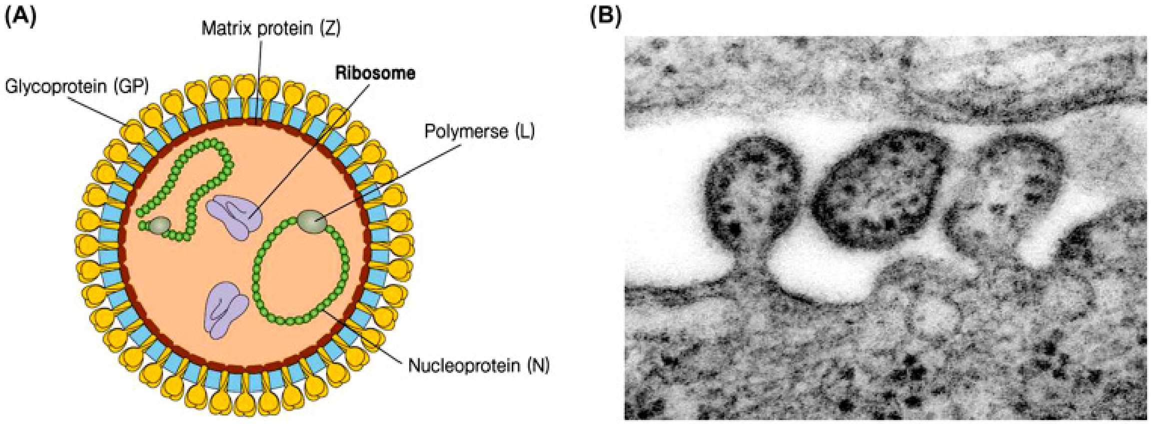
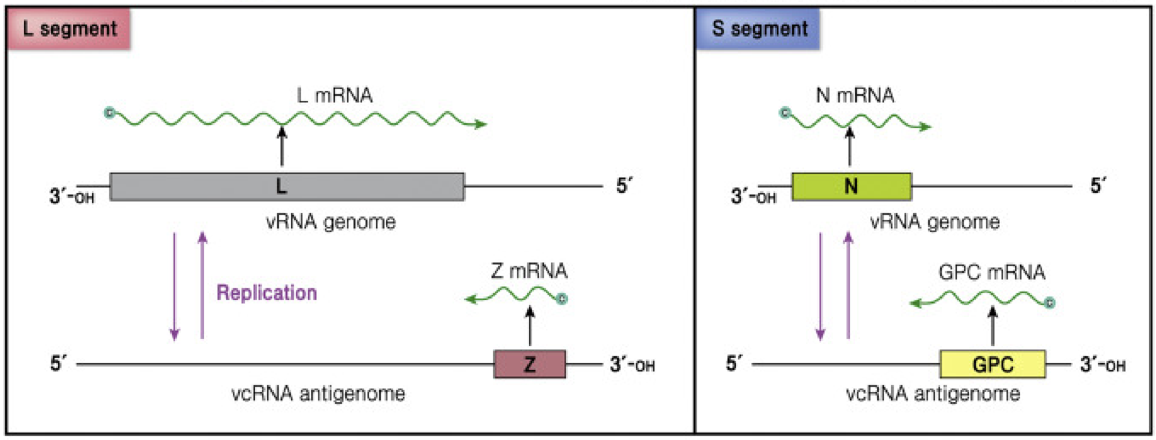
Background

