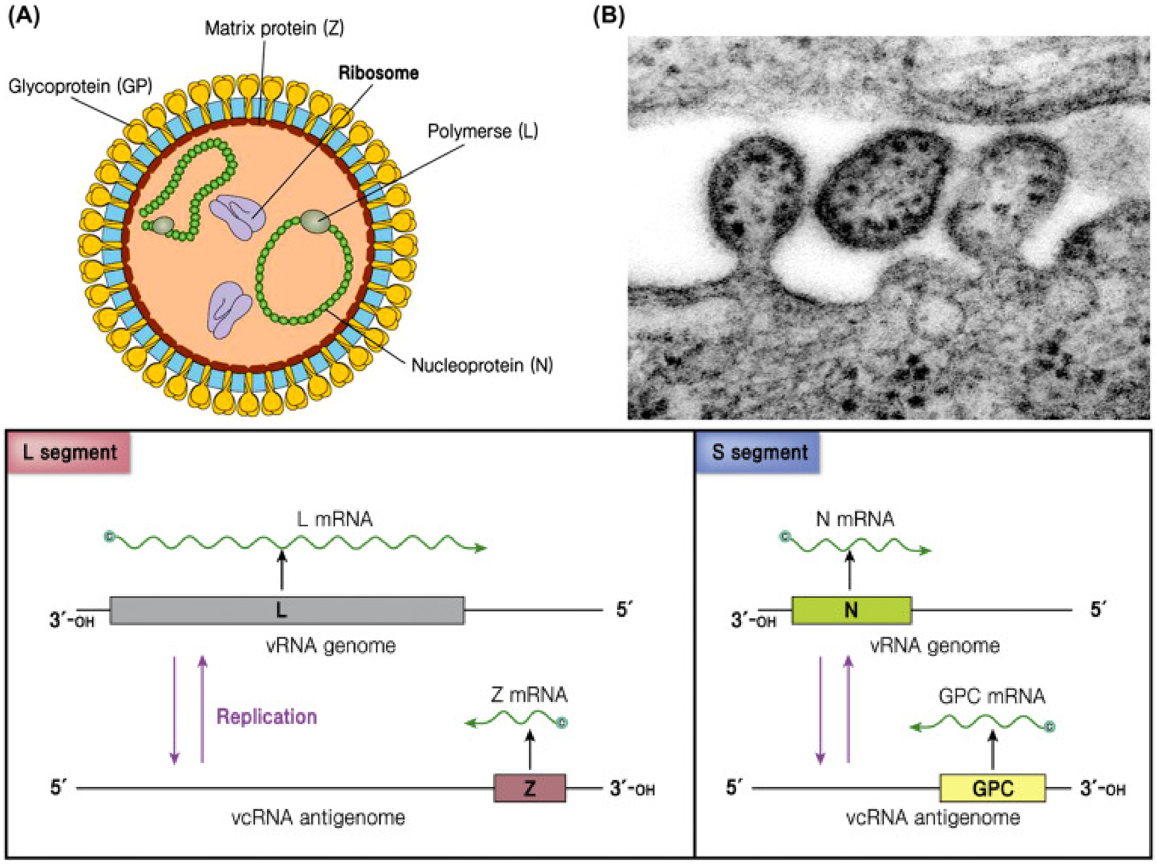
