David Armas

Ryan Schwarzkopf

February 23, 2023

**Analysis of Processing Common Words in SARS-CoV2 Genome**

**Abstract**

The process of finding common words within genomes is an essential tool for researching a given genome. Common words can bring new insight into how the genome is used throughout the given organism. Throughout this project, two different methods of common word processing were implemented and compared. Counting and hashing were the two main methods utilized. Their runtimes were compared using the same input and the results were plotted to visualize the differences. The hashing method was then utilized to report frequent words from a given genome.

**Introduction**

SARS-CoV2, commonly known as a member of coronavirus, is a virus that causes a contagious respiratory disease. The current total recorded cases of coronavirus in the United States is 102,998,014. Total deaths in the United States are 1,113,254. (CDC, 2023) Vaccines have been able to prevent a great number of infections however, there are many variants that stem from the common ancestor of coronavirus, because of a continuous evolution of the genetic code. Because of the severity of the illness, the genetic lineages of coronavirus are monitored in the labs in preparation of an outbreak. The rapid emergence of new strains requires quickness in the development of new vaccines. Therefore, preparation is key to keeping people safe. Key proteins in the coronavirus are the spike proteins which bind to a receptor angiotensin-converting enzyme 2 which allows it to hack into host cells. (Huang, 2020) The genomes of these variants are sequenced and assembled, then K-mers can be identified in individual strains and compared to its ancestors and the other current variants. The DNA sequence can be compared to the sequences of other more aggressive strains to assess where the significant proteins lie and how to combat them. The variants may then be classified into groups: Variants Being Monitored, Variants of Interest, Variants of Concern, and Variants of High Consequence. (CDC, 2022) The Wuhan-Hu-1 variant will be assessed for its common words. The results can be used to further biological research among coronavirus variants.

**Methods**

R was the programming language utilized throughout the project. Final output was generated using Quarto document language. See attached HTML and QMD file. The following packages were utilized, “hash”, “rbenchmark”, “ggplot2”, and “readr”. Uncomment install code in Task 0a for installation if needed.

The genome file utilized in Task 5 was generated from the provided URL (NCBI, 2022). The genome was copied and pasted then the “/n” characters removed. The result was saved as “genome.txt”, in order to run the code change the “setwd” code in Task 5b to match your file location. The file is attached with the supplementary files. With these modifications, you can run the RMD file and generate a new HTML output file on your machine using RStudio.

*Task 1*

The purpose of frequentWords is to find the most common k-mer in a genome. The frequentWords function for every k-mer, of length n, searches a given string for the frequency. FrequentWords calls patternCount for every k-mer starting at index zero. PatternCount returns the frequency of the k-mer in the string. Finally, frequentWords finds the max value and gets rid of duplicates.

*Task 2*

To implement hashing, the r package “hash” was installed and used. A hash list was created using the code “count = as.list.hash(c())”. In order to add to the list the pattern was used as the key “count[pattern]”. The result from patternCount was saved in this location. The max count was checked after each pattern count to determine the largest count. A simple query was used to determine the frequent words “names(count[count == max\_count])”, this was the returned result.

*Task 3*

The purpose of randomSequence is to create short sequences for testing the validity of other methods. The randomSequence function takes in a desired length and returns a string consisting of ACGT in an even multinomial distribution. The probability of A, C, G, and T are all even at 0.25. Rmultinom() creates the multinomial array, apply() names every element of x, and paste() contacts that result all into one string.

*Task 4*

Comparing the frequent word run times was done using the package “rbenchmark”. Each function was run using the same parameters with the following code “benchmark(frequentWords(text, k), replications = 1)”. This package returned a parameter with the elapsed time, this was used to determine the runtime.

To obtain all of the results for visualization, a function was created to save the results of a list of given parameters. From this table of results “ggplot” was utilized to visualize the runtimes based on k values and methods. See results for generated plots. The k values utilized were 3, 6, 9, 12, and 15. The lengths were 250, 500, 750, 1000, and 1500.

