Splice SItes Scoring

Nadzeya Boyeva

2024-11-29

```
library(ggseqlogo)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
readTxt <- function(path) {
  read.table(file=path,
             sep="\t",
             header=TRUE,
             quote="\"",
             as.is=TRUE)
}

fd <- '/home/nadzeya/praktika'
```

```
ucsc <- readTxt(file.path(fd, "ucsc_nas_common.txt"))
eejs <- readTxt(file.path(fd, "eej_nas_common.txt"))
```

SSs from RNA-Seq Derived EEJ matrices

Select sites with SNVs in conservative dinucleotides:

```
dinucs <- eejs %>%
  mutate(
    conservative = case_when(
      ss == "5" ~ substr(refseq, 4, 5),
      ss == "3" ~ substr(refseq, 19, 20)
    ) %>% as.character()
  )
```

Calculate stats on conservative dinucleotides:

```
# Calculate statistics for "5" splice sites
stats_5 <- dinucs %>%
  filter(ss == "5") %>%
  group_by(conservative) %>%
  summarise(n = n()) %>%
  mutate(percentage = n / sum(n) * 100)

# Calculate statistics for "3" splice sites
stats_3 <- dinucs %>%
  filter(ss == "3") %>%
  group_by(conservative) %>%
  summarise(n = n()) %>%
  mutate(percentage = n / sum(n) * 100)

# Print the statistics
print("Statistics for 5' splice sites:")
```

```
## [1] "Statistics for 5’ splice sites:"
```

```
print(stats_5)
```

```
## # A tibble: 16 x 3
##    conservative     n percentage
##    <chr>        <int>      <dbl>
##  1 AA               3      0.713
##  2 AC               4      0.950
##  3 AG               9      2.14
##  4 AT               2      0.475
##  5 CA               9      2.14
##  6 CC              15      3.56
##  7 CG               7      1.66
##  8 CT               7      1.66
##  9 GA               9      2.14
## 10 GC               7      1.66
## 11 GG              28      6.65
## 12 GT             297     70.5
## 13 TA               2      0.475
## 14 TC               8      1.90
## 15 TG               6      1.43
## 16 TT               8      1.90
```

```
print("Statistics for 3' splice sites:")
```

```
## [1] "Statistics for 3’ splice sites:"
```

```
print(stats_3)
```

```
## # A tibble: 16 x 3
##    conservative     n percentage
```