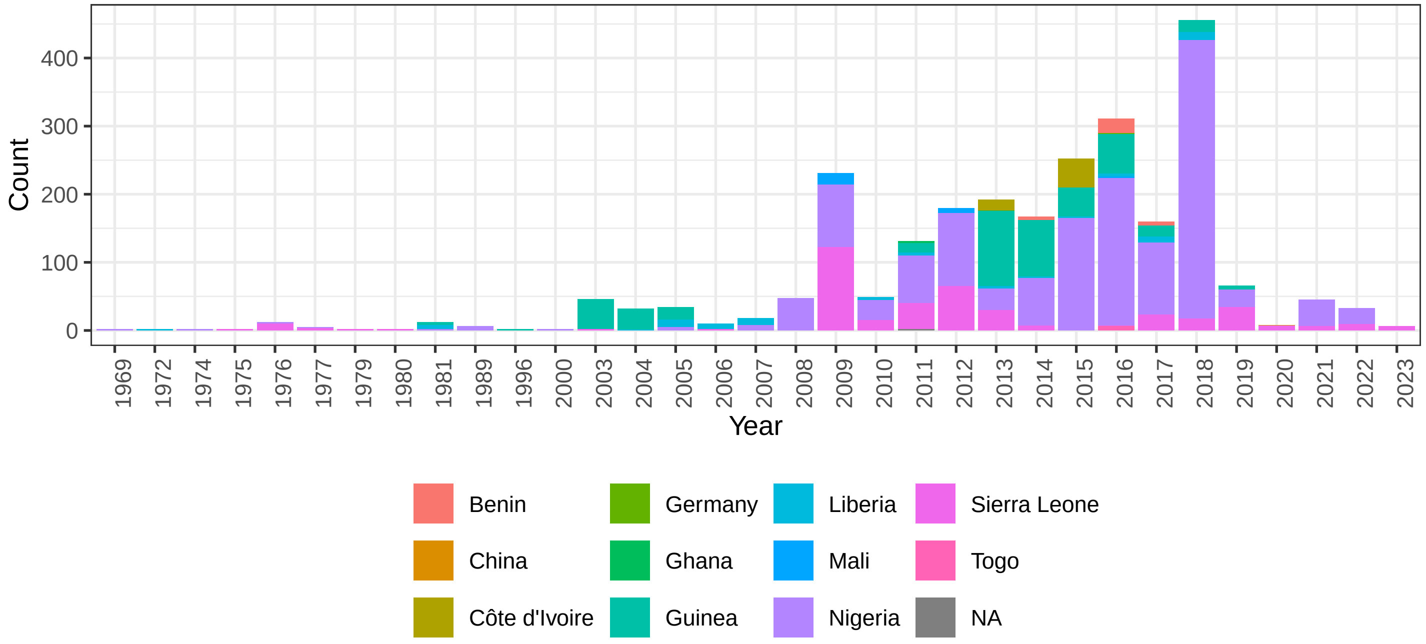
Background