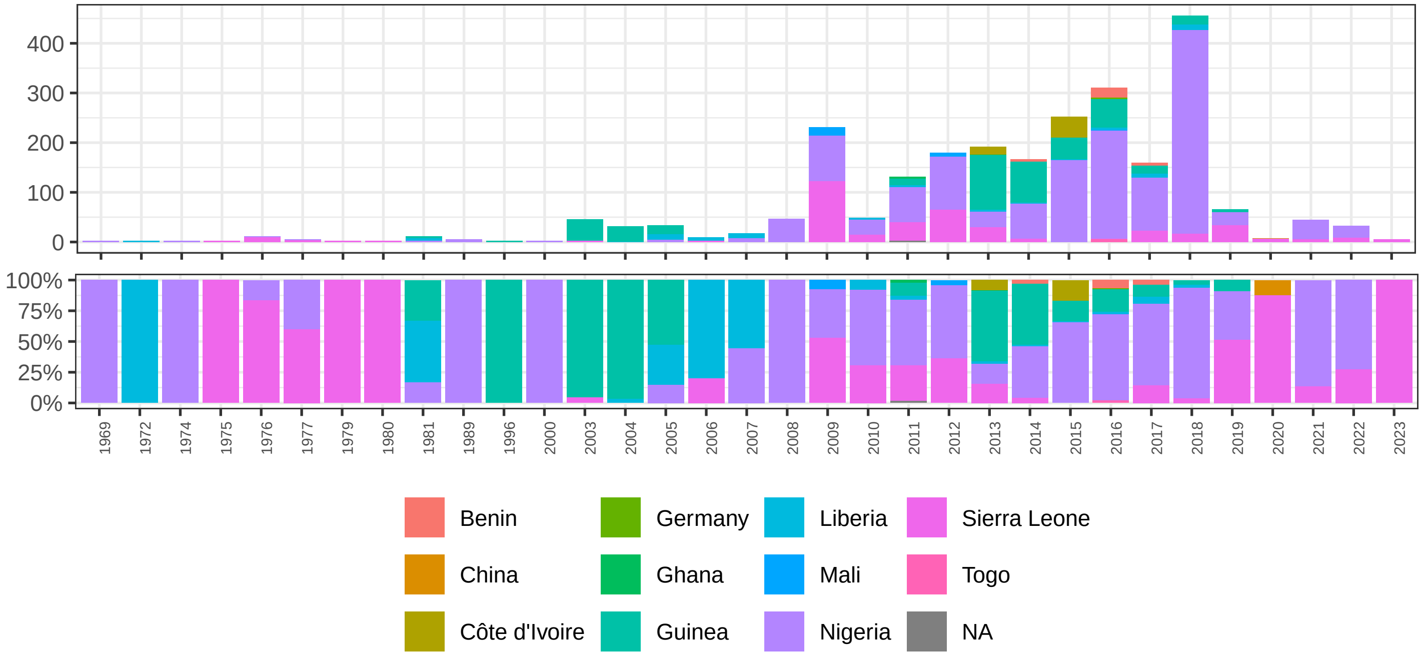
Lassa has two segments ”L” and ”S” from Chapter 16 of ”Molecular Virology of Human Pathogenic Viruses” by Wang-Shick Ryu



