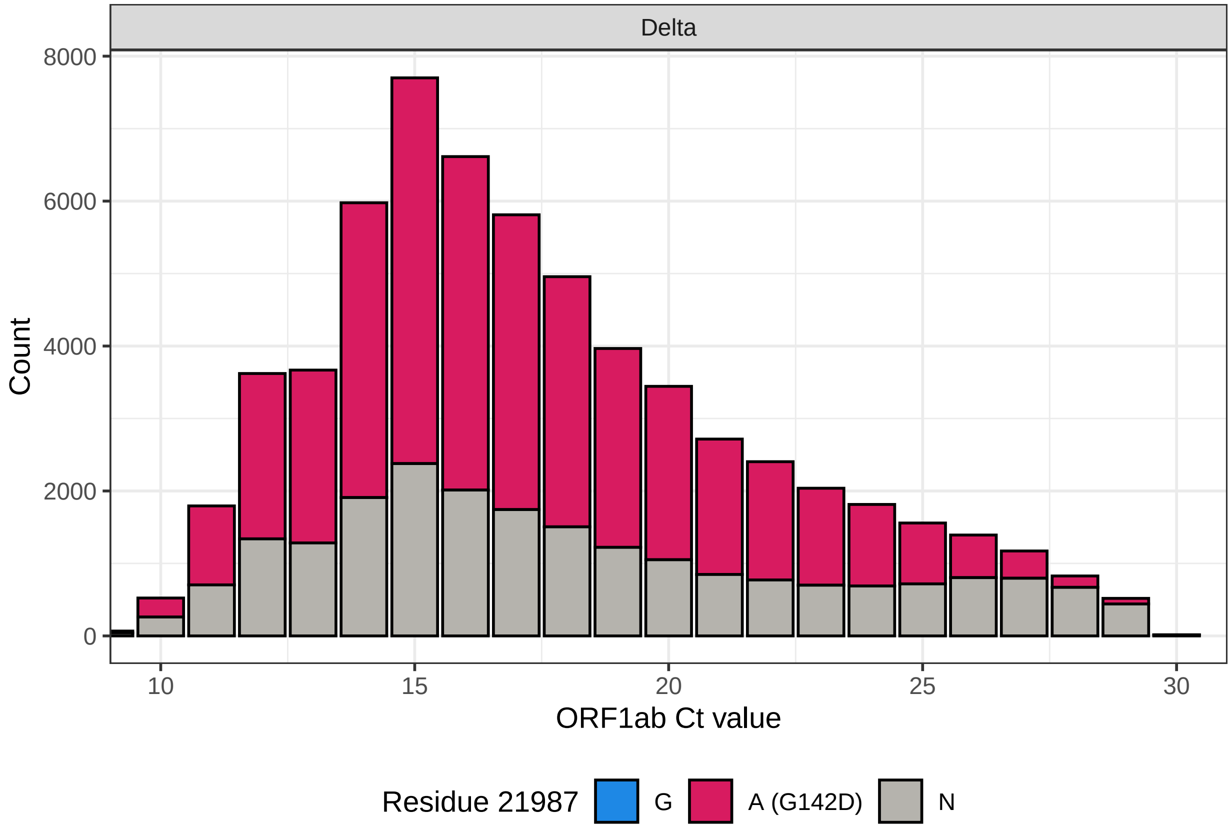
```
caption <- "G142D is fixed in Delta, with almost all Delta sequences where the nucelotide at position 2

cat(caption, file = "./Figures/residue21987.caption", sep = "\n")

ggsave("./Figures/residue21987.pdf", width = 3.5, height = 3)
```

```
# This file is prefiltered to 2021-03-01 to 2021-07-01, and rows huffled randomly w.r.t. to starting da
stripped_ct <- read_csv("./data/stripped_ct_data.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Ch1Cq = col_double(),
##   Ch1Target = col_character(),
##   Ch2Cq = col_double(),
##   Ch2Target = col_character(),
##   Ch3Cq = col_double(),
##   Ch3Target = col_character(),
##   scorpio_call = col_character(),
##   res_21987 = col_character()
## )
```

```
ggplot(stripped_ct %>% mutate(short_lineage = gsub(" \\(.+\\)", "", scorpio_call)) %>% filter(short_lin
  geom_bar(color = "black") +
  theme_bw() +
  labs(fill = "Residue 21987", x = "ORF1ab Ct value", y = "Count") +
  facet_wrap(~short_lineage) +
  coord_cartesian(xlim = c(10, 30)) +
  theme(legend.position = "bottom") +
  scale_fill_manual(values = c("G" = color_ref, "A (G142D)" = color_mut, "N" = color_n))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_count).
```