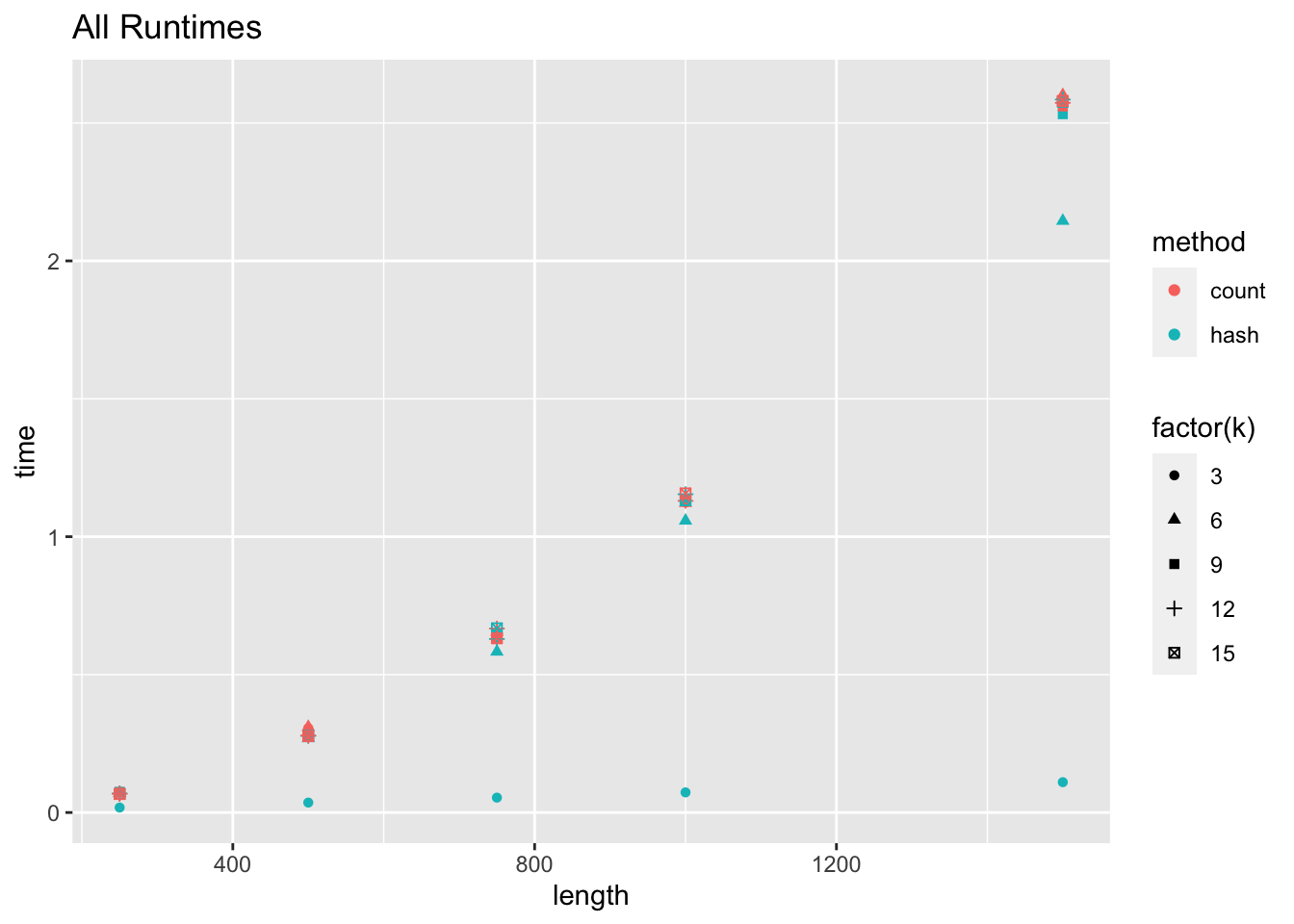
*Task 5*

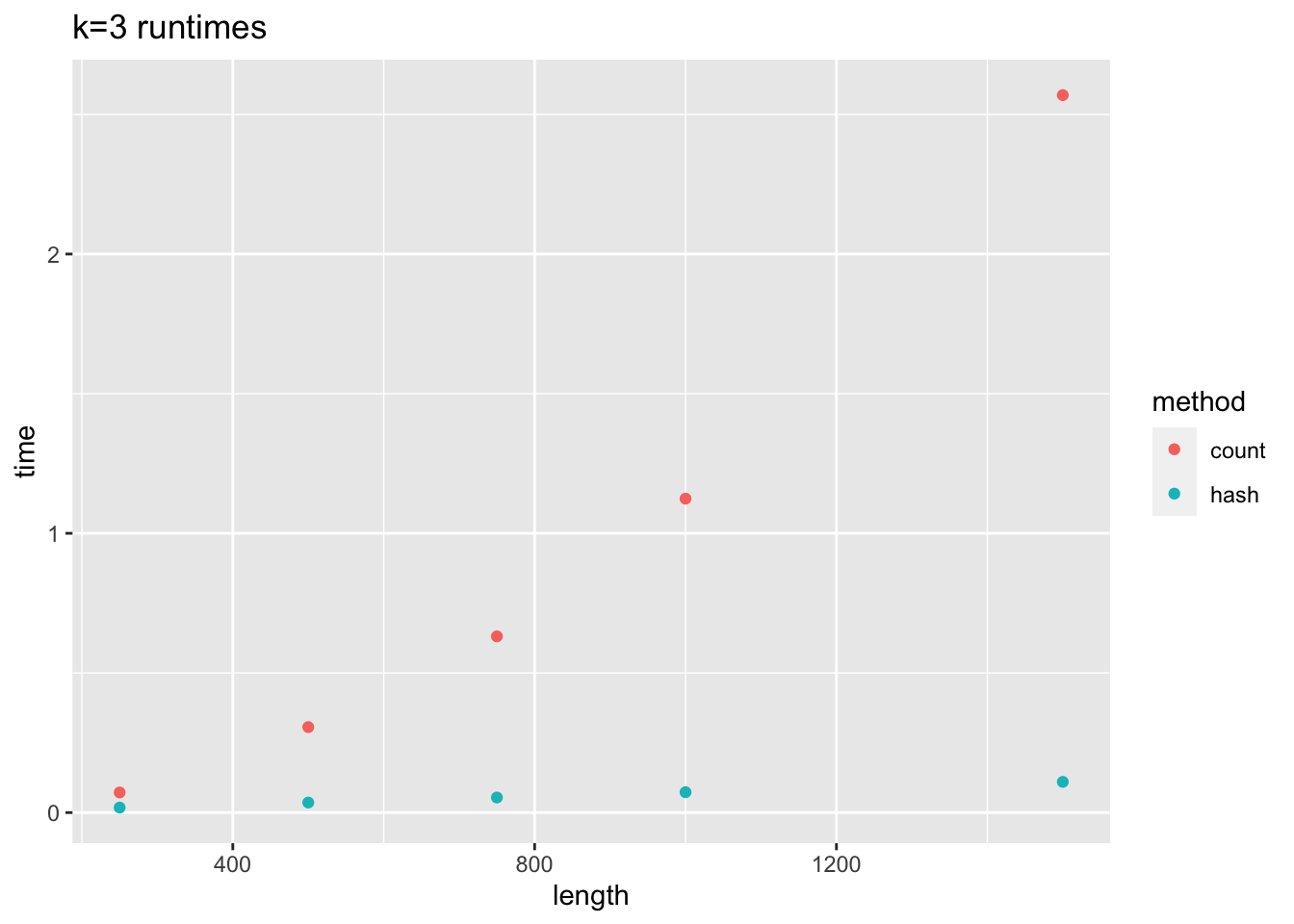
In order to report the frequent words from a genome, a function was created. This function took in a list of k values and the genome text. This returned a table with the frequent words for each k value. The values were then printed to report the output.

