Is Ct for N gene different for 21A (Delta) ?

Stevin Wilson

12/22/2021

Contents

1 Outline

1. Extract clade assignment results for the sequenced samples.
2. Extract sample collection information for COVID +ve samples, and exclude samples with missing collection date or patient ID.
3. Retain only the first COVID +ve sample collected from a patient (Prior infection could potentially affect Ct for subsequent reinfections.)

The following steps are done A) considering samples from all order priority statuses, or B) only retaining surveillance samples.

4. Draw a stacked bar plot to represent clade composition among sequenced samples collected during every month of 2021.
5. Determine mean Ct values for N and RNase P genes. If Ct RNase P is missing, impute median values for samples belonging to the same clade.
6. Retain only samples classified with the following clades:

```
21A (Delta)
20I (Alpha, V1)
20J (Gamma, V3)
20G
```

7. Perform statistical tests to compare Ct values for N and RNase P genes between different clades.

2 Clade Assignment Results

Samples were sent to Premier Medical Services and LabCorp during the course of the year 2021, to get clade and lineage assigned to the samples.

Since these two vendors used different pipelines, in order to maintain consistency, CUGBF reran the SARS-CoV-2 sequence analysis on all samples using the `nf-core/viralrecon` pipeline.

A table `viralrecon_clade` was prepared with `testkit_ids` and `clade` assignment results from `nf-core/viralrecon`.

```
viralrecon_table <- viralrecon_table %>%
  mutate(run_date_time = date(as_datetime(run_date_time))) %>%
  filter(run_date_time != "2021-12-18")

glimpse(viralrecon_table)
```

```
## Rows: 1,981
## Columns: 19
## $ testkit_id              <chr> "117M18D54A87FDE9Y8", "117M18D54A8819F81W", "1~
## $ num_input_reads         <dbl> 1803950, 1793232, 1887908, 2360264, 1435168, 4~
## $ num_trimmed_reads_fastp <dbl> 1121492, 1211678, 1259134, 1681220, 915724, 28~
## $ pc_non_host_read        <dbl> 81.97383, 37.66710, 30.45458, 87.74176, 29.085~
## $ pc_mapped_reads         <dbl> 55.69, 0.04, 0.15, 82.65, 3.65, 99.21, 72.28, ~
## $ num_mapped_reads        <dbl> 624539, 543, 1871, 1389586, 33399, 2796671, 20~
## $ num_trimmed_reads_ivar  <dbl> 618974, NA, NA, 1375419, 32764, 2757078, 19976~
## $ median_coverage         <dbl> NA, NA, NA, 1843, NA, 3919, 2948, 1839, 2453, ~
## $ pc_coverage_gt1x        <dbl> 28, NA, NA, 100, 13, 100, 100, 99, 100, NA, 63~
## $ pc_coverage_gt10x       <dbl> 24, NA, NA, 99, 4, 100, 99, 97, 99, NA, 30, 93~
## $ num_snps                <dbl> 11, NA, NA, 18, 2, 40, 40, 39, 19, NA, 19, 27,~
## $ num_indels              <dbl> NA, NA, NA, NA, NA, 3, 3, 3, NA, NA, 5, 1, 1, ~
## $ num_missense_var        <dbl> 6, NA, NA, 10, 1, 26, 28, 25, 5, NA, 15, 18, 1~
## $ Ns_per_100kb            <dbl> 75992.38, NA, NA, 1337.66, 96090.69, 421.49, 1~
## $ lineage                 <chr> NA, NA, NA, "B.1.243", NA, "B.1.617.2", "B.1.6~
## $ clade                   <chr> "21A (Delta)", NA, NA, "20A", "19B", "21A (Del~
## $ variant_caller          <chr> "iVar", "iVar", "iVar", "iVar", "iVar", "iVar"~
## $ viralrecon_version      <dbl> 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2~
## $ run_date_time           <date> 2021-11-14, 2021-10-23, 2021-10-23, 2021-10-2~
```

```
clade_assignments <- viralrecon_table %>%
  select(testkit_id, clade) %>%
  drop_na() %>%
  mutate(pipeline = "nf-core/viralrecon") %>%
  distinct()

glimpse(clade_assignments)
```

```
## Rows: 1,599
## Columns: 3
## $ testkit_id <chr> "117M18D54A87FDE9Y8", "117M18D5796486135Z", "117M18D6B4187C~
## $ clade      <chr> "21A (Delta)", "20A", "19B", "21A (Delta)", "21A (Delta)", ~
## $ pipeline   <chr> "nf-core/viralrecon", "nf-core/viralrecon", "nf-core/viralr~
```

Are there any `testkit_ids` with multiple clade assignments ?

```
clade_assignments %>%
  group_by(testkit_id) %>%
  filter(n() > 1)
```

```
## # A tibble: 0 x 3
## # Groups:   testkit_id [0]
## # ... with 3 variables: testkit_id <chr>, clade <chr>, pipeline <chr>
```

2.1 Exclude Sample without `patient_id` or `collection_date`, and only retain COVID +ve `testkit_ids`

In the upcoming steps, we only include the first COVID +ve sample isolated from each patient, and in order to not compromise that step, `testkit_ids` with missing `patient_id` and `collection_date` were excluded from further analysis. Also, the `gender` information is required.

```r
sample_collection_without_missing <- sample_collection_table %>%
  filter(
    rymedi_result == "POSITIVE",
    !(is.na(collection_date) | is.na(patient_id)),
    !(is.na(gender))
  ) %>%
  mutate(
    collection_date = as_datetime(collection_date),
    result_date = as_date(result_date)
  )

glimpse(sample_collection_without_missing)
```

```
## Rows: 12,351
## Columns: 20
## $ testkit_id          <chr> "117M18C0D709B3E3IM", "117M18C0D7255382DL", "117M1~
## $ rymedi_result       <chr> "POSITIVE", "POSITIVE", "POSITIVE", "POSITIVE", "P~
## $ population          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ order_priority      <chr> "SURVEILLANCE", "SURVEILLANCE", "SURVEILLANCE", "S~
## $ collection_date     <dttm> 2020-09-11 11:52:37, 2020-09-11 13:58:06, 2020-09~
## $ result_date         <date> 2020-09-13, 2020-09-13, 2020-09-14, 2020-09-14, 2~
## $ gender              <chr> "F", "F", "F", "F", "M", "F", "M", "M", "M", "F", ~
## $ pregnancy_status    <chr> "NO", "NO", "NO", "NO", NA, "NO", NA, NA, NA, "NO"~
## $ zip_code            <chr> "20872", "01867", "78746", "75244", "29631", "0882~
## $ city                <chr> "DAMASCUS", "READING", "AUSTIN", "DALLAS", "CLEMSO~
## $ county              <chr> "MONTGOMERY COUNTY", "MIDDLESEX COUNTY", "TRAVIS C~
## $ state               <chr> "MD", "MA", "TX", "TX", "SC", "NJ", "SC", "NC", "S~
## $ country             <chr> "US", "US", "US", "US", "US", "US", "US", "US", "U~
## $ zip_code_user_input <chr> "20872", "1867", "78746", "75244", "29631", "8825"~
## $ city_user_input     <chr> "DAMASCUS", "READING", "AUSTIN", "DALLAS", "CLEMSO~
## $ state_user_input    <chr> "MD", "MA", "TX", "TX", "SC", "NJ", "SC", "NC", "S~
## $ patient_id          <chr> "fcb108e79831904e198b62a1", "bd23fd30b7fae0727949b~
## $ teskit_sku          <chr> "G7-PCR-DTPM", "G7-PCR-DTPM", "G7-PCR-DTPM", "G7-P~
## $ performing_facility <chr> "CLEMSON UNIVERSITY", "CLEMSON UNIVERSITY", "CLEMS~
## $ testing_facility    <chr> "DAVID STEFANICH", "DAVID STEFANICH", "DAVID STEFA~
```

2.2 Retain only the 1st COVID19 +ve test from a Patient.

Multiple samples were tested from numerous patients during the course of their infection, and many such samples were sequenced. Having multiple COVID19 +ve from a single patient during a single infection event, could affect assumption of independence used for statistical analysis later. Also, prior infection could potentially affect Ct for subsequent reinfections.

```r
sample_collection_without_missing %>%
  group_by(patient_id) %>%
  filter(n() > 1) %>%
  arrange(patient_id, collection_date) %>%
  select(testkit_id, collection_date, patient_id) %>%
  ungroup() %>%
  slice(1:15) %>%
  kbl() %>%
  kable_classic_2(
```

```
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

testkit_id	collection_date	patient_id
117M18C1FB4AC5D522	2020-09-16 19:41:00	00154fa7450a8feec9a5fcb9
117M18C2044154BCWT	2020-09-16 19:41:10	00154fa7450a8feec9a5fcb9
117M18CE12F66182FL	2020-11-19 19:02:13	001c0ca5a8857313549f8afd
117M18CE5A7D02B4DL	2020-11-20 18:58:15	001c0ca5a8857313549f8afd
117M18C08906F6A4HY	2020-09-17 16:23:51	001c38fd878a98a109a3e27f
117M18C1FB4803D1S5	2020-09-17 16:24:07	001c38fd878a98a109a3e27f
117M1911ECC1D0DDXP	2021-09-08 10:09:47	00254cb7ea55304cf70b4c0d
117M1911C551F75FUV	2021-09-14 13:48:03	00254cb7ea55304cf70b4c0d
117M19142B956BCESK	2021-09-15 12:23:50	00254cb7ea55304cf70b4c0d
117M1913EB586BDBHF	2021-10-01 12:23:29	00476fb5ca57c934bae7b45b
117M191AD76C6BB8JR	2021-10-04 13:42:21	00476fb5ca57c934bae7b45b
117M18CED6D6F135G5	2021-01-02 15:03:18	00fe63ab514eb3ec5b4afd7c
117M18D6B40131CCNY	2021-01-04 14:52:22	00fe63ab514eb3ec5b4afd7c
117M18C1FB774C40AB	2020-09-17 16:49:45	0110f10a54eab5bf92346a06
117M18C1FB4640EEH9	2020-09-17 16:50:05	0110f10a54eab5bf92346a06

```
not_first_covid_positive_samples <- sample_collection_without_missing %>%
  filter(rymedi_result == "POSITIVE") %>%
  group_by(patient_id) %>%
  filter(n() > 1) %>%
  select(testkit_id, patient_id, collection_date) %>%
  arrange(patient_id, collection_date) %>%
  slice(2:n()) %>%
  pull(testkit_id)

glimpse(not_first_covid_positive_samples)
```

```
##  chr [1:1268] "117M18C2044154BCWT" "117M18CE5A7D02B4DL" ...
```

Therefore, only the first COVID19 +ve sample from patients are considered for further analysis.

```
sample_collection_without_missing <- sample_collection_without_missing %>%
  filter(!(testkit_id %in% not_first_covid_positive_samples))

glimpse(sample_collection_without_missing)
```

```
## Rows: 11,083
## Columns: 20
## $ testkit_id          <chr> "117M18C0D709B3E3IM", "117M18C0D7255382DL", "117M1~
## $ rymedi_result       <chr> "POSITIVE", "POSITIVE", "POSITIVE", "POSITIVE", "P~
## $ population          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ order_priority      <chr> "SURVEILLANCE", "SURVEILLANCE", "SURVEILLANCE", "S~
```

```
## $ collection_date     <dttm> 2020-09-11 11:52:37, 2020-09-11 13:58:06, 2020-09~
## $ result_date         <date> 2020-09-13, 2020-09-13, 2020-09-14, 2020-09-14, 2~
## $ gender              <chr> "F", "F", "F", "F", "M", "F", "M", "M", "M", "F", ~
## $ pregnancy_status    <chr> "NO", "NO", "NO", "NO", NA, "NO", NA, NA, NA, "NO"~
## $ zip_code            <chr> "20872", "01867", "78746", "75244", "29631", "0882~
## $ city                <chr> "DAMASCUS", "READING", "AUSTIN", "DALLAS", "CLEMSO~
## $ county              <chr> "MONTGOMERY COUNTY", "MIDDLESEX COUNTY", "TRAVIS C~
## $ state               <chr> "MD", "MA", "TX", "TX", "SC", "NJ", "SC", "NC", "S~
## $ country             <chr> "US", "US", "US", "US", "US", "US", "US", "US", "U~
## $ zip_code_user_input <chr> "20872", "1867", "78746", "75244", "29631", "8825"~
## $ city_user_input     <chr> "DAMASCUS", "READING", "AUSTIN", "DALLAS", "CLEMSO~
## $ state_user_input    <chr> "MD", "MA", "TX", "TX", "SC", "NJ", "SC", "NC", "S~
## $ patient_id          <chr> "fcb108e79831904e198b62a1", "bd23fd30b7fae0727949b~
## $ teskit_sku          <chr> "G7-PCR-DTPM", "G7-PCR-DTPM", "G7-PCR-DTPM", "G7-P~
## $ performing_facility <chr> "CLEMSON UNIVERSITY", "CLEMSON UNIVERSITY", "CLEMS~
## $ testing_facility    <chr> "DAVID STEFANICH", "DAVID STEFANICH", "DAVID STEFA~
```

2.3 Join sample_collection Information

In this step, we join the sample collection information with the `clade_assignments` table described previously, and we name the output table as `clades_and_collection`.

```
clades_and_collection <- clade_assignments %>%
  inner_join(sample_collection_without_missing,
    by = "testkit_id"
  ) %>%
  mutate(collection_date = date(collection_date)) %>%
  select(
    patient_id, testkit_id, collection_date, clade, population, order_priority, gender,
    pregnancy_status, pipeline, rymedi_result
  ) %>%
  arrange(patient_id, collection_date)

clades_and_collection <- clades_and_collection %>%
  mutate(
    population = factor(population,
      levels = c(
        "UNIVERSITY",
        "ATHLETICS",
        "COMMUNITY",
        "TRICOUNTY"
      )
    ),
    order_priority = factor(order_priority,
      levels = c(
        "SURVEILLANCE",
        "SYMPTOMATIC",
        "EXPOSED",
        "ONE DAY"
      )
    ),
    gender = factor(gender,
      levels = c("M", "F")
```

```r
    ),
    pregnancy_status = factor(pregnancy_status,
      levels = c("YES", "NO")
    ),
    clade = as_factor(clade),
    rymedi_result = factor(rymedi_result,
      levels = c("POSITIVE")
    ),
    pipeline = as_factor(pipeline)
  ) %>%
  arrange(collection_date, order_priority, population)

glimpse(clades_and_collection)
```

```
## Rows: 1,458
## Columns: 10
## $ patient_id       <chr> "0a6f1c09b23ae1ef7b322a8f", "1133ae22349307de010cdeb4~
## $ testkit_id       <chr> "117M18DCE7D400229H", "117M18DCE7D40EDCMA", "117M18DC~
## $ collection_date  <date> 2021-01-13, 2021-01-13, 2021-01-13, 2021-01-13, 2021~
## $ clade            <fct> "20A", "21C (Epsilon)", "20G", "20G", "20C", "20G", "~
## $ population       <fct> UNIVERSITY, UNIVERSITY, UNIVERSITY, UNIVERSITY, ATHLE~
## $ order_priority   <fct> SURVEILLANCE, SURVEILLANCE, SURVEILLANCE, SURVEILLANC~
## $ gender           <fct> M, F, F, M, M, M, M, F, M, F, F, M, F, F, F, M, F, F,~
## $ pregnancy_status <fct> NA, NO, NO, NA, NA, NA, NA, NO, NA, NO, NO, NA, NO, N~
## $ pipeline         <fct> nf-core/viralrecon, nf-core/viralrecon, nf-core/viral~
## $ rymedi_result    <fct> POSITIVE, POSITIVE, POSITIVE, POSITIVE, POSITIVE, POS~
```

The following are the range of responses to columns in the `clades_and_collection` table.

```r
map(
  clades_and_collection %>% select(
    population, order_priority, gender,
    pregnancy_status, rymedi_result, clade, pipeline
  ),
  unique
)
```

```
## $population
## [1] UNIVERSITY ATHLETICS  COMMUNITY
## Levels: UNIVERSITY ATHLETICS COMMUNITY TRICOUNTY
##
## $order_priority
## [1] SURVEILLANCE SYMPTOMATIC  EXPOSED      ONE DAY
## Levels: SURVEILLANCE SYMPTOMATIC EXPOSED ONE DAY
##
## $gender
## [1] M F
## Levels: M F
##
## $pregnancy_status
## [1] <NA> NO   YES
## Levels: YES NO
```

```
##
## $rymedi_result
## [1] POSITIVE
## Levels: POSITIVE
##
## $clade
##  [1] 20A              21C (Epsilon)   20G              20C
##  [5] 20B              20H (Beta, V2)  20I (Alpha, V1) 21F (Iota)
##  [9] 21D (Eta)        20J (Gamma, V3) 21A (Delta)     19A
## [13] 19B              21B (Kappa)
## 14 Levels: 20I (Alpha, V1) 20G 21A (Delta) 21F (Iota) 20C 21C (Epsilon) ... 21D (Eta)
##
## $pipeline
## [1] nf-core/viralrecon
## Levels: nf-core/viralrecon
```

3 Exploratory Data Analysis

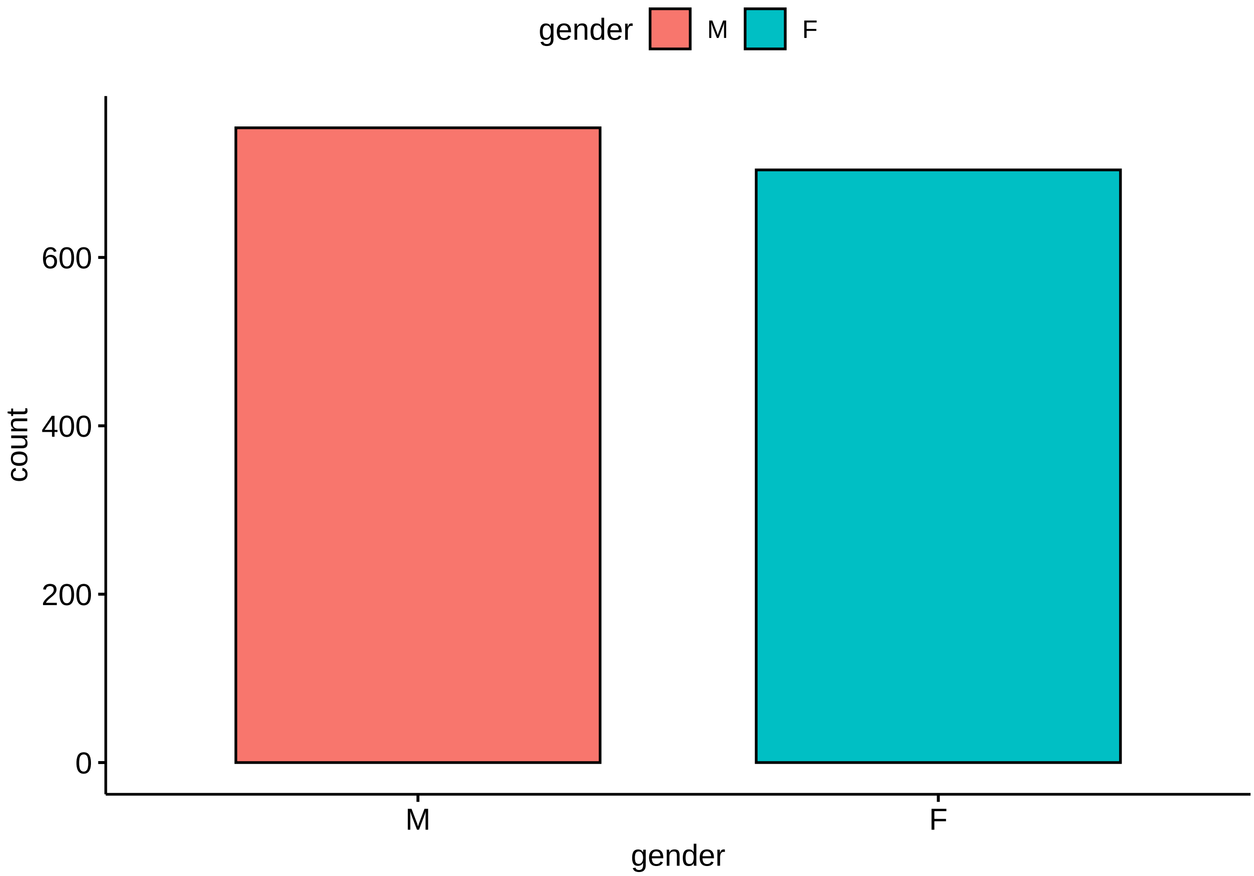
In this step we take a visual glimpse of the `clades_and_collection` table.

```
summary(clades_and_collection)
```

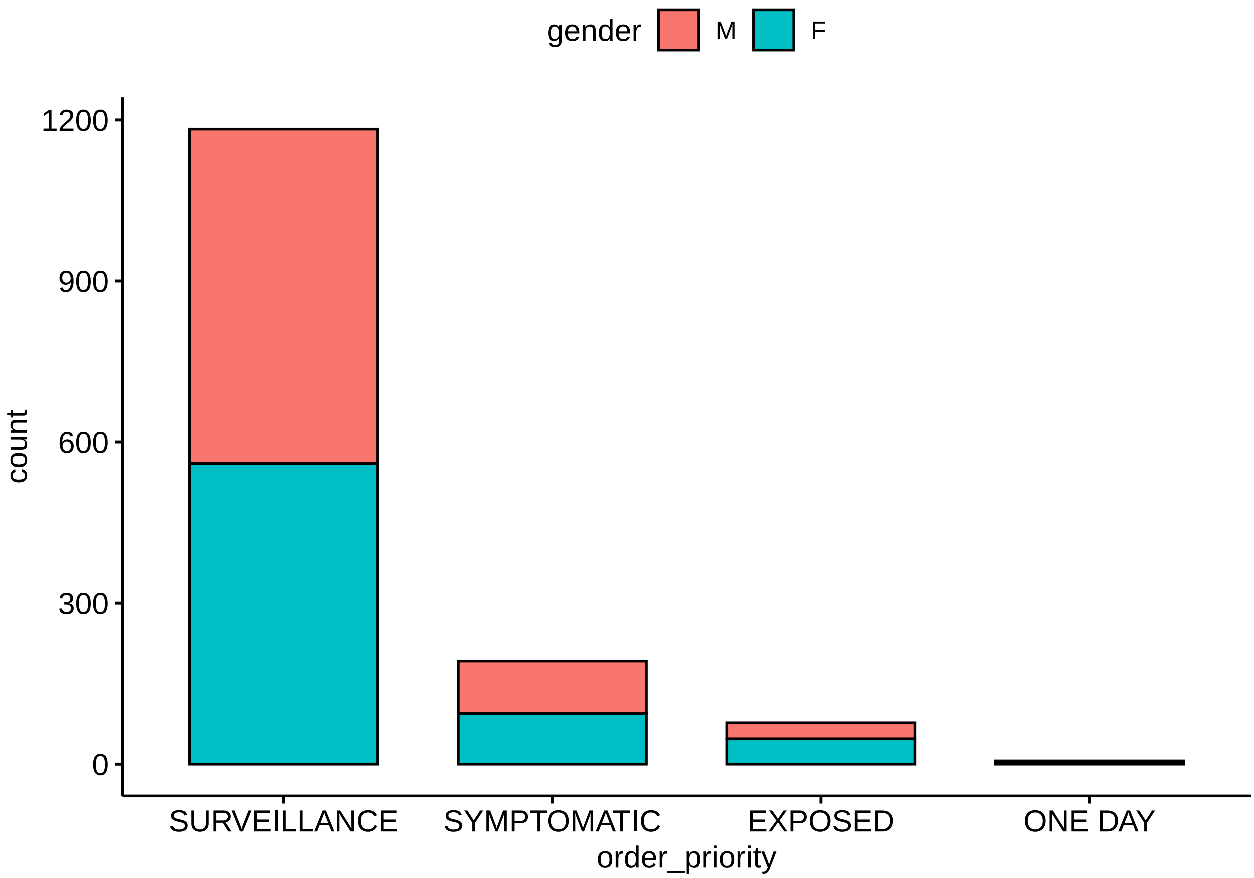
```
##   patient_id         testkit_id        collection_date
##  Length:1458        Length:1458        Min.   :2021-01-13
##  Class :character   Class :character   1st Qu.:2021-03-29
##  Mode  :character   Mode  :character   Median :2021-08-05
##                                        Mean   :2021-06-26
##                                        3rd Qu.:2021-09-14
##                                        Max.   :2021-11-09
##
##              clade          population       order_priority gender
##  21A (Delta)    :789   UNIVERSITY:898   SURVEILLANCE:1183   M:754
##  20I (Alpha, V1):264   ATHLETICS : 27   SYMPTOMATIC : 192   F:704
##  20G            :159   COMMUNITY :533   EXPOSED     :  77
##  20J (Gamma, V3): 87   TRICOUNTY :  0   ONE DAY     :   6
##  20A            : 65
##  20B            : 29
##  (Other)        : 65
##  pregnancy_status              pipeline     rymedi_result
##  YES :  5         nf-core/viralrecon:1458   POSITIVE:1458
##  NO  :698
##  NA's:755
##
##
##
##
```

```
clades_and_collection %>%
  group_by(gender) %>%
  summarise(count = n()) %>%
  ggbarplot(
    x = "gender",
```

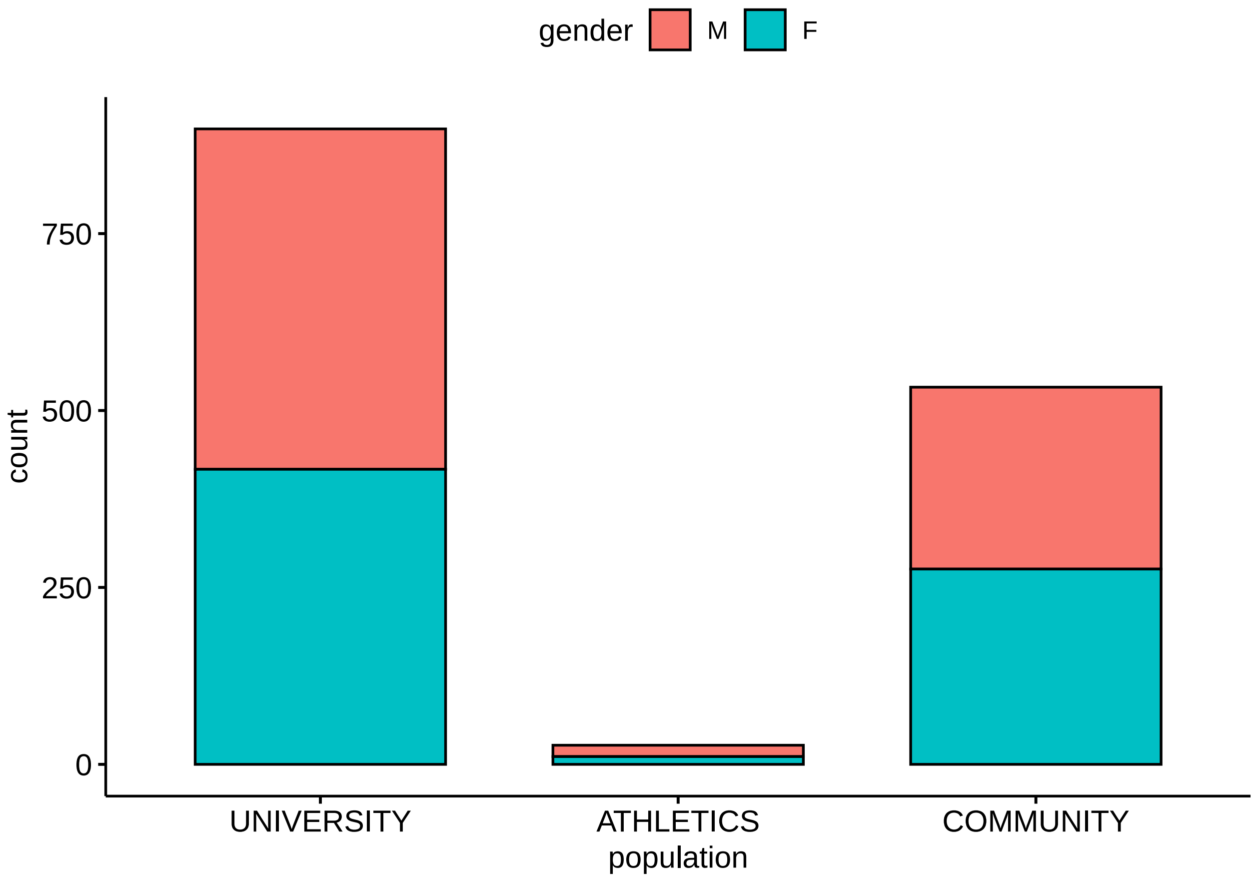
```
    y = "count",
    fill = "gender"
  )
```



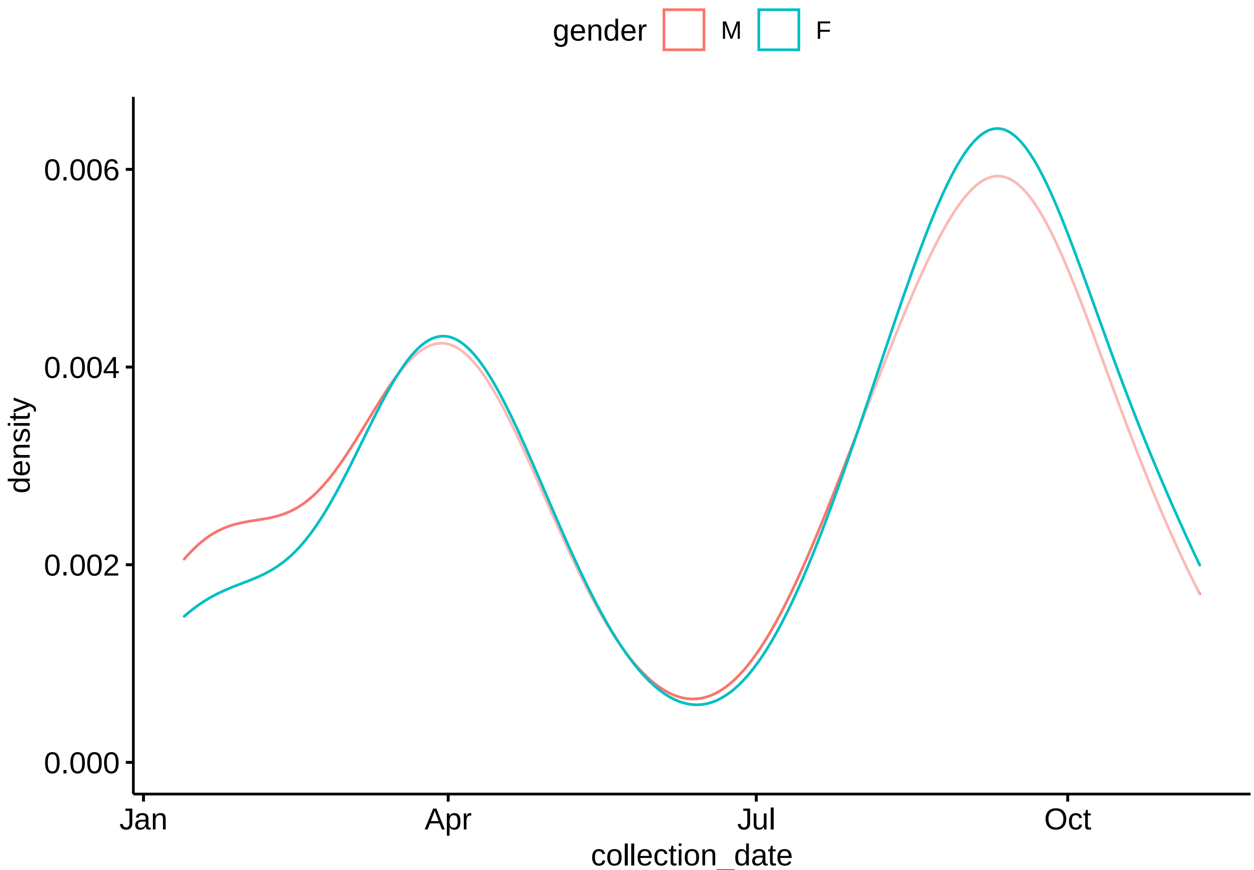