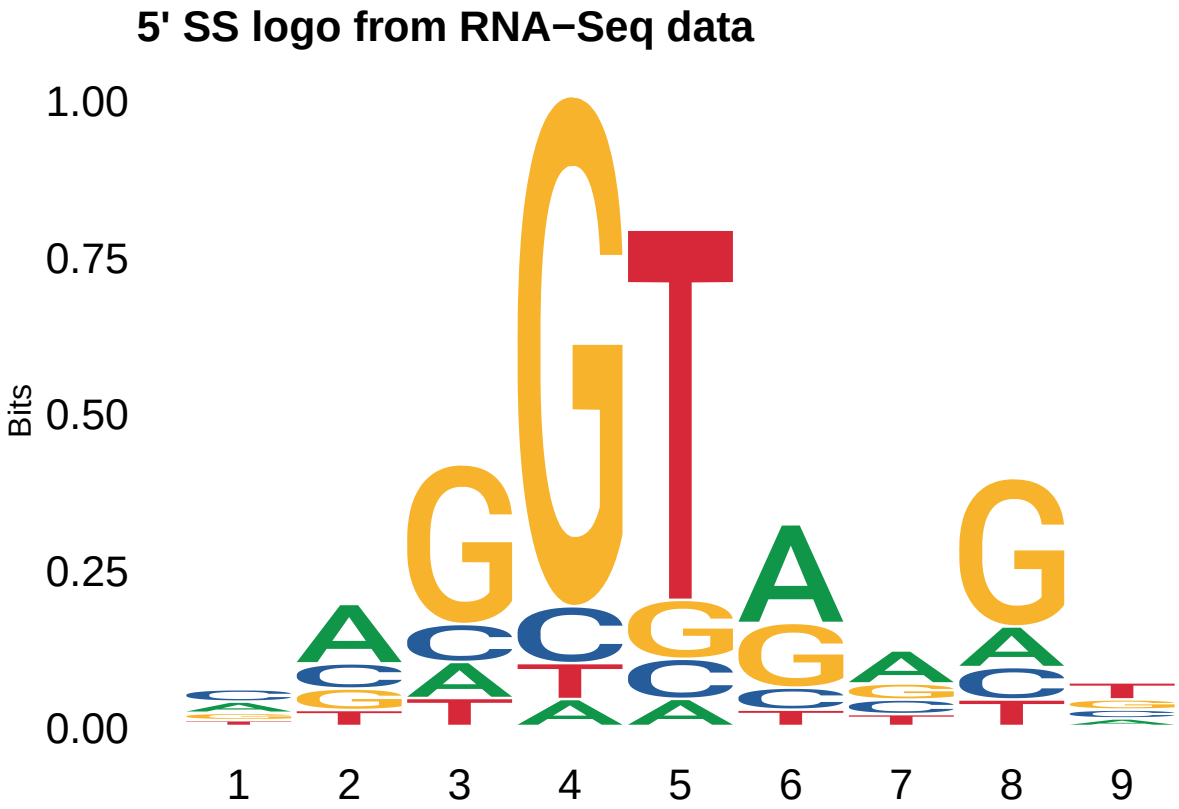
```
##    <chr>        <int>      <dbl>
##  1 AA              16      0.982
##  2 AC              26      1.60
##  3 AG            1259     77.3
##  4 AT              14      0.859
##  5 CA              32      1.96
##  6 CC              50      3.07
##  7 CG              14      0.859
##  8 CT              32      1.96
##  9 GA              16      0.982
## 10 GC              29      1.78
## 11 GG              40      2.46
## 12 GT              18      1.10
## 13 TA              18      1.10
## 14 TC              21      1.29
## 15 TG              28      1.72
## 16 TT              16      0.982
```

```r
# Create sequence logos
# Filter sequences for "5" and "3" splice sites
sequences_5 <- dinucs %>% filter(ss == "5") %>% pull(refseq)
sequences_3 <- dinucs %>% filter(ss == "3") %>% pull(refseq)
```