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

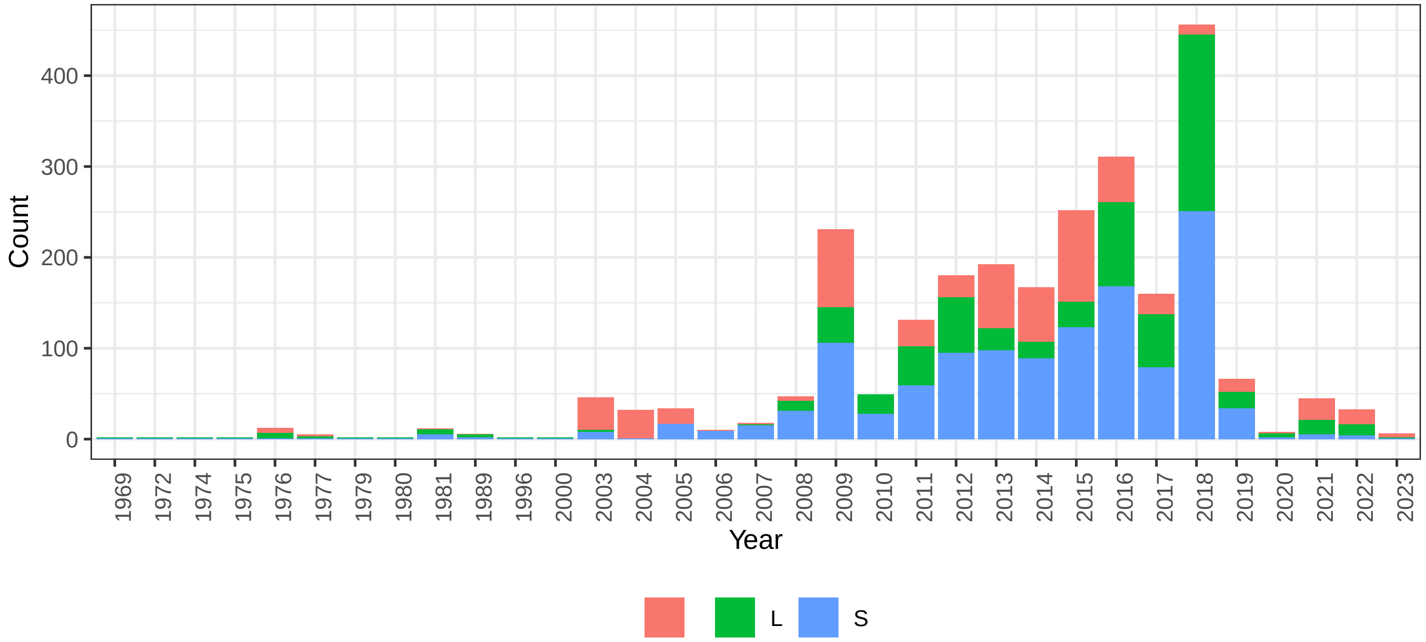


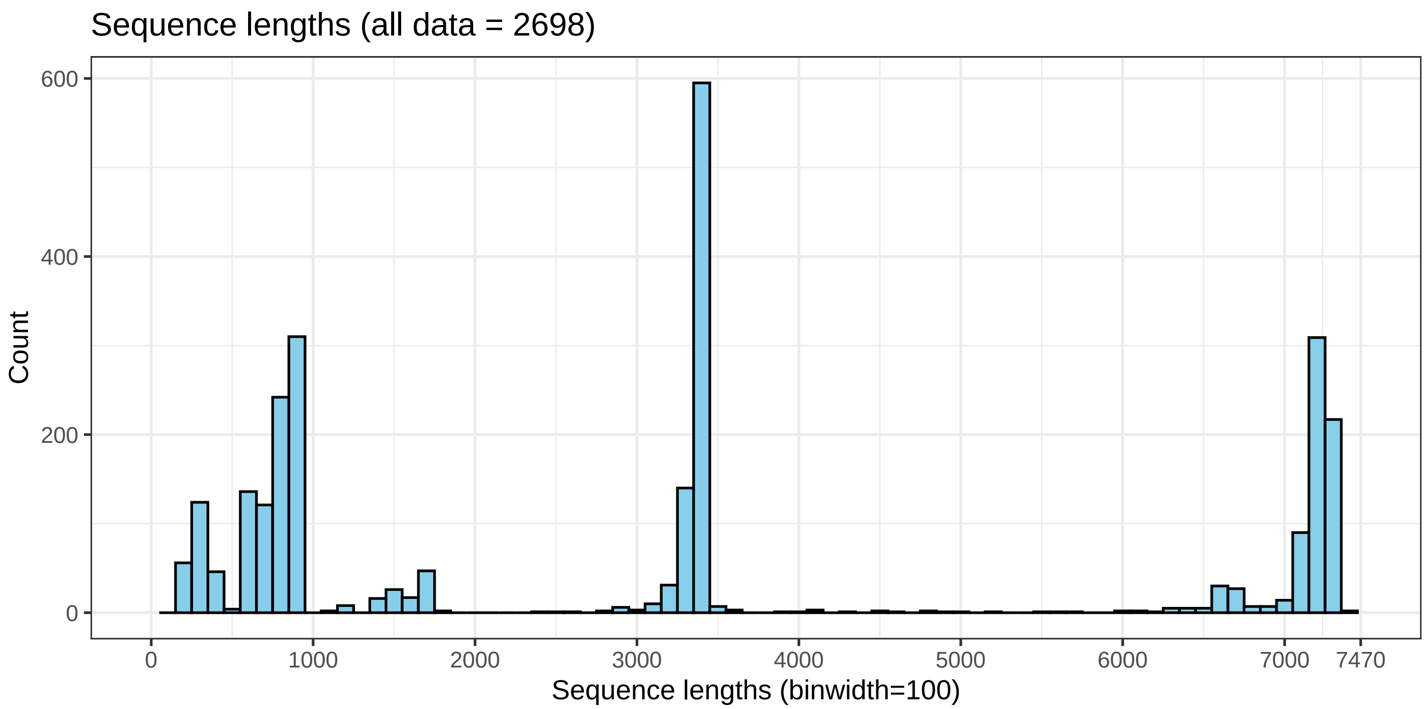
Exploratory Graphics