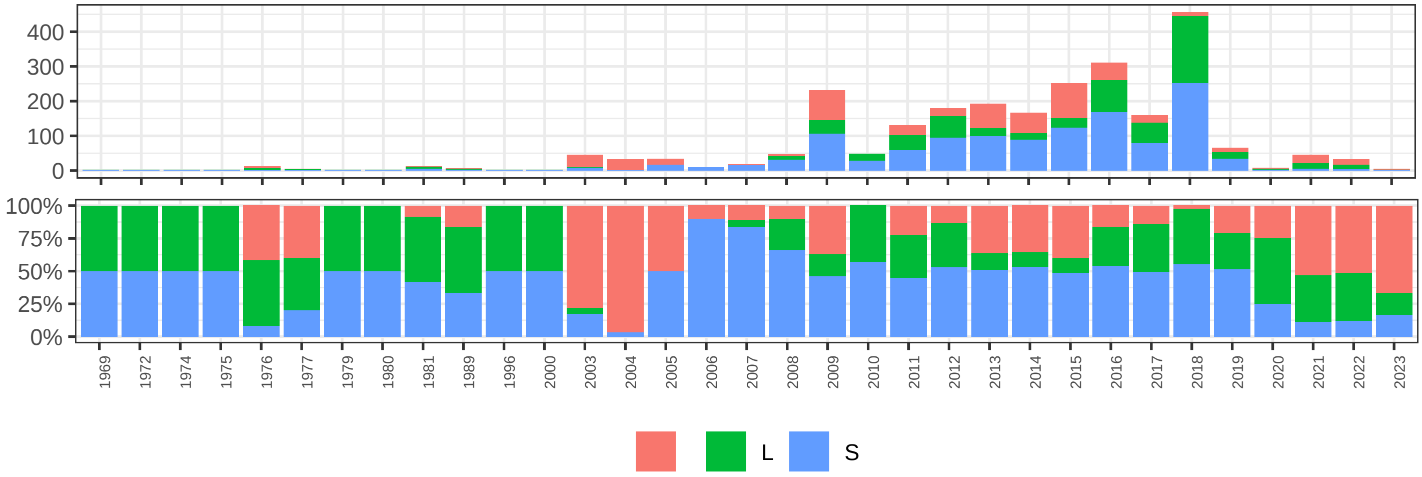


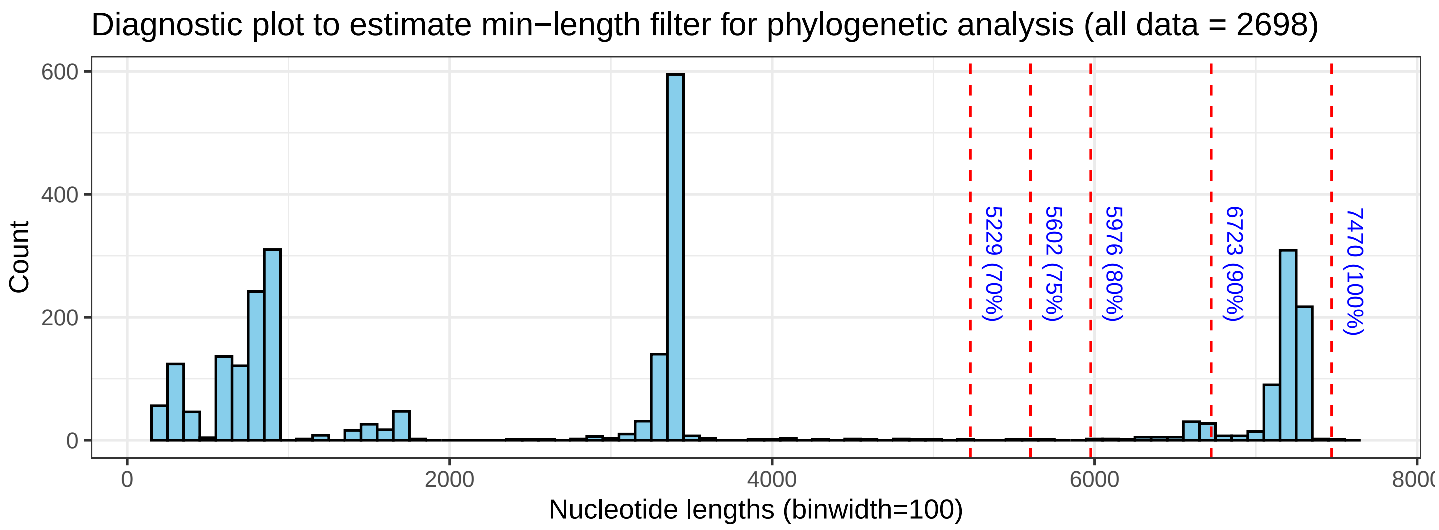
Exploratory Graphics

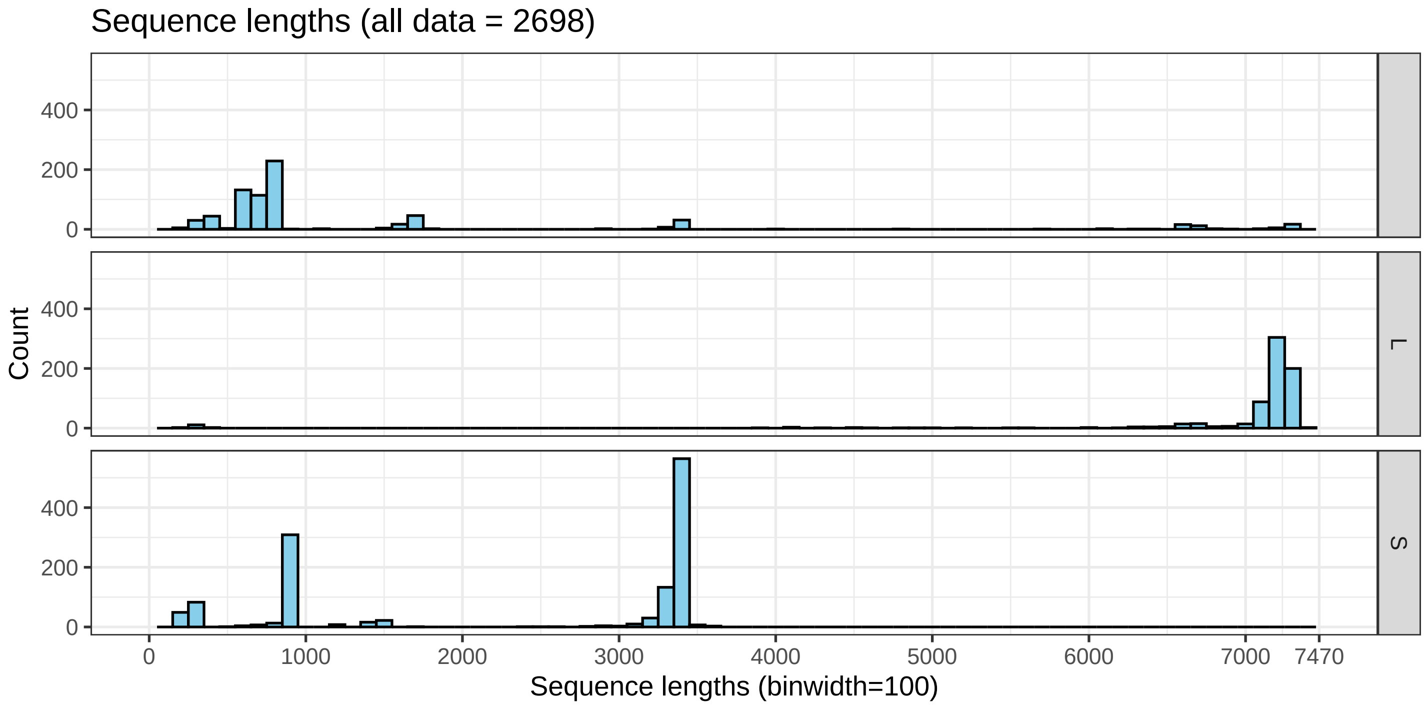
Frequency and proportion (n=2525)