```
clades_and_collection %>%
  group_by(order_priority, gender) %>%
  summarise(count = n()) %>%
  ggbarplot(
    x = "order_priority",
    y = "count",
    fill = "gender"
  )
```



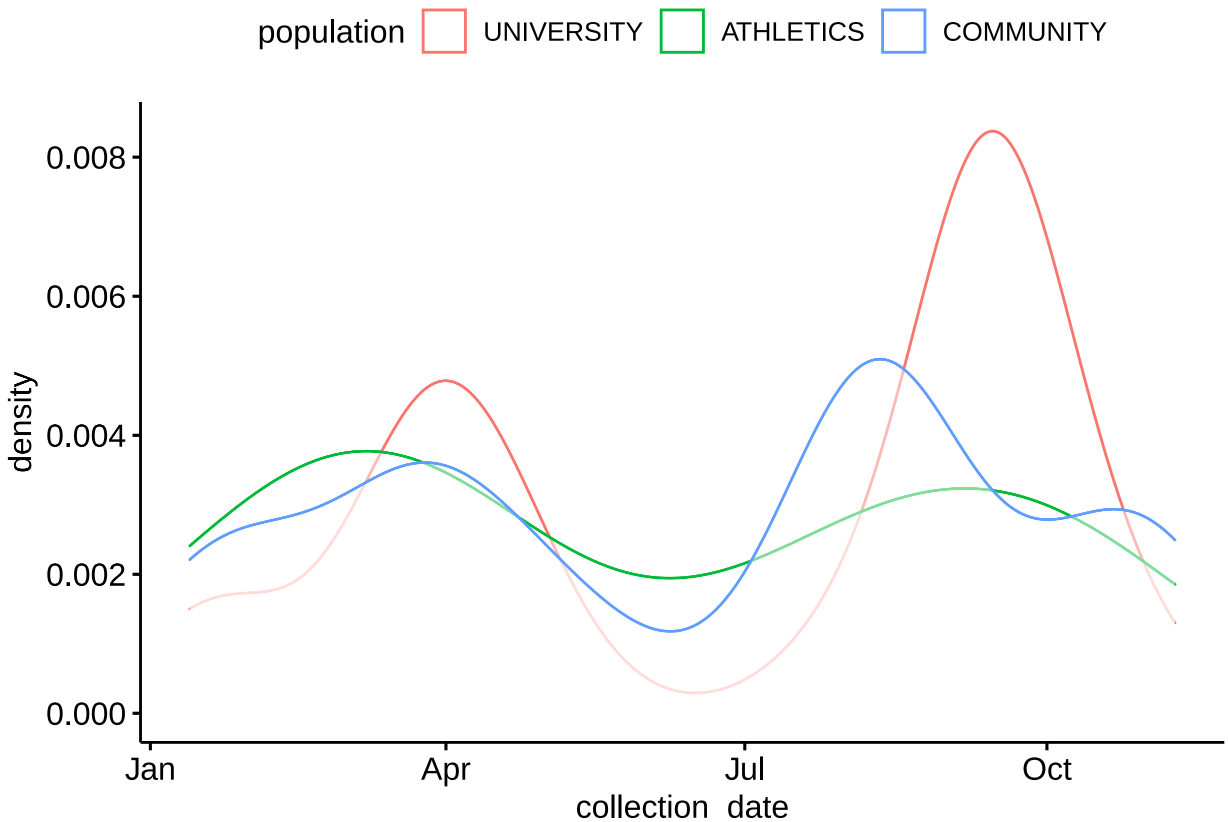

```
clades_and_collection %>%
  group_by(population, gender) %>%
  summarise(count = n()) %>%
  ggbarplot(
    x = "population",
    y = "count",
    fill = "gender"
  )
```



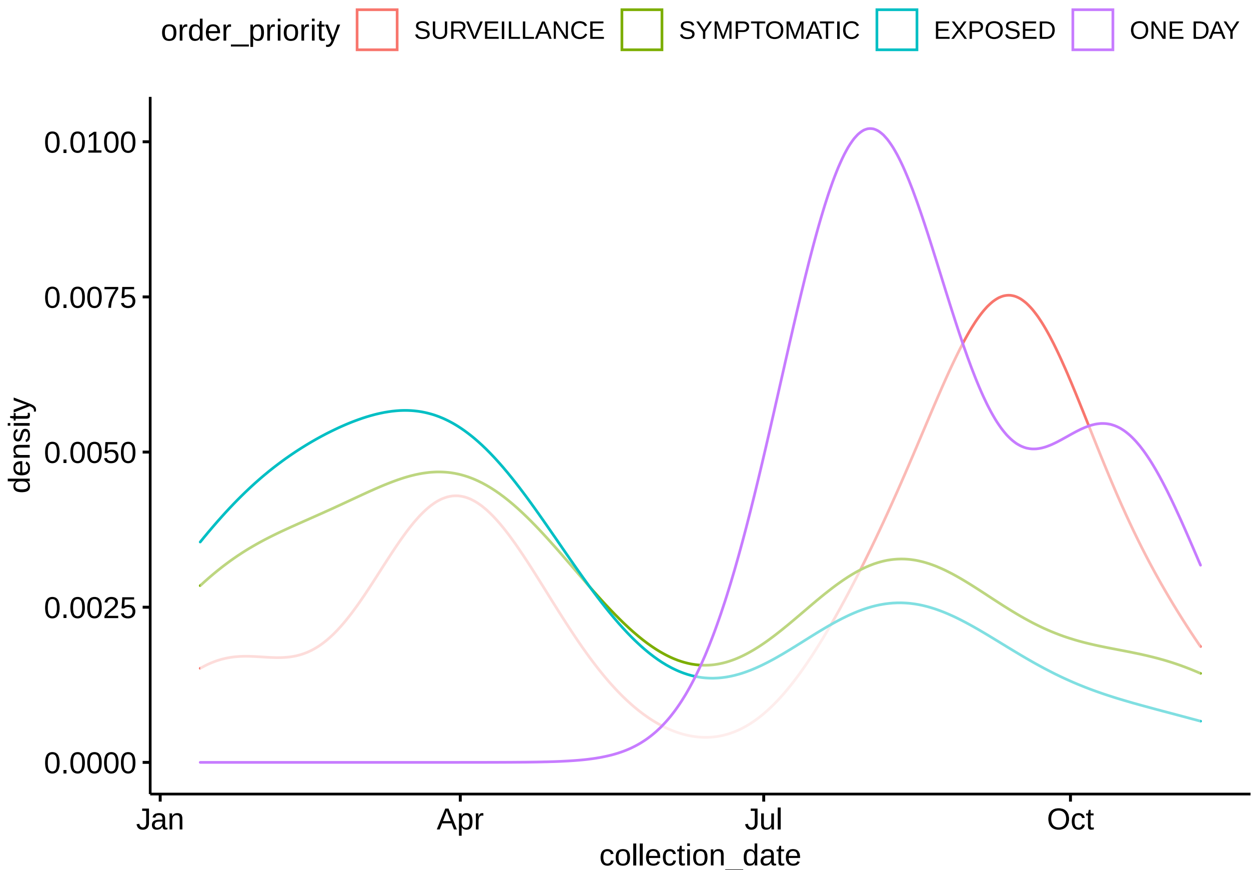

```
ggdensity(clades_and_collection,
  x = "collection_date",
  color = "gender"
)
```




```
ggdensity(clades_and_collection,
  x = "collection_date",
  color = "population"
)
```




```
ggdensity(clades_and_collection,
  x = "collection_date",
  color = "order_priority"
)
```



4 Clade definition from NextClade

Since clades in this study follow the nomenclature set by the NextClade team, it is useful to learn the phylogenetic relations between the clades.

The upcoming steps are performed

A) considering samples from all order priority statuses, or

B) only retaining surveillance samples.

5 Pooling Samples across `order_priority` Statuses

5.1 Bar Plots with Monthly Count for each Clade

The clades present in `clades_and_collection` when all `order_priority` statuses are considered are as follows:

```
clades_factor_level <- clades_and_collection %>%
  group_by(clade) %>%
  summarize(count = n()) %>%
  pull(clade)

(clades_factor_level <- sort(as.character(clades_factor_level)))
```

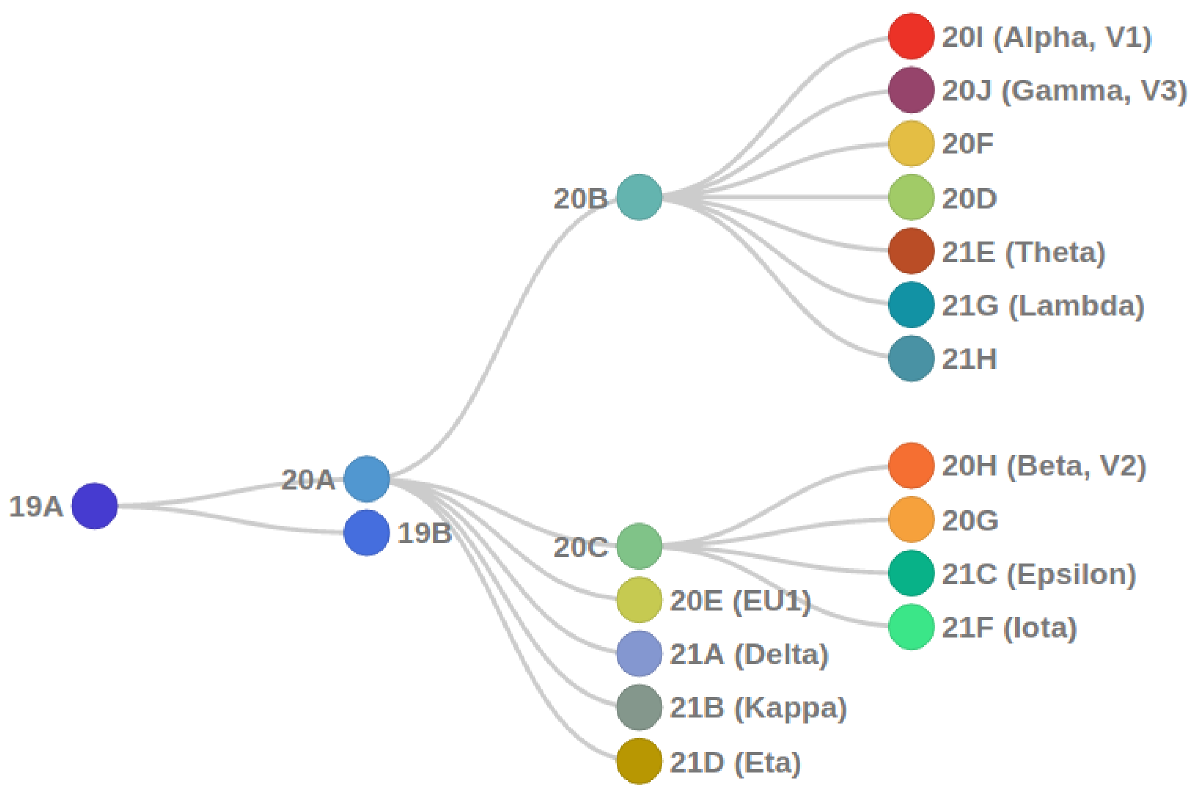


Figure 1: From clades.nextstrain.org

```
##  [1] "19A"              "19B"              "20A"              "20B"
##  [5] "20C"              "20G"              "20H (Beta, V2)"  "20I (Alpha, V1)"
##  [9] "20J (Gamma, V3)" "21A (Delta)"      "21B (Kappa)"      "21C (Epsilon)"
## [13] "21D (Eta)"        "21F (Iota)"
```

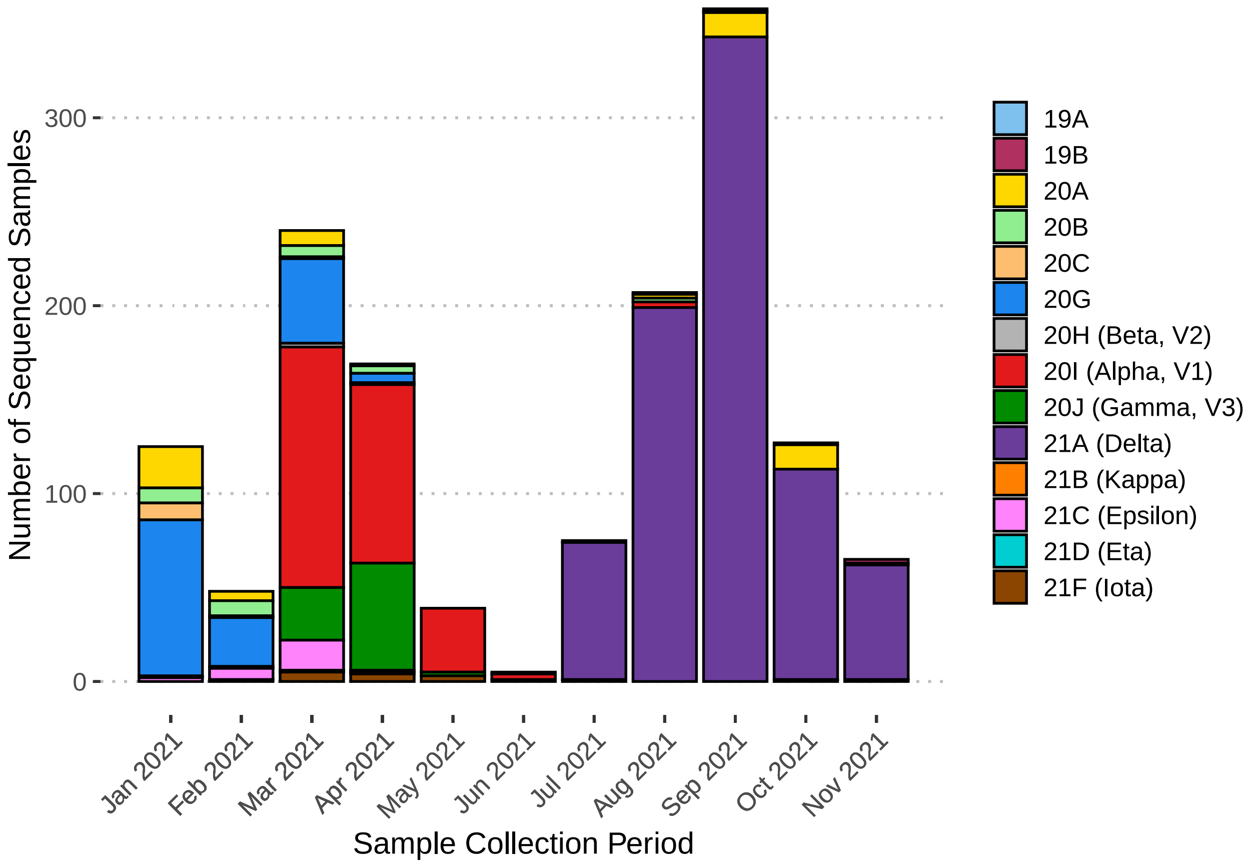
From `clades_and_collection`, we tabulate the count of each clade among sequenced samples collected in each month of 2021.

```
monthly_clade_date <- clades_and_collection %>%
  mutate(
    collection_period = as.yearmon(collection_date),
    clade = factor(clade,
      levels = clades_factor_level
    )
  ) %>%
  group_by(collection_period, clade) %>%
  summarize(count = n())

glimpse(monthly_clade_date)
```

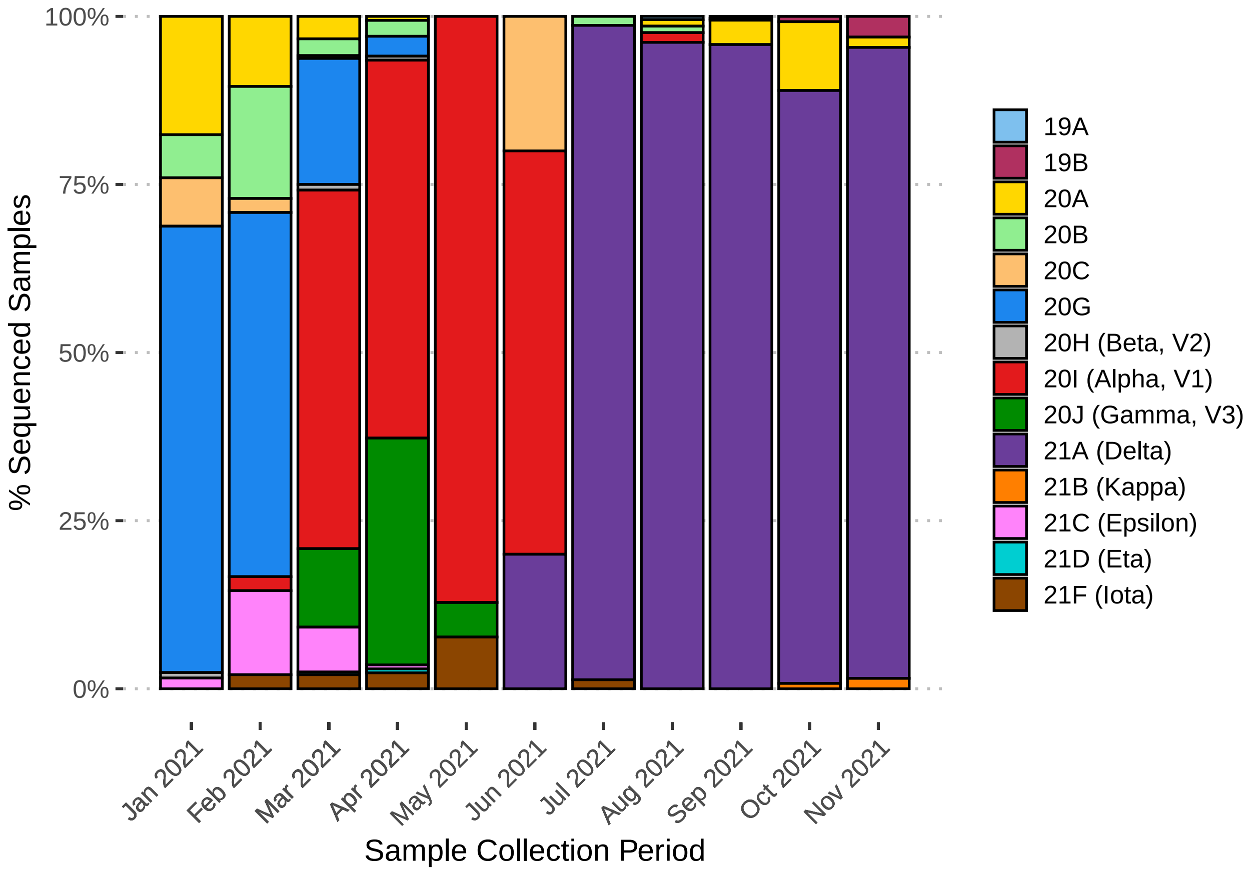
```
## Rows: 58
## Columns: 3
## Groups: collection_period [11]
## $ collection_period <yearmon> Jan 2021, Jan 2021, Jan 2021, Jan 2021, Jan 2021~
## $ clade             <fct> "20A", "20B", "20C", "20G", "20H (Beta, V2)", "21C (~
## $ count             <int> 22, 8, 9, 83, 1, 2, 5, 8, 1, 26, 1, 6, 1, 8, 6, 1, 4~
```

Visualizing the table above using a stacked bar plot.

```
ggplot(
  monthly_clade_date,
  aes(
    x = collection_period,
    y = count,
    fill = clade
  )
) +
  geom_bar(colour = "black", position = "stack", stat = "identity") +
  scale_x_yearmon(breaks = monthly_clade_date$collection_period) +
  scale_fill_manual(values = color_tbl %>%
    filter(clade %in% monthly_clade_date$clade) %>%
    pull(color)) +
  labs(
    y = "Number of Sequenced Samples",
    x = "Sample Collection Period"
  ) +
  theme_pubclean() +
  theme(
    legend.position = "right",
    legend.title = element_blank(),
    legend.key.size = unit(0.5, "cm"),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)
  )
```




```
ggplot(
  monthly_clade_date,
  aes(
    x = collection_period,
    y = count,
    fill = clade
  )
) +
  geom_col(colour = "black", position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  scale_x_yearmon(breaks = monthly_clade_date$collection_period) +
  scale_fill_manual(values = color_tbl %>%
    filter(clade %in% monthly_clade_date$clade) %>%
    pull(color)) +
  labs(
    y = "% Sequenced Samples",
    x = "Sample Collection Period"
  ) +
  theme_pubclean() +
  theme(
    legend.position = "right",
    legend.title = element_blank(),
    legend.key.size = unit(0.5, "cm"),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)
  )
```



5.2 Preparing diagnostics data

Our goal is to determine whether 21A (Delta) shows lower Ct values for N gene when compared with the clades 20I (Alpha, V1), 20G, and 20J (Gamma, V3). In order to know that, we use the Ct values for N gene from the REDDI lab dataset. Since two replicates were done for each qRT-PCR reaction, we will compute the mean Ct for N gene and RNase P control for each `testkit_id`

```
# diagnostics_data
diagnostics_data <- diagnostics_table %>%
  filter(testkit_id %in% clades_and_collection$testkit_id) %>%
  select(testkit_id, ct_rnasep_rep1, ct_rnasep_rep2, ct_N_rep1, ct_N_rep2) %>%
  arrange(testkit_id) %>%
  mutate(
    average_ct_rnasep = rowMeans(.[, c("ct_rnasep_rep1", "ct_rnasep_rep2")], na.rm = TRUE),
    average_ct_N = rowMeans(.[, c("ct_N_rep1", "ct_N_rep2")], na.rm = TRUE)
  ) %>%
  select(-c(
    ct_rnasep_rep1, ct_rnasep_rep2,
    ct_N_rep1, ct_N_rep2
  )) %>%
  group_by(testkit_id) %>%
  summarise(
    ct_RNaseP = mean(average_ct_rnasep, na.rm = TRUE),
    ct_N = mean(average_ct_N, na.rm = TRUE)
```

```
  )

diagnostics_data %>%
  filter(!complete.cases(.)) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

testkit_id	ct_RNaseP	ct_N
117M18DBFE8BB0A1JJ	NaN	21.29940
117M18DBFE8BB18FTW	NaN	26.99776
117M18DBFE8C4DF29V	NaN	30.18376
117M18DBFEE11FC1HA	NaN	23.52962
117M18DCB750444B1K	NaN	29.40396
117M18DCB750672AZL	NaN	25.19355
117M18DCB750EF7DKW	NaN	23.28848
117M18DCB7CE53ED5Q	NaN	22.70748
117M18DCB7CE6AC0J1	NaN	27.95281
117M18DCB7CE7001AZ	NaN	22.73960
117M18DCB7CE700F6S	NaN	25.49053
117M18DCB7CE70B07G	NaN	19.39017
117M18DCB7CE7F0BO1	NaN	17.95333
117M18DCB7CE81040Q	NaN	24.27213
117M18DCB7CE9BEE6F	NaN	25.83106
117M18DCB7CE9BFF04	NaN	24.22209
117M18DD9F7D8DF0LH	NaN	23.69066
117M18DD9F7DF503YP	NaN	16.45756
117M18DD9F7E062D0Q	NaN	24.82307
117M18DD9F7E67D57T	NaN	26.20155
117M18DDC2E0AD0A70	NaN	29.20234
117M18DDD30852A7D3	NaN	16.67494
117M18DDD3085427A5	NaN	30.80809
117M18DDD308C25FZY	NaN	21.30878
117M18DDD308C63FEZ	NaN	25.31193
117M18DDD4896537BX	NaN	18.88141
117M18DDD48F66A8GX	NaN	26.61771
117M18DDD48FDEBD87	NaN	26.39463
117M18DDD490489F80	NaN	26.09010
117M18DDD4916F9CG2	NaN	21.62621
117M18DE172DA3990T	NaN	22.35571
117M18DE172DA3E12G	NaN	27.27337
117M18DE172DA66AEW	NaN	23.83437
117M18DE1733C0D5PY	NaN	26.53400
117M18E355CFD320PV	NaN	29.74500

35 `testkit_id`s do not have a ct value for RNase P. We will use imputation to deal with the missing values.

5.3 Join clade_assignments and diagnostics_data

We join clades_and_collection and the diagnostics data to get the clade_collection_diagnostics table.

```r
clade_collection_diagnostics <- clades_and_collection %>%
  inner_join(diagnostics_data, by = "testkit_id")

glimpse(clade_collection_diagnostics)
```

```
## Rows: 1,456
## Columns: 12
## $ patient_id       <chr> "0a6f1c09b23ae1ef7b322a8f", "1133ae22349307de010cdeb4~
## $ testkit_id       <chr> "117M18DCE7D400229H", "117M18DCE7D40EDCMA", "117M18DC~
## $ collection_date  <date> 2021-01-13, 2021-01-13, 2021-01-13, 2021-01-13, 2021~
## $ clade            <fct> "20A", "21C (Epsilon)", "20G", "20G", "20C", "20G", "~
## $ population       <fct> UNIVERSITY, UNIVERSITY, UNIVERSITY, UNIVERSITY, ATHLE~
## $ order_priority   <fct> SURVEILLANCE, SURVEILLANCE, SURVEILLANCE, SURVEILLANC~
## $ gender           <fct> M, F, F, M, M, M, M, F, M, F, F, M, F, F, F, M, F, F,~
## $ pregnancy_status <fct> NA, NO, NO, NA, NA, NA, NA, NO, NA, NO, NO, NA, NO, N~
## $ pipeline         <fct> nf-core/viralrecon, nf-core/viralrecon, nf-core/viral~
## $ rymedi_result    <fct> POSITIVE, POSITIVE, POSITIVE, POSITIVE, POSITIVE, POS~
## $ ct_RNaseP        <dbl> 16.59912, 17.39000, 21.62271, 24.57796, 19.51531, 18.~
## $ ct_N             <dbl> 22.26138, 23.00548, 28.37183, 23.92611, 22.17447, 25.~
```

Since many samples had missing Ct values for RNase P, we impute median value of ct_RNaseP for each clade to the missing values.

```r
get_median_ct <- function(input_clade) {
  median_RNAseP <- clade_collection_diagnostics %>%
    filter(clade == input_clade) %>%
    pull(ct_RNaseP) %>%
    median(na.rm = TRUE)

  return(median_RNAseP)
}
```

Median Ct RNase P - 20I (Alpha, V1)

```r
# 20I (Alpha, V1)
(median_RNAseP_20I <- get_median_ct("20I (Alpha, V1)"))
```

```
## [1] 18.65
```

Median Ct RNase P - 21A (Delta)

```r
# 21A (Delta)
(median_RNAseP_21A <- get_median_ct("21A (Delta)"))
```

```
## [1] 18.9
```

Median Ct RNase P - 20G

```r
# 20G
(median_RNAseP_20G <- get_median_ct("20G"))
```

```
## [1] 19.39105
```

Median Ct RNase P - 20J (Gamma, V3)

```r
# 20J (Gamma, V3)
(median_RNAseP_20J <- get_median_ct("20J (Gamma, V3)"))
```

```
## [1] 19.16812
```

Summary of the `clade_collection_diagnostics` table after imputation

```r
clade_collection_diagnostics <- clade_collection_diagnostics %>%
  mutate(
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20I (Alpha, V1)")), median_RNAseP_20I
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "21A (Delta)")), median_RNAseP_21A
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20G")), median_RNAseP_20G
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20J (Gamma, V3)")), median_RNAseP_20J
    )
  )

summary(clade_collection_diagnostics)
```

```
##   patient_id         testkit_id        collection_date
##  Length:1456        Length:1456        Min.   :2021-01-13
##  Class :character   Class :character   1st Qu.:2021-03-29
##  Mode  :character   Mode  :character   Median :2021-08-05
##                                        Mean   :2021-06-26
##                                        3rd Qu.:2021-09-14
##                                        Max.   :2021-11-09
##
##              clade         population       order_priority gender
##  21A (Delta)    :787   UNIVERSITY:898   SURVEILLANCE:1183   M:753
##  20I (Alpha, V1):264   ATHLETICS : 27   SYMPTOMATIC : 190   F:703
##  20G            :159   COMMUNITY :531   EXPOSED     :  77
##  20J (Gamma, V3): 87   TRICOUNTY :  0   ONE DAY     :   6
##  20A            : 65
##  20B            : 29
```

```
##  (Other)        : 65
##  pregnancy_status               pipeline     rymedi_result    ct_RNaseP
##  YES :  5         nf-core/viralrecon:1456   POSITIVE:1456   Min.   :14.12
##  NO  :697                                                   1st Qu.:17.71
##  NA's:754                                                   Median :18.91
##                                                             Mean   :19.47
##                                                             3rd Qu.:20.57
##                                                             Max.   :34.35
##                                                             NA's   :15
##       ct_N
##  Min.   : 7.98
##  1st Qu.:20.36
##  Median :23.49
##  Mean   :23.22
##  3rd Qu.:26.37
##  Max.   :33.45
##
```

The counts of different clades in the `clade_collection_diagnostics` table are as follows:

```r
clade_collection_diagnostics %>%
  group_by(clade) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

clade	count
21A (Delta)	787
20I (Alpha, V1)	264
20G	159
20J (Gamma, V3)	87
20A	65
20B	29
21C (Epsilon)	25
21F (Iota)	14
20C	12
20H (Beta, V2)	4
19B	4
21B (Kappa)	2
19A	2
21D (Eta)	2

In order to not affect our assumption of phylogenetic independence, we are only going to compare Ct values for variants at terminal nodes of NextClade tree. We extract data from `clade_collection_diagnostics` table and create a new table `limited_clade_collection_diagnostics` with only data for the following clades:

```
21A (Delta)
20I (Alpha, V1)
20J (Gamma, V3)
20G
```

```
limited_clade_collection_diagnostics <- clade_collection_diagnostics %>%
  filter(clade %in% c(
    "21A (Delta)",
    "20I (Alpha, V1)",
    "20J (Gamma, V3)",
    "20G"
  )) %>%
  mutate(clade = factor(clade,
    levels = c(
      "20G",
      "20I (Alpha, V1)",
      "20J (Gamma, V3)",
      "21A (Delta)"
    )
  )) %>%
  select(testkit_id, clade, ct_RNaseP, ct_N) %>%
  ungroup() %>%
  drop_na()

glimpse(limited_clade_collection_diagnostics)
```

```
## Rows: 1,297
## Columns: 4
## $ testkit_id <chr> "117M18DCE7D410C0MX", "117M18D7B495AED7RY", "117M18DBD66167~
## $ clade      <fct> "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20~
## $ ct_RNaseP  <dbl> 21.62271, 24.57796, 18.01706, 18.35299, 17.96138, 21.27448,~
## $ ct_N       <dbl> 28.37183, 23.92611, 25.77516, 21.04739, 20.37994, 19.55871,~
```

Summary of the `limited_clade_collection_diagnostics` table :

```
summary(limited_clade_collection_diagnostics)
```

```
##   testkit_id                    clade       ct_RNaseP          ct_N
##  Length:1297        20G            :159   Min.   :14.48   Min.   : 7.98
##  Class :character   20I (Alpha, V1):264   1st Qu.:17.71   1st Qu.:20.29
##  Mode  :character   20J (Gamma, V3): 87   Median :18.95   Median :23.28
##                     21A (Delta)    :787   Mean   :19.45   Mean   :23.08
##                                           3rd Qu.:20.54   3rd Qu.:26.18
##                                           Max.   :34.35   Max.   :33.27
```

Median and range of patient age whose samples are in the `limited_clade_collection_diagnostics` table as follows:

```
sample_collection_table %>%
  mutate(collection_date = year(as_datetime(collection_date))) %>%
  filter(testkit_id %in% limited_clade_collection_diagnostics$testkit_id) %>%
  left_join(demographics_table, by = "patient_id") %>%
```