Lassa entries with collection date (n = 2525)

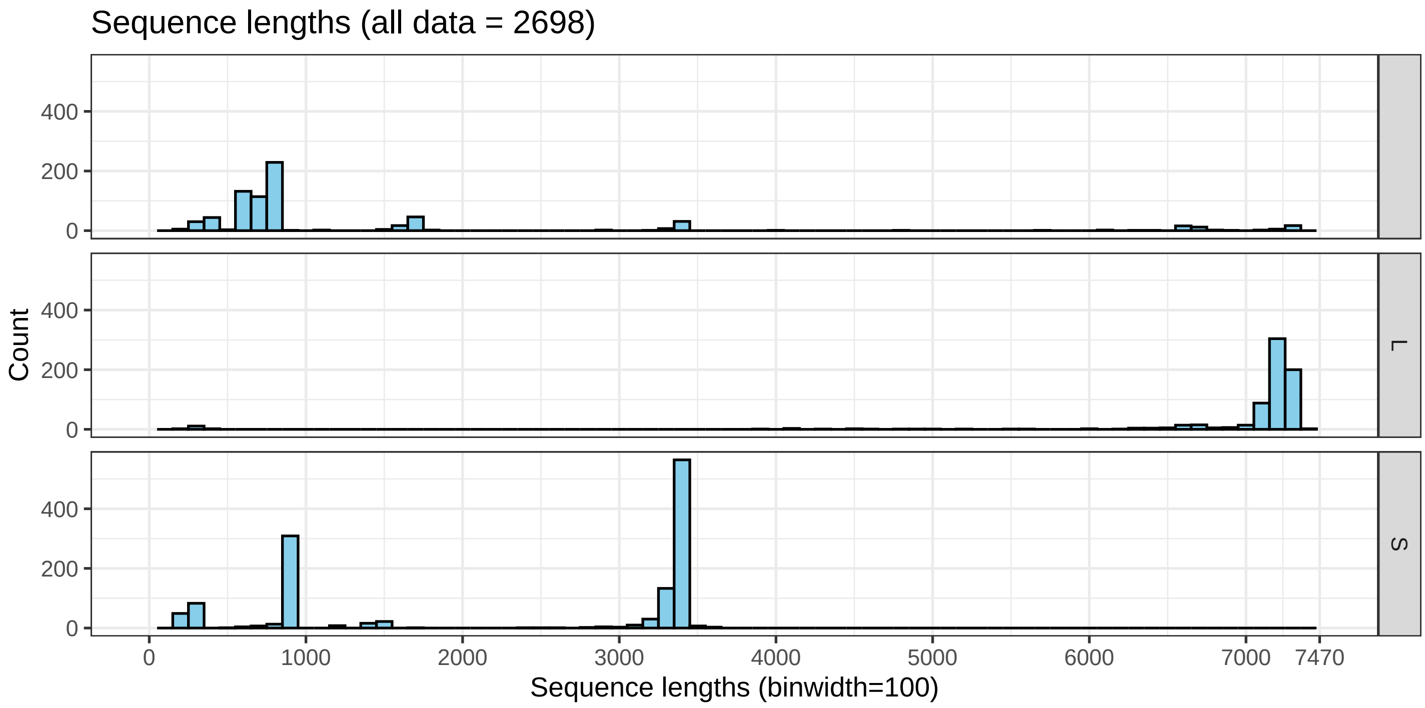


Lassa entries with collection date (n = 2525) colored by Segment





Sequence lengths (all data = 2698)



