Study the parsed fasta file for read lengths

### Read the data

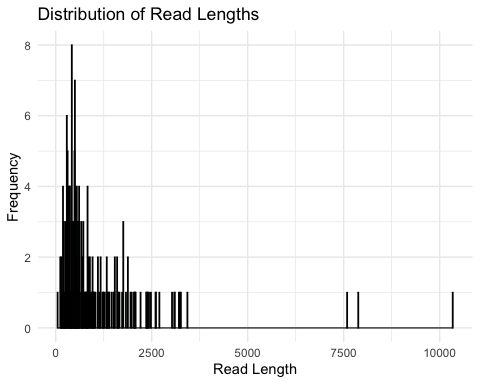
# Load the necessary libraries  
library(ggplot2)  
library(data.table)

## Warning: package 'data.table' was built under R version 4.4.1

# Read the TSV file generated by the Python script  
data <- fread("/Users/valishashah/Desktop/all\_projects/HaLab\_assessment/Sequencing\_analysis\_assignment/parsed\_fasta\_data.tsv", header = TRUE, sep = "\t")

# Step 1: Plot the distribution of read lengths as a histogram

ggplot(data, aes(x = length)) +  
 geom\_histogram(binwidth = 10, fill = "skyblue", color = "black", alpha = 0.7) +  
 labs(title = "Distribution of Read Lengths", x = "Read Length", y = "Frequency") +  
 theme\_minimal()



### A lot of short reads   
### Very few long reads

### Step 2: Calculate the cumulative distribution of 2-mer frequencies

# Extract the columns corresponding to 2-mer counts (columns 5 to 20)  
two\_mer\_columns <- colnames(data)[5:20]  
  
# Define a function to calculate 2-mer frequencies for a sequence  
calculate\_two\_mer\_frequencies <- function(row) {  
 # Get the counts of each 2-mer  
 counts <- as.numeric(row[two\_mer\_columns])  
   
 # Add 1 to each count to obtain pseudo-counts  
 pseudo\_counts <- counts + 1  
   
 # Calculate the total number of pseudo-counts  
 total\_count <- sum(pseudo\_counts)  
   
 # Calculate the frequency of each 2-mer  
 frequencies <- pseudo\_counts / total\_count  
   
 return(frequencies)  
}  
  
# Apply the function to each row to get the frequencies of 2-mers  
two\_mer\_frequencies <- t(apply(data, 1, calculate\_two\_mer\_frequencies))  
  
# Create a data frame for plotting  
two\_mer\_df <- data.frame(two\_mer\_frequencies)  
two\_mer\_df$sequence <- 1:nrow(data) # Adding a sequence number for tracking  
  
# Reshape the data frame to long format for plotting  
library(tidyr)  
long\_data <- gather(two\_mer\_df, key = "two\_mer", value = "frequency", -sequence)  
  
# Plot the cumulative distribution of 2-mer frequencies  
ggplot(long\_data, aes(x = frequency, color = two\_mer)) +  
 stat\_ecdf(geom = "step") +  
 labs(title = "Cumulative Distribution of 2-mer Frequencies", x = "2-mer Frequency", y = "Cumulative Probability") +  
 theme\_minimal() +  
 theme(legend.position = "none")