```
  mutate(age_at_sample_collection = (collection_date - birth_year)) %>%
  select(testkit_id, patient_id, age_at_sample_collection) %>%
  summarize(
    median_age_at_sample_collection = median(age_at_sample_collection),
    lowest_age_at_sample_collection = min(age_at_sample_collection),
    highest_age_at_sample_collection = max(age_at_sample_collection)
  ) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

median_age_at_sample_collection	lowest_age_at_sample_collection	highest_age_at_sample_collection
21	0	91

The counts of each gender in the `limited_clade_collection_diagnostics` table are as follows:

```
sample_collection_table %>%
  filter(testkit_id %in% limited_clade_collection_diagnostics$testkit_id) %>%
  group_by(gender) %>%
  summarize(count = n()) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

gender	count
F	627
M	670

The median and range of Ct values for each clade in the `limited_clade_collection_diagnostics` table are as follows:

```
limited_clade_collection_diagnostics %>%
  group_by(clade) %>%
  summarise(
    median_ct_N = round(median(ct_N), 3),
    min_ct_N = round(range(ct_N)[1], 3),
    max_ct_N = round(range(ct_N)[2], 3),
    median_ct_RNaseP = round(median(ct_RNaseP), 3),
    min_ct_RNaseP = round(range(ct_RNaseP)[1], 3),
    max_ct_RNaseP = round(range(ct_RNaseP)[2], 3)
  ) %>%
  kbl() %>%
```

```
  kable_classic_2(
    latex_options = c(
      "hold_position",
      "striped",
      "scale_down"
    )
  )
```

clade	median_ct_N	min_ct_N	max_ct_N	median_ct_RNaseP	min_ct_RNaseP	max_ct_RNaseP
20G	25.210	12.87	33.274	19.391	15.714	28.890
20I (Alpha, V1)	23.925	12.24	32.968	18.650	15.330	34.350
20J (Gamma, V3)	24.740	13.31	31.212	19.168	15.590	28.708
21A (Delta)	22.615	7.98	32.355	18.900	14.485	33.280

The median and range of Ct values for each `clade` - `order_priority` combination in the `limited_clade_collection_diagnos`
table are as follows:

```
limited_clade_collection_diagnostics %>%
  inner_join(sample_collection_without_missing %>%
    select(testkit_id, order_priority),
  by = "testkit_id"
  ) %>%
  group_by(clade, order_priority) %>%
  summarise(
    count = n(),
    median_ct_N = round(median(ct_N), 3),
    min_ct_N = round(range(ct_N)[1], 3),
    max_ct_N = round(range(ct_N)[2], 3),
    median_ct_RNaseP = round(median(ct_RNaseP), 3),
    min_ct_RNaseP = round(range(ct_RNaseP)[1], 3),
    max_ct_RNaseP = round(range(ct_RNaseP)[2], 3)
  ) %>%
  kbl() %>%
  kable_classic_2(
    latex_options = c(
      "hold_position",
      "striped",
      "scale_down"
    )
  )
```

clade	order_priority	count	median_ct_N	min_ct_N	max_ct_N	median_ct_RNaseP	min_ct_RNaseP	max_ct_RNaseP
20G	EXPOSED	25	25.210	17.965	30.832	19.391	17.165	24.660
20G	SURVEILLANCE	95	25.752	17.735	33.274	19.391	15.714	25.753
20G	SYMPTOMATIC	39	23.165	12.870	29.890	19.391	16.280	28.890
20I (Alpha, V1)	EXPOSED	25	25.455	16.900	31.495	17.700	16.210	23.520
20I (Alpha, V1)	SURVEILLANCE	181	23.810	12.240	32.968	18.772	15.330	34.350
20I (Alpha, V1)	SYMPTOMATIC	58	23.230	13.225	32.505	18.555	16.150	29.920
20J (Gamma, V3)	SURVEILLANCE	86	24.688	13.310	31.212	19.167	15.590	28.708
20J (Gamma, V3)	SYMPTOMATIC	1	28.035	28.035	28.035	23.030	23.030	23.030
21A (Delta)	EXPOSED	21	22.390	11.943	31.935	18.765	16.535	23.515
21A (Delta)	ONE DAY	5	22.790	22.300	30.130	21.605	17.735	23.995
21A (Delta)	SURVEILLANCE	691	22.560	10.675	32.355	18.935	14.485	32.160
21A (Delta)	SYMPTOMATIC	70	22.985	7.980	29.980	18.540	14.690	33.280

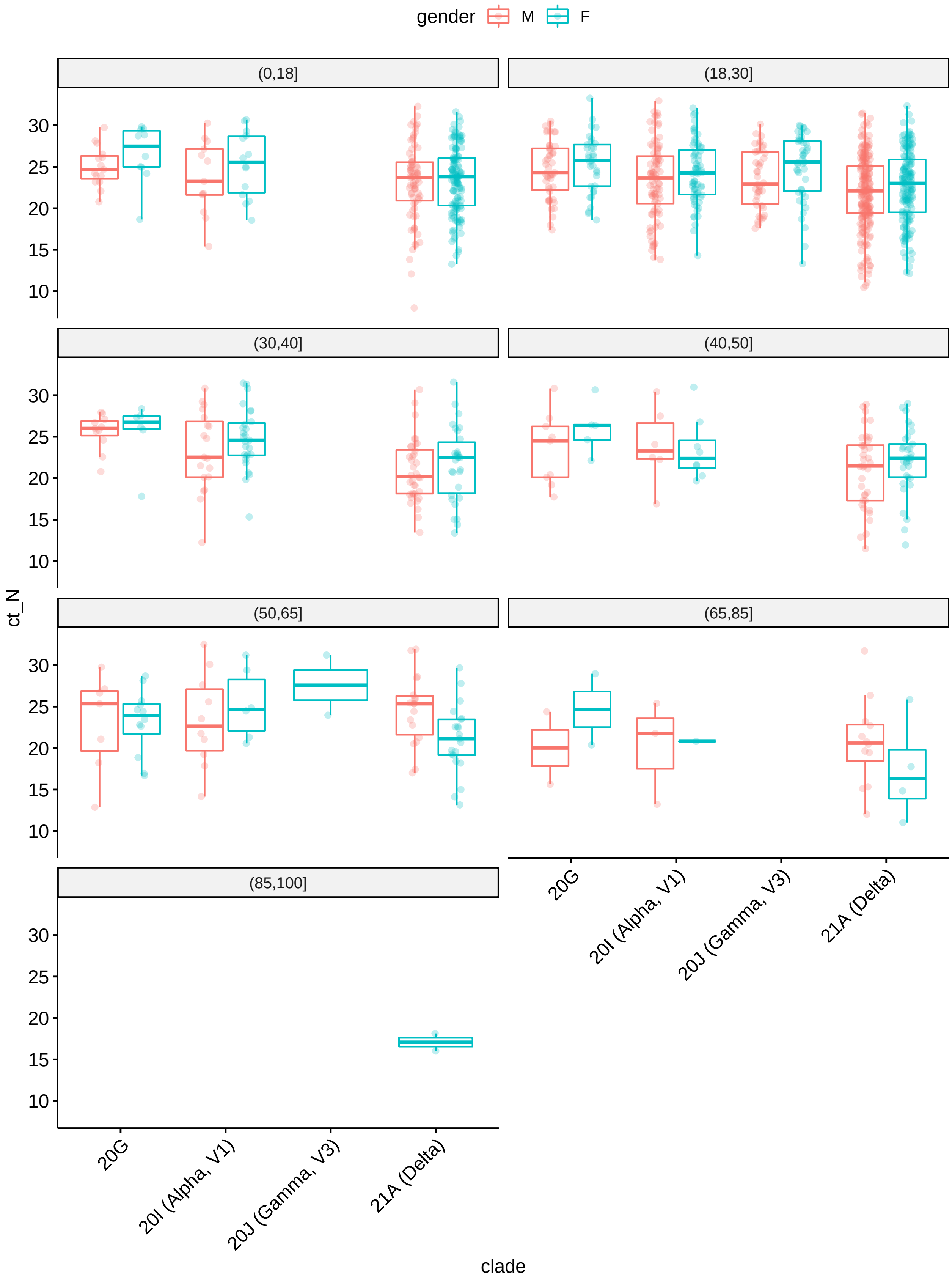
Exploring if Ct values have any relation with gender

```r
lc_age_gender <- limited_clade_collection_diagnostics %>%
  inner_join(sample_collection_table %>%
    select(testkit_id, patient_id, gender, collection_date),
  by = "testkit_id"
  ) %>%
  inner_join(demographics_table %>%
    select(patient_id, birth_year),
  by = "patient_id"
  ) %>%
  mutate(
    collection_date = as_datetime(collection_date),
    gender = factor(gender, levels = c("M", "F")),
    age = year(collection_date) - birth_year,
    age_group = cut(x = age, breaks = c(0, 18, 30, 40, 50, 65, 85, 100))
  )

summary(lc_age_gender)
```

```
##   testkit_id                    clade        ct_RNaseP          ct_N
##  Length:1297        20G           :159   Min.   :14.48   Min.   : 7.98
##  Class :character   20I (Alpha, V1):264   1st Qu.:17.71   1st Qu.:20.29
##  Mode  :character   20J (Gamma, V3): 87   Median :18.95   Median :23.28
##                     21A (Delta)    :787   Mean   :19.45   Mean   :23.08
##                                           3rd Qu.:20.54   3rd Qu.:26.18
##                                           Max.   :34.35   Max.   :33.27
##
##   patient_id        gender  collection_date                 birth_year
##  Length:1297        M:670   Min.   :2021-01-13 13:28:00   Min.   :1930
##  Class :character   F:627   1st Qu.:2021-04-02 11:43:33   1st Qu.:1990
##  Mode  :character           Median :2021-08-10 16:45:09   Median :2000
##                             Mean   :2021-07-05 03:41:03   Mean   :1995
##                             3rd Qu.:2021-09-14 18:23:30   3rd Qu.:2002
##                             Max.   :2021-11-09 15:41:20   Max.   :2021
##
##       age          age_group
##  Min.   : 0.00   (18,30]:703
##  1st Qu.:19.00   (0,18] :262
##  Median :21.00   (30,40]:129
##  Mean   :26.24   (40,50]: 98
##  3rd Qu.:31.00   (50,65]: 76
##  Max.   :91.00   (Other): 26
##                  NA's   :  3
```

```r
p <- lc_age_gender %>%
  filter(
    !is.na(gender),
    !is.na(age_group)
  ) %>%
  ggboxplot(
    x = "clade",
    y = "ct_N",
    color = "gender",
    alpha = 0.3,
```

```
    add = c("jitter"),
    add.params = list(alpha = 0.25),
  )

p <- facet(p,
  facet.by = "age_group",
  ncol = 2
)

p + rotate_x_text(45)
```



gender
M
F
(0,18]
(18,30]
(30,40]
(40,50]
(50,65]
(65,85]
(85,100]
ct_N
10
15
20
25
30
20G
20I (Alpha, V1)
20J (Gamma, V3)
21A (Delta)
clade

5.4 Compilation of Supplementary Table with Database Accession Numbers

5.4.1 Clade and Lineage Info

```
glimpse(viralrecon_table)
```

```
## Rows: 1,981
## Columns: 19
## $ testkit_id              <chr> "117M18D54A87FDE9Y8", "117M18D54A8819F81W", "1~
## $ num_input_reads         <dbl> 1803950, 1793232, 1887908, 2360264, 1435168, 4~
## $ num_trimmed_reads_fastp <dbl> 1121492, 1211678, 1259134, 1681220, 915724, 28~
## $ pc_non_host_read        <dbl> 81.97383, 37.66710, 30.45458, 87.74176, 29.085~
## $ pc_mapped_reads         <dbl> 55.69, 0.04, 0.15, 82.65, 3.65, 99.21, 72.28, ~
## $ num_mapped_reads        <dbl> 624539, 543, 1871, 1389586, 33399, 2796671, 20~
## $ num_trimmed_reads_ivar  <dbl> 618974, NA, NA, 1375419, 32764, 2757078, 19976~
## $ median_coverage         <dbl> NA, NA, NA, 1843, NA, 3919, 2948, 1839, 2453, ~
## $ pc_coverage_gt1x        <dbl> 28, NA, NA, 100, 13, 100, 100, 99, 100, NA, 63~
## $ pc_coverage_gt10x       <dbl> 24, NA, NA, 99, 4, 100, 99, 97, 99, NA, 30, 93~
## $ num_snps                <dbl> 11, NA, NA, 18, 2, 40, 40, 39, 19, NA, 19, 27,~
## $ num_indels              <dbl> NA, NA, NA, NA, NA, 3, 3, 3, NA, NA, 5, 1, 1, ~
## $ num_missense_var        <dbl> 6, NA, NA, 10, 1, 26, 28, 25, 5, NA, 15, 18, 1~
## $ Ns_per_100kb            <dbl> 75992.38, NA, NA, 1337.66, 96090.69, 421.49, 1~
## $ lineage                 <chr> NA, NA, NA, "B.1.243", NA, "B.1.617.2", "B.1.6~
## $ clade                   <chr> "21A (Delta)", NA, NA, "20A", "19B", "21A (Del~
## $ variant_caller          <chr> "iVar", "iVar", "iVar", "iVar", "iVar", "iVar"~
## $ viralrecon_version      <dbl> 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2.2, 2~
## $ run_date_time           <date> 2021-11-14, 2021-10-23, 2021-10-23, 2021-10-2~
```

```
clades_lineage_table <- viralrecon_table %>%
  select(testkit_id, lineage, clade) %>%
  filter(!(is.na(lineage) & is.na(clade))) %>%
  distinct() %>%
  arrange(testkit_id)


glimpse(clades_lineage_table)
```

```
## Rows: 1,599
## Columns: 3
## $ testkit_id <chr> "117M18D54A87FDE9Y8", "117M18D5796486135Z", "117M18D6B4187C~
## $ lineage    <chr> NA, "B.1.243", NA, "B.1.617.2", "B.1.617.2", "B.1.617.2", "~
## $ clade      <chr> "21A (Delta)", "20A", "19B", "21A (Delta)", "21A (Delta)", ~
```

```
selected_clades_lineage_table <- limited_clade_collection_diagnostics %>%
  select(testkit_id) %>%
  left_join(clades_lineage_table, by = "testkit_id") %>%
  select(testkit_id, lineage, clade) %>%
  arrange(testkit_id)

glimpse(selected_clades_lineage_table)
```

```
## Rows: 1,297
## Columns: 3
## $ testkit_id <chr> "117M18D54A87FDE9Y8", "117M18D6B5444CCFEJ", "117M18D6B544FA~
## $ lineage    <chr> NA, "B.1.617.2", "B.1.617.2", "B.1.617.2", NA, "B.1.2", "B.~
## $ clade      <chr> "21A (Delta)", "21A (Delta)", "21A (Delta)", "21A (Delta)",~
```

5.4.2 SC DHEC Info

```
glimpse(dhec_table)
```

```
## Rows: 1,412
## Columns: 4
## $ testkit_id      <chr> "117M18D5796486135Z", "117M18D6B5444CCFEJ", "117M18D6B~
## $ dhec_accession  <chr> "5796486135Z", "6B5444CCFEJ", "6B544FA90Y1", "6B545ED0~
## $ pipeline        <chr> "nf-core/viralrecon", "nf-core/viralrecon", "nf-core/v~
## $ submission_date <dbl> 18948, 18948, 18948, 18948, 18948, 18948, 18948, 18948~
```

```
selected_clades_lineage_table <- selected_clades_lineage_table %>%
  left_join(dhec_table %>%
    filter(pipeline == "nf-core/viralrecon"),
  by = "testkit_id"
  ) %>%
  select(testkit_id, dhec_accession, clade, lineage)

glimpse(selected_clades_lineage_table)
```

```
## Rows: 1,297
## Columns: 4
## $ testkit_id     <chr> "117M18D54A87FDE9Y8", "117M18D6B5444CCFEJ", "117M18D6B5~
## $ dhec_accession <chr> NA, "6B5444CCFEJ", "6B544FA90Y1", "6B545ED0BHB", NA, "7~
## $ clade          <chr> "21A (Delta)", "21A (Delta)", "21A (Delta)", "21A (Delt~
## $ lineage        <chr> NA, "B.1.617.2", "B.1.617.2", "B.1.617.2", NA, "B.1.2",~
```

5.4.3 GenBank Info

```
glimpse(genbank_table)
```

```
## Rows: 942
## Columns: 5
## $ testkit_id        <chr> "117M18E004BA5990II", "117M18E004BA5BDFBK", "117M18E~
## $ genbank_sample_id <chr> "004BA5990II", "004BA5BDFBK", "004BA5C10SU", "004BA5~
## $ genbank_accession <chr> "OK340962", "OK340963", "OK340964", "OK340965", "OK3~
## $ pipeline_used     <chr> "labcorp", "labcorp", "labcorp", "labcorp", "labcorp~
## $ submission_date   <dbl> 18900, 18900, 18900, 18900, 18900, 18900, 18900, 189~
```

```
unique(genbank_table$pipeline_used)
```

```
## [1] "labcorp"                "nf-core/viralrecon 2.2"
```

```r
selected_clades_lineage_table <- selected_clades_lineage_table %>%
  left_join(genbank_table %>%
    filter(pipeline_used == "nf-core/viralrecon 2.2"),
  by = "testkit_id"
  ) %>%
  select(testkit_id, dhec_accession, genbank_accession, clade, lineage)

glimpse(selected_clades_lineage_table)
```

```
## Rows: 1,297
## Columns: 5
## $ testkit_id        <chr> "117M18D54A87FDE9Y8", "117M18D6B5444CCFEJ", "117M18D~
## $ dhec_accession    <chr> NA, "6B5444CCFEJ", "6B544FA90Y1", "6B545ED0BHB", NA,~
## $ genbank_accession <chr> NA, NA, NA, NA, NA, "OL709445", "OL709477", NA, NA, ~
## $ clade             <chr> "21A (Delta)", "21A (Delta)", "21A (Delta)", "21A (D~
## $ lineage           <chr> NA, "B.1.617.2", "B.1.617.2", "B.1.617.2", NA, "B.1.~
```

5.4.4 GISAID Info

```r
glimpse(gisaid_table)
```

```
## Rows: 604
## Columns: 5
## $ testkit_id       <chr> "117M18DBD3F633A6VM", "117M18DBD66163CENK", "117M18DE~
## $ gisaid_sample_id <chr> "hCoV-19/USA/SC-REDDI-BD3F633A6VM/2021", "hCoV-19/USA~
## $ gisaid_epi_isl   <chr> "EPI_ISL_7155707", "EPI_ISL_7155714", "EPI_ISL_715571~
## $ pipeline_used    <chr> "nf-core/viralrecon 2.2", "nf-core/viralrecon 2.2", "~
## $ submission_date  <dbl> 18965, 18965, 18965, 18965, 18965, 18965, 18965, 1896~
```

```r
selected_clades_lineage_table <- selected_clades_lineage_table %>%
  left_join(gisaid_table %>%
    filter(pipeline_used == "nf-core/viralrecon 2.2"),
  by = "testkit_id"
  ) %>%
  select(testkit_id, dhec_accession, genbank_accession, gisaid_epi_isl, clade, lineage)

glimpse(selected_clades_lineage_table)
```

```
## Rows: 1,297
## Columns: 6
## $ testkit_id        <chr> "117M18D54A87FDE9Y8", "117M18D6B5444CCFEJ", "117M18D~
## $ dhec_accession    <chr> NA, "6B5444CCFEJ", "6B544FA90Y1", "6B545ED0BHB", NA,~
## $ genbank_accession <chr> NA, NA, NA, NA, NA, "OL709445", "OL709477", NA, NA, ~
## $ gisaid_epi_isl    <chr> NA, NA, NA, NA, NA, "EPI_ISL_7155960", "EPI_ISL_7156~
## $ clade             <chr> "21A (Delta)", "21A (Delta)", "21A (Delta)", "21A (D~
## $ lineage           <chr> NA, "B.1.617.2", "B.1.617.2", "B.1.617.2", NA, "B.1.~
```

```r
selected_clades_lineage_table <- selected_clades_lineage_table %>%
  mutate(sample_id = str_sub(testkit_id,
    start = 8L
```

```
  )) %>%
  relocate(sample_id, .after = testkit_id)

glimpse(selected_clades_lineage_table)
```

```
## Rows: 1,297
## Columns: 7
## $ testkit_id        <chr> "117M18D54A87FDE9Y8", "117M18D6B5444CCFEJ", "117M18D~
## $ sample_id         <chr> "54A87FDE9Y8", "6B5444CCFEJ", "6B544FA90Y1", "6B545E~
## $ dhec_accession    <chr> NA, "6B5444CCFEJ", "6B544FA90Y1", "6B545ED0BHB", NA,~
## $ genbank_accession <chr> NA, NA, NA, NA, NA, "OL709445", "OL709477", NA, NA, ~
## $ gisaid_epi_isl    <chr> NA, NA, NA, NA, NA, "EPI_ISL_7155960", "EPI_ISL_7156~
## $ clade             <chr> "21A (Delta)", "21A (Delta)", "21A (Delta)", "21A (D~
## $ lineage           <chr> NA, "B.1.617.2", "B.1.617.2", "B.1.617.2", NA, "B.1.~
```

5.4.5 Sample Collection Info

```
imp_sc_info <- sample_collection_table %>%
  filter(testkit_id %in% selected_clades_lineage_table$testkit_id) %>%
  select(testkit_id, collection_date, order_priority) %>%
  mutate(collection_date = date(as_datetime(collection_date)))

glimpse(imp_sc_info)
```

```
## Rows: 1,297
## Columns: 3
## $ testkit_id      <chr> "117M18DBD66165C8BE", "117M18DBD66167A8RC", "117M18DBD~
## $ collection_date <date> 2021-01-13, 2021-01-13, 2021-01-13, 2021-01-13, 2021--~
## $ order_priority  <chr> "EXPOSED", "SURVEILLANCE", "SYMPTOMATIC", "SURVEILLANC~
```

```
selected_clades_lineage_table <- selected_clades_lineage_table %>%
  left_join(imp_sc_info,
    by = "testkit_id"
  ) %>%
  mutate(collection_period = as.yearmon(collection_date)) %>%
  select(testkit_id, sample_id, collection_period, order_priority, collection_date, dhec_accession, genb
  arrange(collection_period, order_priority, sample_id)

glimpse(selected_clades_lineage_table)
```

```
## Rows: 1,297
## Columns: 10
## $ testkit_id        <chr> "117M18DBD5B6F749CQ", "117M18DBD5B70AA1EI", "117M18D~
## $ sample_id         <chr> "BD5B6F749CQ", "BD5B70AA1EI", "BD5C92943TH", "BD5C92~
## $ collection_period <yearmon> Jan 2021, Jan 2021, Jan 2021, Jan 2021, Jan 2021~
## $ order_priority    <chr> "EXPOSED", "EXPOSED", "EXPOSED", "EXPOSED", "EXPOSED~
## $ collection_date   <date> 2021-01-19, 2021-01-19, 2021-01-19, 2021-01-19, 202~
## $ dhec_accession    <chr> NA, "BD5B70AA1EI", "BD5C92943TH", "BD5C92CA4AQ", "BD~
## $ genbank_accession <chr> NA, "OL709487", NA, "OL709493", "OL709499", "OL70953~
## $ gisaid_epi_isl    <chr> NA, "EPI_ISL_7156011", "EPI_ISL_7156946", "EPI_ISL_7~
```

```
## $ clade             <chr> "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20~
## $ lineage           <chr> NA, "B.1.2", "B.1.2", "B.1.2", "B.1.2", "B.1.2", "B.~
```

```r
selected_clades_lineage_table <- selected_clades_lineage_table %>%
  select(-c(testkit_id, collection_date))
```

```r
selected_clades_lineage_table %>%
  group_by(sample_id) %>%
  filter(n() > 1)
```

```
## # A tibble: 0 x 8
## # Groups:   sample_id [0]
## # ... with 8 variables: sample_id <chr>, collection_period <yearmon>,
## #   order_priority <chr>, dhec_accession <chr>, genbank_accession <chr>,
## #   gisaid_epi_isl <chr>, clade <chr>, lineage <chr>
```

```r
glimpse(selected_clades_lineage_table)
```

```
## Rows: 1,297
## Columns: 8
## $ sample_id         <chr> "BD5B6F749CQ", "BD5B70AA1EI", "BD5C92943TH", "BD5C92~
## $ collection_period <yearmon> Jan 2021, Jan 2021, Jan 2021, Jan 2021, Jan 2021~
## $ order_priority    <chr> "EXPOSED", "EXPOSED", "EXPOSED", "EXPOSED", "EXPOSED~
## $ dhec_accession    <chr> NA, "BD5B70AA1EI", "BD5C92943TH", "BD5C92CA4AQ", "BD~
## $ genbank_accession <chr> NA, "OL709487", NA, "OL709493", "OL709499", "OL70953~
## $ gisaid_epi_isl    <chr> NA, "EPI_ISL_7156011", "EPI_ISL_7156946", "EPI_ISL_7~
## $ clade             <chr> "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20~
## $ lineage           <chr> NA, "B.1.2", "B.1.2", "B.1.2", "B.1.2", "B.1.2", "B.~
```

5.4.6 Final Formatting

```r
selected_clades_lineage_table %>%
  mutate(collection_period = as.character(collection_period)) %>%
  rename(
    `Sample ID` = "sample_id",
    `Collection Period` = "collection_period",
    `Priority Status` = "order_priority",
    `SC DHEC Accession` = "dhec_accession",
    `GenBank Accession` = "genbank_accession",
    `GISAID Accession` = "gisaid_epi_isl",
    Clade = "clade",
    Lineage = "lineage"
  ) %>%
  write_csv("supplementary_table_1.csv")
```

```r
selected_clades_lineage_table %>%
  mutate(
    collection_period = as.character(collection_period),
    across(.cols = c(
      "order_priority", "collection_period",
```

```
      "clade", "lineage"
    ), as_factor)
  ) %>%
  summary()
```

```
##    sample_id          collection_period      order_priority dhec_accession
##  Length:1297        Sep 2021:343      EXPOSED      :  71   Length:1297
##  Class :character   Aug 2021:202      SURVEILLANCE:1053   Class :character
##  Mode  :character   Mar 2021:201      SYMPTOMATIC : 168   Mode  :character
##                     Apr 2021:157      ONE DAY     :   5
##                     Oct 2021:111
##                     Jan 2021: 83
##                     (Other) :200
##  genbank_accession  gisaid_epi_isl                  clade          lineage
##  Length:1297        Length:1297        20G             :159   B.1.617.2:536
##  Class :character   Class :character   20I (Alpha, V1):264   B.1.1.7  :248
##  Mode  :character   Mode  :character   20J (Gamma, V3): 87   B.1.2    :131
##                                        21A (Delta)    :787   P.1      : 84
##                                                               AY.3     : 25
##                                                               (Other)  : 19
##                                                               NA's     :254
```

5.5 N gene

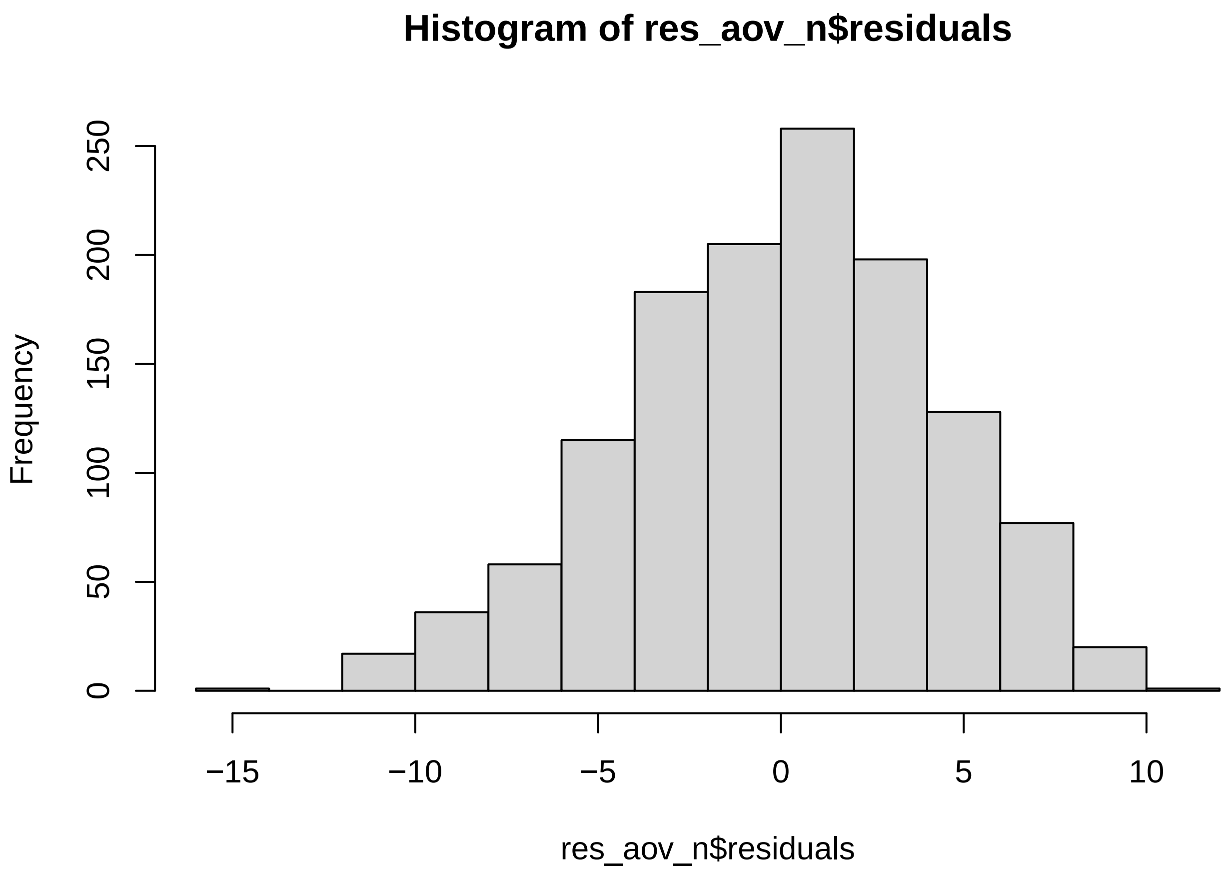
5.5.1 Assumptions

5.5.1.1 Independence No two samples came from the same `patient_id`. The samples are from Clemson University's COVID19 testing program. The CU REDDI lab chose samples that were sent for sequencing.

5.5.1.2 Normality :

```
res_aov_n <- aov(ct_N ~ clade,
  data = limited_clade_collection_diagnostics
)

hist(res_aov_n$residuals)
```

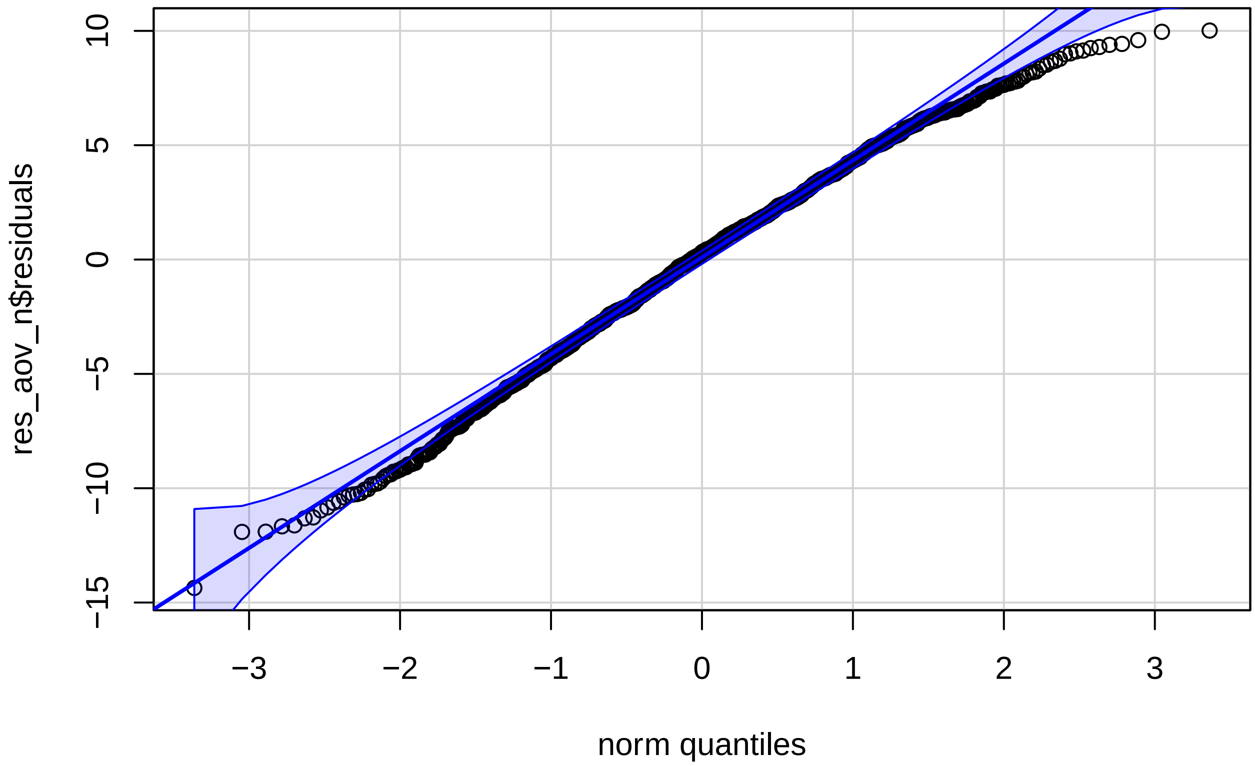
5.5.1.2.1 Histogram of Residuals



The histogram shows a slightly left skewed distribution.