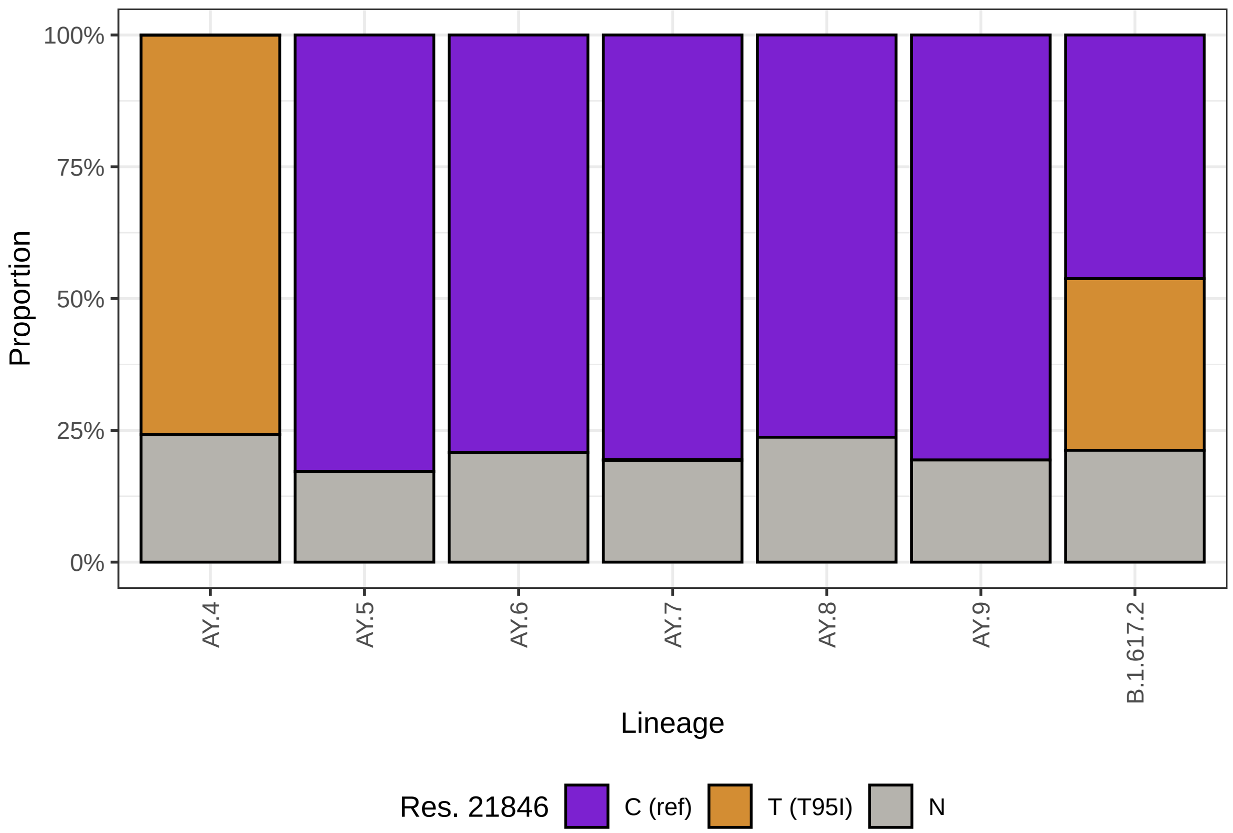
```
caption <- "Relationship of Ct value and residue at position 21987 for COG-UK Delta samples until 30 Ju

cat(caption, file = "Figures/ct.caption", sep = "\n")

ggsave("Figures/ct.pdf", width = 3.5, height = 3)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_count).
```

```
table(everything$scorpio_call)
```

```
##
##                A.23.1-like             A.23.1-like+E484K
##                         10                             3
##         Alpha (B.1.1.7-like)                    AV.1-like
##                      90903                            63
##                B.1.1.318-like            B.1.1.7-like+E484K
##                           175                           122
##                B.1.617.1-like                B.1.617.3-like
##                           234                             7
##                  B.1.621-like           Beta (B.1.351-like)
##                            11                           302
##        Delta (B.1.617.2-like) Delta (B.1.617.2-like) +K417N
##                        192052                            90
##    Epsilon (B.1.427/429-like)            Eta (B.1.525-like)
##                             3                           182
##              Gamma (P.1-like)           Iota (B.1.526-like)
##                            75                             7
##            Lambda (C.37-like)              Theta (P.3-like)
##                             2                             2
##               Zeta (P.2-like)
##                            13
```

```
common_lineages <- delta %>%
  group_by(lineage) %>%
  summarise(n = n()) %>%
  filter(n > 500)


ggplot(delta %>% filter(lineage %in% common_lineages$lineage, residue21846 != "Y", residue21846 != "G",
  geom_bar(color = "black", position = "fill") +
  theme_bw() +
  labs(x = "Lineage", fill = "Res. 21846", y = "Proportion") +
  scale_y_continuous(label = scales::percent) +
  theme(legend.position = "bottom") +
  scale_fill_manual(values = c("C (ref)" = "#7c21d0", "T (T95I)" = "#d38d33", "N" = color_n)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