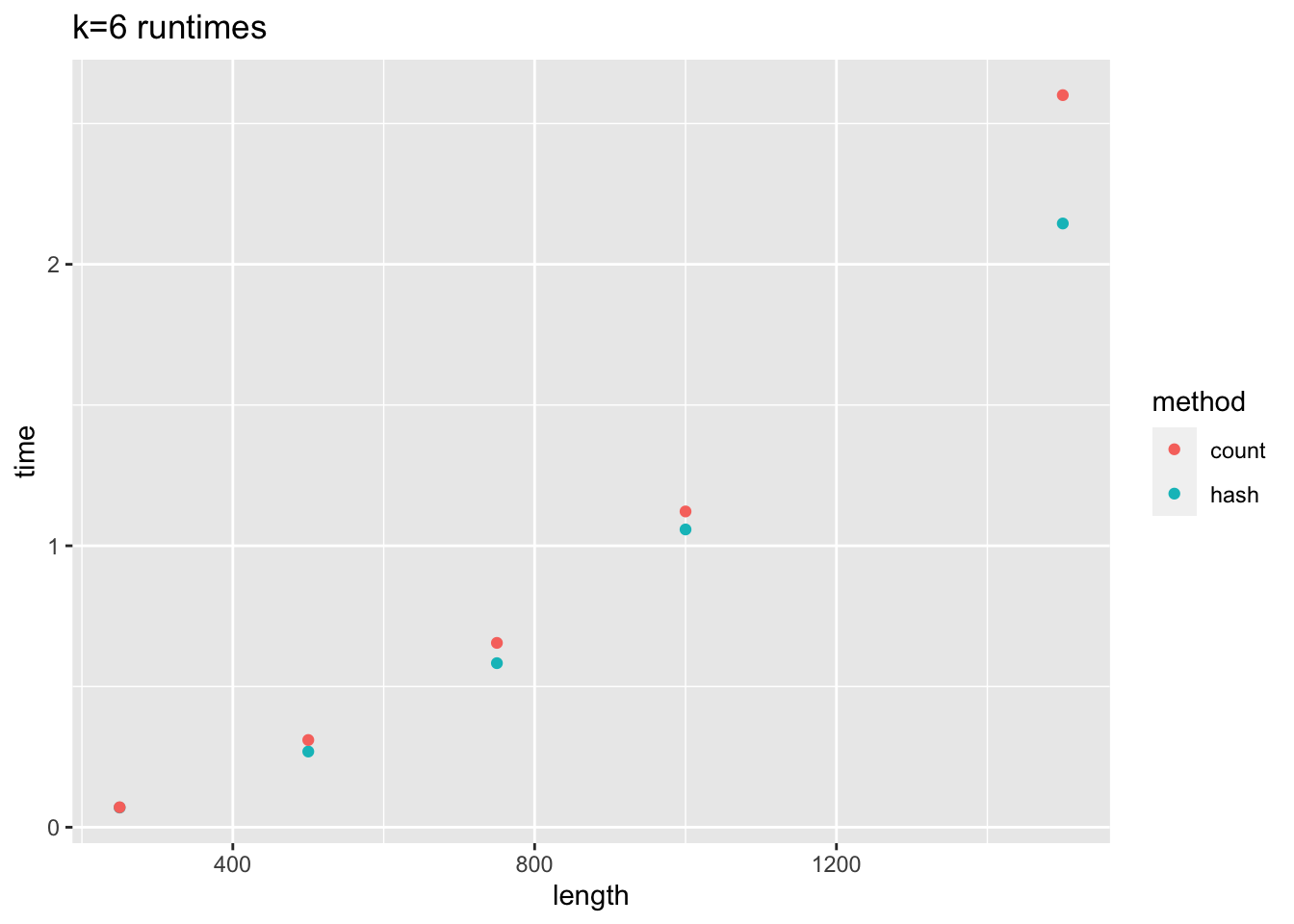
*Task 6*

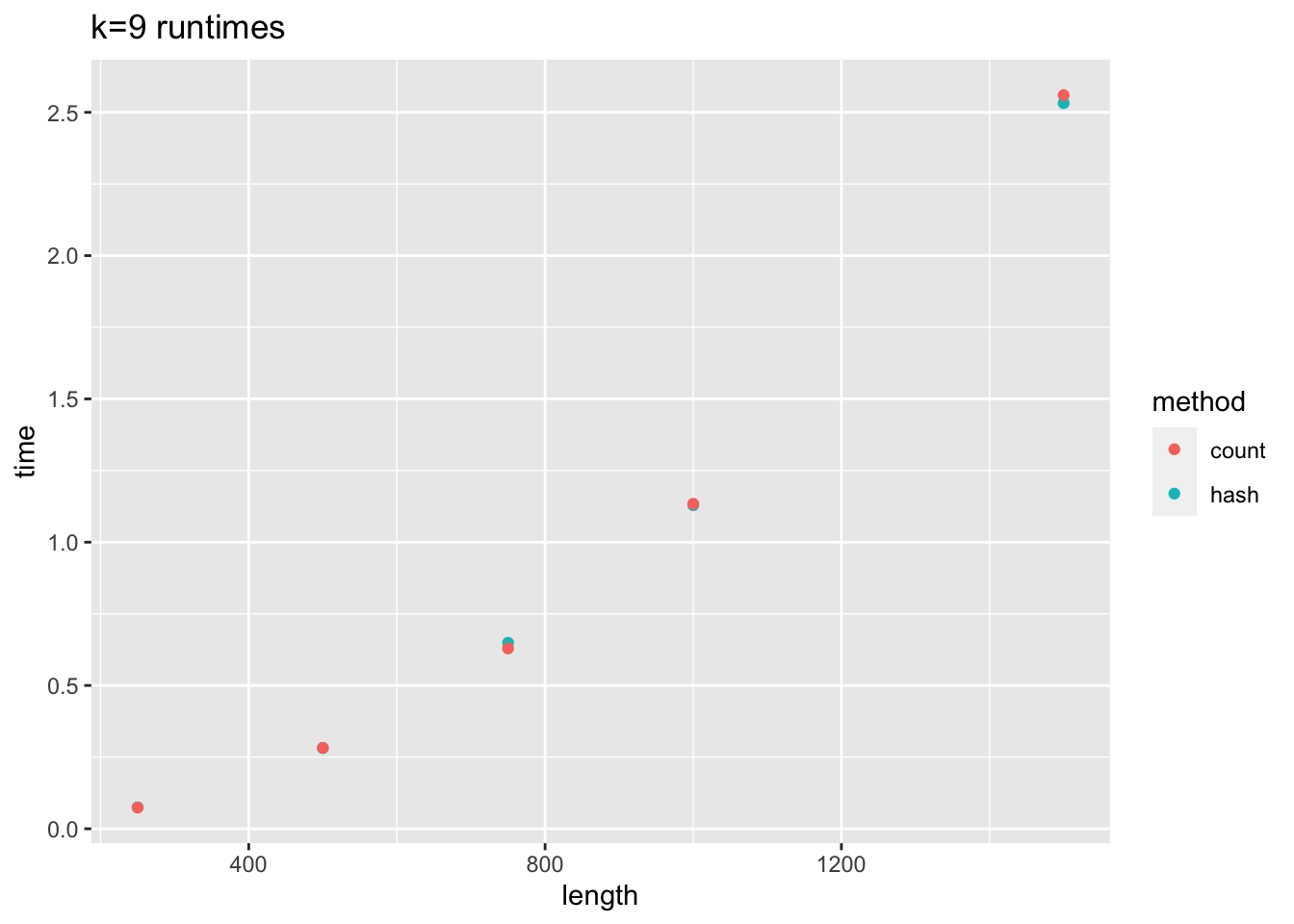
The most frequent words reported by hashing frequent words in SARS-CoV2 genome are searched online for biological insight. “NCBI” + common word was searched in Google for a previous experiment regarding the word.