```
qqPlot(res_aov_n$residuals,
  id = FALSE
)
```

5.5.1.2.2 QQ-Plot of Residuals



5.5.1.2.3 Shapiro-Wilk

Null Hypothesis: Data comes from a normal distribution.

Alternate Hypothesis: Data does not come from a normal distribution.

```
shapiro.test(res_aov_n$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov_n$residuals
## W = 0.99307, p-value = 9.203e-06
```

Since p-value < 0.05, we reject the null hypothesis. The data does not follow a normal distribution.

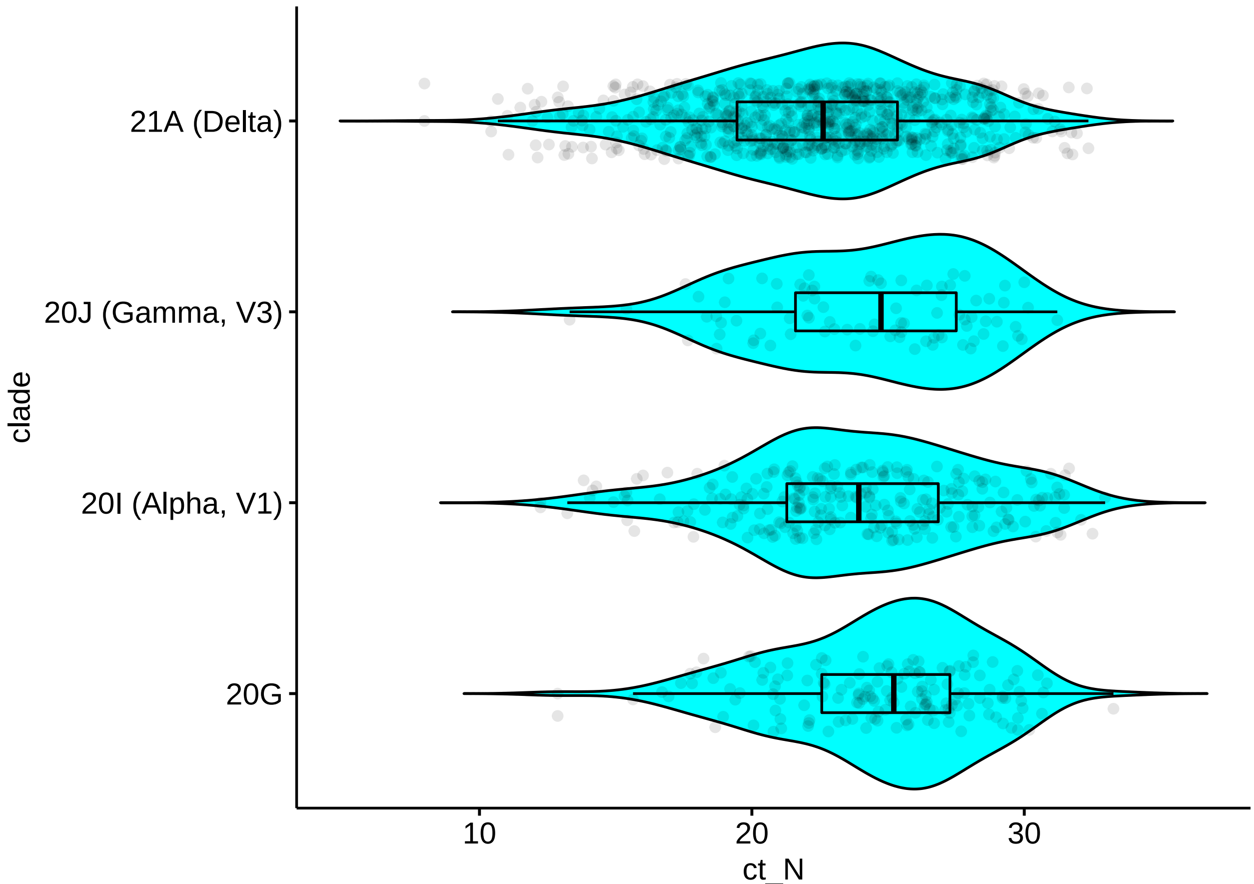
We will perform Kruskal-Wallis Test to compare the Ct values for N gene between the different SARS-CoV-2 clades.

5.5.1.3 Equality of Variances

```
p <- limited_clade_collection_diagnostics %>%
  ggviolin(
    x = "clade",
```

```
    y = "ct_N",
    fill = "cyan",
    add = c("jitter", "boxplot"),
    add.params = list(alpha = 0.1),
    notch = TRUE
  )

ggpar(p, orientation = "horiz")
```

5.5.1.3.1 Box Plot



The box plot appears to show groups that have different variances when compared with the other groups.

5.5.1.3.2 Levene's Test

Null Hypothesis : Variances are equal.

Alternate Hypothesis : At least one variance is different.

```
leveneTest(ct_N ~ clade,
  data = limited_clade_collection_diagnostics
)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
## group    3  2.7112 0.04377 *
##       1293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value < 0.05, we reject the null hypothesis. Equality of Variances assumption is not met.

5.5.2 Kruskal-Wallis Test for Stochastic Dominance

Null Hypothesis :

$H_0 : P(Xi > Xj) = 0.5$ for all groups i and j from 1 to k

Alternate Hypothesis :

$H_A : P(Xi > Xj) \neq 0.5$ for at least one group $i \neq j$

From Non-normal distribution even with Kruskal-Wallis test.

```
kruskal.test(ct_N ~ clade,
  data = limited_clade_collection_diagnostics
)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ct_N by clade
## Kruskal-Wallis chi-squared = 61.758, df = 3, p-value = 2.476e-13
```

Since p-value < 0.05, we reject the null hypothesis. The groups are sampled from populations with different distributions.

5.5.3 Kruskal-Wallis Effect Size

```
kruskal_effsize(ct_N ~ clade,
  data = limited_clade_collection_diagnostics
) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

5.5.4 Dunn's Test of Multiple Comparisons

.y.	n	effsize	method	magnitude
ct_N	1297	0.0454428	eta2[H]	small

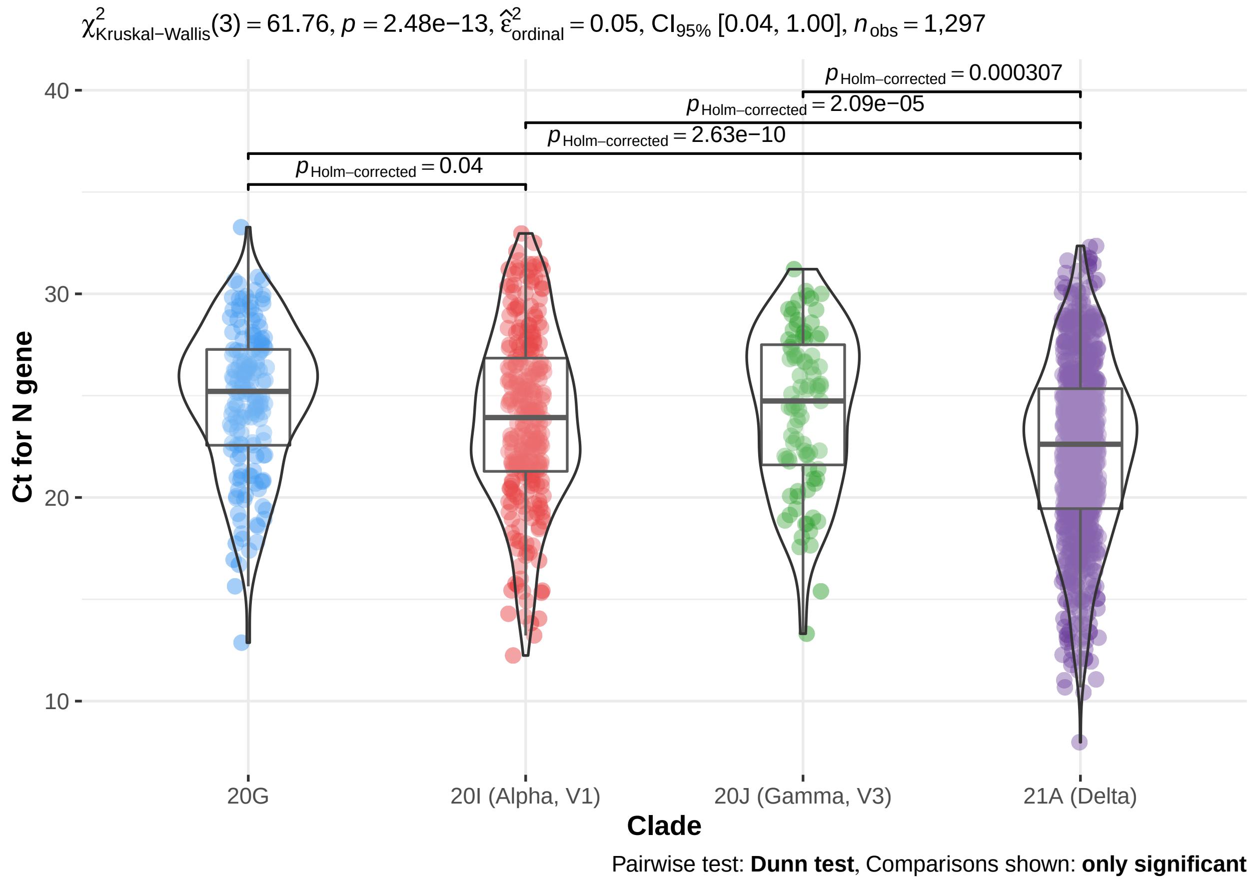
```
dunn_test(ct_N ~ clade,
  data = limited_clade_collection_diagnostics,
  p.adjust.method = "holm"
) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

.y.	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
ct_N	20G	20I (Alpha, V1)	159	264	-2.4477602	0.0143747	0.0431242	*
ct_N	20G	20J (Gamma, V3)	159	87	-0.9466712	0.3438064	0.6676212	ns
ct_N	20G	21A (Delta)	159	787	-6.5905363	0.0000000	0.0000000	****
ct_N	20I (Alpha, V1)	20J (Gamma, V3)	264	87	0.9664669	0.3338106	0.6676212	ns
ct_N	20I (Alpha, V1)	21A (Delta)	264	787	-4.6020787	0.0000042	0.0000209	****
ct_N	20J (Gamma, V3)	21A (Delta)	87	787	-3.9545344	0.0000767	0.0003067	***

There is a significant difference in Ct values of N gene between 21A (Delta) and other clades in the study. Additionally, there is a significant difference in Ct values of N gene between 20I (Alpha, V1) and 20G.

5.5.5 KW Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics,
  x = clade,
  y = ct_N,
  type = "nonparametric",
  xlab = "Clade",
  ylab = "Ct for N gene",
  var.equal = FALSE,
  plot.type = "boxviolin",
  pairwise.comparisons = TRUE,
  pairwise.display = "significant",
  centrality.plotting = FALSE,
  bf.message = FALSE,
  p.adjust.method = "holm"
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```



5.5.6 Welch's ANOVA

```
welch_anova_test(
  formula = ct_N ~ clade,
  data = limited_clade_collection_diagnostics
)
```

```
## # A tibble: 1 x 7
##   .y.       n statistic   DFn   DFd        p method
## * <chr> <int>     <dbl> <dbl> <dbl>    <dbl> <chr>
## 1 ct_N   1297      24.4     3  294. 4.07e-14 Welch ANOVA
```

5.5.7 Games-Howell Test

```
games_howell_test(ct_N ~ clade,
  data = limited_clade_collection_diagnostics
) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
```

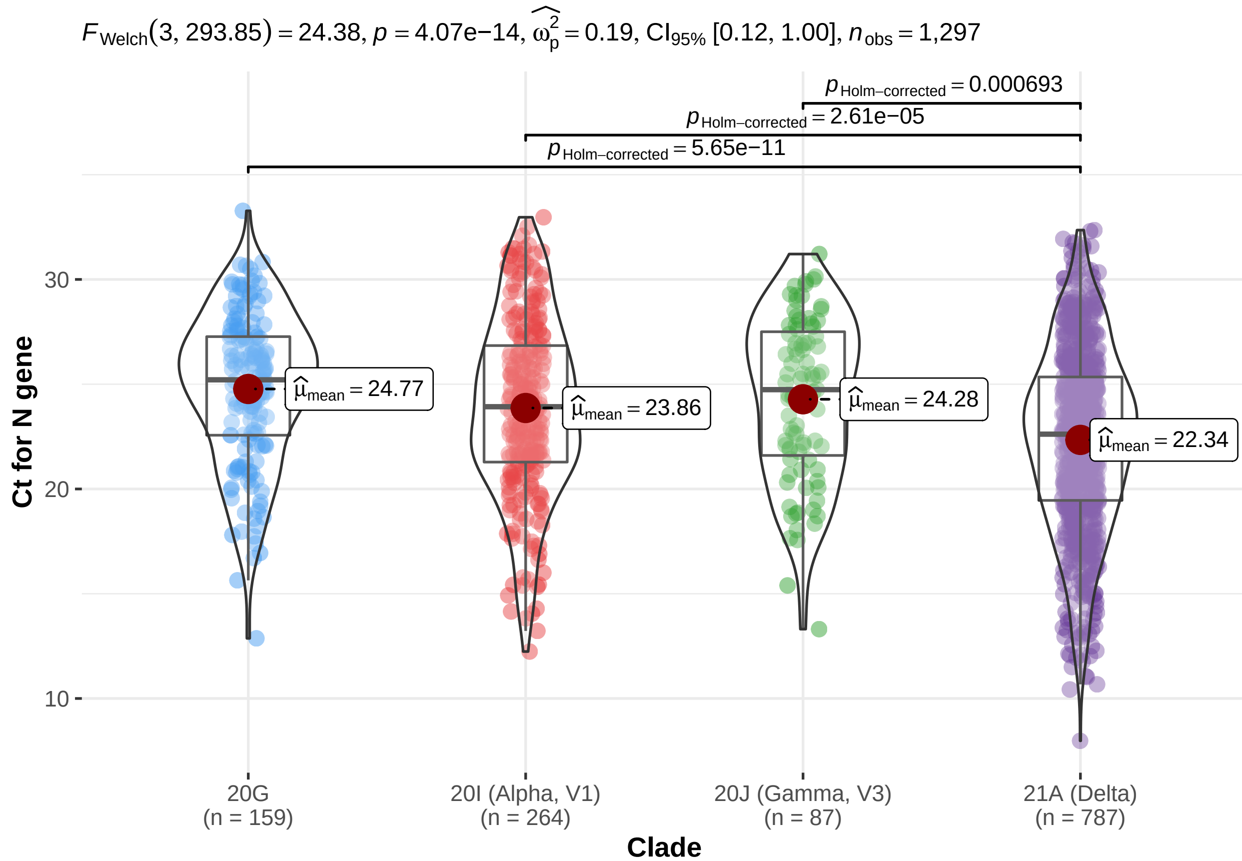
```
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

.y.	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
ct_N	20G	20I (Alpha, V1)	-0.9082844	-1.9145811	0.0980123	9.30e-02	ns
ct_N	20G	20J (Gamma, V3)	-0.4921003	-1.8078835	0.8236829	7.66e-01	ns
ct_N	20G	21A (Delta)	-2.4311226	-3.2769032	-1.5853421	0.00e+00	****
ct_N	20I (Alpha, V1)	20J (Gamma, V3)	0.4161841	-0.8646932	1.6970614	8.34e-01	ns
ct_N	20I (Alpha, V1)	21A (Delta)	-1.5228382	-2.3104644	-0.7352121	5.20e-06	****
ct_N	20J (Gamma, V3)	21A (Delta)	-1.9390223	-3.1005751	-0.7774695	1.73e-04	***

5.5.8 Welch's ANOVA Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics,
  x = clade,
  y = ct_N,
  type = "parametric",
  xlab = "Clade",
  ylab = "Ct for N gene",
  var.equal = FALSE,
  plot.type = "boxviolin"
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



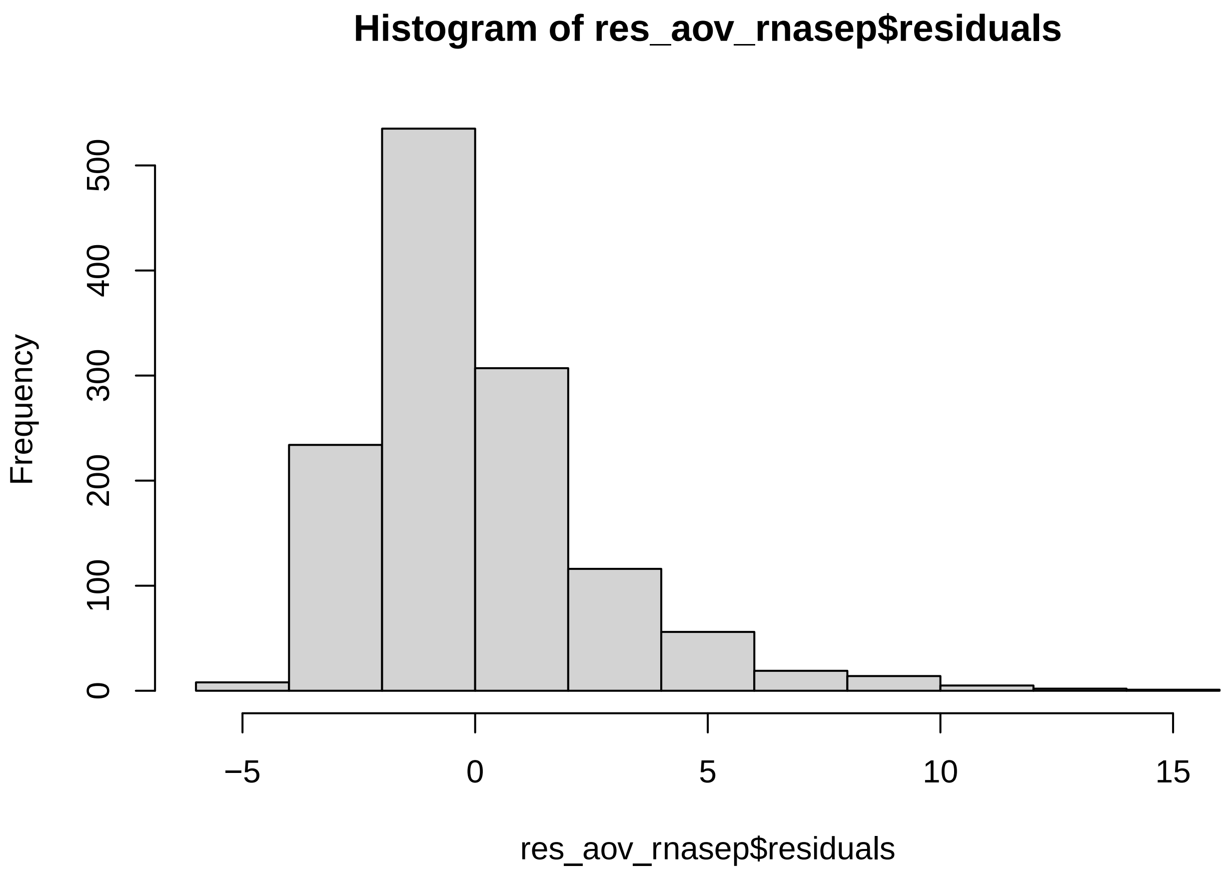
5.6 CONTROL : Compare Ct values for the Human RNase P gene

5.6.1 Assumptions

5.6.1.1 Normality :

```
res_aov_rnasep <- aov(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics
)

hist(res_aov_rnasep$residuals)
```

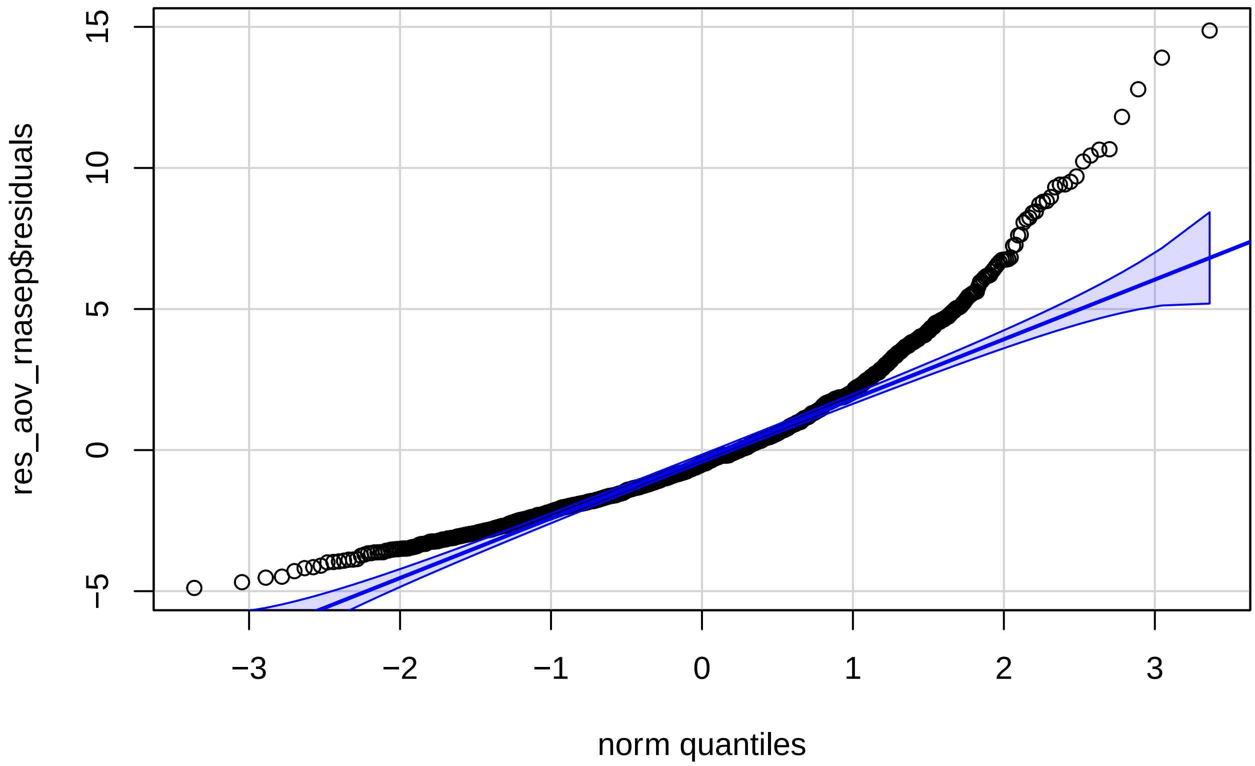
5.6.1.1.1 Histogram of Residuals



There is a strong right skew according to the histogram.

```
qqPlot(res_aov_rnasep$residuals,
  id = FALSE
)
```

5.6.1.1.2 QQ-Plot of Residuals



5.6.1.1.3 Shapiro-Wilk

Null Hypothesis: Data comes from a normal distribution.

Alternate Hypothesis: Data does not come from a normal distribution.

```
shapiro.test(res_aov_rnasep$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov_rnasep$residuals
## W = 0.89981, p-value < 2.2e-16
```

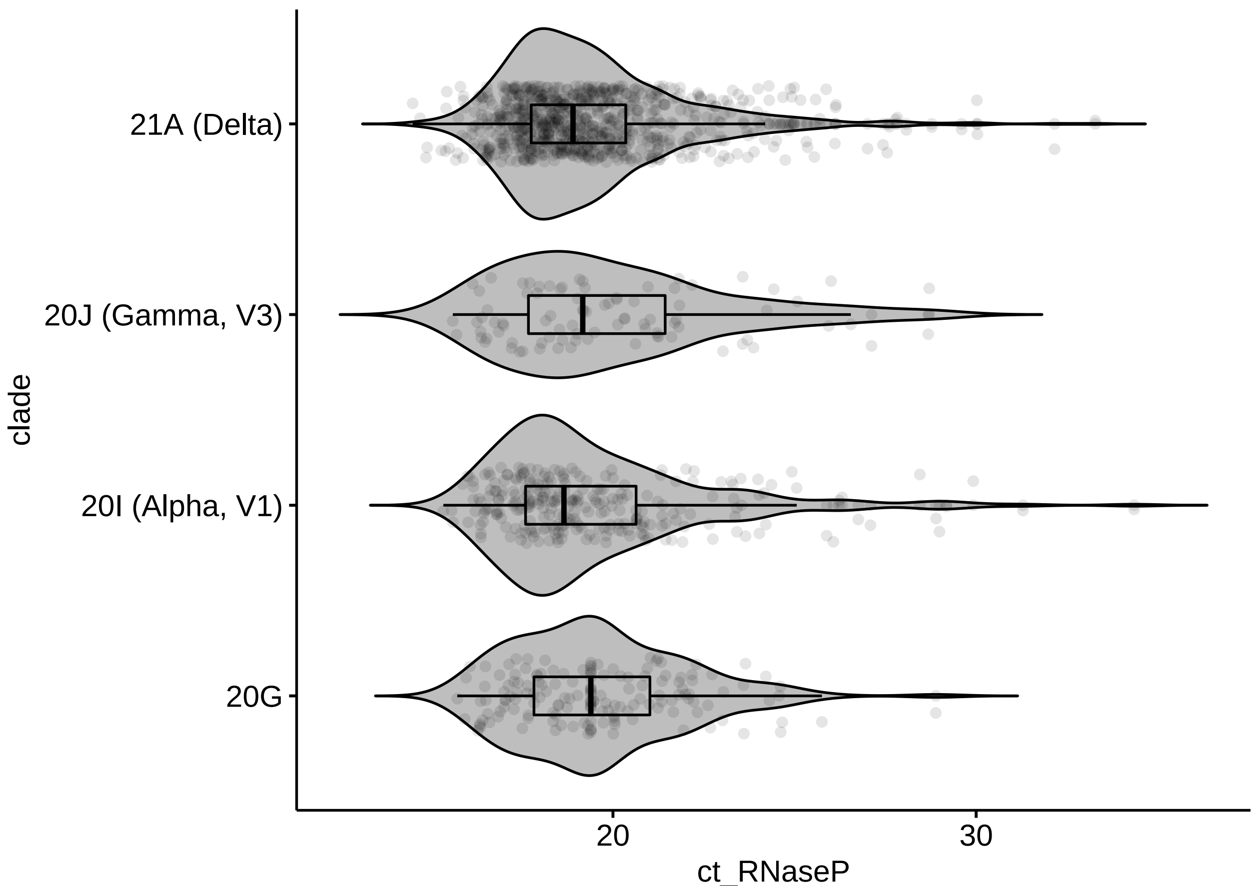
Since p-value < 0.05, we reject the null hypothesis. The data does not follow a normal distribution.

5.6.1.2 Equality of Variances

```
p <- limited_clade_collection_diagnostics %>%
  ggviolin(
    x = "clade",
    y = "ct_RNaseP",
    fill = "grey",
```

```r
    add = c("jitter", "boxplot"),
    add.params = list(alpha = 0.1),
    notch = TRUE
  )

ggpar(p, orientation = "horiz")
```

5.6.1.2.1 Box Plot



5.6.1.2.2 Levene's Test

Null Hypothesis : Population variances are equal.

Alternate Hypothesis : At least one population variance is different.

```r
leveneTest(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics
)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
## group    3  3.2996 0.01975 *
```

```
##       1293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value < 0.05, we reject the null hypothesis. Equality of Variances assumption is not met.

5.6.2 Kruskal-Wallis Test for Stochastic Dominance

Null Hypothesis :

$H_0 : P(Xi > Xj) = 0.5$ for all groups i and j from 1 to k

Alternate Hypothesis :

$H_A : P(Xi > Xj) \neq 0.5$ for at least one group $i \neq j$

From Non-normal distribution even with Kruskal-Wallis test.

```
kruskal.test(ct_RNaseP ~ clade, data = limited_clade_collection_diagnostics)
```

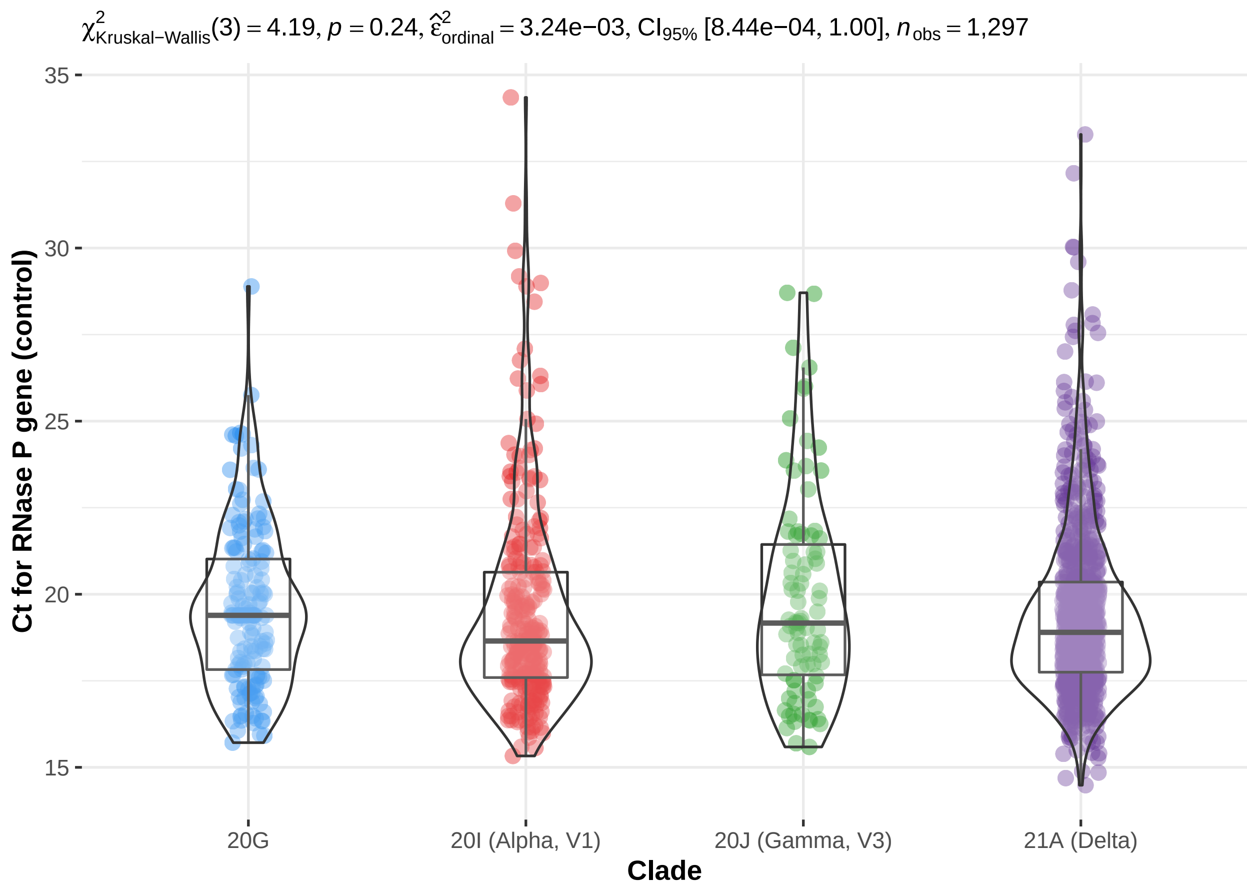
```
##
##  Kruskal-Wallis rank sum test
##
## data:  ct_RNaseP by clade
## Kruskal-Wallis chi-squared = 4.1948, df = 3, p-value = 0.2412
```

Since the p-value is > 0.05, we cannot reject the null hypothesis. There is no significant difference in Ct values for RNase P gene between the different clades in this study.