Sequence lengths (all data = 2698)

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

Lassa strain name

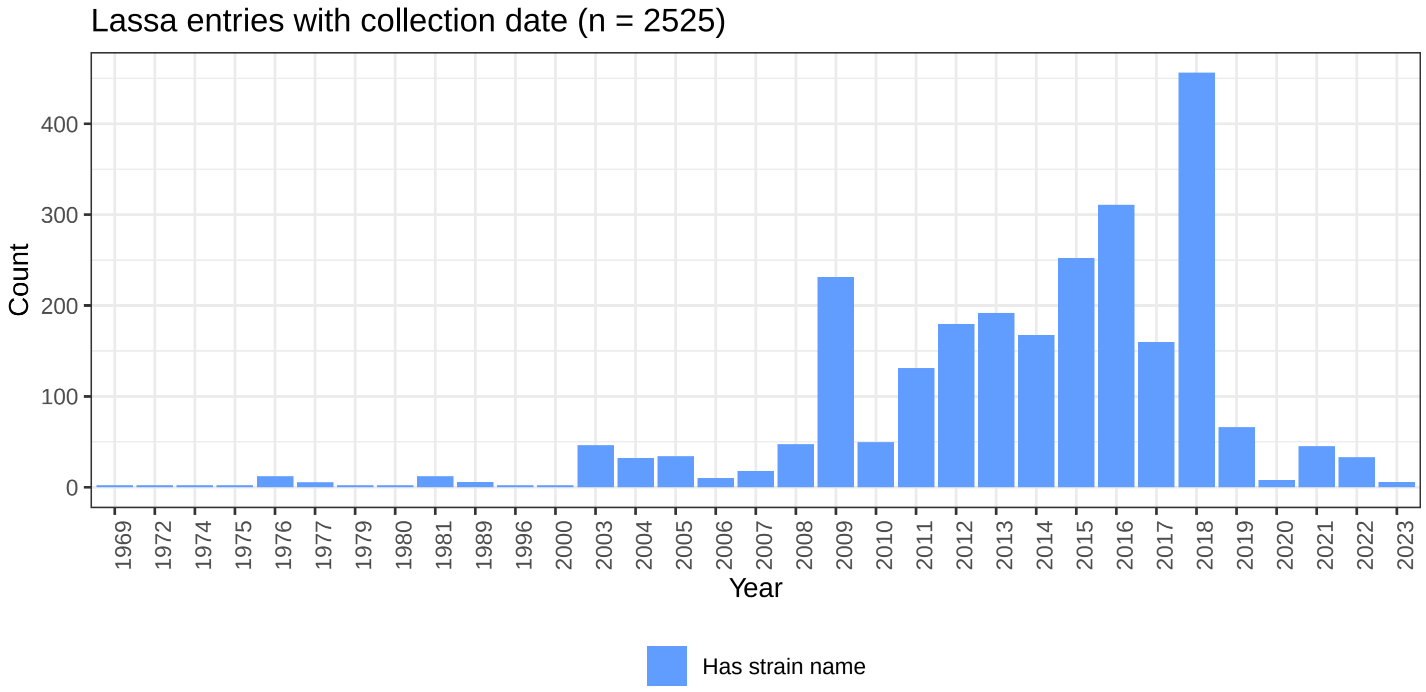
Is setting the strain name helpful for lassa? It looks like a majority of the strain names are the GenBank accession anyways.

```
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%)