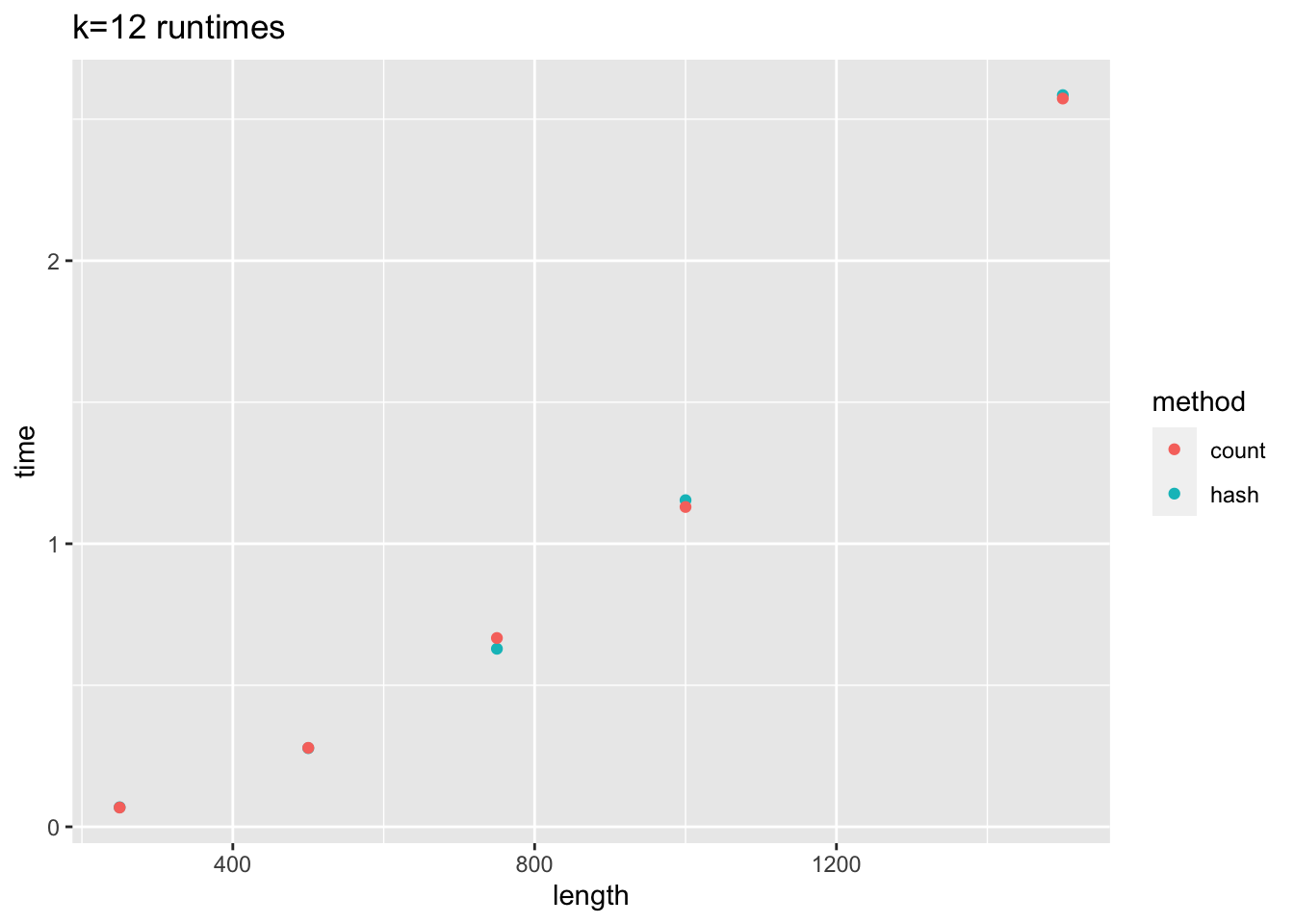
**Results**

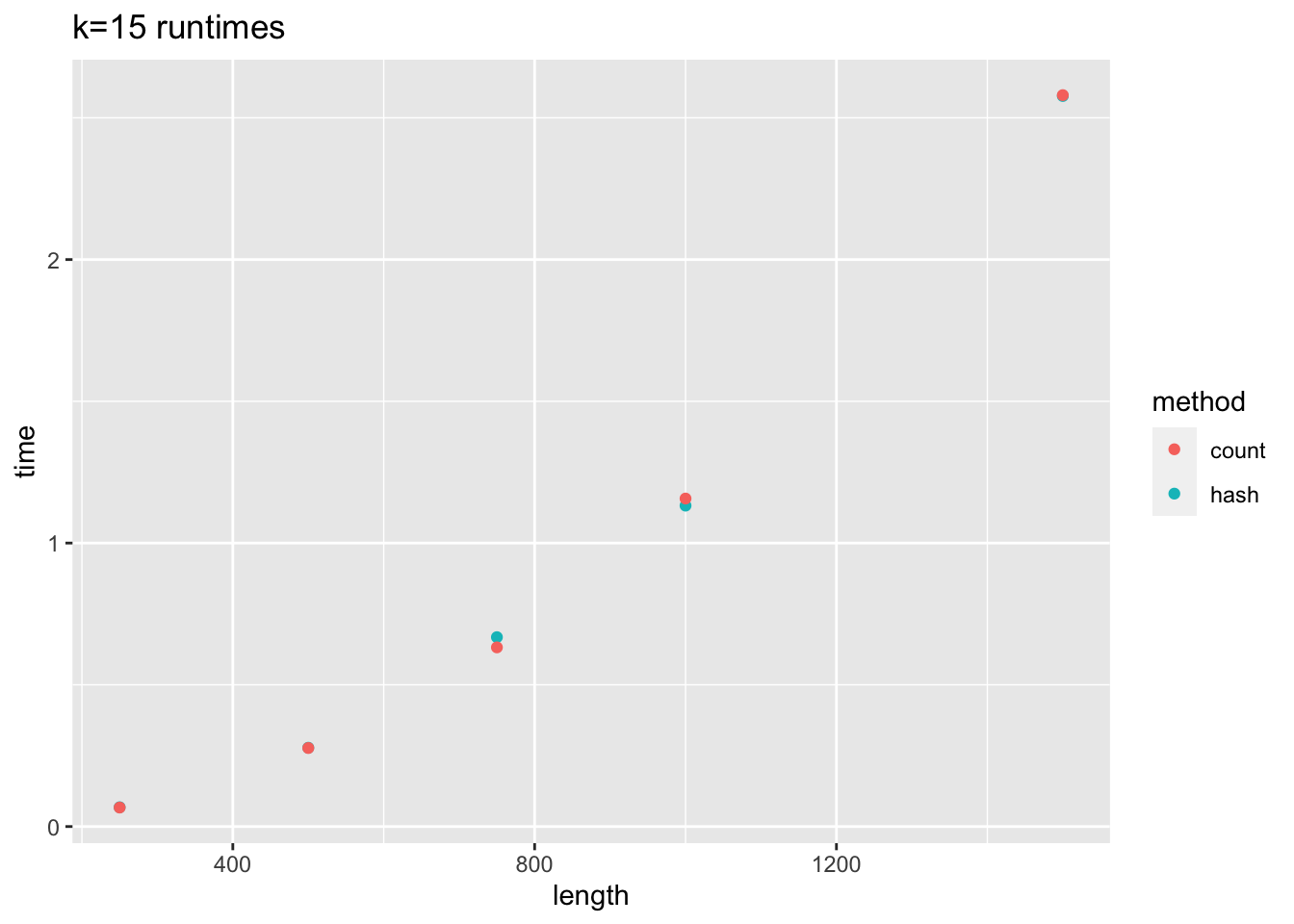
****











*Frequent words from genome.txt*

[1] "k-mer 3 = TTT"

[1] "k-mer 6 = TTGTTA"

[1] "k-mer 9 = TAAACGAAC”

[1] "k-mer 12 = GTTGATGGTGTT"

[1] "k-mer 15 = c(\"ATCAGACAACTACTA\", \"TCAGACAACTACTAT\", \"CAGACAACTACTATT\", \"CAATTATTATAAGAA\", \"AATTATTATAAGAAA\", \"ATTATTATAAGAAAG\", \"TTGCAGAGTGGTTTT\", \"AAAGTTGATGGTGTT\", \"AAGTTGATGGTGTTG\", \"TAAACGAACATGAAA\")"

**Discussion**

The resulting plots comparing the runtimes between hashing and counting show significant differences for k values 3 and 6. The hashing method was faster than counting in these groups of tests. K values 9, 12, and 15 show little to no difference. The hashing method was similar in many of the tests. For k value 9, hashing was slower for length of 750, but faster for length of 1500. For k value 12 hashing was faster for length of 750, but slower for length of 1000. K value 15 also showed hashing slower for length of 750, but faster for length 1000.

