Cow Antibody NGS Pipeline

Jeremy Haakenson

2023-03-30

R Markdown

Load packages.

```
library(Biostrings)
library(dplyr)
library(ggplot2)
```

Read in file of unique nt seqs.

```
dna1 = read.csv('DNA1_Unique_Seq.csv')
```

Trim off first nt to put seqs in correct reading frame.

```
dna1$Sequence=substr(dna1$Sequence, 2, nchar(dna1$Sequence))
```

Isolate DNA seqs.

```
seq = dna1$Sequence
```

Convert to DNA string set.

```
dna = DNAStringSet(seq)
```

Translate.

```
AA2 = translate(dna, no.init.codon = TRUE, if.fuzzy.codon = 'X')
head(AA2)
```

```
## AAStringSet object of length 6:
##     width seq
## [1]   124 VRQAPGKALEWLGGIDTGGSTGYNPGLKSRLSIT...CCRNRSYNAADLGHCTDYTSVYAYEFYVDTWGQ
## [2]     0
## [3]   119 VRQAPGKSLEWLGSIDTGGSTGYNPGLKSRLSVT...FGCSRDGCCRSRGTCVDDTIRYTYDWYVDAWGQ
## [4]   108 VRQAPGKALEWLGGIDTGGSTGYNPGLKSRLSIT...NVRLVVVVALAVGAVVLQSIFILTNTTSRPGAK
## [5]   124 VRQAPGKALEWLGSIDTSGSTGYNPGLKSRLSIT...VVVSVVCGLILVVGSVVIR*LMLTNGTSMPGAK
## [6]   125 IRQAPGKALEWLGSIDTSGTTGYNPGLKTRLSIT...GCLGCDPDRGWAYNWRSYTHTNSYQFHVDAWGQ
```

Add translated column to dataframe.

```
dna1 = cbind.data.frame(dna1, AA2)
```

Remove junk seqs, which are defined as amino acid seqs that do not begin with _R_A and nt seqs less than 220 bp.

```
dna2 = dna1 %>%
  filter(grepl('^.?R.?A.*', AA2)) %>%
  filter(nchar(Sequence) >= 220)
```

Filter for functional seqs, which are defined as amino acid seqs that do not contain a premature stop codon and that contain the conserved WG motif in the JH region.

```
dna3 = dna2 %>%
  filter(!grepl('\\*', AA2)) %>%
  filter(grepl('.*WG.*', substr(AA2, -18, nchar(AA2))))
```

Isolate CDR3s.

```
cdr3 = substr(dna3$AA2, 59, nchar(dna3$AA2) - 2)
head(cdr3)
```

```
## [1] "CTTVLQITHTKKSCPDDYQYNCGSLGRGCTGRDCCRNRSYNAADLGHCTDYTSVYAYEFYVDTW"
## [2] "CXTVHQRTSQRRDCPXGYDANSGAVCSLFGCSRDGCCRSRGTCVDDTIRYTYDWYVDAW"
## [3] "CTAVHQKTETIRSCPDGYTDCSTCSYTRRDCSAGGCLGCDPDRGWAYNWRSYTHTNSYQFHVDAW"
## [4] "CTTVVQQTHRTCPTPTGGDIDCRIGFVPWSYNYEWYIDAW"
## [5] "CTTVHQETIQKRGCPSGCINNGGCGSGCCCRHCWTSRPQCTTYISSITYEVHVDAW"
## [6] "CTTVYQKTHRNCPDGDEYVQIWNRCRYRGTITYTYEWHIDAW"
```

Add CDR3s to dataframe.

```
dna3_cdr3 = cbind.data.frame(dna3, cdr3)
```

Make a table of CDR3s.

```
dna3_cdr3table = as.data.frame(table(dna3_cdr3$cdr3))

dna3_cdr3table = dna3_cdr3table %>%
  arrange(desc(Freq))

head(dna3_cdr3table)
```

```
##                                                          Var1 Freq
## 1                                CTTVHPGGYGYGGYGCYGYGYGYGYVDAW   50
## 2                                CTTVHQMVVMVMVVMVVMVMVMAYYVDAW   43
## 3                  CTTVHQETTRNCPVAYVWRSDHACCWHAWNGCTSSNSYKYEWYIDAW   30
## 4                                CTTVHLMVVMVMVVMVVMVMVMVMDYVDAW   29
## 5               CTTVHQETRKSCPDGYPYQCGAGCQTYSCRYTGRITQYIYTYEHHIEAW   26
## 6 CTTVHQKTQRGCPDGYSYGCGCSESSFICCAYGCWPSNNVNYLGYYYGIPTDSHTYTYEFHVDAW   26
```

Replace Var1 column name with CDRH3.

```
colnames(dna3_cdr3table) = c('CDRH3', 'Freq')
```

Write a CSV file of the CDR3 table.

```
write.csv(dna3_cdr3table, 'dna3_cdr3table.csv')
```

Isolate ultralong antibodies.

```
ultra = cdr3[which(nchar(cdr3) >= 42)]
```

Calculate percent ultralong.

```
64936/73827 * 100
```

```
## [1] 87.95698
```

88.0% of seqs were ultralong.