```
caption <- "Distribution of nucleotides at 21846 (nucleotide T encodes T95I) for different sublineages w

cat(caption, file = "./Figures/t95i.caption", sep = "\n")

ggsave("./Figures/t95i.pdf", width = 3.5, height = 3)
```

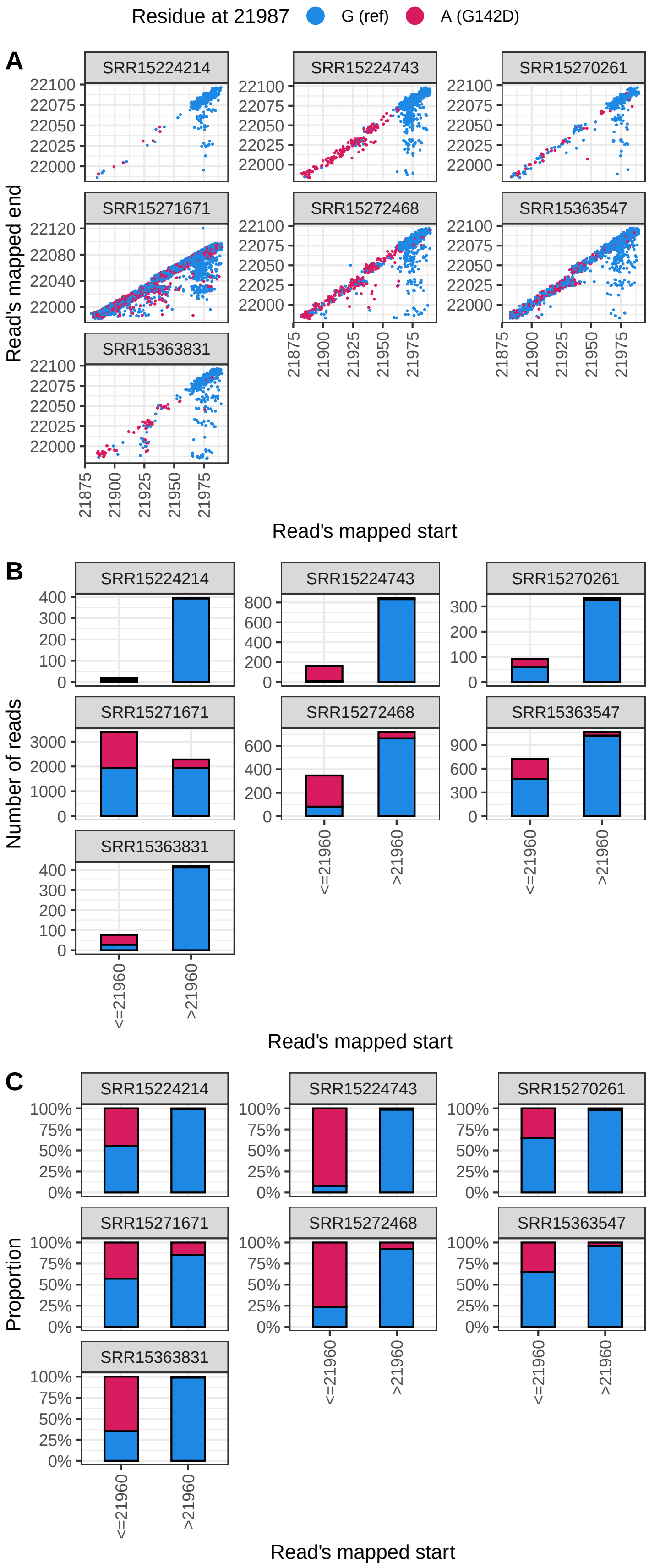
```
library(tidyverse)
palette <- c("G (ref)" = color_ref, "A (G142D)" = color_mut)
data <- read_tsv("./data/pileups.tsv", col_names = c("file", "pos", "res", "read_start", "read_end")) %>
  filter(res %in% c("G", "A")) %>%
  mutate(val = case_when(res == "G" ~ "G (ref)", res == "A" ~ "A (G142D)"))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   file = col_character(),
##   pos = col_double(),
##   res = col_character(),
##   read_start = col_double(),
##   read_end = col_double()
## )
```