The “TTGTTA” hexanucleotide was the most common 6-mer. “TTGTTA” with “TAAAAT” is part of a promoter region in the alpha-amylase gene in Bacillus amyloliquefaciens. (L. U. P. 1984) TAAACGAAC nine-mer is the synthesis starter in transcription regulating sequences in major cases (start sites of S, ORF3, M, ORF7, and N) and subsequence ACGAAC for cases (start site of E) in the coronavirus genome. (W. Z. et al. 2021) The reason this 9-mer is found so many times is because coronavirus has a unique strategy of discontinuous RNA transcriptional synthesis which means that it needs multiple different sites to start RNA transcription. The replication is continuous so it may have a lower frequency. There is not much insight regarding the 12-mer “GTTGATGGTGTT” and all 15-mers, only found one other genome sequence. (Dzinleski)

*The reported frequent words from task 5 were as follows:*

* K-mer 3 “TTT” was found 709 times evenly throughout the genome.
* K-mer 6 “TTGTTA” was found 42 times. Pretty evenly scattered at first and near the end of the genome there are gaps between the clusters. The largest cluster between gaps was near the end and had 7 occurrences starting at 22089.
* K-mer 9 “TAAACGAAC” was found 7 times. 5 at the very end of the sequence starting at 26107 and the last k-mer at 29066
* K-mer 12 “GTTGATGGTGTT” was found three times; two times together at nucleotides 20285 and 20742 and one time before at 14853
* K-mer 15

"ATCAGACAACTACTA"

"TCAGACAACTACTAT"

"CAGACAACTACTATT"

"CAATTATTATAAGAA"

"AATTATTATAAGAAA"

"ATTATTATAAGAAAG"

"TTGCAGAGTGGTTTT"

"AAAGTTGATGGTGTT"

"AAGTTGATGGTGTTG"

"TAAACGAACATGAAA"

All of the 15 length k-mers found had a count of 2.

**Distribution of Work**

|  | David | Ryan |
| --- | --- | --- |
| Task 1 |  | X |
| Task2 | X |  |
| Task 3 |  | X |
| Task 4 | X |  |
| Task 5 | X |  |
| Task 6 | X | X |
| Title | X |  |
| Abstract | X |  |
| Introduction |  | X |
| Methods | X | X |
| Results | X | X |
| Discussion | X | X |

**References**

1. NCBI, N. L. M. (2022). *Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-hu-1, co - nucleotide - NCBI*. National Center for Biotechnology Information. Retrieved February 22, 2023, from https://www.ncbi.nlm.nih.gov/nuccore/NC\_045512.2?report=fasta
2. Online Character Counter: Calculate String Length. Retrieved February 22, 2023 from <https://string-functions.com/length.aspx>
3. CDC (2022). *SARS-CoV-2 Variant Classifications and Definitions*. Centers for Disease Control and Prevention. Retrieved February 22, 2023 from https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html
4. CDC (2022). *COVID Data Tracker*. Centers for Disease Control and Prevention. Retrieved February 22, 2023 from [https://covid.cdc.gov/covid-data-tracker/#datatracker-hom](https://covid.cdc.gov/covid-data-tracker/#datatracker-home)e
5. Huang, Y., Yang, C., Xu, Xf. *et al.* Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* **41**, 1141–1149 (2020). <https://doi.org/10.1038/s41401-020-0485-4>

# Lehtovaara P, Ulmanen I, Palva I. *In vivo transcription initiation and termination sites of an alpha-amylase gene from Bacillus amyloliquefaciens cloned in Bacillus subtilis.* Gene. 1984 Oct;30(1-3):11-6. doi: 10.1016/0378-1119(84)90099-4. PMID: 6210229.

1. Wang X, Zhao Y, Yan F, Wang T, Sun W, Feng N, Wang W, Wang H, He H, Yang S, Xia X, Gao Y. *Viral and Host Transcriptomes in SARS-CoV-2-Infected Human Lung Cells.* J Virol. 2021 Aug 25;95(18):e0060021. doi: 10.1128/JVI.00600-21. Epub 2021 Aug 25. PMID: 34106002; PMCID: PMC8387032.

###### Dzinleski J. Genetic sequence binary factor routines. Jasenko work. Retrieved February 23, 2023 from <https://jasenko.work/>

# Sola I, Almazan F, Zuniga S, Enjuanes L (2015). Annual Reviews. *Continuous and Discontinuous RNA Synthesis in Coronaviruses* Retrieved February 23, 2023 from https://www.annualreviews.org/doi/10.1146/annurev-virology-100114-055218