5.6.3 KW Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics,
  x = clade,
  y = ct_RNaseP,
  type = "nonparametric",
  xlab = "Clade",
  ylab = "Ct for RNase P gene (control)",
  var.equal = FALSE,
  plot.type = "boxviolin",
  pairwise.comparisons = FALSE,
  centrality.plotting = FALSE,
  bf.message = FALSE,
  p.adjust.method = "holm"
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



5.7 Conclusions

We analysed the results from 1297 SARS-CoV-2 infected patients across from the COVID19 testing program in the university area (median age 21 years [<1 year to 91 years], 670 Male : 627 Female).

The clade composition among the sequenced samples were as follows:

```
20G            :159
20I (Alpha, V1):264
20J (Gamma, V3): 87
21A (Delta)    :787
```

Statistical analyses were performed in an R environment (Kruskal-Wallis test followed by Dunn's test of multiple comparisons). There were significant differences in the Ct values for N gene between 21A (Delta) [median: 22.615, range: 7.98-32.355] and other clades [20G : 25.210 (12.87-33.274), 20I (Alpha, V1) : 23.925 (12.24-32.968) , 20J (Gamma, V3): 24.740 (13.31-31.212)]. Additionally, there was a significant difference in Ct values for N gene between 20I (Alpha, V1) and 20G.

Concurrently, the Ct values for the RNase P control did not exhibit a statistically significant difference among these clades [20G : 19.391 (15.714-28.890), 20I (Alpha, V1) : 18.650 (15.330-34.350), 20J (Gamma, V3): 19.168 (15.590-28.708), 21A (Delta) : 18.900 (14.485-33.280)].

These results suggests significant differences in viral load of 21A (Delta) relative to the other clades.

6 surveillance Samples Only

In the following steps, we are limiting our study to samples in the `surveillance` category. We filter the `clades_and_collection` for such samples and name the output table as `clades_and_collection_surveillance`.

```
clades_and_collection_surveillance <- clades_and_collection %>%
  filter(order_priority == "SURVEILLANCE")

glimpse(clades_and_collection_surveillance)
```

```
## Rows: 1,183
## Columns: 10
## $ patient_id       <chr> "0a6f1c09b23ae1ef7b322a8f", "1133ae22349307de010cdeb4~
## $ testkit_id       <chr> "117M18DCE7D400229H", "117M18DCE7D40EDCMA", "117M18DC~
## $ collection_date  <date> 2021-01-13, 2021-01-13, 2021-01-13, 2021-01-13, 2021~
## $ clade            <fct> "20A", "21C (Epsilon)", "20G", "20G", "20C", "20G", "~
## $ population       <fct> UNIVERSITY, UNIVERSITY, UNIVERSITY, UNIVERSITY, ATHLE~
## $ order_priority   <fct> SURVEILLANCE, SURVEILLANCE, SURVEILLANCE, SURVEILLANC~
## $ gender           <fct> M, F, F, M, M, M, M, F, M, F, M, M, M, M, F, M, M, F,~
## $ pregnancy_status <fct> NA, NO, NO, NA, NA, NA, NA, NO, NA, NO, NA, NA, NA, N~
## $ pipeline         <fct> nf-core/viralrecon, nf-core/viralrecon, nf-core/viral~
## $ rymedi_result    <fct> POSITIVE, POSITIVE, POSITIVE, POSITIVE, POSITIVE, POS~
```

The summary of the `clades_and_collection_surveillance` table is as follows:

```
summary(clades_and_collection_surveillance)
```

```
##   patient_id         testkit_id        collection_date
##  Length:1183        Length:1183        Min.   :2021-01-13
##  Class :character   Class :character   1st Qu.:2021-04-01
##  Mode  :character   Mode  :character   Median :2021-08-17
##                                        Mean   :2021-07-06
##                                        3rd Qu.:2021-09-15
##                                        Max.   :2021-11-09
##
##              clade          population       order_priority gender
##  21A (Delta)    :691   UNIVERSITY:887   SURVEILLANCE:1183   M:623
##  20I (Alpha, V1):181   ATHLETICS : 26   SYMPTOMATIC :   0   F:560
##  20G            : 95   COMMUNITY :270   EXPOSED     :   0
##  20J (Gamma, V3): 86   TRICOUNTY :  0   ONE DAY     :   0
##  20A            : 55
##  21C (Epsilon)  : 23
##  (Other)        : 52
##  pregnancy_status               pipeline     rymedi_result
##  YES :  3         nf-core/viralrecon:1183   POSITIVE:1183
##  NO  :557
##  NA's:623
##
##
##
##
```

6.1 Bar Plots with Monthly Count for each Clade : `surveillance` Samples Only

The clades present in `clades_and_collection_surveillance` when only `surveillance` samples are considered are as follows:

```
clades_factor_level <- clades_and_collection_surveillance %>%
  pull(clade) %>%
  unique()

(clades_factor_level <- sort(as.character(clades_factor_level)))
```

```
##  [1] "19A"             "19B"             "20A"             "20B"
##  [5] "20C"             "20G"             "20H (Beta, V2)"  "20I (Alpha, V1)"
##  [9] "20J (Gamma, V3)" "21A (Delta)"     "21B (Kappa)"     "21C (Epsilon)"
## [13] "21D (Eta)"       "21F (Iota)"
```

From `clades_and_collection_surveillance`, we tabulate the count of each clade among sequenced samples collected in each month of 2021.

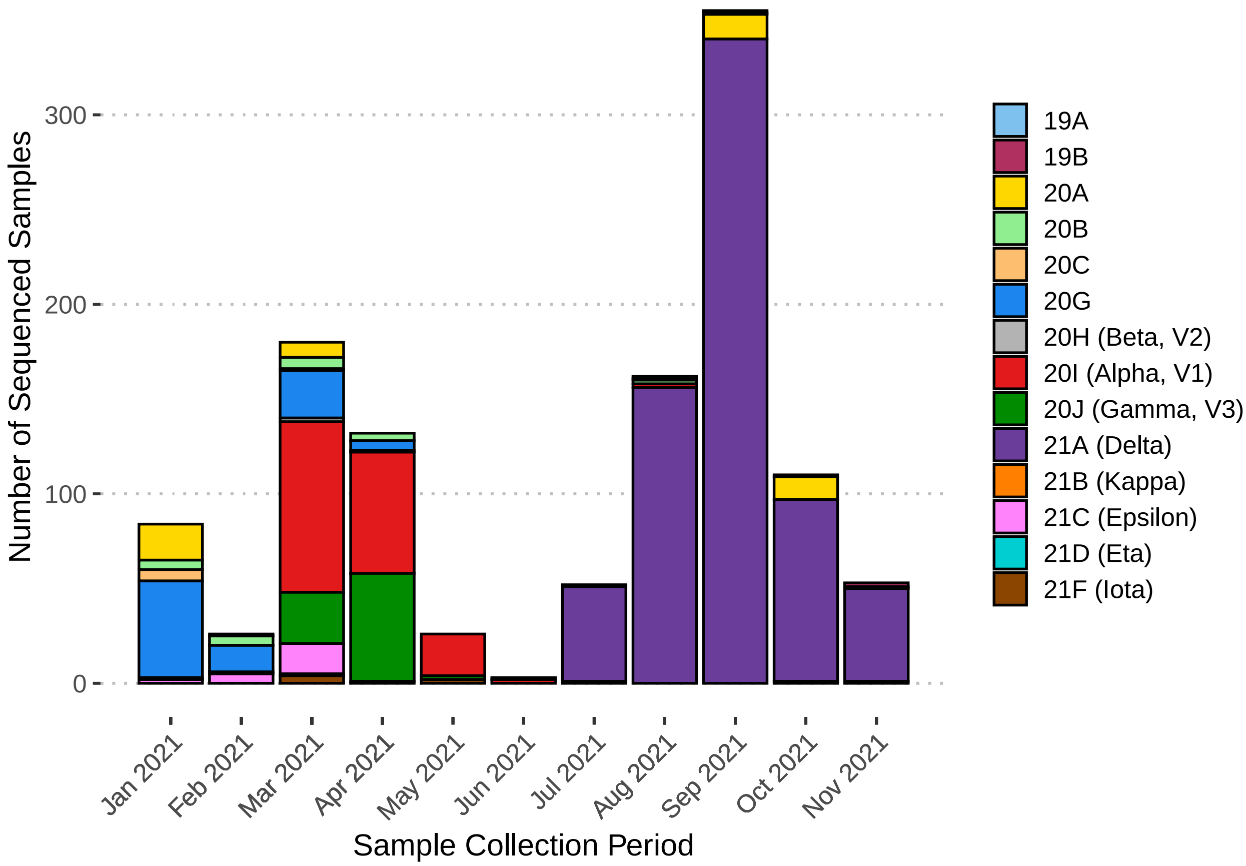
```
monthly_clade_date_surveillance <- clades_and_collection_surveillance %>%
  mutate(
    collection_period = as.yearmon(collection_date),
    clade = factor(clade,
      levels = clades_factor_level
    )
  ) %>%
  group_by(collection_period, clade) %>%
  summarize(count = n())

glimpse(monthly_clade_date_surveillance)
```

```
## Rows: 52
## Columns: 3
## Groups: collection_period [11]
## $ collection_period <yearmon> Jan 2021, Jan 2021, Jan 2021, Jan 2021, Jan 2021~
## $ clade             <fct> "20A", "20B", "20C", "20G", "20H (Beta, V2)", "21C (~
## $ count             <int> 19, 5, 6, 51, 1, 2, 1, 5, 14, 1, 5, 8, 6, 1, 25, 2, ~
```

Visualizing the table above using a stacked bar plot.

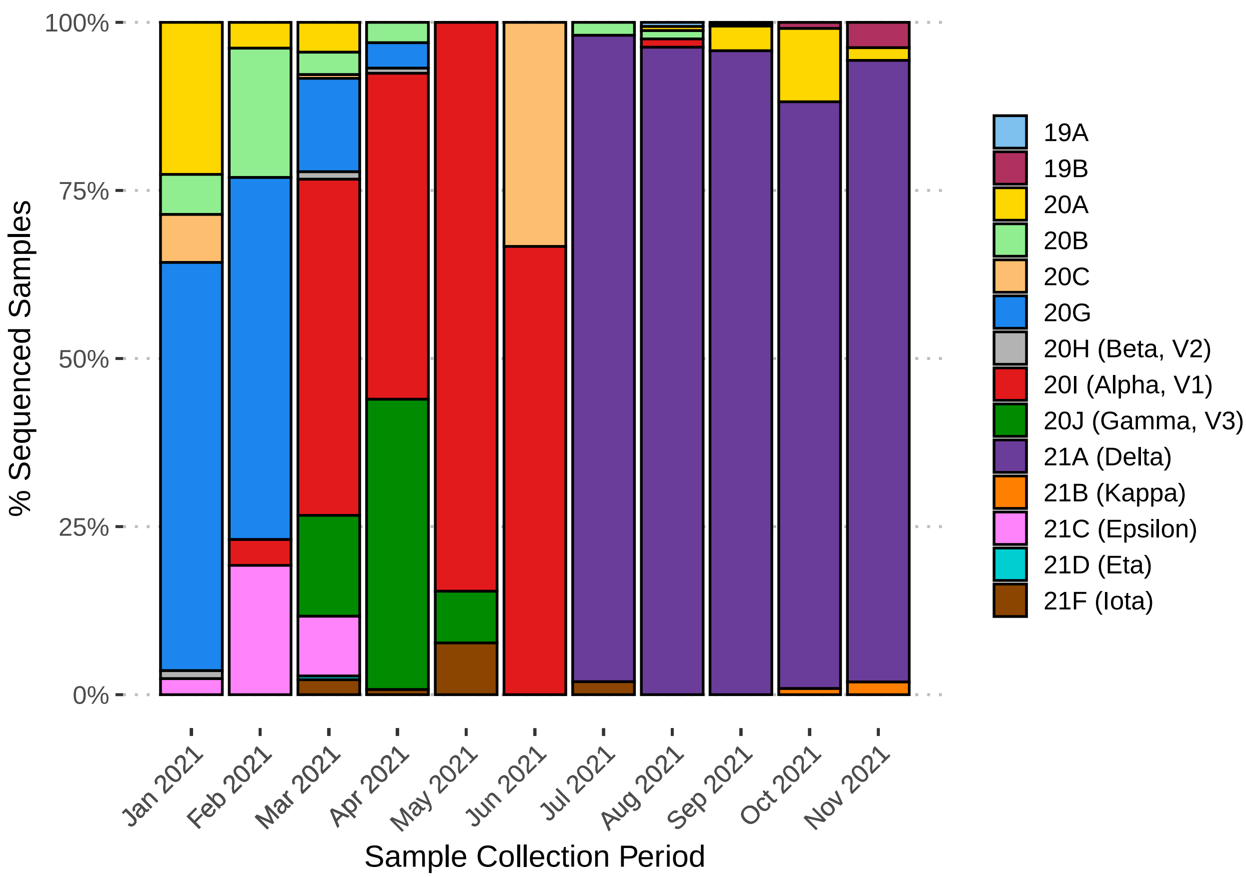
```
ggplot(
  monthly_clade_date_surveillance,
  aes(
    x = collection_period,
    y = count,
    fill = clade
  )
) +
  geom_bar(colour = "black", position = "stack", stat = "identity") +
  scale_x_yearmon(breaks = monthly_clade_date$collection_period) +
  scale_fill_manual(values = color_tbl %>%
    filter(clade %in% monthly_clade_date_surveillance$clade) %>%
```

```
    pull(color)) +
  labs(
    y = "Number of Sequenced Samples",
    x = "Sample Collection Period"
  ) +
  theme_pubclean() +
  theme(
    legend.position = "right",
    legend.title = element_blank(),
    legend.key.size = unit(0.5, "cm"),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)
  )
```




```
ggplot(
  monthly_clade_date_surveillance,
  aes(
    x = collection_period,
    y = count,
    fill = clade
  )
) +
  geom_col(colour = "black", position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  scale_x_yearmon(breaks = monthly_clade_date$collection_period) +
```

```
scale_fill_manual(values = color_tbl %>%
  filter(clade %in% monthly_clade_date_surveillance$clade) %>%
  pull(color)) +
labs(
  y = "% Sequenced Samples",
  x = "Sample Collection Period"
) +
theme_pubclean() +
theme(
  legend.position = "right",
  legend.title = element_blank(),
  legend.key.size = unit(0.5, "cm"),
  axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)
)
```



6.2 Preparing `diagnostics` data for surveillance-only set of samples

Our goal is to determine whether 21A (Delta) shows lower Ct values for N gene when compared with the clades 20I (Alpha, V1), 20G, and 20J (Gamma, V3) *when only the surveillance samples are considered.* In order to know that, we use the Ct values for N gene from the REDDI lab dataset. Since two replicates were done for each qRT-PCR reaction, we will compute the mean Ct for N gene and RNase P control for each `testkit_id`

```
diagnostics_data_surveillance <- diagnostics_table %>%
  filter(testkit_id %in% clades_and_collection_surveillance$testkit_id) %>%
  select(testkit_id, ct_rnasep_rep1, ct_rnasep_rep2, ct_N_rep1, ct_N_rep2) %>%
  arrange(testkit_id) %>%
  mutate(
    average_ct_rnasep = rowMeans(.[, c("ct_rnasep_rep1", "ct_rnasep_rep2")], na.rm = TRUE),
    average_ct_N = rowMeans(.[, c("ct_N_rep1", "ct_N_rep2")], na.rm = TRUE)
  ) %>%
  select(-c(
    ct_rnasep_rep1, ct_rnasep_rep2,
    ct_N_rep1, ct_N_rep2
  )) %>%
  group_by(testkit_id) %>%
  summarise(
    ct_RNaseP = mean(average_ct_rnasep, na.rm = TRUE),
    ct_N = mean(average_ct_N, na.rm = TRUE)
  )

diagnostics_data_surveillance %>%
  filter(!complete.cases(.)) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

26 `testkit_id`s do not have a ct value for RNase P. We will use imputation to deal with the missing values.

6.3 Join `clade_assignments` and `diagnostics_data`

We join `clades_and_collection_surveillance` and the diagnostics data to get the `clade_collection_diagnostics_surve` table.

```
clade_collection_diagnostics_surveillance <- clades_and_collection_surveillance %>%
  inner_join(diagnostics_data_surveillance, by = "testkit_id")

glimpse(clade_collection_diagnostics_surveillance)
```

```
## Rows: 1,183
## Columns: 12
## $ patient_id       <chr> "0a6f1c09b23ae1ef7b322a8f", "1133ae22349307de010cdeb4~
## $ testkit_id       <chr> "117M18DCE7D400229H", "117M18DCE7D40EDCMA", "117M18DC~
## $ collection_date  <date> 2021-01-13, 2021-01-13, 2021-01-13, 2021-01-13, 2021~
## $ clade            <fct> "20A", "21C (Epsilon)", "20G", "20G", "20C", "20G", "~
## $ population       <fct> UNIVERSITY, UNIVERSITY, UNIVERSITY, UNIVERSITY, ATHLE~
## $ order_priority   <fct> SURVEILLANCE, SURVEILLANCE, SURVEILLANCE, SURVEILLANC~
## $ gender           <fct> M, F, F, M, M, M, M, F, M, F, M, M, M, M, F, M, M, F,~
## $ pregnancy_status <fct> NA, NO, NO, NA, NA, NA, NA, NO, NA, NO, NA, NA, NA, N~
## $ pipeline         <fct> nf-core/viralrecon, nf-core/viralrecon, nf-core/viral~
```

testkit_id	ct_RNaseP	ct_N
117M18DBFE8BB0A1JJ	NaN	21.29940
117M18DBFE8BB18FTW	NaN	26.99776
117M18DBFE8C4DF29V	NaN	30.18376
117M18DBFEE11FC1HA	NaN	23.52962
117M18DCB750444B1K	NaN	29.40396
117M18DCB750672AZL	NaN	25.19355
117M18DCB750EF7DKW	NaN	23.28848
117M18DCB7CE9BFF04	NaN	24.22209
117M18DD9F7D8DF0LH	NaN	23.69066
117M18DD9F7DF503YP	NaN	16.45756
117M18DD9F7E062D0Q	NaN	24.82307
117M18DD9F7E67D57T	NaN	26.20155
117M18DDC2E0AD0A70	NaN	29.20234
117M18DDD30852A7D3	NaN	16.67494
117M18DDD3085427A5	NaN	30.80809
117M18DDD308C25FZY	NaN	21.30878
117M18DDD308C63FEZ	NaN	25.31193
117M18DDD4896537BX	NaN	18.88141
117M18DDD48F66A8GX	NaN	26.61771
117M18DDD48FDEBD87	NaN	26.39463
117M18DDD490489F80	NaN	26.09010
117M18DDD4916F9CG2	NaN	21.62621
117M18DE172DA3990T	NaN	22.35571
117M18DE172DA3E12G	NaN	27.27337
117M18DE172DA66AEW	NaN	23.83437
117M18DE1733C0D5PY	NaN	26.53400

```
## $ rymedi_result    <fct> POSITIVE, POSITIVE, POSITIVE, POSITIVE, POSITIVE, POS~
## $ ct_RNaseP        <dbl> 16.59912, 17.39000, 21.62271, 24.57796, 19.51531, 18.~
## $ ct_N             <dbl> 22.26138, 23.00548, 28.37183, 23.92611, 22.17447, 25.~
```

Since many samples had missing Ct values for RNase P, we impute median value of `ct_RNaseP` for each clade to the missing values.

```r
get_median_ct_surveillance <- function(input_clade) {
  median_RNAseP <- clade_collection_diagnostics_surveillance %>%
    filter(clade == input_clade) %>%
    pull(ct_RNaseP) %>%
    median(na.rm = TRUE)

  return(median_RNAseP)
}
```

Median Ct RNase P - 20I (Alpha, V1)

```r
# 20I (Alpha, V1)
(median_RNAseP_20I_surveillance <- get_median_ct_surveillance("20I (Alpha, V1)"))
```

```
## [1] 18.77233
```

Median Ct RNase P - 21A (Delta)

```r
# 21A (Delta)
(median_RNAseP_21A_surveillance <- get_median_ct_surveillance("21A (Delta)"))
```

```
## [1] 18.935
```

Median Ct RNase P - 20G

```r
# 20G
(median_RNAseP_20G_surveillance <- get_median_ct_surveillance("20G"))
```

```
## [1] 19.18
```

Median Ct RNase P - 20J (Gamma, V3)

```r
# 20J (Gamma, V3)
(median_RNAseP_20J_surveillance <- get_median_ct_surveillance("20J (Gamma, V3)"))
```

```
## [1] 19.16656
```

Summary of the `clade_collection_diagnostics` table after imputation

```r
clade_collection_diagnostics_surveillance <- clade_collection_diagnostics_surveillance %>%
  mutate(
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20I (Alpha, V1)")),
      median_RNAseP_20I_surveillance
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "21A (Delta)")),
      median_RNAseP_21A_surveillance
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20G")),
      median_RNAseP_20G_surveillance
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20J (Gamma, V3)")),
      median_RNAseP_20J_surveillance
    )
  )

summary(clade_collection_diagnostics_surveillance)
```

```
##   patient_id         testkit_id        collection_date
##  Length:1183        Length:1183        Min.   :2021-01-13
##  Class :character   Class :character   1st Qu.:2021-04-01
```

```
##  Mode  :character   Mode  :character   Median :2021-08-17
##                                        Mean   :2021-07-06
##                                        3rd Qu.:2021-09-15
##                                        Max.   :2021-11-09
##
##              clade          population       order_priority gender
##  21A (Delta)    :691   UNIVERSITY:887   SURVEILLANCE:1183   M:623
##  20I (Alpha, V1):181   ATHLETICS : 26   SYMPTOMATIC :   0   F:560
##  20G            : 95   COMMUNITY :270   EXPOSED     :   0
##  20J (Gamma, V3): 86   TRICOUNTY :  0   ONE DAY     :   0
##  20A            : 55
##  21C (Epsilon)  : 23
##  (Other)        : 52
##  pregnancy_status               pipeline     rymedi_result    ct_RNaseP
##  YES :  3         nf-core/viralrecon:1183   POSITIVE:1183   Min.   :14.48
##  NO  :557                                                   1st Qu.:17.76
##  NA's:623                                                   Median :18.96
##                                                             Mean   :19.50
##                                                             3rd Qu.:20.59
##                                                             Max.   :34.35
##                                                             NA's   :12
##       ct_N
##  Min.   :10.68
##  1st Qu.:20.31
##  Median :23.44
##  Mean   :23.17
##  3rd Qu.:26.36
##  Max.   :33.45
##
```

The counts of different clades in the `clade_collection_diagnostics` table are as follows:

```
clade_collection_diagnostics_surveillance %>%
  group_by(clade) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

As in the previous analysis, in order to not affect our assumption of phylogenetic independence, we are only going to compare Ct values for variants at terminal nodes of NextClade tree. We extract data from `clade_collection_diagnostics_surveillance` table and create a new table `limited_clade_collection_diagnostics_surveillance` with only data for the following clades:

```
21A (Delta)
20I (Alpha, V1)
20J (Gamma, V3)
20G
```

clade	count
21A (Delta)	691
20I (Alpha, V1)	181
20G	95
20J (Gamma, V3)	86
20A	55
21C (Epsilon)	23
20B	23
21F (Iota)	8
20C	8
20H (Beta, V2)	4
19B	4
21B (Kappa)	2
19A	2
21D (Eta)	1

```
limited_clade_collection_diagnostics_surveillance <- clade_collection_diagnostics_surveillance %>%
  filter(clade %in% c(
    "21A (Delta)",
    "20I (Alpha, V1)",
    "20J (Gamma, V3)",
    "20G"
  )) %>%
  mutate(clade = factor(clade,
    levels = c(
      "20G",
      "20I (Alpha, V1)",
      "20J (Gamma, V3)",
      "21A (Delta)"
    )
  )) %>%
  select(testkit_id, clade, ct_RNaseP, ct_N) %>%
  ungroup()

glimpse(limited_clade_collection_diagnostics_surveillance)
```

```
## Rows: 1,053
## Columns: 4
## $ testkit_id <chr> "117M18DCE7D410C0MX", "117M18D7B495AED7RY", "117M18DBD66167~
## $ clade      <fct> "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20~
## $ ct_RNaseP  <dbl> 21.62271, 24.57796, 18.01706, 18.35299, 17.96138, 24.60967,~
## $ ct_N       <dbl> 28.37183, 23.92611, 25.77516, 21.04739, 20.37994, 22.67064,~
```

Summary of the `limited_clade_collection_diagnostics_surveillance` table :

```
summary(limited_clade_collection_diagnostics_surveillance)
```

```
##   testkit_id                      clade        ct_RNaseP          ct_N
##  Length:1053        20G             : 95   Min.   :14.48   Min.   :10.68
##  Class :character   20I (Alpha, V1):181   1st Qu.:17.77   1st Qu.:20.20
##  Mode  :character   20J (Gamma, V3): 86   Median :18.98   Median :23.20
```

```
##                    21A (Delta)    :691   Mean   :19.48   Mean   :23.02
##                                          3rd Qu.:20.57   3rd Qu.:26.15
##                                          Max.   :34.35   Max.   :33.27
```

Median and range of patient age whose samples are in the `limited_clade_collection_diagnostics_surveillance` table as follows:

```
sample_collection_table %>%
  mutate(collection_date = year(as_datetime(collection_date))) %>%
  filter(testkit_id %in% limited_clade_collection_diagnostics_surveillance$testkit_id) %>%
  left_join(demographics_table, by = "patient_id") %>%
  mutate(age_at_sample_collection = (collection_date - birth_year)) %>%
  select(testkit_id, patient_id, age_at_sample_collection) %>%
  summarize(
    median_age_at_sample_collection = median(age_at_sample_collection),
    lowest_age_at_sample_collection = min(age_at_sample_collection),
    highest_age_at_sample_collection = max(age_at_sample_collection)
  ) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

median_age_at_sample_collection	lowest_age_at_sample_collection	highest_age_at_sample_collection
21	0	91

The counts of each gender in the `limited_clade_collection_diagnostics_surveillance` table are as follows:

```
sample_collection_table %>%
  filter(testkit_id %in% limited_clade_collection_diagnostics_surveillance$testkit_id) %>%
  group_by(gender) %>%
  summarize(count = n()) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

gender	count
F	500
M	553

The median and range of Ct values for each clade in the `limited_clade_collection_diagnostics_surveillance` table are as follows:

```r
limited_clade_collection_diagnostics_surveillance %>%
  group_by(clade) %>%
  summarise(
    count = n(),
    median_ct_N = round(median(ct_N), 3),
    min_ct_N = round(range(ct_N)[1], 3),
    max_ct_N = round(range(ct_N)[2], 3),
    median_ct_RNaseP = round(median(ct_RNaseP), 3),
    min_ct_RNaseP = round(range(ct_RNaseP)[1], 3),
    max_ct_RNaseP = round(range(ct_RNaseP)[2], 3)
  ) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped",
      "scale_down"
    )
  )
```

clade	count	median_ct_N	min_ct_N	max_ct_N	median_ct_RNaseP	min_ct_RNaseP	max_ct_RNaseP
20G	95	25.752	17.735	33.274	19.180	15.714	25.753
20I (Alpha, V1)	181	23.810	12.240	32.968	18.772	15.330	34.350
20J (Gamma, V3)	86	24.688	13.310	31.212	19.167	15.590	28.708
21A (Delta)	691	22.560	10.675	32.355	18.935	14.485	32.160

6.4 N gene

6.4.1 Assumptions

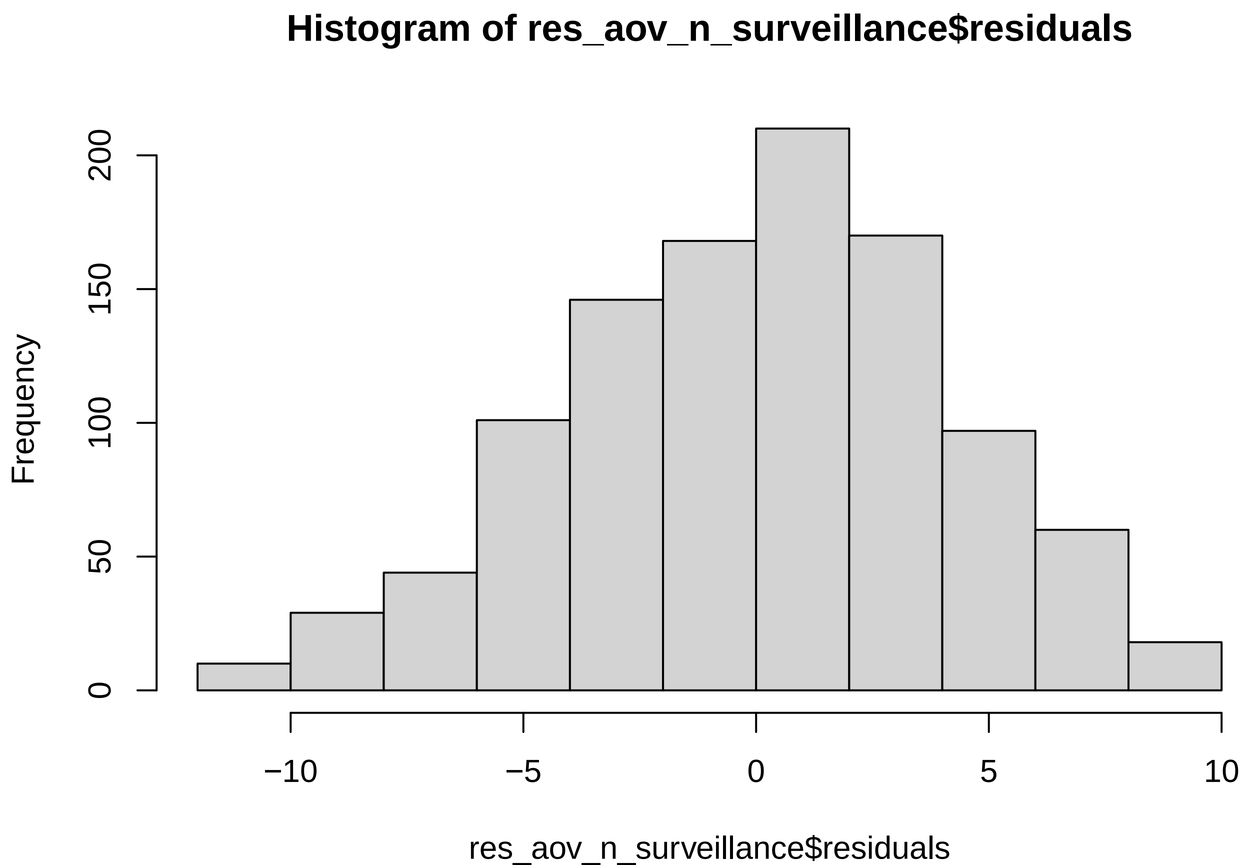
6.4.1.1 Independence

No two samples came from the same `patient_id`. The samples are from Clemson University's COVID19 testing program and the order priority of these samples were labeled as `surveillance`. The CU REDDI lab chose samples that were sent for sequencing.