Lassa entries with collection date (n=2525)







- Total lassa records = 2698.
- Total L lassa records = 694
- Total S lassa records = 1272
- Non L or S lassa records = 732

Lassa strain name

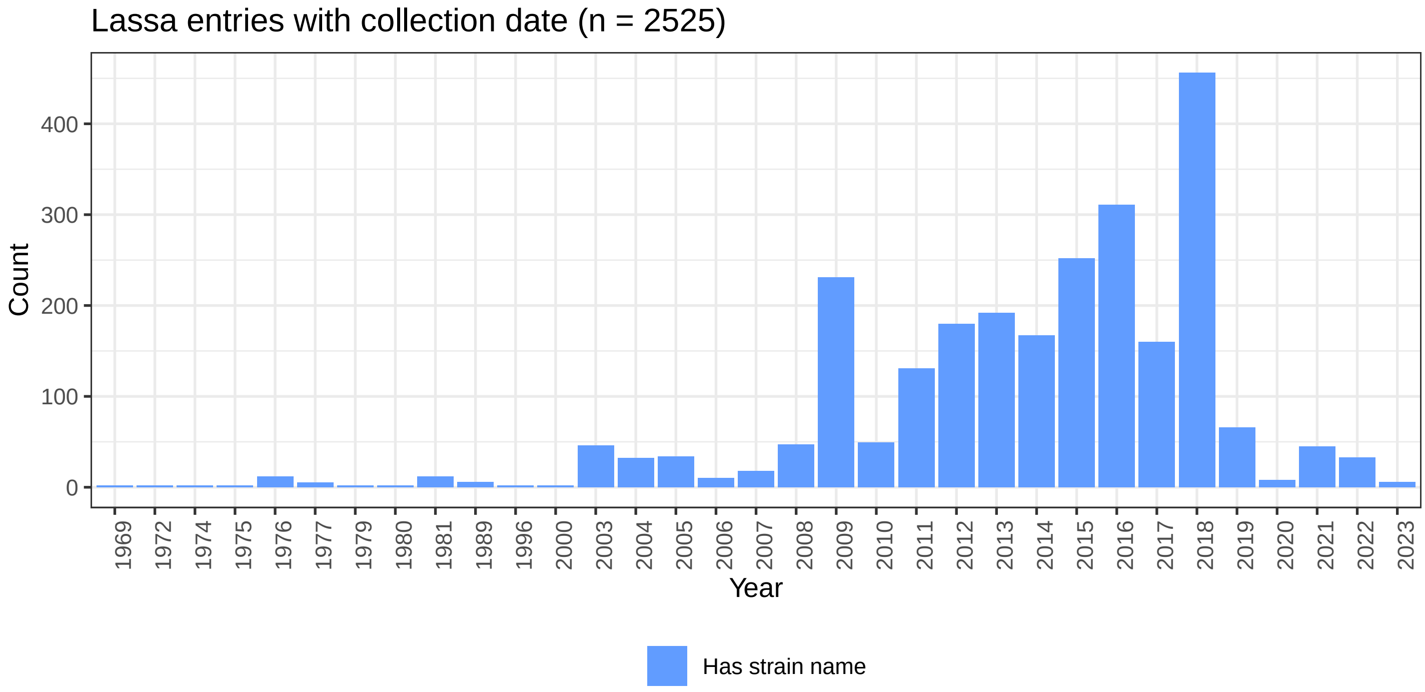
Is setting the strain name helpful for lassa? It looks like a majority of the strain names are the GenBank accession anyways.

```
nrow(data)

## [1] 2698

strainset=data %>%
  subset(strain != accession) %>%
  nrow(.)

cdata %>%
  ggplot(., aes(x=col_year, fill=strain != accession)) +
  geom_bar() +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle=90, vjust=1, hjust=1),
    legend.position = "bottom",
    legend.title = element_blank()
  ) +
  labs(
    title=paste("Lassa entries with collection date (n = ",nrow(cdata),")", sep=""),
    x="Year", y="Count"
  ) +
  scale_fill_manual(values = c("TRUE" = "#619CFF", "FALSE" = "#F8766D"),
                    labels = c("TRUE" = "Has strain name", "FALSE" = "Uses accession"))
```