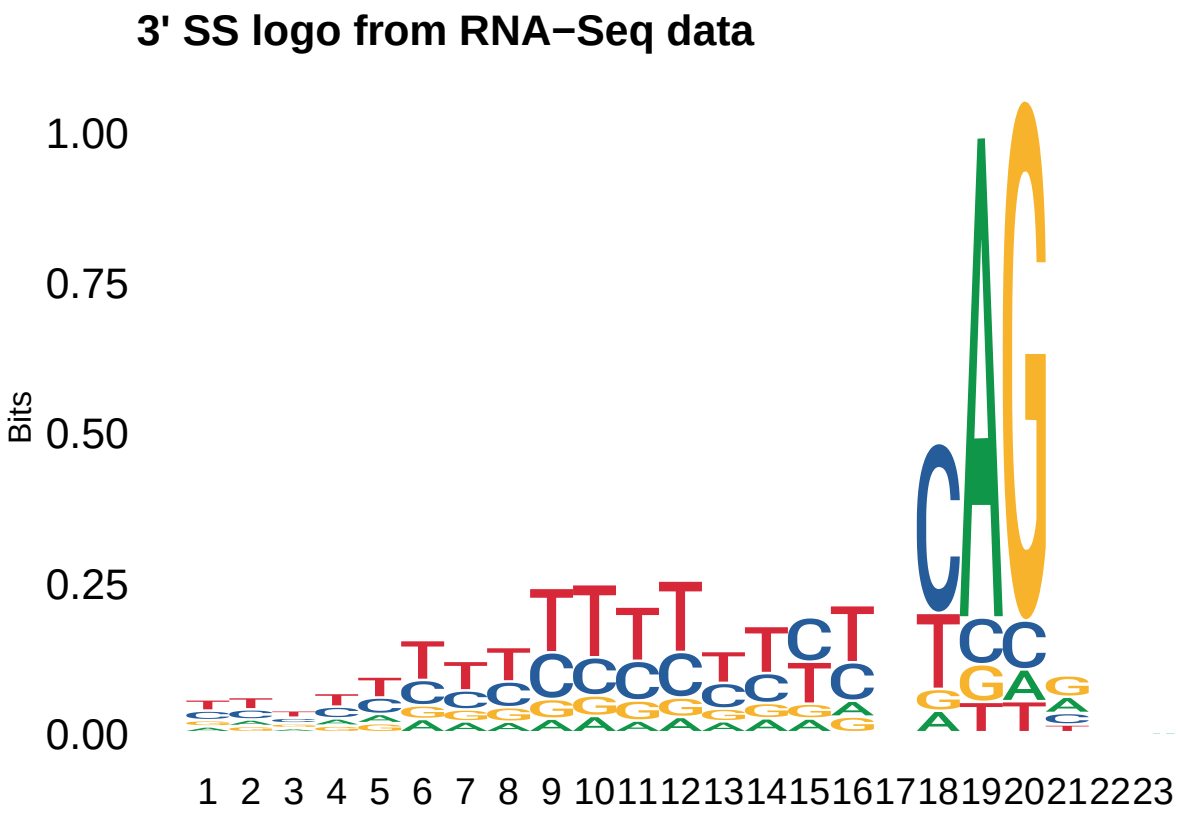
Plot logos:

```r
ggseqlogo(sequences_5, stack_width = 0.95) +
  labs(title = "5' SS logo from RNA-Seq data") +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.text.x = element_text(size = 16),
    axis.text.y = element_text(size = 16)
  )
```

```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the ggseqlogo package.
##   Please report the issue at <https://github.com/omarwagih/ggseqlogo/issues>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

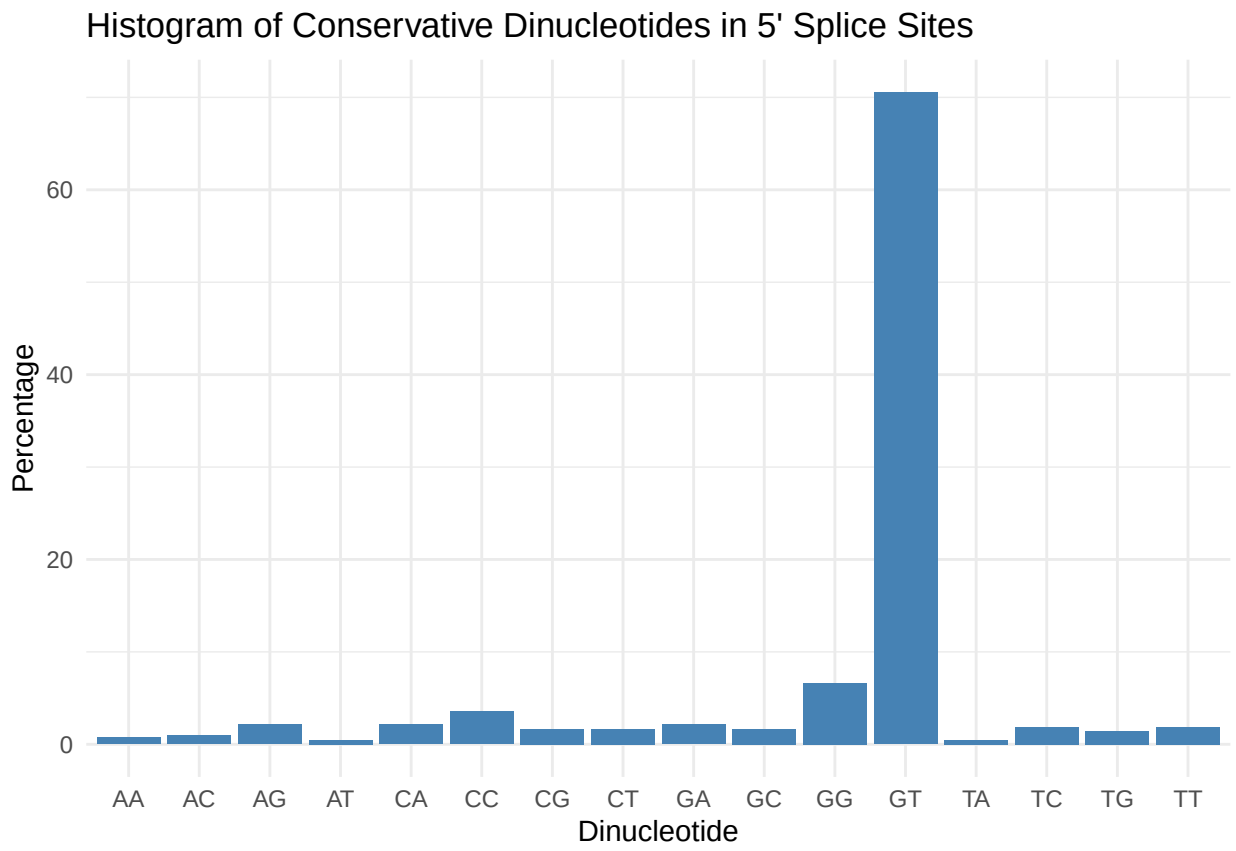