```
data$is_73LEFT <- case_when(data$read_start > 21960 ~ ">21960", TRUE ~ "<=21960")

unique(data$file)
```

```
## [1] "SRR15224214.bam.sorted.bam" "SRR15224743.bam.sorted.bam"
## [3] "SRR15270261.bam.sorted.bam" "SRR15271671.bam.sorted.bam"
## [5] "SRR15272468.bam.sorted.bam" "SRR15363547.bam.sorted.bam"
## [7] "SRR15363831.bam.sorted.bam"
```

```
p1 <- ggplot(data, aes(color = val, x = read_start, y = read_end)) +
  geom_jitter(width = 4, height = 4, size = 0.1, alpha = 1) +
  theme_bw() +
  labs(color = "Residue at 21987", x = "Read’s mapped start", y = "Read’s mapped end") +
  facet_wrap(~ gsub(".bam.sorted.bam", "", file), scales = "free_y") +
  scale_color_manual(values = palette) +
  theme(legend.position = "bottom") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  guides(colour = guide_legend(override.aes = list(size = 4)))

p2 <- ggplot(data, aes(color = val, x = is_73LEFT, fill = val)) +
  geom_bar(color = "black", position = "stack", width = 0.5) +
  theme_bw() +
  labs(color = "Residue at 21987", x = "Read’s mapped start", y = "Number of reads") +
  facet_wrap(~ gsub(".bam.sorted.bam", "", file), scales = "free_y") +
  scale_fill_manual(values = palette) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  theme(legend.position = "bottom")


p3 <- ggplot(data, aes(color = val, x = is_73LEFT, fill = val)) +
  geom_bar(color = "black", position = "fill", width = 0.5) +
  theme_bw() +
  labs(color = "Residue at 21987", x = "Read’s mapped start", y = "Proportion") +
  facet_wrap(~ gsub(".bam.sorted.bam", "", file), scales = "free_y") +
  scale_fill_manual(values = palette) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  scale_y_continuous(labels = scales::percent) +
  theme(legend.position = "top")




library(ggpubr)

ggarrange(p1, p2, p3, ncol = 1, nrow = 3, common.legend = TRUE, legend = "top", labels = "AUTO")
```


Residue at 21987
● G (ref)
● A (G142D)
A
SRR15224214
SRR15224743
SRR15270261
SRR15271671
SRR15272468
SRR15363547
SRR15363831
Read's mapped end
22000
22025
22050
22075
22100
22120
22080
22040
21875
21900
21925
21950
21975
Read's mapped start
B
Number of reads
0
100
200
300
400
600
800
900
1000
2000
3000
<=21960
>21960
Read's mapped start
C
Proportion
0%
25%
50%
75%
100%
<=21960
>21960
Read's mapped start

```r
ggsave("./Figures/sra.pdf", width = 5, height = 10)
```