Make a table of CDR3 lengths.

```
cdr3len = nchar(as.character(dna3_cdr3table$CDRH3)) - 2
len.table = as.data.frame(table(cdr3len))
head(len.table)
```

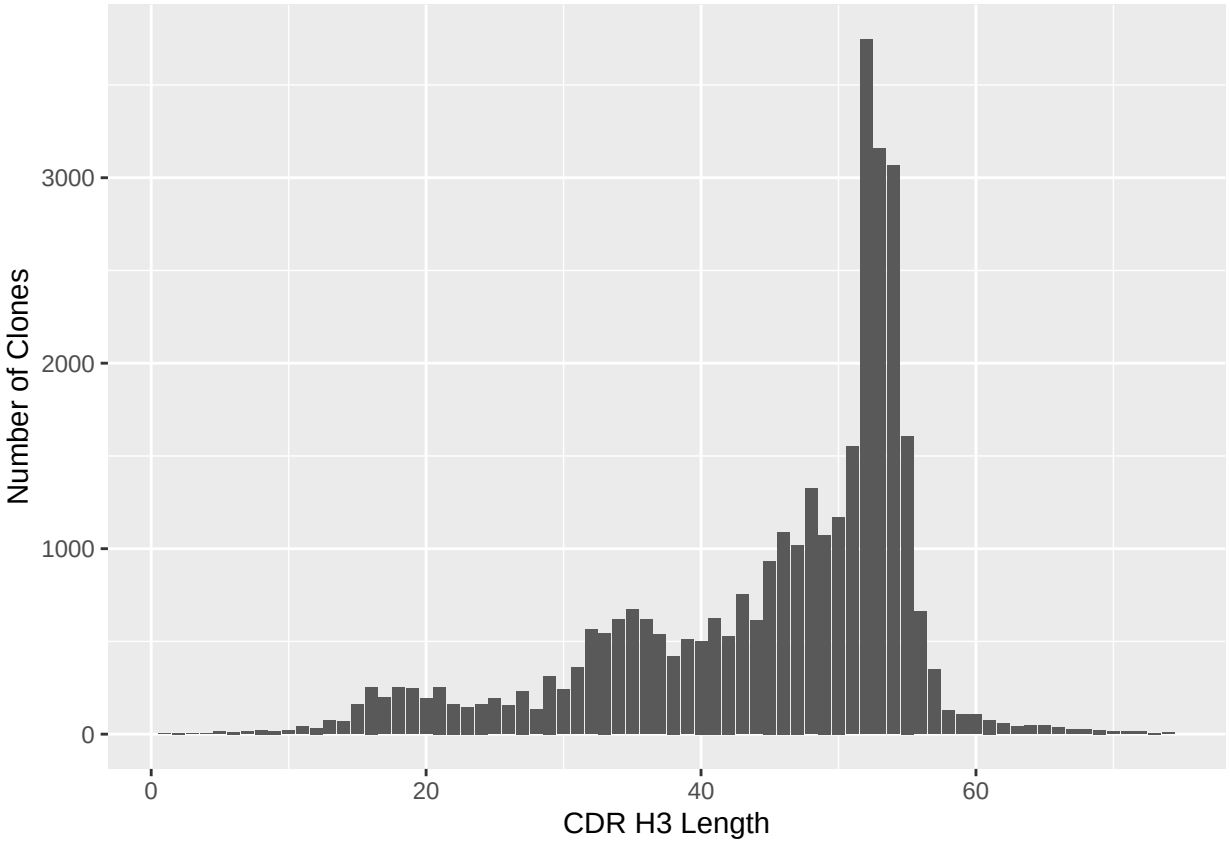
```
##   cdr3len Freq
## 1      11    5
## 2      12    6
## 3      14    5
## 4      15    3
## 5      16   15
## 6      17   12
```

Make a figure of the CDR3 length distribution.

```
len.table$cdr3len = as.numeric(len.table$cdr3len)

cdr3fig = ggplot(len.table, aes(x = cdr3len, y = Freq)) + geom_col() +
  xlab('CDR H3 Length') + ylab('Number of Clones')

cdr3fig
```




```
range(len.table$cdr3len)
```

```
## [1]  1 74
```

The CDR H3s in this data set ranged from 1-74 amino acids long, with peaks at ~19, 35, and 52, the latter representing ultralong cow antibodies.