6.4.1.2 Normality :

```r
res_aov_n_surveillance <- aov(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
)

hist(res_aov_n_surveillance$residuals)
```

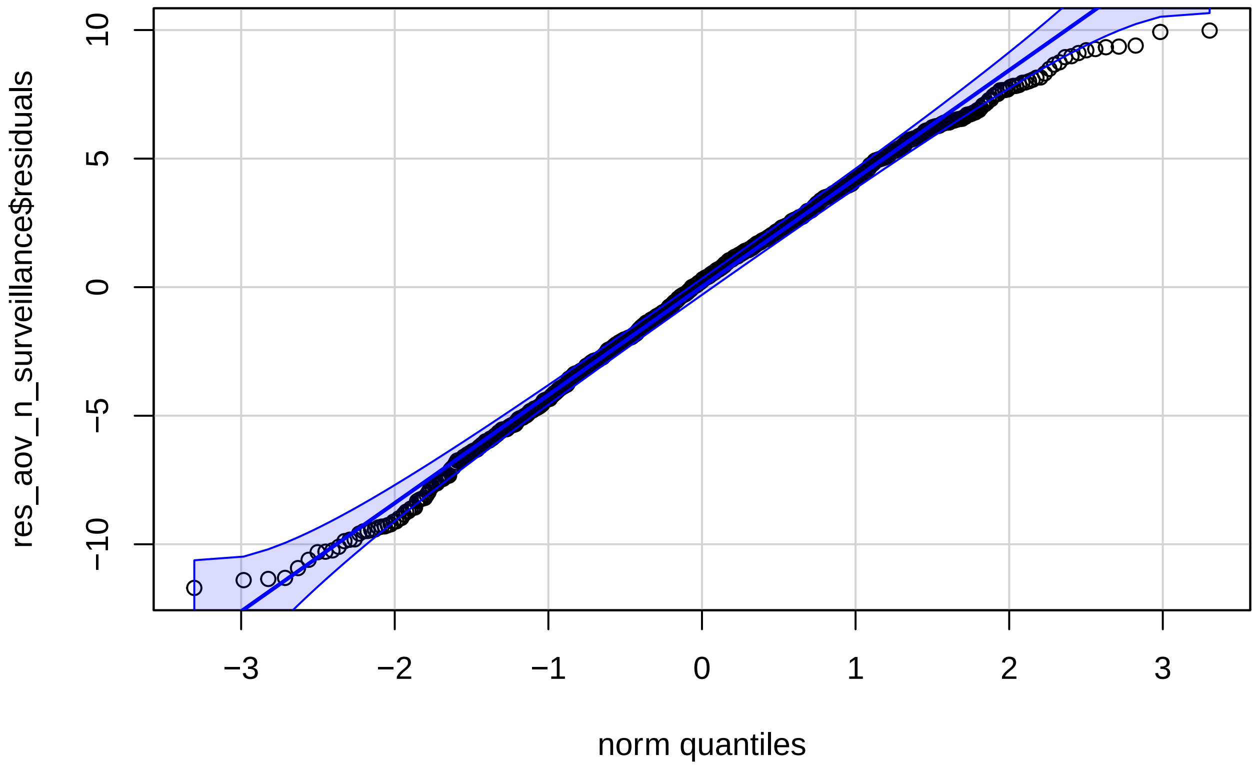
6.4.1.2.1 Histogram of Residuals



The histogram shows a slightly left skewed distribution.

```
qqPlot(res_aov_n_surveillance$residuals,
  id = FALSE
)
```

6.4.1.2.2 QQ-Plot of Residuals



6.4.1.2.3 Shapiro-Wilk

Null Hypothesis: Data comes from a normal distribution.

Alternate Hypothesis: Data does not come from a normal distribution.

```
shapiro.test(res_aov_n_surveillance$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov_n_surveillance$residuals
## W = 0.99412, p-value = 0.0003742
```

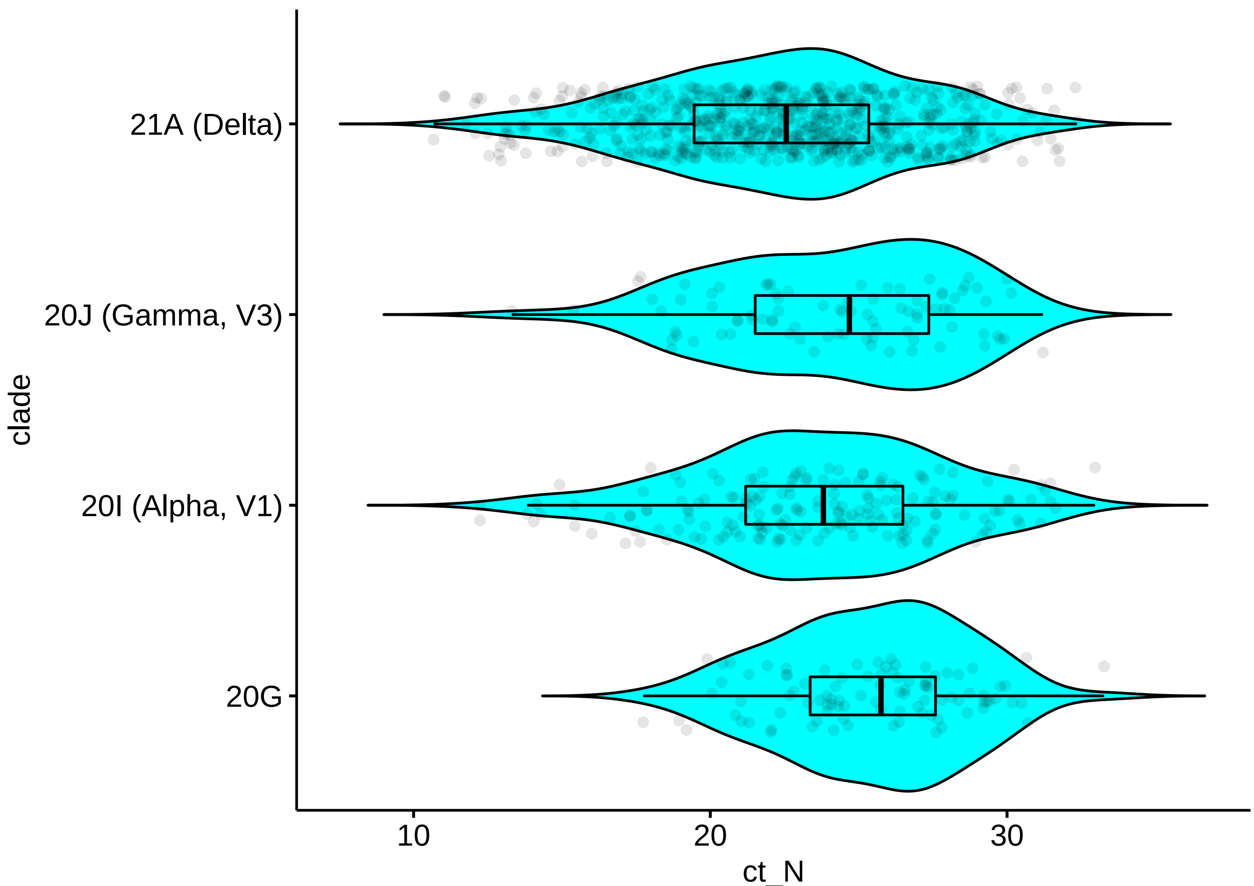
Since p-value < 0.05, we reject the null hypothesis. The data not follow a normal distribution.

6.4.1.3 Equality of Variances

```
p <- limited_clade_collection_diagnostics_surveillance %>%
  ggviolin(
    x = "clade",
    y = "ct_N",
    fill = "cyan",
```

```
    add = c("jitter", "boxplot"),
    add.params = list(alpha = 0.1),
    notch = TRUE
  )

ggpar(p, orientation = "horiz")
```

6.4.1.3.1 Box Plot



6.4.1.3.2 Levene's Test

Null Hypothesis : Variances are equal.

Alternate Hypothesis : At least one variance is different.

```
leveneTest(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value   Pr(>F)
## group    3  4.0785 0.006822 **
```

```
##       1049
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value < 0.05, we reject the null hypothesis. Equality of Variances assumption is not met.

6.4.2 Kruskal-Wallis Test for Stochastic Dominance

Null Hypothesis :

$H_0 : P(Xi > Xj) = 0.5$ for all groups i and j from 1 to k

Alternate Hypothesis :

$H_A : P(Xi > Xj) \neq 0.5$ for at least one group $i \neq j$

From Non-normal distribution even with Kruskal-Wallis test.

```
kruskal.test(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ct_N by clade
## Kruskal-Wallis chi-squared = 56.708, df = 3, p-value = 2.967e-12
```

Since p-value < 0.05, we reject the null hypothesis. The groups are sampled from populations with different distributions.

6.4.3 Kruskal-Wallis Effect Size

```
kruskal_effsize(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

.y.	n	effsize	method	magnitude
ct_N	1053	0.0511989	eta2[H]	small

6.4.4 Dunn's Test of Multiple Comparisons

```
dunn_test(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_surveillance,
  p.adjust.method = "holm"
) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

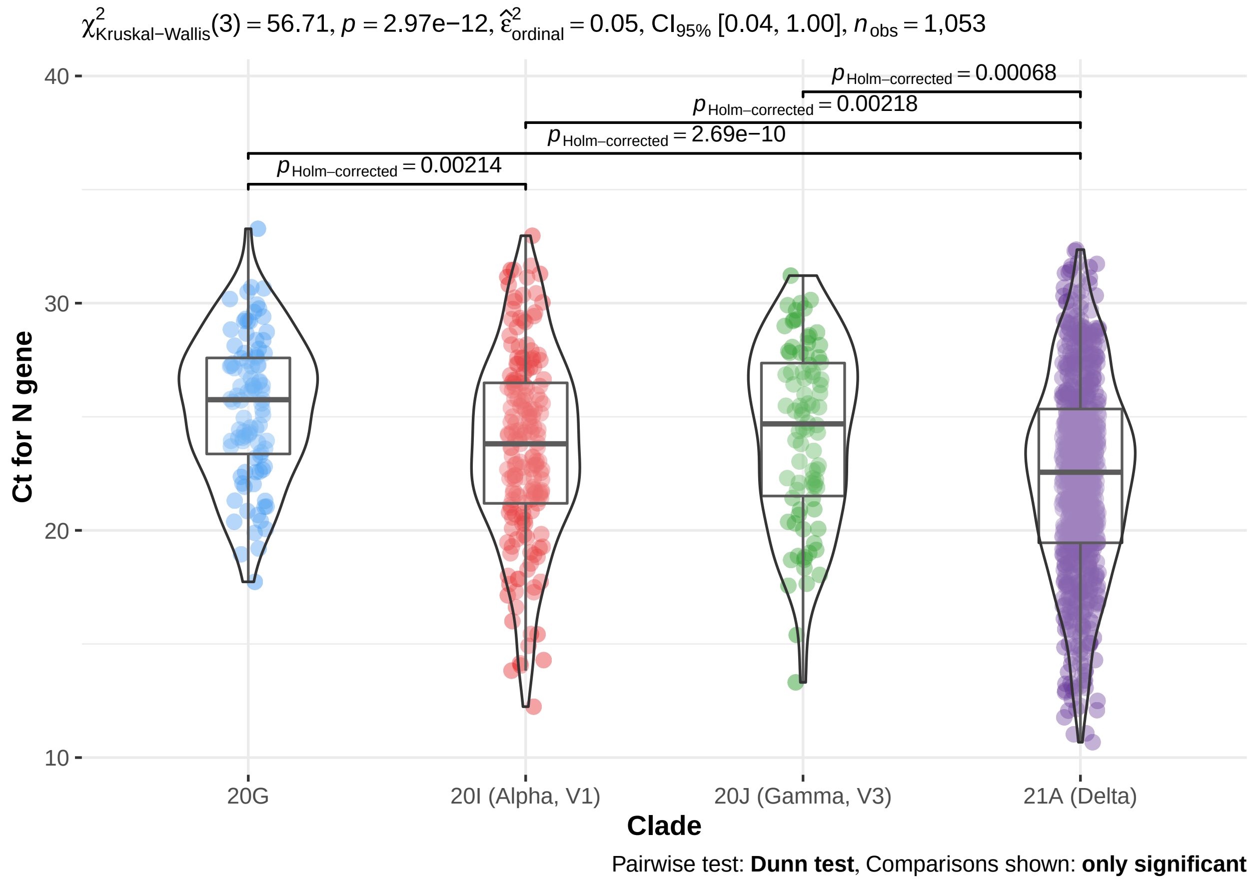
.y.	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
ct_N	20G	20I (Alpha, V1)	95	181	-3.462600	0.0005350	0.0021399	**
ct_N	20G	20J (Gamma, V3)	95	86	-1.911782	0.0559042	0.1118084	ns
ct_N	20G	21A (Delta)	95	691	-6.587392	0.0000000	0.0000000	****
ct_N	20I (Alpha, V1)	20J (Gamma, V3)	181	86	1.176868	0.2392480	0.2392480	ns
ct_N	20I (Alpha, V1)	21A (Delta)	181	691	-3.378820	0.0007280	0.0021839	**
ct_N	20J (Gamma, V3)	21A (Delta)	86	691	-3.815252	0.0001360	0.0006802	***

There are significant differences in Ct values of N gene between 21A (Delta) and other clades in the study. Additionally, there is a significant difference in Ct values of N gene between 20I (Alpha, V1) and 20G.

6.4.5 KW Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics_surveillance,
  x = clade,
  y = ct_N,
  type = "nonparametric",
  xlab = "Clade",
  ylab = "Ct for N gene",
  var.equal = FALSE,
  plot.type = "boxviolin",
  pairwise.comparisons = TRUE,
  pairwise.display = "significant",
  centrality.plotting = FALSE,
  bf.message = FALSE,
  p.adjust.method = "holm"
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



6.4.6 Welch's ANOVA

```
welch_anova_test(
  formula = ct_N ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
)
```

```
## # A tibble: 1 x 7
##   .y.       n statistic   DFn   DFd         p method
## * <chr> <int>     <dbl> <dbl> <dbl>     <dbl> <chr>
## 1 ct_N   1053      26.8     3  233.  6.19e-15 Welch ANOVA
```

6.4.7 Games-Howell Test

```
games_howell_test(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
```

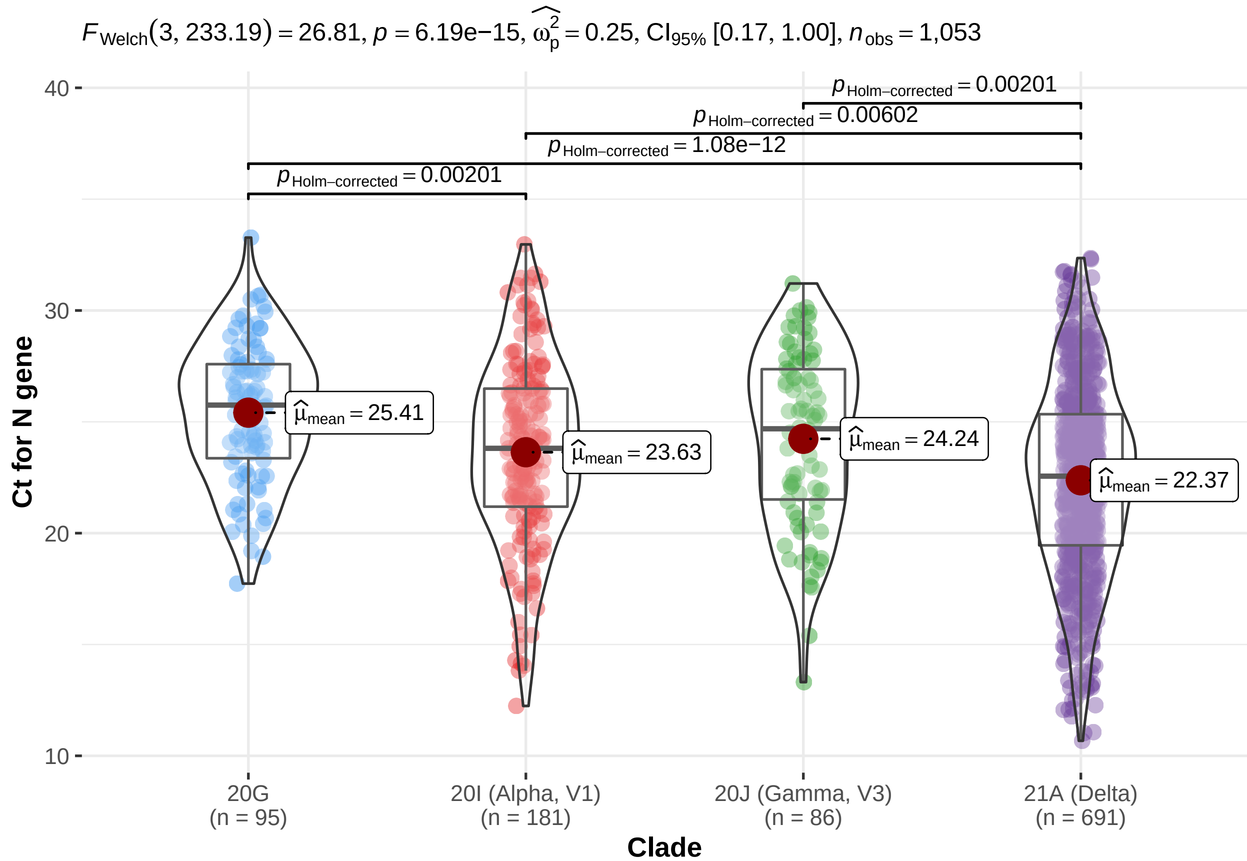
```
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

.y.	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
ct_N	20G	20I (Alpha, V1)	-1.7752235	-2.9246345	-0.6258126	0.000495	***
ct_N	20G	20J (Gamma, V3)	-1.1734858	-2.5446256	0.1976541	0.122000	ns
ct_N	20G	21A (Delta)	-3.0359345	-3.9716066	-2.1002624	0.000000	****
ct_N	20I (Alpha, V1)	20J (Gamma, V3)	0.6017378	-0.7486234	1.9520990	0.655000	ns
ct_N	20I (Alpha, V1)	21A (Delta)	-1.2607109	-2.1622121	-0.3592098	0.002000	**
ct_N	20J (Gamma, V3)	21A (Delta)	-1.8624487	-3.0384119	-0.6864856	0.000402	***

6.4.8 Welch's ANOVA Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics_surveillance,
  x = clade,
  y = ct_N,
  type = "parametric",
  xlab = "Clade",
  ylab = "Ct for N gene",
  var.equal = FALSE,
  plot.type = "boxviolin"
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



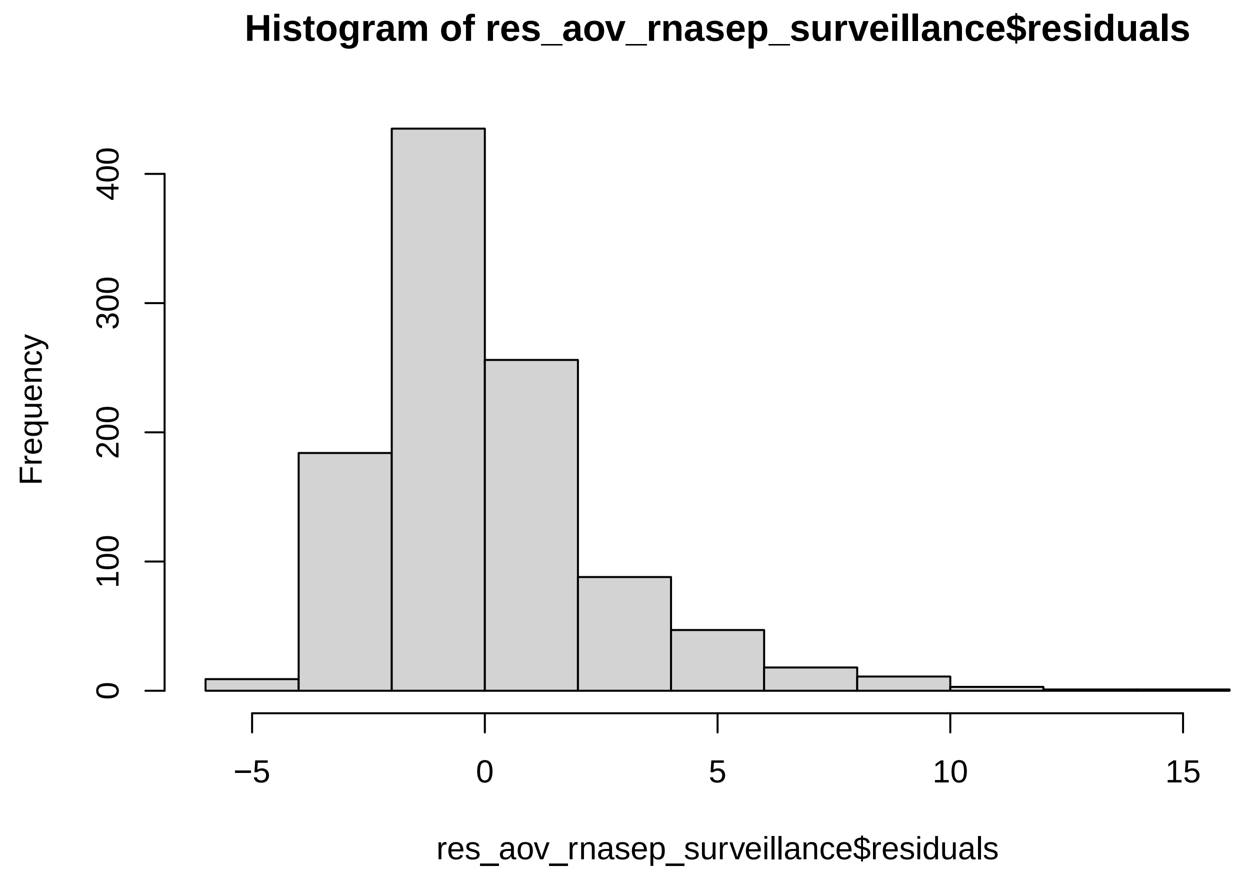
6.5 CONTROL : Compare Ct values for the Human RNase P gene

6.5.1 Assumptions

6.5.1.1 Normality :

```
res_aov_rnasep_surveillance <- aov(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
)

hist(res_aov_rnasep_surveillance$residuals)
```

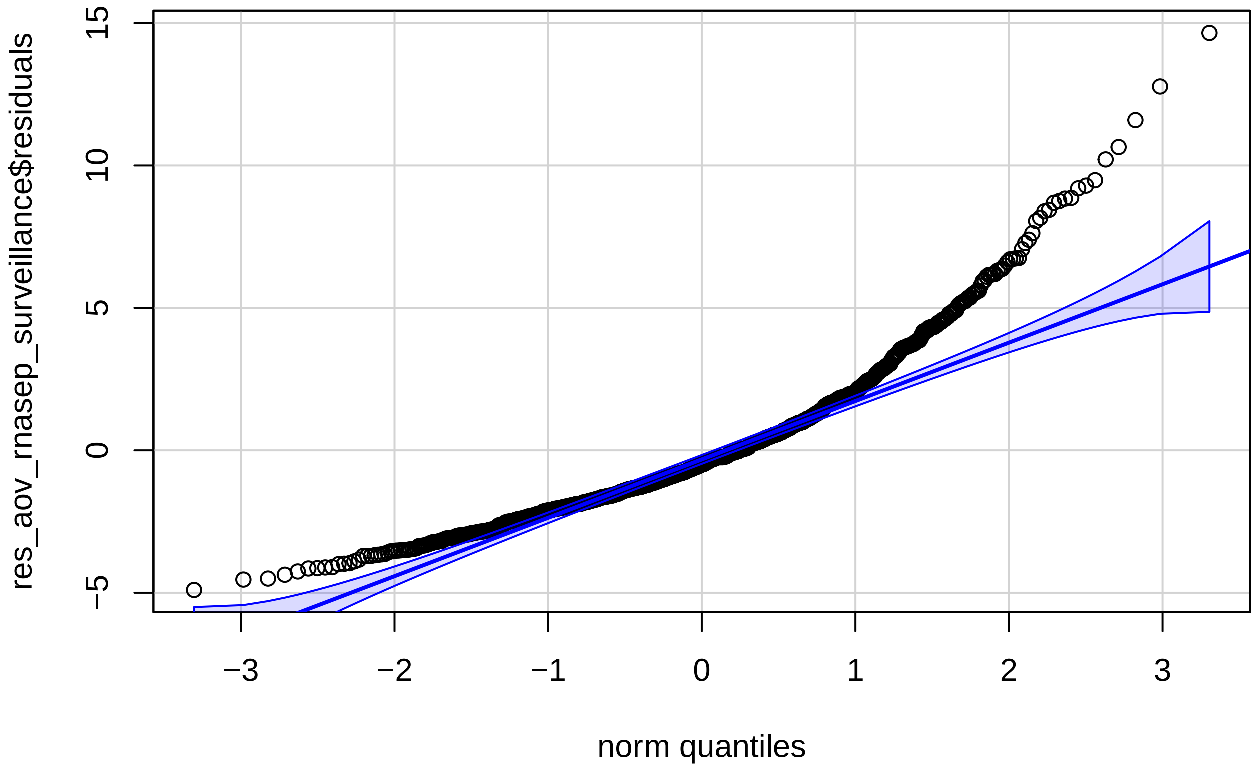
6.5.1.1.1 Histogram of Residuals



There is a strong right skew according to the histogram.

```
qqPlot(res_aov_rnasep_surveillance$residuals,
  id = FALSE
)
```

6.5.1.1.2 QQ-Plot of Residuals



6.5.1.1.3 Shapiro-Wilk

Null Hypothesis: Data comes from a normal distribution.

Alternate Hypothesis: Data does not come from a normal distribution.

```
shapiro.test(res_aov_rnasep_surveillance$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov_rnasep_surveillance$residuals
## W = 0.90703, p-value < 2.2e-16
```

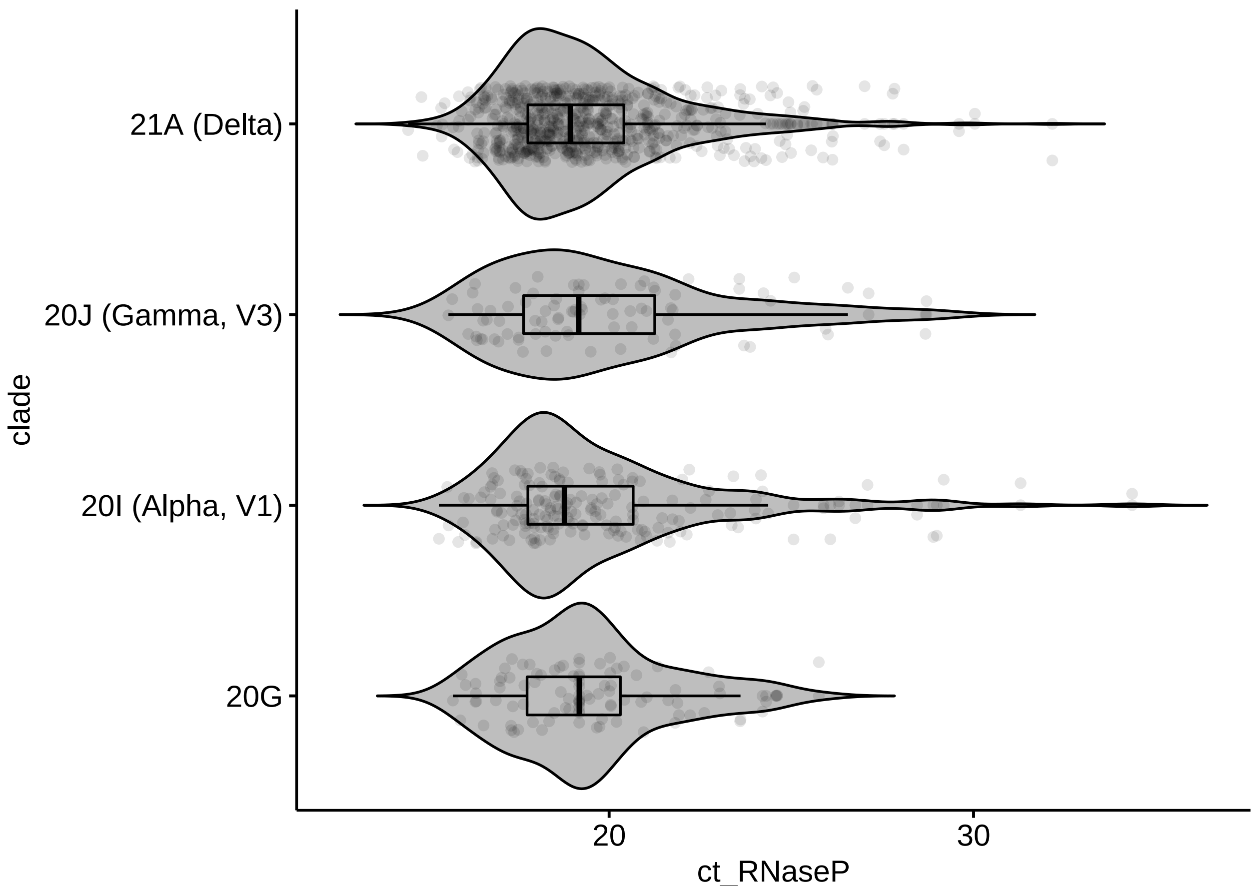
Since p-value < 0.05, we reject the null hypothesis. The data does not follow a normal distribution.

6.5.1.2 Equality of Variances

```
p <- limited_clade_collection_diagnostics_surveillance %>%
  ggviolin(
    x = "clade",
    y = "ct_RNaseP",
    fill = "grey",
```

```r
    add = c("jitter", "boxplot"),
    add.params = list(alpha = 0.1),
    notch = TRUE
  )

ggpar(p, orientation = "horiz")
```

6.5.1.2.1 Box Plot



6.5.1.2.2 Levene's Test

Null Hypothesis : Variances are equal.

Alternate Hypothesis : At least one variance is different.

```r
leveneTest(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value   Pr(>F)
## group    3  3.9001 0.008719 **
```

```
##       1049
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value $<$ 0.05, we reject the null hypothesis. Equality of Variances assumption is not met.

6.5.2 Kruskal-Wallis Test for Stochastic Dominance

Null Hypothesis :

$H_0 : P(Xi > Xj) = 0.5$ for all groups i and j from 1 to k

Alternate Hypothesis :

$H_A : P(Xi > Xj) \neq 0.5$ for at least one group $i \neq j$

From Non-normal distribution even with Kruskal-Wallis test.

