CS 471 Project 3 Report:
Genome-Scale Comparisons for COVID-19 Strains

Gage Unruh

April 21, 2025

1 System Configuration

- **CPU:** Intel Core i5-12600KF (12th Gen, 10 cores (6P/4E) & 16 logical threads)
- **Clock Rate:** 3.70 GHz (Base Clock)
- **RAM:** 32 GB DDR5 @ 5600 MT/s
- **L2 Cache:** 9.5 MB (9728 KB)
- **L3 Cache:** 20 MB (20480 KB)
- **Motherboard:** Gigabyte B760M DS3H AX (LGA 1700)
- **Operating System:** Windows 11 Home (64-bit) - Build 22631
- **Rust / Cargo Compiler Version:** RustC 1.84.1 / Cargo 1.84.1

2 Quality

2.1 Similarity Matrix

Table 1: For each pair of sequences (s_i, s_j), the value $D[i, j]$ represents the total number of matching base pairs. This is calculated by taking the LCS + best prefix + best suffix

	C-Aus	C-Bra	C-Ind	C-USA	C-Wuh	M-12	M-14K	M-14U	S-03	S-17
C-Aus	29893	29861	29836	29866	29890	2556	2556	2557	23803	23877
C-Bra	29861	29876	29845	29869	29872	2555	2555	2556	23814	23888
C-Ind	29836	29845	29854	29846	29849	2554	2554	2555	23812	23879
C-USA	29866	29869	29846	29882	29879	2556	2556	2557	23811	23885
C-Wuh	29890	29872	29849	29879	29903	2556	2556	2557	23814	23888
M-12	2556	2555	2554	2556	2556	30055	30005	30003	529	529
M-14K	2556	2555	2554	2556	2556	30005	30123	30053	995	995
M-14U	2557	2556	2555	2557	2557	30003	30053	30123	995	995
S-03	23803	23814	23812	23811	23814	529	995	995	29644	29629
S-17	23877	23888	23879	23885	23888	529	995	995	29629	29727

2.2 LCS Length Matrix

Table 2: Longest Common Substring (LCS) Length. Calculated by finding the length of the longest identical sequence shared between each pair, found using a Generalized Suffix Tree (GST).

	C-Aus	C-Bra	C-Ind	C-USA	C-Wuh	M-12	M-14K	M-14U	S-03	S-17
C-Aus	29893	11082	7961	13980	19064	23	23	23	104	104
C-Bra	11082	29876	4620	8896	11082	23	23	23	104	104
C-Ind	7961	4620	29854	7961	7961	23	23	23	104	104
C-USA	13980	8896	7961	29882	23769	23	23	23	104	104
C-Wuh	19064	11082	7961	23769	29903	23	23	23	104	104
M-12	23	23	23	23	23	30055	2890	3182	20	20
M-14K	23	23	23	23	23	2890	30123	3094	20	20
M-14U	23	23	23	23	23	3182	3094	30123	20	20
S-03	104	104	104	104	104	20	20	20	29644	7878
S-17	104	104	104	104	104	20	20	20	7878	29727

2.3 Discussion

What does the similarity matrix D tell you about the relationship between these strains?

Looking at the similarity matrix 1 we can observe higher similarity scores between all strains of the same virus (as we should expect). Covid strains match well with other covid strains, Mers with Mers, and Sars with Sars. Also of note is that the Sars strains analyzed share a significant similarity with the Covid strains. This should also not be surprising as it is well known that Covid-19 (full name: SARS-CoV-2) and SARs both coronaviruses.

Interestingly the MERS viruses also seem to be much more similar to the Covid virsuses than they are to the SARS viruses.

Did computing the longest common substring first help reduce computation time for you

Yes this version seems to run much faster (especially when compiled) than previous project. Though it is worth noting that I changed some code from the McCreight algorithm project, and also did not test the speed of the code without computing the LCS first, so there are confounding factors at play.

3 Performance

3.1 Total Time

The total time spent on each major computational task across all 55 pairwise comparisons was:

- **Total GST Construction & LCS Search Time:** 4.450 s
- **Total Alignment Time (Prefixes & Suffixes):** 267.922 s

The sum of these components is approximately 272.372 s. The total execution time was 274.998 s. The difference (˜2.6 s) is from time spent on file I/O, and printing results.

The results shows that, even with the LCS optimization, the alignment phase was the most computationally intensive part, consuming 97.6% of the total computational time (267.922 s / 272.372 s). The GST/LCS part was relatively fast. The timings for each non-trivial pair are shown in Table 3.