```
ggseqlogo(sequences_3, stack_width = 1) +
  labs(title = "3' SS logo from RNA-Seq data") +
    theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.text.x = element_text(size = 14),
    axis.text.y = element_text(size = 16)
  )
```

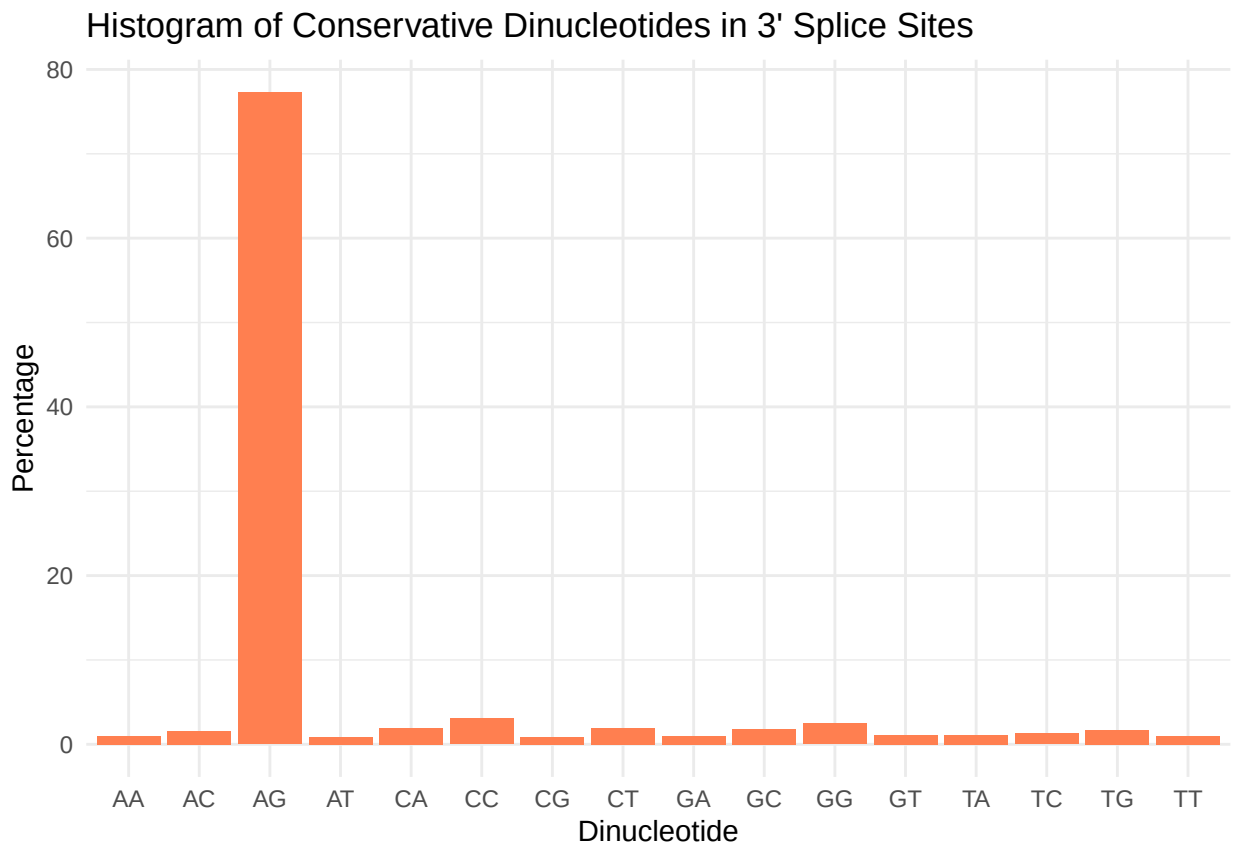


Plot stats:

```
ggplot(stats_5, aes(x = conservative, y = percentage)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(
    title = "Histogram of Conservative Dinucleotides in 5' Splice Sites",
    x = "Dinucleotide",
    y = "Percentage"
  ) +
  theme_minimal()
```




```
ggplot(stats_3, aes(x = conservative, y = percentage)) +
  geom_bar(stat = "identity", fill = "coral") +
  labs(
    title = "Histogram of Conservative Dinucleotides in 3' Splice Sites",
    x = "Dinucleotide",
    y = "Percentage"
  ) +
  theme_minimal()
```



Unique SSs from UCSC Annotation

```
dinucs <- ucsc %>%
  mutate(
    conservative = case_when(
      ss == "5" ~ substr(refseq, 4, 5),
      ss == "3" ~ substr(refseq, 19, 20)
    ) %>% as.character()
  )
```

```
stats_5 <- dinucs %>%
  filter(ss == "5") %>%
  group_by(conservative) %>%
  summarise(n = n()) %>%
  mutate(percentage = n / sum(n) * 100)

# Calculate statistics for "3" splice sites
stats_3 <- dinucs %>%
  filter(ss == "3") %>%
  group_by(conservative) %>%
  summarise(n = n()) %>%
  mutate(percentage = n / sum(n) * 100)
```

```r
# Print the statistics
print("Statistics for 5' splice sites:")
```

```
## [1] "Statistics for 5’ splice sites:"
```

```r
print(stats_5)
```

```
## # A tibble: 2 x 3
##   conservative     n percentage
##   <chr>        <int>      <dbl>
## 1 GC               9       27.3
## 2 GT              24       72.7
```

```r
print("Statistics for 3' splice sites:")
```

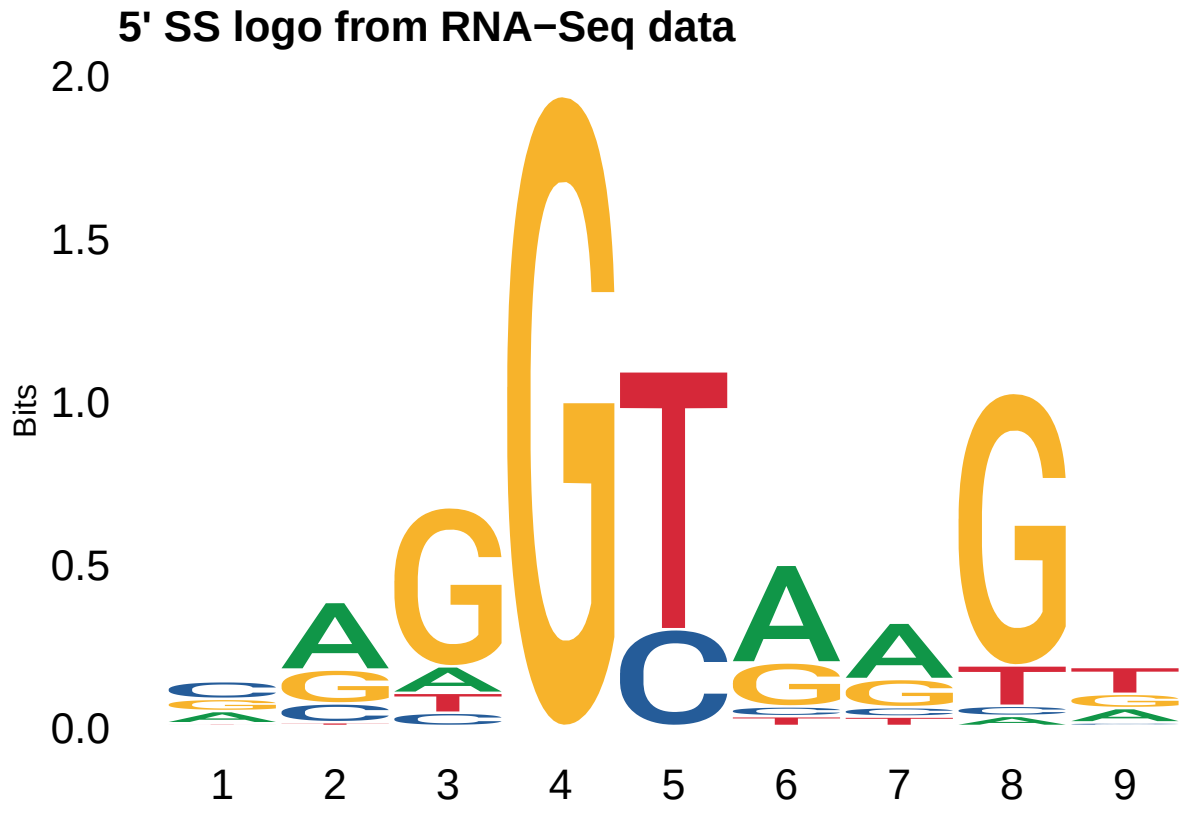
```
## [1] "Statistics for 3’ splice sites:"
```