```
kruskal.test(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics_surveillance
)
```

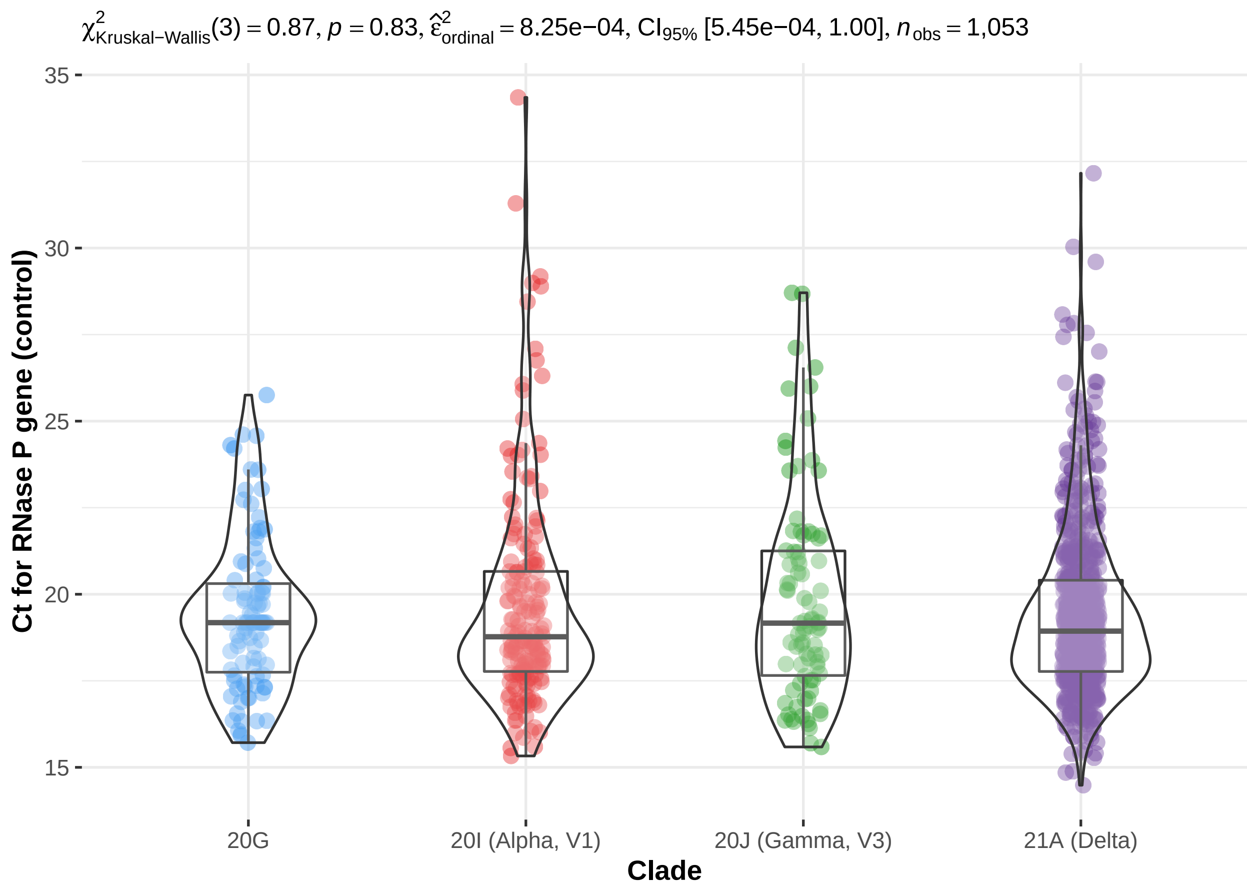
```
##
##  Kruskal-Wallis rank sum test
##
## data:  ct_RNaseP by clade
## Kruskal-Wallis chi-squared = 0.86757, df = 3, p-value = 0.8332
```

Since the p-value is $>$ 0.05, we cannot reject the null hypothesis. There is no significant difference in Ct values for RNase P gene between the different clades in this study.

6.5.3 KW Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics_surveillance,
  x = clade,
  y = ct_RNaseP,
  type = "nonparametric",
  xlab = "Clade",
  ylab = "Ct for RNase P gene (control)",
  var.equal = FALSE,
  plot.type = "boxviolin",
  pairwise.comparisons = FALSE,
  centrality.plotting = FALSE,
  bf.message = FALSE,
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



6.6 Conclusions

We analysed the results from 1053 SARS-CoV-2 infected patients from the COVID19 surveillance testing program in the university area (median age 21 years [<1 year to 91 years], 553 Male : 500 Female).

The clade composition among the sequenced samples were as follows:

```
20G            : 95
20I (Alpha, V1):181
20J (Gamma, V3): 86
21A (Delta)    :691
```

Statistical analyses were performed in an R environment (Kruskal-Wallis test followed by Dunn's test of multiple comparisons). There were significant differences in the Ct values for N gene between 21A (Delta) [median: 22.560, range: 10.675-32.355] and other clades [20G : 25.752 (17.735-33.274), 20I (Alpha, V1) : 23.810 (12.240-32.968), 20J (Gamma, V3): 24.688 (13.310-31.212)]. Additionally, there was a significant difference in Ct values for N gene between 20I (Alpha, V1) and 20G.

Concurrently, the Ct values for the RNase P control did not exhibit a statistically significant difference among these clades [20G : 19.180 (15.714-25.753), 20I (Alpha, V1) : 18.772 (15.330-34.350), 20J (Gamma, V3): 19.167 (15.590-28.708), 21A (Delta) : 18.935 (14.485-32.160)].

These results suggests significant differences in the viral load of 21A (Delta) relative to the other clades.

7 `symptomatic` Samples Only

In the following steps, we are limiting our study to samples in the `symptomatic` category. We filter the `clades_and_collection` for such samples and name the output table as `clades_and_collection_symptomatic`.

```
clades_and_collection_symptomatic <- clades_and_collection %>%
  filter(order_priority == "SYMPTOMATIC")

glimpse(clades_and_collection_symptomatic)
```

```
## Rows: 192
## Columns: 10
## $ patient_id       <chr> "6e7e2e2183a368ea0ae832df", "7ca4cbe296aea284c226cd16~
## $ testkit_id       <chr> "117M18DCD4E4B5AECE", "117M18DBD6617396JO", "117M18DB~
## $ collection_date  <date> 2021-01-13, 2021-01-13, 2021-01-13, 2021-01-13, 2021~
## $ clade            <fct> "20G", "20G", "20G", "20B", "20G", "20G", "20C", "20G~
## $ population       <fct> UNIVERSITY, COMMUNITY, COMMUNITY, COMMUNITY, COMMUNIT~
## $ order_priority   <fct> SYMPTOMATIC, SYMPTOMATIC, SYMPTOMATIC, SYMPTOMATIC, S~
## $ gender           <fct> F, F, M, F, F, F, M, M, F, M, M, M, F, F, M, M, M, M,~
## $ pregnancy_status <fct> NO, NO, NA, NO, NO, NO, NA, NA, NO, NA, NA, NA, NO, N~
## $ pipeline         <fct> nf-core/viralrecon, nf-core/viralrecon, nf-core/viral~
## $ rymedi_result    <fct> POSITIVE, POSITIVE, POSITIVE, POSITIVE, POSITIVE, POS~
```

The summary of the `clades_and_collection_symptomatic` table is as follows:

```
summary(clades_and_collection_symptomatic)
```

```
##   patient_id         testkit_id        collection_date
##  Length:192         Length:192         Min.   :2021-01-13
##  Class :character   Class :character   1st Qu.:2021-03-10
##  Mode  :character   Mode  :character   Median :2021-04-20
##                                        Mean   :2021-05-18
##                                        3rd Qu.:2021-08-10
##                                        Max.   :2021-11-09
##
##              clade         population       order_priority gender
##  21A (Delta)    :72   UNIVERSITY:  6   SURVEILLANCE:  0    M:98
##  20I (Alpha, V1):58   ATHLETICS :  1   SYMPTOMATIC :192    F:94
##  20G            :39   COMMUNITY :185   EXPOSED     :  0
##  20A            : 7   TRICOUNTY :  0   ONE DAY     :  0
##  20B            : 5
##  21F (Iota)     : 4
##  (Other)        : 7
##  pregnancy_status               pipeline    rymedi_result
##  YES : 2          nf-core/viralrecon:192   POSITIVE:192
##  NO  :92
##  NA's:98
##
##
##
##
```

7.1 Bar Plots with Monthly Count for each Clade : `symptomatic` Samples Only

The clades present in `clades_and_collection_symptomatic` when only `symptomatic` samples are considered are as follows:

```
clades_and_collection_symptomatic %>%
  pull(clade) %>%
  unique()
```

```
##  [1] 20G              20B              20C              20A
##  [5] 21C (Epsilon)    20I (Alpha, V1) 21F (Iota)       20J (Gamma, V3)
##  [9] 21D (Eta)        21A (Delta)
## 14 Levels: 20I (Alpha, V1) 20G 21A (Delta) 21F (Iota) 20C 21C (Epsilon) ... 21D (Eta)
```

```
(clades_factor_level <- sort(as.character(clades_factor_level)))
```

```
##  [1] "19A"             "19B"             "20A"             "20B"
##  [5] "20C"             "20G"             "20H (Beta, V2)"  "20I (Alpha, V1)"
##  [9] "20J (Gamma, V3)" "21A (Delta)"     "21B (Kappa)"     "21C (Epsilon)"
## [13] "21D (Eta)"       "21F (Iota)"
```

From `clades_and_collection_symptomatic`, we tabulate the count of each clade among sequenced samples collected in each month of 2021.

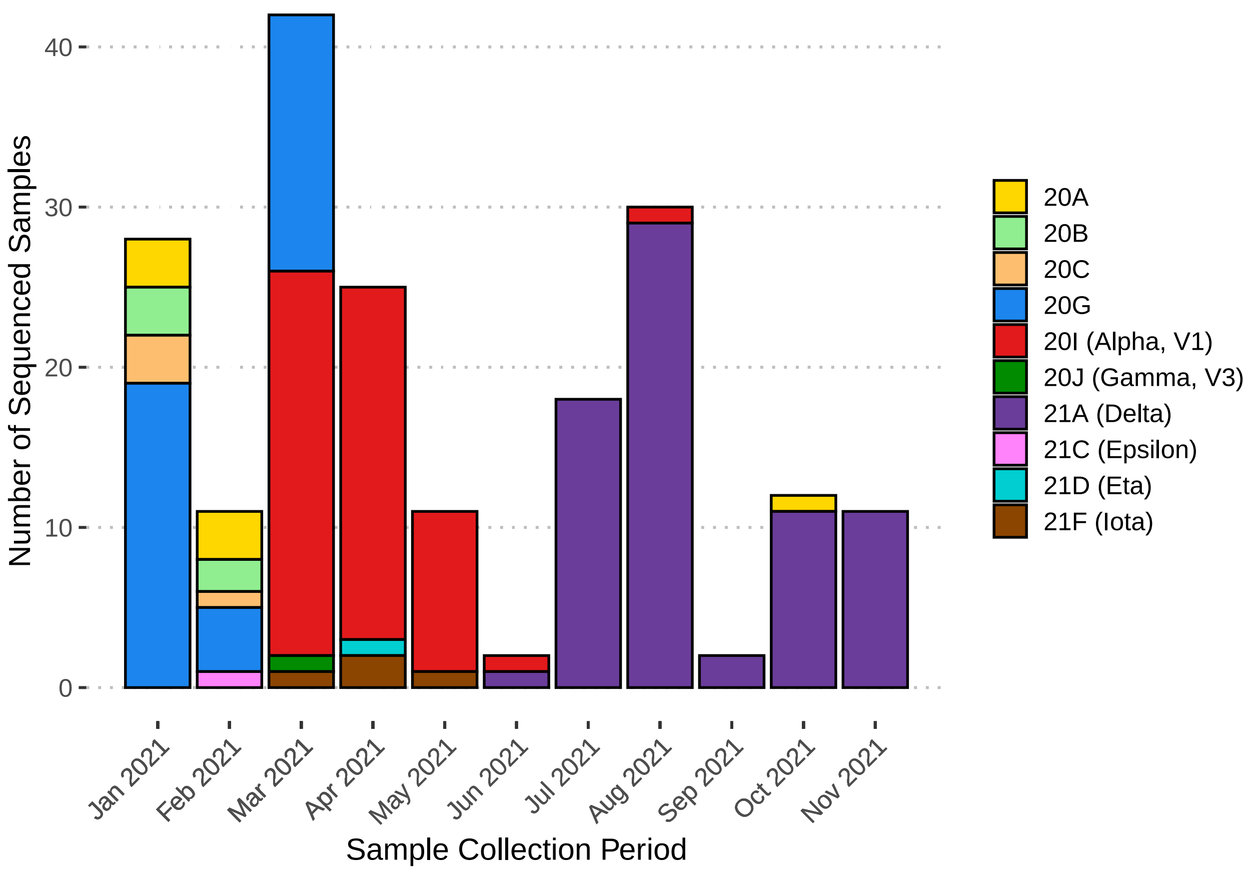
```
monthly_clade_date_symptomatic <- clades_and_collection_symptomatic %>%
  mutate(
    collection_period = as.yearmon(collection_date),
    clade = factor(clade,
      levels = clades_factor_level
    )
  ) %>%
  group_by(collection_period, clade) %>%
  summarize(count = n())

glimpse(monthly_clade_date_symptomatic)
```

```
## Rows: 27
## Columns: 3
## Groups: collection_period [11]
## $ collection_period <yearmon> Jan 2021, Jan 2021, Jan 2021, Jan 2021, Feb 2021~
## $ clade             <fct> "20A", "20B", "20C", "20G", "20A", "20B", "20C", "20~
## $ count             <int> 3, 3, 3, 19, 3, 2, 1, 4, 1, 16, 24, 1, 1, 22, 1, 2, ~
```

Visualizing the table above using a stacked bar plot.

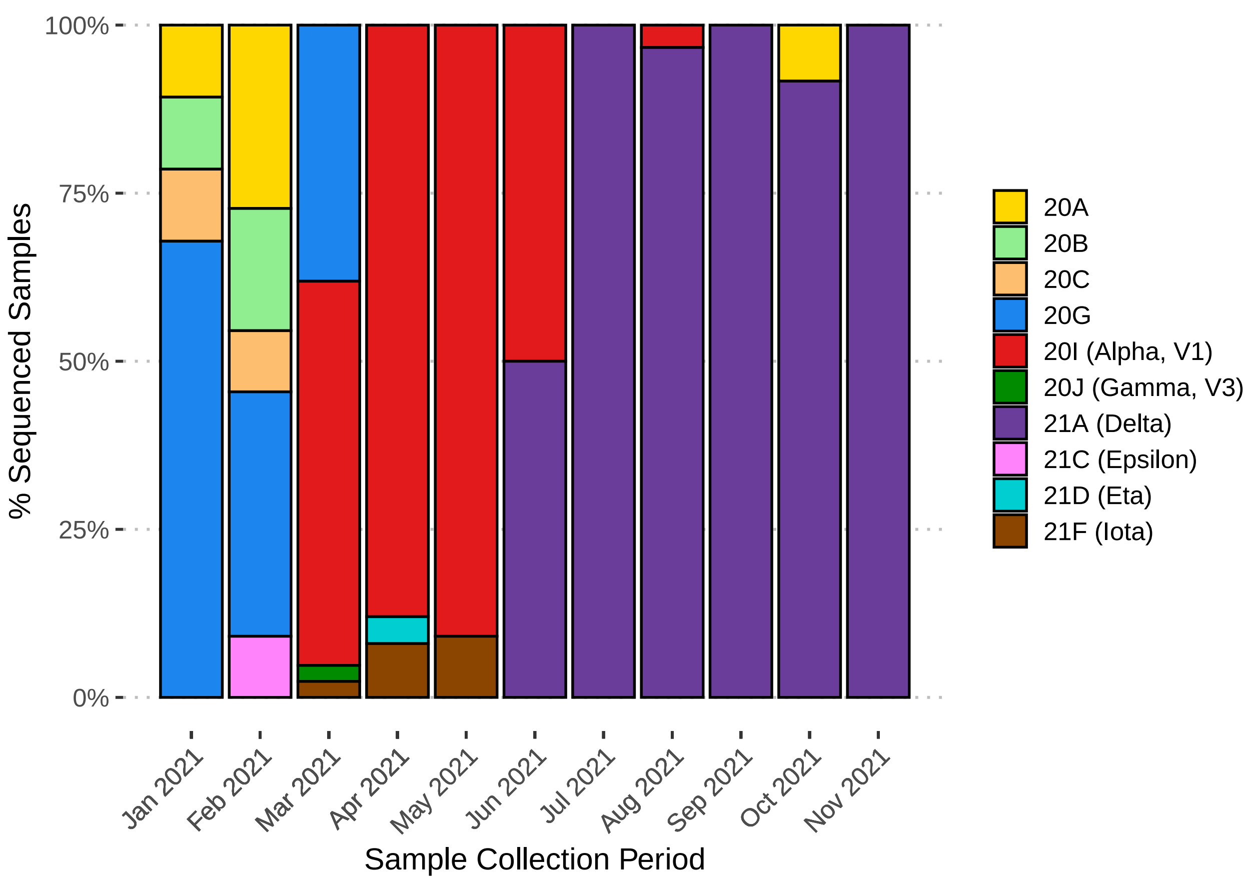
```
ggplot(
  monthly_clade_date_symptomatic,
  aes(
    x = collection_period,
    y = count,
    fill = clade
```

```
    )
  ) +
  geom_bar(colour = "black", position = "stack", stat = "identity") +
  scale_x_yearmon(breaks = monthly_clade_date$collection_period) +
  scale_fill_manual(values = color_tbl %>%
    filter(clade %in% monthly_clade_date_symptomatic$clade) %>%
    pull(color)) +
  labs(
    y = "Number of Sequenced Samples",
    x = "Sample Collection Period"
  ) +
  theme_pubclean() +
  theme(
    legend.position = "right",
    legend.title = element_blank(),
    legend.key.size = unit(0.5, "cm"),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)
  )
```




```
ggplot(
  monthly_clade_date_symptomatic,
  aes(
    x = collection_period,
    y = count,
```

```
    fill = clade
  )
) +
  geom_col(colour = "black", position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  scale_x_yearmon(breaks = monthly_clade_date$collection_period) +
  scale_fill_manual(values = color_tbl %>%
    filter(clade %in% monthly_clade_date_symptomatic$clade) %>%
    pull(color)) +
  labs(
    y = "% Sequenced Samples",
    x = "Sample Collection Period"
  ) +
  theme_pubclean() +
  theme(
    legend.position = "right",
    legend.title = element_blank(),
    legend.key.size = unit(0.5, "cm"),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)
  )
```



7.2 Preparing `diagnostics` data for `symptomatic`-only set of samples

Our goal is to determine whether 21A (Delta) shows lower Ct values for N gene when compared with the clades 20I (Alpha, V1), 20G, and 20J (Gamma, V3) *when only the `symptomatic` samples are considered*. In order to know that, we use the Ct values for N gene from the REDDI lab dataset. Since two replicates were done for each qRT-PCR reaction, we will compute the mean Ct for N gene and RNase P control for each `testkit_id`

```
diagnostics_data_symptomatic <- diagnostics_table %>%
  filter(testkit_id %in% clades_and_collection_symptomatic$testkit_id) %>%
  select(testkit_id, ct_rnasep_rep1, ct_rnasep_rep2, ct_N_rep1, ct_N_rep2) %>%
  arrange(testkit_id) %>%
  mutate(
    average_ct_rnasep = rowMeans(.[, c("ct_rnasep_rep1", "ct_rnasep_rep2")], na.rm = TRUE),
    average_ct_N = rowMeans(.[, c("ct_N_rep1", "ct_N_rep2")], na.rm = TRUE)
  ) %>%
  select(-c(
    ct_rnasep_rep1, ct_rnasep_rep2,
    ct_N_rep1, ct_N_rep2
  )) %>%
  group_by(testkit_id) %>%
  summarise(
    ct_RNaseP = mean(average_ct_rnasep, na.rm = TRUE),
    ct_N = mean(average_ct_N, na.rm = TRUE)
  )

diagnostics_data_symptomatic %>%
  filter(!complete.cases(.)) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

testkit_id	ct_RNaseP	ct_N
117M18DCB7CE53ED5Q	NaN	22.70748
117M18DCB7CE7001AZ	NaN	22.73960
117M18DCB7CE7F0BO1	NaN	17.95333
117M18DCB7CE81040Q	NaN	24.27213
117M18E355CFD320PV	NaN	29.74500

5 `testkit_id`s do not have a ct value for RNase P. We will use imputation to deal with the missing values.

7.3 Join `clade_assignments` and `diagnostics_data`

We join `clades_and_collection_symptomatic` and the diagnostics data to get the `clade_collection_diagnostics_sympto` table.

```r
clade_collection_diagnostics_symptomatic <- clades_and_collection_symptomatic %>%
  inner_join(diagnostics_data_symptomatic, by = "testkit_id")

glimpse(clade_collection_diagnostics_symptomatic)
```

```
## Rows: 190
## Columns: 12
## $ patient_id       <chr> "6e7e2e2183a368ea0ae832df", "7ca4cbe296aea284c226cd16~
## $ testkit_id       <chr> "117M18DCD4E4B5AECE", "117M18DBD6617396JO", "117M18DB~
## $ collection_date  <date> 2021-01-13, 2021-01-13, 2021-01-13, 2021-01-13, 2021~
## $ clade            <fct> "20G", "20G", "20G", "20B", "20G", "20G", "20C", "20G~
## $ population       <fct> UNIVERSITY, COMMUNITY, COMMUNITY, COMMUNITY, COMMUNIT~
## $ order_priority   <fct> SYMPTOMATIC, SYMPTOMATIC, SYMPTOMATIC, SYMPTOMATIC, S~
## $ gender           <fct> F, F, M, F, F, F, M, M, F, M, M, M, F, F, M, M, M, M,~
## $ pregnancy_status <fct> NO, NO, NA, NO, NO, NO, NA, NA, NO, NA, NA, NA, NO, N~
## $ pipeline         <fct> nf-core/viralrecon, nf-core/viralrecon, nf-core/viral~
## $ rymedi_result    <fct> POSITIVE, POSITIVE, POSITIVE, POSITIVE, POSITIVE, POS~
## $ ct_RNaseP        <dbl> 21.27448, 21.00248, 17.12404, 19.25705, 18.04183, 22.~
## $ ct_N             <dbl> 19.55871, 26.42359, 17.39554, 28.29983, 22.63608, 25.~
```

Since many samples had missing Ct values for RNase P, we impute median value of `ct_RNaseP` for each clade to the missing values.

```r
get_median_ct_symptomatic <- function(input_clade) {
  median_RNAseP <- clade_collection_diagnostics_symptomatic %>%
    filter(clade == input_clade) %>%
    pull(ct_RNaseP) %>%
    median(na.rm = TRUE)

  return(median_RNAseP)
}
```

Median Ct RNase P - 20I (Alpha, V1)

```r
# 20I (Alpha, V1)
(median_RNAseP_20I_symptomatic <- get_median_ct_symptomatic("20I (Alpha, V1)"))
```

```
## [1] 18.555
```

Median Ct RNase P - 21A (Delta)

```r
# 21A (Delta)
(median_RNAseP_21A_symptomatic <- get_median_ct_symptomatic("21A (Delta)"))
```

```
## [1] 18.54
```

Median Ct RNase P - 20G

```r
# 20G
(median_RNAseP_20G_symptomatic <- get_median_ct_symptomatic("20G"))
```

```
## [1] 19.32
```

Median Ct RNase P - 20J (Gamma, V3)

```r
# 20J (Gamma, V3)
(median_RNAseP_20J_symptomatic <- get_median_ct_symptomatic("20J (Gamma, V3)"))
```

```
## [1] 23.03
```

Summary of the `clade_collection_diagnostics` table after imputation

```r
clade_collection_diagnostics_symptomatic <- clade_collection_diagnostics_symptomatic %>%
  mutate(
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20I (Alpha, V1)")),
      median_RNAseP_20I_symptomatic
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "21A (Delta)")),
      median_RNAseP_21A_symptomatic
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20G")),
      median_RNAseP_20G_symptomatic
    ),
    ct_RNaseP = replace(
      ct_RNaseP,
      (is.na(ct_RNaseP) & (clade == "20J (Gamma, V3)")),
      median_RNAseP_20J_symptomatic
    )
  )

summary(clade_collection_diagnostics_symptomatic)
```

```
##   patient_id         testkit_id        collection_date
##  Length:190         Length:190         Min.   :2021-01-13
##  Class :character   Class :character   1st Qu.:2021-03-10
##  Mode  :character   Mode  :character   Median :2021-04-20
##                                        Mean   :2021-05-17
##                                        3rd Qu.:2021-08-09
##                                        Max.   :2021-11-09
##
##              clade         population       order_priority gender
##  21A (Delta)    :70   UNIVERSITY:  6   SURVEILLANCE:  0    M:97
##  20I (Alpha, V1):58   ATHLETICS :  1   SYMPTOMATIC :190    F:93
##  20G            :39   COMMUNITY :183   EXPOSED     :  0
##  20A            : 7   TRICOUNTY :  0   ONE DAY     :  0
##  20B            : 5
##  21F (Iota)     : 4
```

```
##   (Other)        : 7
##   pregnancy_status               pipeline    rymedi_result   ct_RNaseP
##   YES : 2          nf-core/viralrecon:190   POSITIVE:190   Min.   :14.12
##   NO  :91                                                  1st Qu.:17.51
##   NA's:97                                                  Median :18.67
##                                                            Mean   :19.34
##                                                            3rd Qu.:20.47
##                                                            Max.   :33.28
##                                                            NA's   :3
##        ct_N
##   Min.   : 7.98
##   1st Qu.:20.25
##   Median :23.11
##   Mean   :22.96
##   3rd Qu.:26.15
##   Max.   :32.51
##
```

The counts of different clades in the `clade_collection_diagnostics` table are as follows:

```
clade_collection_diagnostics_symptomatic %>%
  group_by(clade) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

clade	count
21A (Delta)	70
20I (Alpha, V1)	58
20G	39
20A	7
20B	5
21F (Iota)	4
20C	4
21C (Epsilon)	1
20J (Gamma, V3)	1
21D (Eta)	1

As in the previous analysis, in order to not affect our assumption of phylogenetic independence, we are only going to compare Ct values for variants at terminal nodes of NextClade tree. We extract data from `clade_collection_diagnostics_symptomatic` table and create a new table `limited_clade_collection_diagnostics_symptomatic` with only data for the following clades:

```
21A (Delta)
20I (Alpha, V1)
20G
```

```r
limited_clade_collection_diagnostics_symptomatic <- clade_collection_diagnostics_symptomatic %>%
  filter(clade %in% c(
    "21A (Delta)",
    "20I (Alpha, V1)",
    "20G"
  )) %>%
  mutate(clade = factor(clade,
    levels = c(
      "20G",
      "20I (Alpha, V1)",
      "21A (Delta)"
    )
  )) %>%
  select(testkit_id, clade, ct_RNaseP, ct_N) %>%
  ungroup()

glimpse(limited_clade_collection_diagnostics_symptomatic)
```

```
## Rows: 167
## Columns: 4
## $ testkit_id <chr> "117M18DCD4E4B5AECE", "117M18DBD6617396J0", "117M18DBD3F633~
## $ clade      <fct> "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20G", "20~
## $ ct_RNaseP  <dbl> 21.27448, 21.00248, 17.12404, 18.04183, 22.68423, 21.34041,~
## $ ct_N       <dbl> 19.55871, 26.42359, 17.39554, 22.63608, 25.70162, 15.63790,~
```

Summary of the `limited_clade_collection_diagnostics_symptomatic` table :

```r
summary(limited_clade_collection_diagnostics_symptomatic)
```

```
##   testkit_id                      clade      ct_RNaseP          ct_N
##  Length:167         20G             :39   Min.   :14.69   Min.   : 7.98
##  Class :character   20I (Alpha, V1):58   1st Qu.:17.38   1st Qu.:20.34
##  Mode  :character   21A (Delta)     :70   Median :18.61   Median :23.05
##                                            Mean   :19.26   Mean   :22.84
##                                            3rd Qu.:20.03   3rd Qu.:26.09
##                                            Max.   :33.28   Max.   :32.51
```

Median and range of patient age whose samples are in the `limited_clade_collection_diagnostics_symptomatic` table as follows:

```r
sample_collection_table %>%
  mutate(collection_date = year(as_datetime(collection_date))) %>%
  filter(testkit_id %in% limited_clade_collection_diagnostics_symptomatic$testkit_id) %>%
  left_join(demographics_table, by = "patient_id") %>%
  mutate(age_at_sample_collection = (collection_date - birth_year)) %>%
  select(testkit_id, patient_id, age_at_sample_collection) %>%
  summarize(
    median_age_at_sample_collection = median(age_at_sample_collection),
    lowest_age_at_sample_collection = min(age_at_sample_collection),
    highest_age_at_sample_collection = max(age_at_sample_collection)
  ) %>%
  kbl() %>%
```

```
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

median_age_at_sample_collection	lowest_age_at_sample_collection	highest_age_at_sample_collection
32	5	82

The counts of each gender in the `limited_clade_collection_diagnostics_symptomatic` table are as follows:

```
sample_collection_table %>%
  filter(testkit_id %in% limited_clade_collection_diagnostics_symptomatic$testkit_id) %>%
  group_by(gender) %>%
  summarize(count = n()) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped"
    )
  )
```

gender	count
F	81
M	86

The median and range of Ct values for each clade in the `limited_clade_collection_diagnostics_symptomatic` table are as follows:

```
limited_clade_collection_diagnostics_symptomatic %>%
  group_by(clade) %>%
  summarise(
    median_ct_N = round(median(ct_N), 3),
    min_ct_N = round(range(ct_N)[1], 3),
    max_ct_N = round(range(ct_N)[2], 3),
    median_ct_RNaseP = round(median(ct_RNaseP), 3),
    min_ct_RNaseP = round(range(ct_RNaseP)[1], 3),
    max_ct_RNaseP = round(range(ct_RNaseP)[2], 3)
  ) %>%
  kbl() %>%
  kable_classic_2(
    full_width = F,
    latex_options = c(
      "hold_position",
      "striped",
      "scale_down"
    )
  )
```

clade	median_ct_N	min_ct_N	max_ct_N	median_ct_RNaseP	min_ct_RNaseP	max_ct_RNaseP
20G	23.165	12.870	29.890	19.320	16.28	28.89
20I (Alpha, V1)	23.230	13.225	32.505	18.555	16.15	29.92
21A (Delta)	22.985	7.980	29.980	18.540	14.69	33.28

7.4 N gene

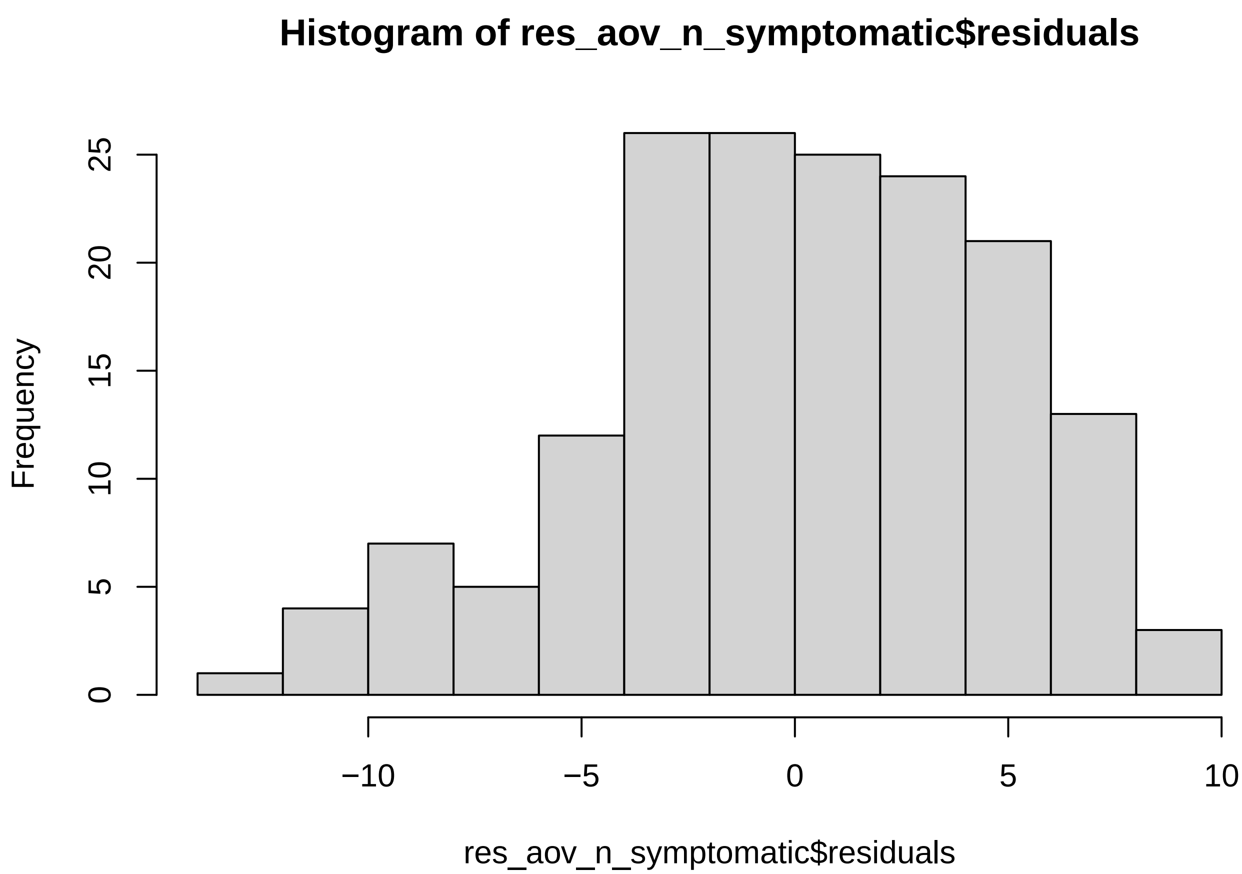
7.4.1 Assumptions

7.4.1.1 Independence No two samples came from the same `patient_id`. The samples are from Clemson University's COVID19 testing program and the order priority of these samples were labeled as `symptomatic`. The CU REDDI lab chose samples that were sent for sequencing.