- Total lassa records = 2698. Notice how this includes samples that lack a collection date.
- Total lassa records where strain does not equal accession = 2692
- Percentage with strain names 99.78%
- Pulling in ”strain=(.*)” filled in more samples (as compared to 49%)

Strain name duplicates

Are there more than 2 strain name duplicates? More than the S and L segments?

```r
library(gt)
count_names <- data %>%
  group_by(strain) %>%
  summarize(n=n(), average_length=mean(length)) %>%
  arrange(desc(n))

more_names <- count_names %>%
  filter(n>2)

# Print top 10
more_names %>%
  head(10) %>%
  gt() %>%
  tab_header(title = "Top 10 Strains Name Duplicates by Count") %>%
  fmt_number(columns = c(n, average_length), decimals = 0) %>%
  cols_label(
    strain = "Strain Name",
    n = "Count Duplicates",
    average_length = "Average Length"
  )
```

Top 10 Strains Name Duplicates by Count

Strain Name	Count Duplicates	Average Length
Josiah	15	3,878
AV	4	5,319
LASV_3523	4	2,508
LASV_3604	4	2,525
LASV_3609	4	2,500
LASV_3625	4	2,524
LASV_3629	4	2,510
LASV_3630	4	2,524
LASV_3706	4	2,516
LASV_3711	4	2,506

- Percentage with strain names 99.78%
- Number of strain names that have 1 sequence record: 576
- Number of strain names that have 2 sequence records: 943
- Number of strain names that have more than 2 sequence records: 70

Table 2: Top 25 most frequent sequence submitters with their region and countries

authors	n	regions	countries
andersen et al.	361	Africa	Sierra Leone, Liberia, Nigeria
siddle et al.	254	Africa	Nigeria
kafetzopoulou et al.	214	Africa	Nigeria
ehichioya et al.	193	Africa	Nigeria, Sierra Leone, Liberia, Guinea
marien et al.	179	Africa	Guinea
adesina et al.	170	Africa	Nigeria
odia et al.	156	Africa	Nigeria
olayemi et al.	143	Africa	Nigeria, Guinea
fichet-calvet et al.	132	Africa	Guinea
leski et al.	106	Africa	Sierra Leone
bangura et al.	68	Africa	Sierra Leone, Guinea
happi et al.	63	Africa	Nigeria
bowen et al.	57	NA	NA
welch et al.	56	Africa	Liberia, Nigeria, Sierra Leone, Guinea
escalera-zamudio et al.	52	Africa	Côte d'Ivoire
oloniniyi et al.	40	Africa	Nigeria
sandi et al.	40	Africa	Sierra Leone
ghersi et al.	36	Africa	Sierra Leone
asogun et al.	35	Africa	Nigeria
lecompte et al.	33	Africa	Guinea
olschlager et al.	32	Africa	Liberia, Guinea
yadouleton et al.	31	Africa	Benin
safronetz et al.	22	Africa	Mali
omilabu et al.	20	Africa	Nigeria, Benin
karan et al.	18	Africa	Guinea