```r
print(stats_3)
```

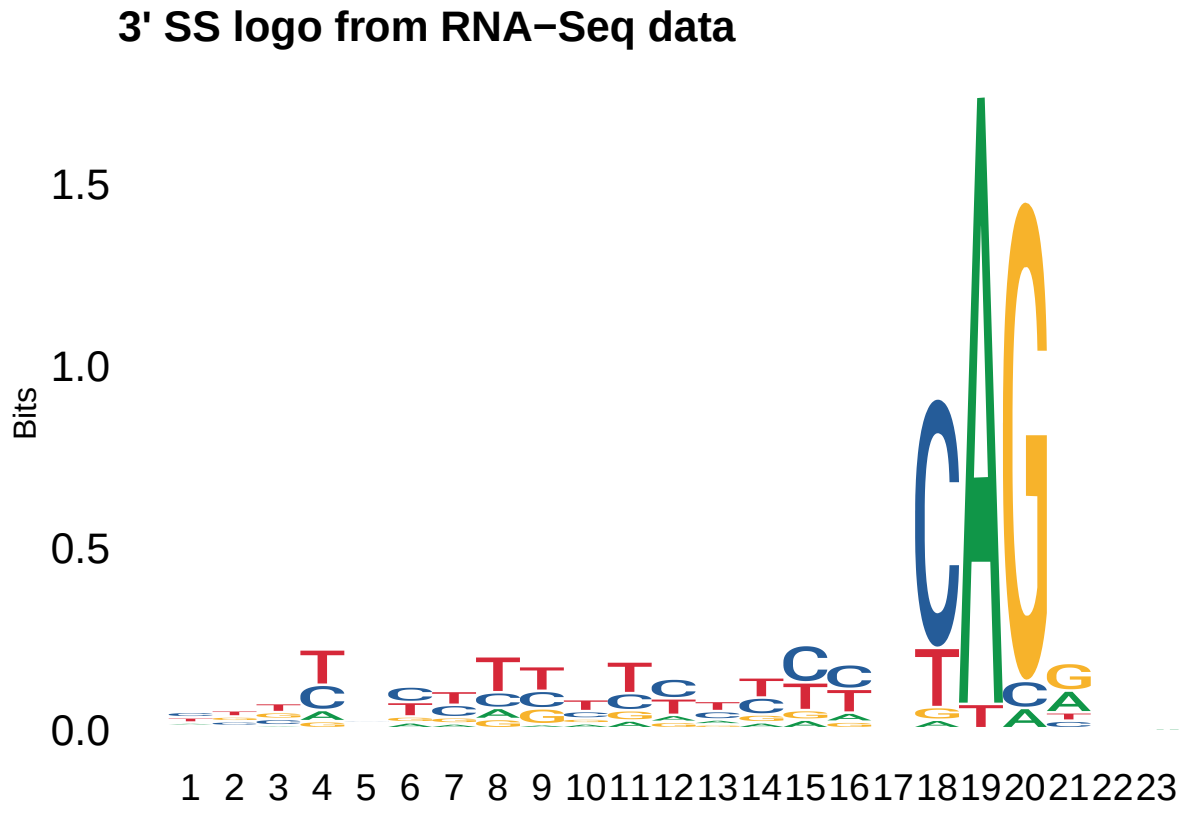
```
## # A tibble: 5 x 3
##   conservative     n percentage
##   <chr>        <int>      <dbl>
## 1 AA               1       1.28
## 2 AC               3       3.85
## 3 AG              71      91.0
## 4 TA               2       2.56
## 5 TC               1       1.28
```