7.4.1.2 Normality :

```
res_aov_n_symptomatic <- aov(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_symptomatic
)

hist(res_aov_n_symptomatic$residuals)
```

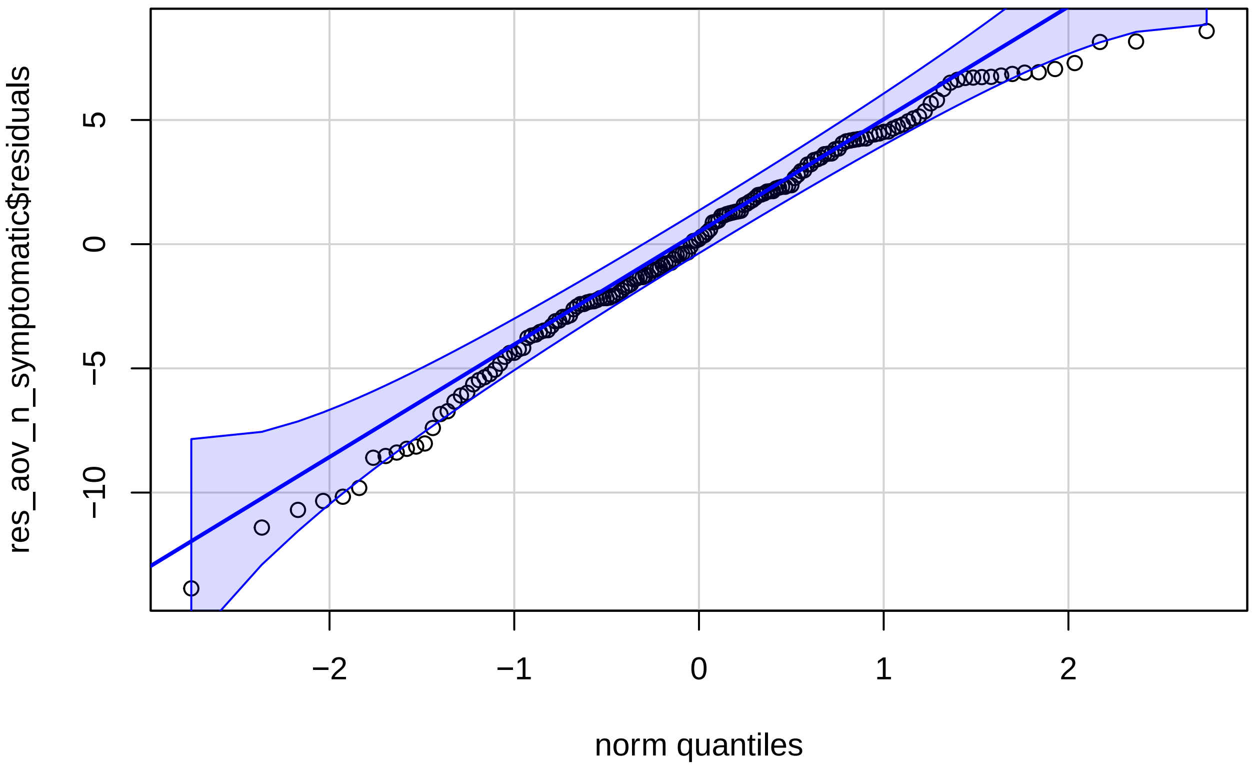
7.4.1.2.1 Histogram of Residuals



The histogram shows a slightly left skewed distribution.

```
qqPlot(res_aov_n_symptomatic$residuals,
  id = FALSE
)
```

7.4.1.2.2 QQ-Plot of Residuals



7.4.1.2.3 Shapiro-Wilk

Null Hypothesis: Data comes from a normal distribution.

Alternate Hypothesis: Data does not come from a normal distribution.

```
shapiro.test(res_aov_n_symptomatic$residuals)
```

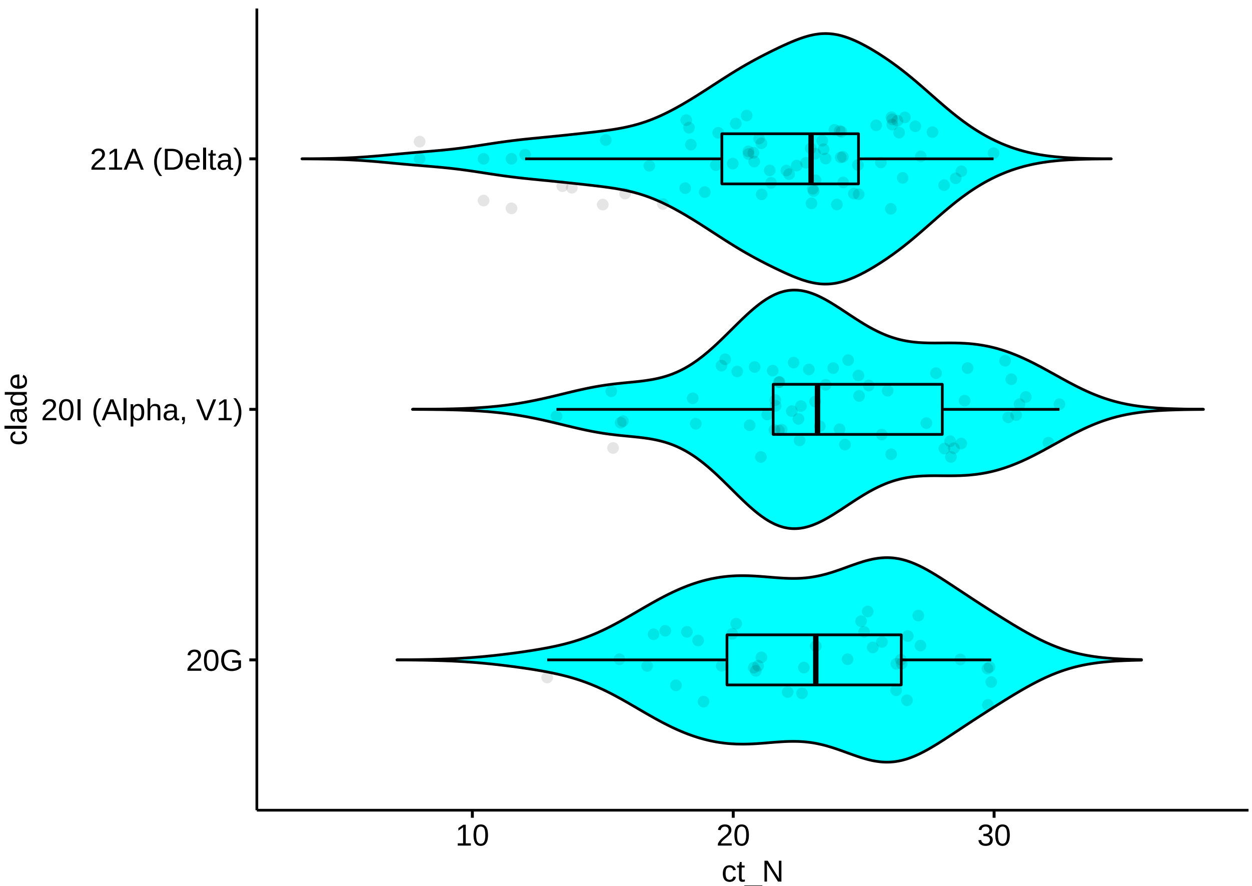
```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov_n_symptomatic$residuals
## W = 0.97964, p-value = 0.01476
```

Since p-value < 0.05, we reject the null hypothesis. The data not follow a normal distribution.

7.4.1.3 Equality of Variances

```
p <- limited_clade_collection_diagnostics_symptomatic %>%
  ggviolin(
    x = "clade",
    y = "ct_N",
    fill = "cyan",
    add = c("jitter", "boxplot"),
    add.params = list(alpha = 0.1),
    notch = TRUE
  )

ggpar(p, orientation = "horiz")
```

7.4.1.3.1 Box Plot



7.4.1.3.2 Levene's Test

Null Hypothesis : Variances are equal.

Alternate Hypothesis : At least one variance is different.

```
leveneTest(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_symptomatic
)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2   0.129 0.8791
##       164
```

Since p-value > 0.05, we cannot reject the null hypothesis. Equality of Variances assumption is met.

7.4.2 Kruskal-Wallis Test for Stochastic Dominance

Null Hypothesis :

$H_0 : P(Xi > Xj) = 0.5$ for all groups i and j from 1 to k

Alternate Hypothesis :

$H_A : P(Xi > Xj) \neq 0.5$ for at least one group $i \neq j$

From Non-normal distribution even with Kruskal-Wallis test.

```
kruskal.test(ct_N ~ clade,
  data = limited_clade_collection_diagnostics_symptomatic
)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ct_N by clade
## Kruskal-Wallis chi-squared = 4.4895, df = 2, p-value = 0.106
```

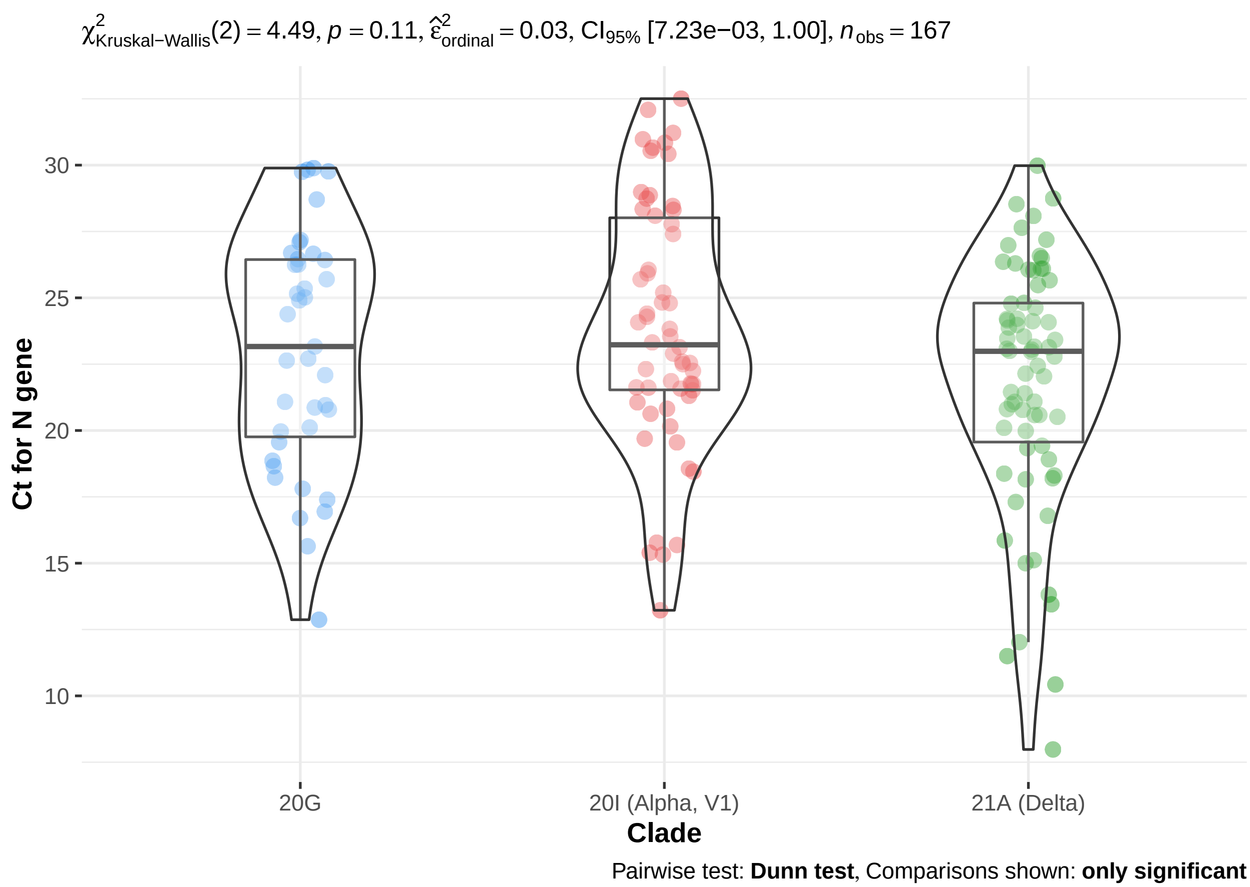
Since the p-value is > 0.05, we cannot reject the null hypothesis. There is no significant difference in Ct values for RNase P gene between the different clades in this study.

7.4.3 KW Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics_symptomatic,
  x = clade,
  y = ct_N,
  type = "nonparametric",
  xlab = "Clade",
  ylab = "Ct for N gene",
  var.equal = TRUE,
  plot.type = "boxviolin",
  pairwise.comparisons = TRUE,
  pairwise.display = "significant",
```

```
  centrality.plotting = FALSE,
  bf.message = FALSE,
  p.adjust.method = "holm"
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



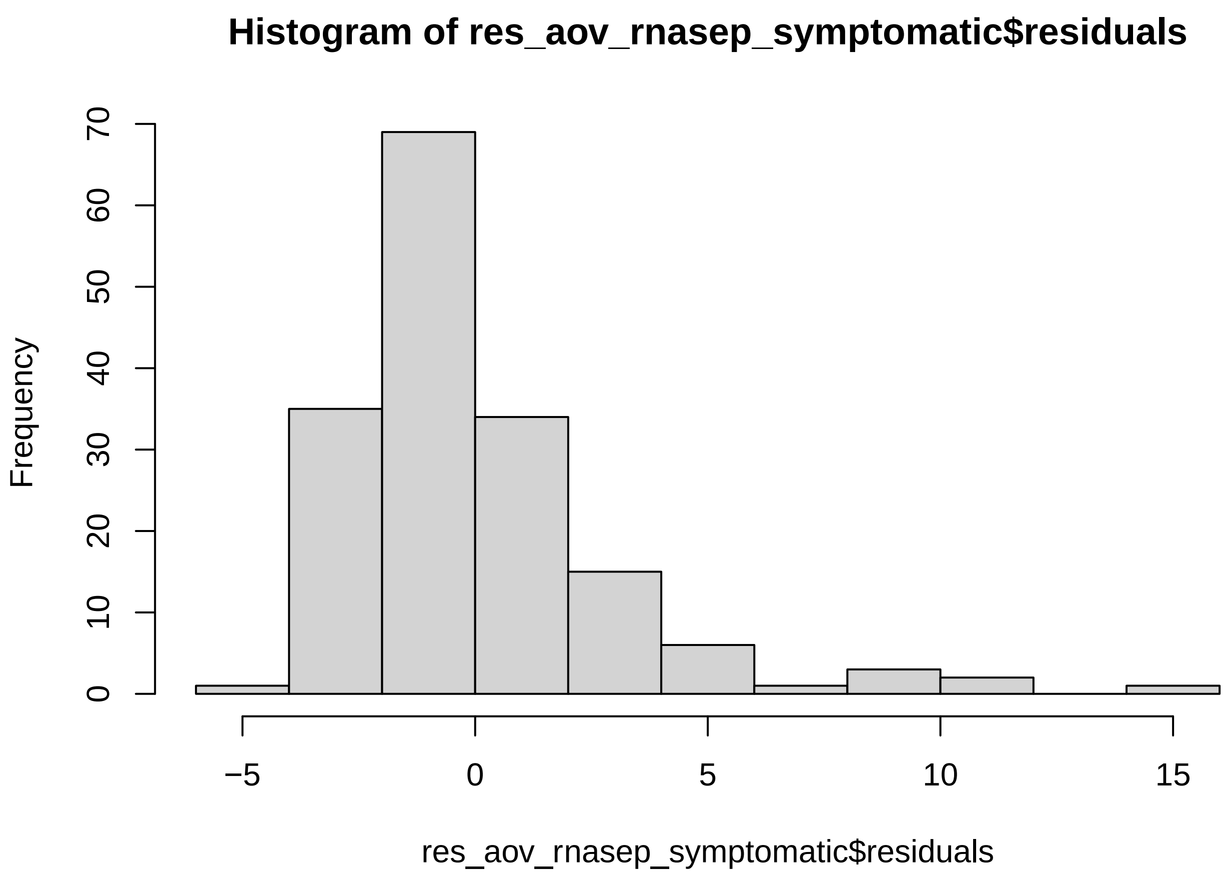
7.5 CONTROL : Compare Ct values for the Human RNase P gene

7.5.1 Assumptions

7.5.1.1 Normality :

```
res_aov_rnasep_symptomatic <- aov(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics_symptomatic
)

hist(res_aov_rnasep_symptomatic$residuals)
```

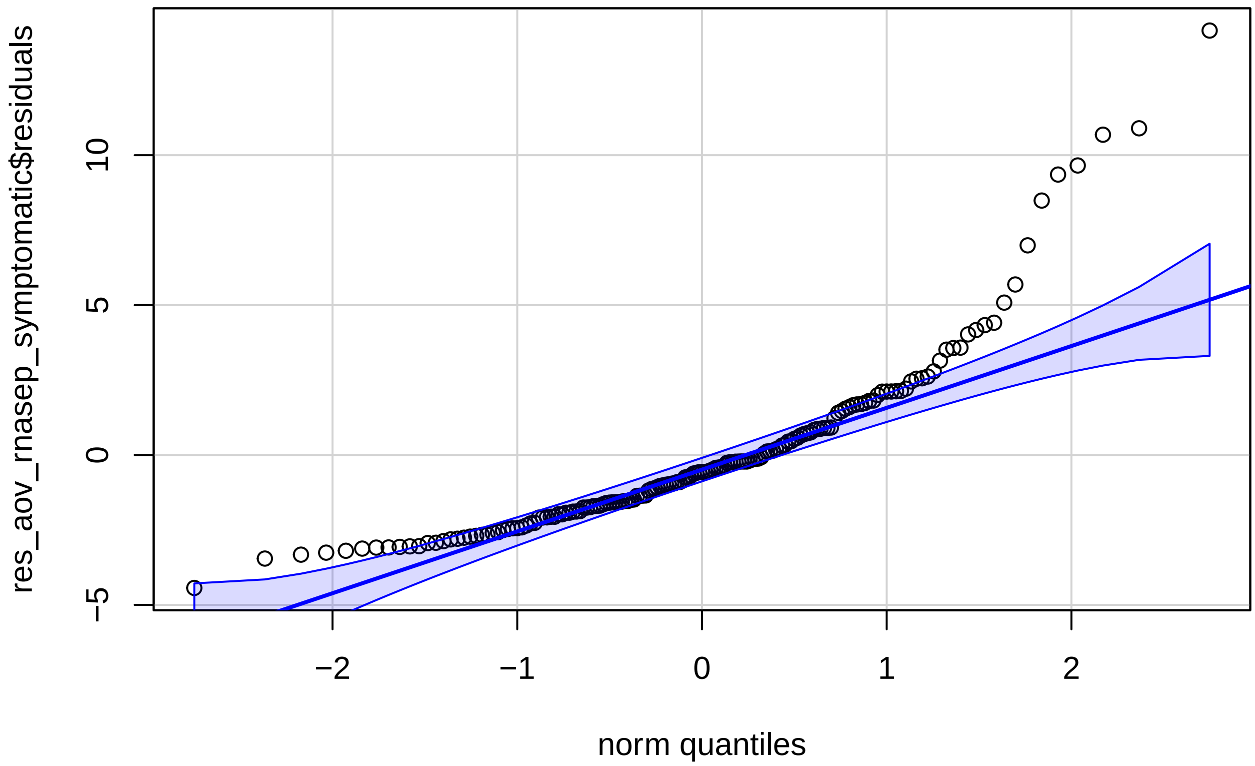
7.5.1.1.1 Histogram of Residuals



There is a strong right skew according to the histogram.

```
qqPlot(res_aov_rnasep_symptomatic$residuals,
  id = FALSE
)
```

7.5.1.1.2 QQ-Plot of Residuals



7.5.1.1.3 Shapiro-Wilk

Null Hypothesis: Data comes from a normal distribution.

Alternate Hypothesis: Data does not come from a normal distribution.

```
shapiro.test(res_aov_rnasep_symptomatic$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov_rnasep_symptomatic$residuals
## W = 0.82929, p-value = 1.076e-12
```

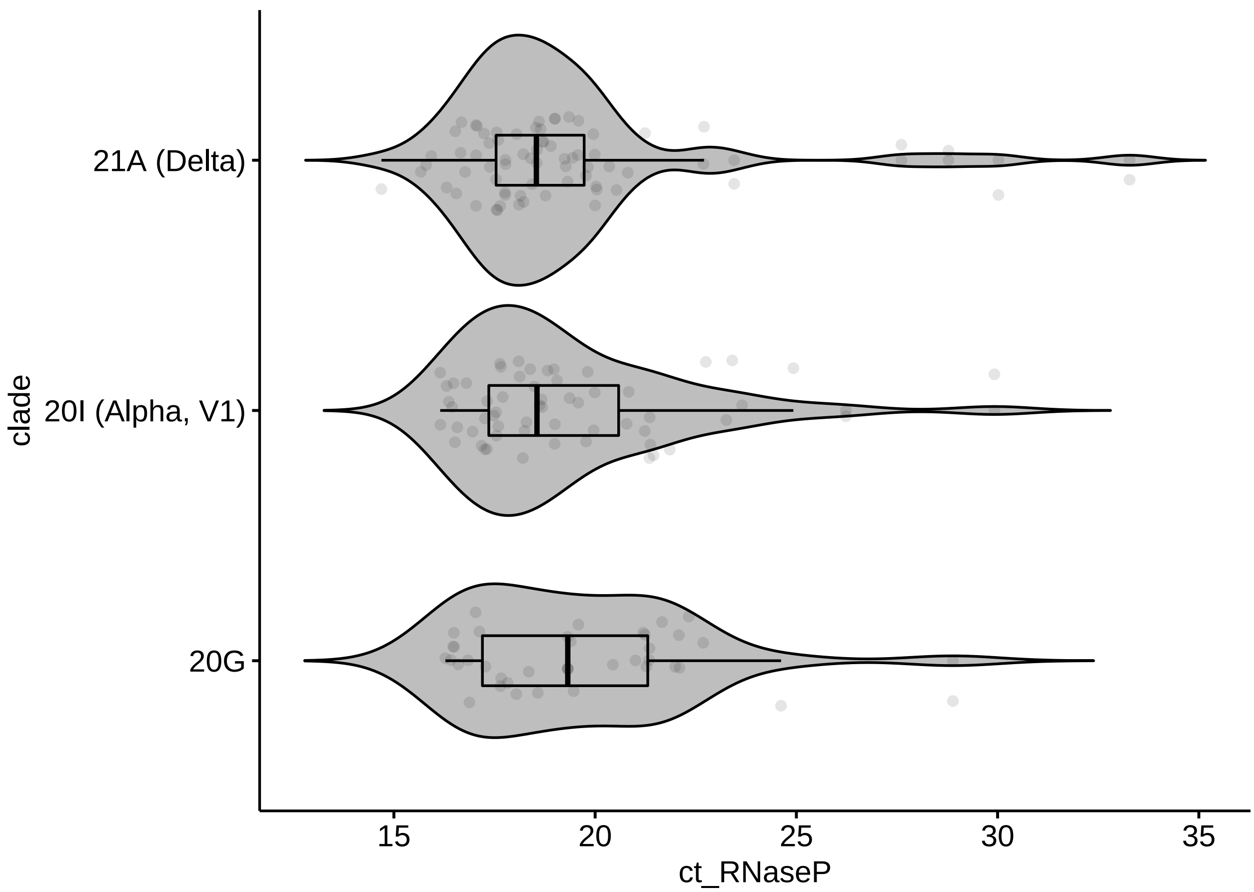
Since p-value < 0.05, we reject the null hypothesis. The data does not follow a normal distribution.

7.5.1.2 Equality of Variances

```
p <- limited_clade_collection_diagnostics_symptomatic %>%
  ggviolin(
    x = "clade",
    y = "ct_RNaseP",
    fill = "grey",
```

```r
    add = c("jitter", "boxplot"),
    add.params = list(alpha = 0.1),
    notch = TRUE
  )

ggpar(p, orientation = "horiz")
```

7.5.1.2.1 Box Plot



7.5.1.2.2 Levene's Test

Null Hypothesis : Variances are equal.

Alternate Hypothesis : At least one variance is different.

```r
leveneTest(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics_symptomatic
)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.1325  0.876
##       164
```

Since p-value > 0.05, we cannot reject the null hypothesis. Equality of Variances assumption is met.

7.5.2 Kruskal-Wallis Test for Stochastic Dominance

Null Hypothesis :

$H_0 : P(Xi > Xj) = 0.5$ for all groups i and j from 1 to k

Alternate Hypothesis :

$H_A : P(Xi > Xj) \neq 0.5$ for at least one group $i \neq j$

From Non-normal distribution even with Kruskal-Wallis test.

```
kruskal.test(ct_RNaseP ~ clade,
  data = limited_clade_collection_diagnostics_symptomatic
)
```

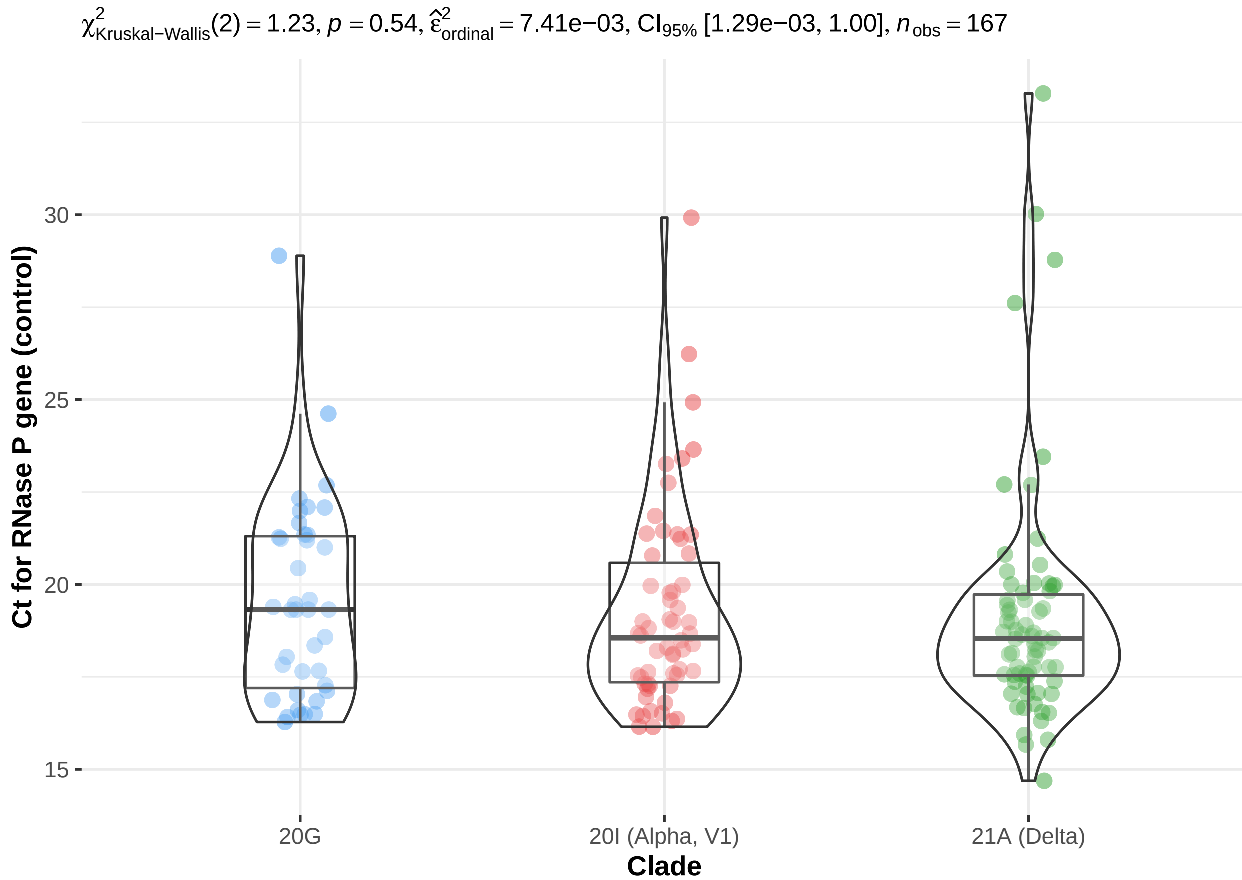
```
##
##  Kruskal-Wallis rank sum test
##
## data:  ct_RNaseP by clade
## Kruskal-Wallis chi-squared = 1.2302, df = 2, p-value = 0.5406
```

Since the p-value is > 0.05, we cannot reject the null hypothesis. There is no significant difference in Ct values for RNase P gene between the different clades in this study.

7.5.3 Visualization

```
ggbetweenstats(
  data = limited_clade_collection_diagnostics_symptomatic,
  x = clade,
  y = ct_RNaseP,
  type = "nonparametric",
  xlab = "Clade",
  ylab = "Ct for RNase P gene (control)",
  var.equal = TRUE,
  plot.type = "boxviolin",
  pairwise.comparisons = FALSE,
  centrality.plotting = FALSE,
  bf.message = FALSE,
  p.adjust.method = "holm"
) +
  scale_color_manual(values = c(
    "dodgerblue2",
    "#E31A1C",
    "green4",
    "#6A3D9A"
  ))
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



7.6 Conclusions

We analysed the results from 167 SARS-CoV-2 infected patients with `symptomatic` order priority status from the COVID19 testing program in the university area (median age 32 years [5 year to 82 years], 86 Male : 81 Female).

The clade composition among the sequenced samples were as follows:

```
20G             : 39
20I (Alpha, V1): 58
21A (Delta)     : 70
```

Statistical analyses were performed in an R environment (Kruskal-Wallis test). The Ct values for the N gene did not exhibit a statistically significant difference among these clades [20G : 23.165 (12.870-29.890), 20I (Alpha, V1) : 23.230 (13.225-32.505), 21A (Delta) : 22.985 (7.980-29.980)].

Concurrently, the Ct values for the RNase P control did not exhibit a statistically significant difference among these clades [20G : 19.320 (16.28-28.89), 20I (Alpha, V1) : 18.555 (16.15-29.92), 21A (Delta) : 18.540 (14.69-33.28)].

These results suggests that there is no difference in the viral load of 21A (Delta) relative to the other clades.