```r
# Create sequence logos
# Filter sequences for "5" and "3" splice sites
sequences_5 <- dinucs %>% filter(ss == "5") %>% pull(refseq)
sequences_3 <- dinucs %>% filter(ss == "3") %>% pull(refseq)
```

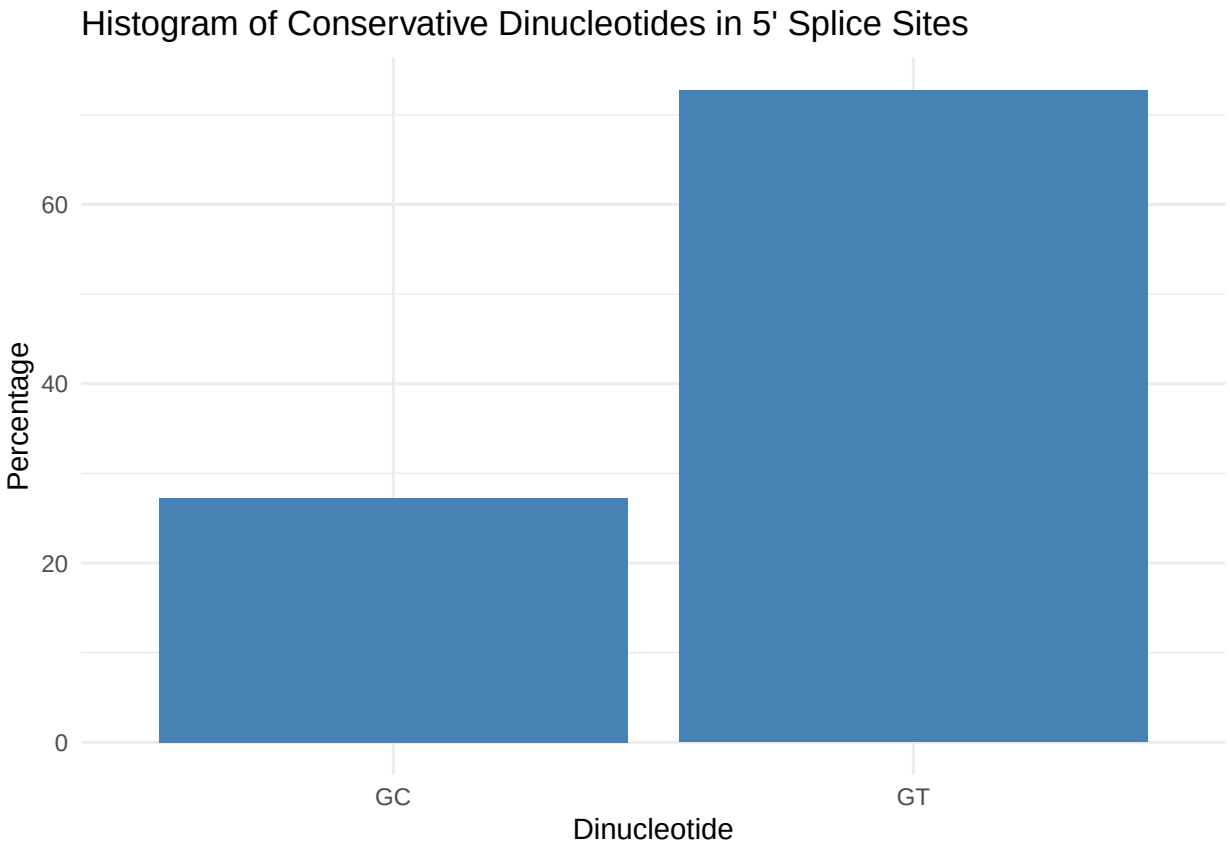
```r
ggseqlogo(sequences_5, stack_width = 0.95) +
  labs(title = "5' SS logo from RNA-Seq data") +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.text.x = element_text(size = 16),
    axis.text.y = element_text(size = 16)
  )
```



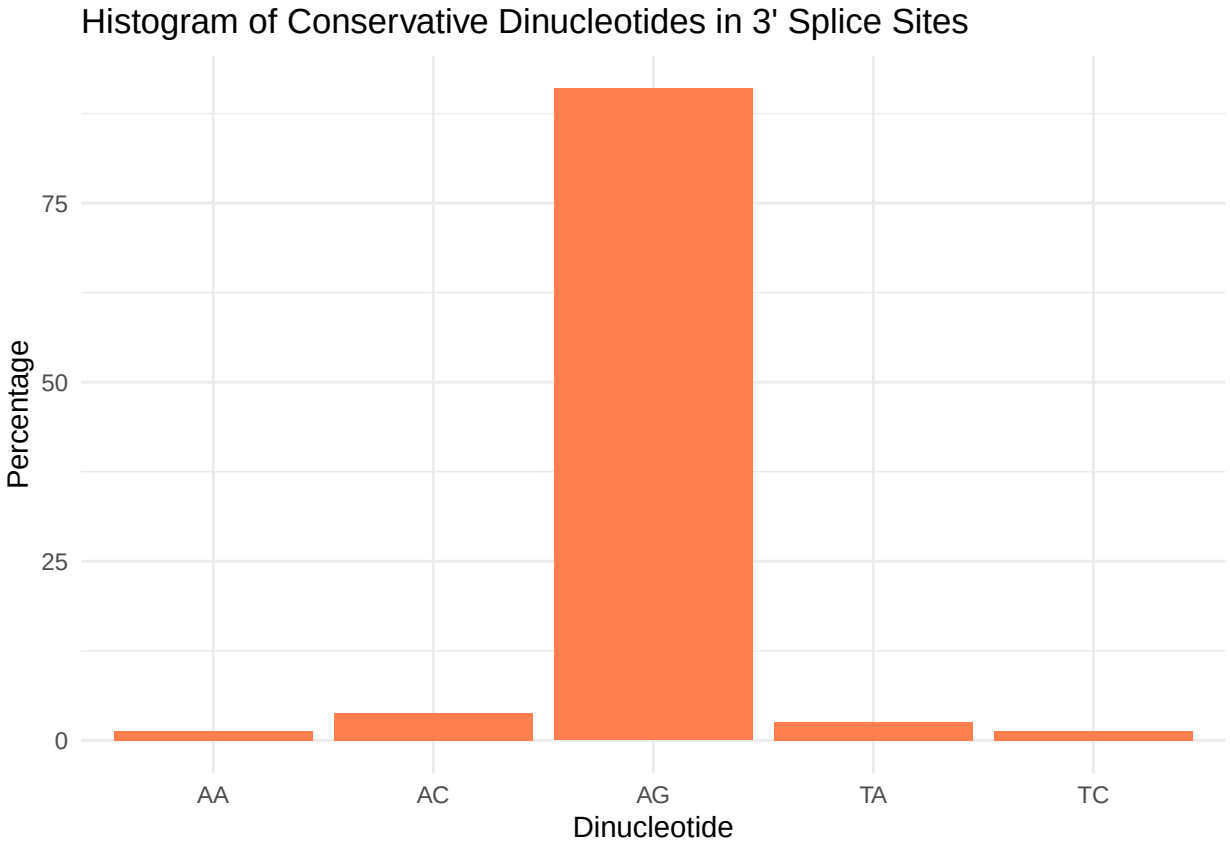

```
ggseqlogo(sequences_3, stack_width = 1) +
  labs(title = "3' SS logo from RNA-Seq data") +
    theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.text.x = element_text(size = 14),
    axis.text.y = element_text(size = 16)
  )
```




```r
ggplot(stats_5, aes(x = conservative, y = percentage)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(
    title = "Histogram of Conservative Dinucleotides in 5' Splice Sites",
    x = "Dinucleotide",
    y = "Percentage"
  ) +
  theme_minimal()
```




```
ggplot(stats_3, aes(x = conservative, y = percentage)) +
  geom_bar(stat = "identity", fill = "coral") +
  labs(
    title = "Histogram of Conservative Dinucleotides in 3' Splice Sites",
    x = "Dinucleotide",
    y = "Percentage"
  ) +
  theme_minimal()
```


Histogram of Conservative Dinucleotides in 3' Splice Sites
Percentage
75
50
25
0
AA
AC
AG
TA
TC
Dinucleotide