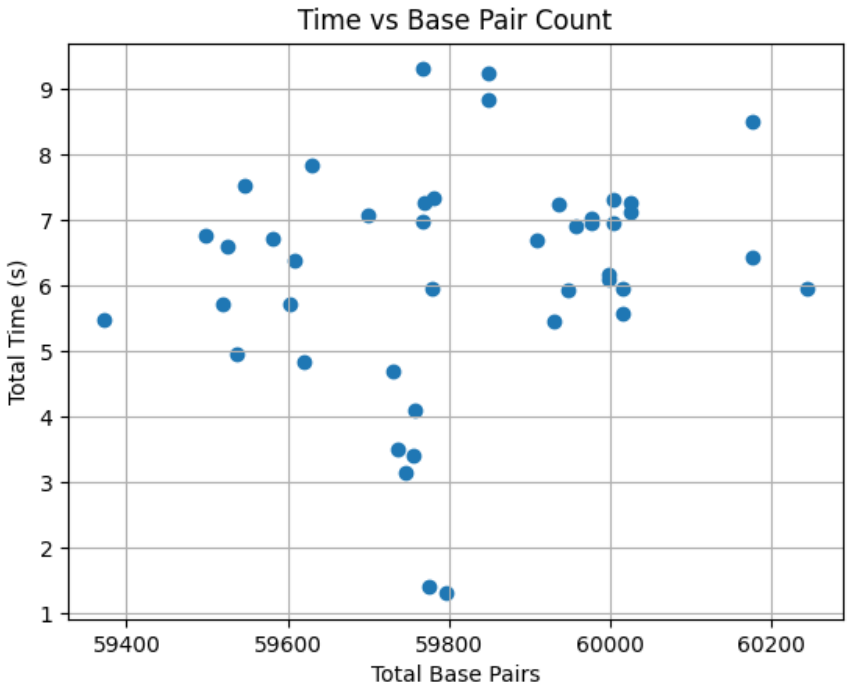

Figure 1: Plot showing the relationship between the total number of base pairs in the pair of sequences against the amount of time it took to calculate thar pairs entry in the similarity matrix. 1

There only seems to be a slight correlation between the total number of base pairs processed and the time taken for computation of a given pairs entry in the similarity matrix. Any trends that could be observed would likely be more prominent given a dataset with a wider range of base pair lengths.

3.2 Memory

For the given sequences (∼30k bp) the program uses a constant ∼1.3GB of memory during whilst calculating each entry of the similarity matrix.

Table 3: Detailed Computation Times for each Pair of Sequences

`Sequence 1` vs `Sequence 2`	GST Time (ms)	Align Time (s)
`C-Aus` vs `C-Bra`	188.359	7.066
`C-Aus` vs `C-Ind`	135.464	3.026

Continued on next page

Sequence 1	vs	Sequence 2	GST Time (ms)	Align Time (s)
C-Aus	vs	C-USA	217.261	1.206
C-Aus	vs	C-Wuh	296.351	1.015
C-Aus	vs	M-12	58.431	5.868
C-Aus	vs	M-14K	60.685	5.905
C-Aus	vs	M-14U	61.556	5.514
C-Aus	vs	S-03	64.134	4.892
C-Aus	vs	S-17	62.812	4.775
C-Bra	vs	C-Ind	134.925	4.553
C-Bra	vs	C-USA	158.624	3.949
C-Bra	vs	C-Wuh	199.463	5.750
C-Bra	vs	M-12	60.676	5.391
C-Bra	vs	M-14K	62.479	6.038
C-Bra	vs	M-14U	61.732	6.108
C-Bra	vs	S-03	63.642	5.667
C-Bra	vs	S-17	61.222	5.656
C-Ind	vs	C-USA	150.899	3.356
C-Ind	vs	C-Wuh	163.562	3.256
C-Ind	vs	M-12	69.568	6.634
C-Ind	vs	M-14K	64.102	6.893
C-Ind	vs	M-14U	61.576	6.970
C-Ind	vs	S-03	62.719	6.703
C-Ind	vs	S-17	61.560	6.655
C-USA	vs	C-Wuh	315.995	1.086
C-USA	vs	M-12	61.035	7.183
C-USA	vs	M-14K	65.059	7.244
C-USA	vs	M-14U	62.826	6.883
C-USA	vs	S-03	62.566	6.539
C-USA	vs	S-17	66.519	6.327
C-Wuh	vs	M-12	63.277	6.855
C-Wuh	vs	M-14K	59.268	7.197
C-Wuh	vs	M-14U	62.811	7.064
C-Wuh	vs	S-03	64.555	7.463
C-Wuh	vs	S-17	61.154	7.771
M-12	vs	M-14K	97.249	6.346
M-12	vs	M-14U	98.784	8.415
M-12	vs	S-03	63.498	7.002
M-12	vs	S-17	61.391	7.271
M-14K	vs	M-14U	108.511	5.838
M-14K	vs	S-03	70.252	6.902
M-14K	vs	S-17	62.544	8.779
M-14U	vs	S-03	66.930	9.235
M-14U	vs	S-17	63.102	9.185
S-03	vs	S-17	146.223	5.336

A Additional Exploratory Visualizations

The following figures were mostly the result of me just wanting to practice data viz in python, but they might still add something of value, so I thought I would add them as an appendix

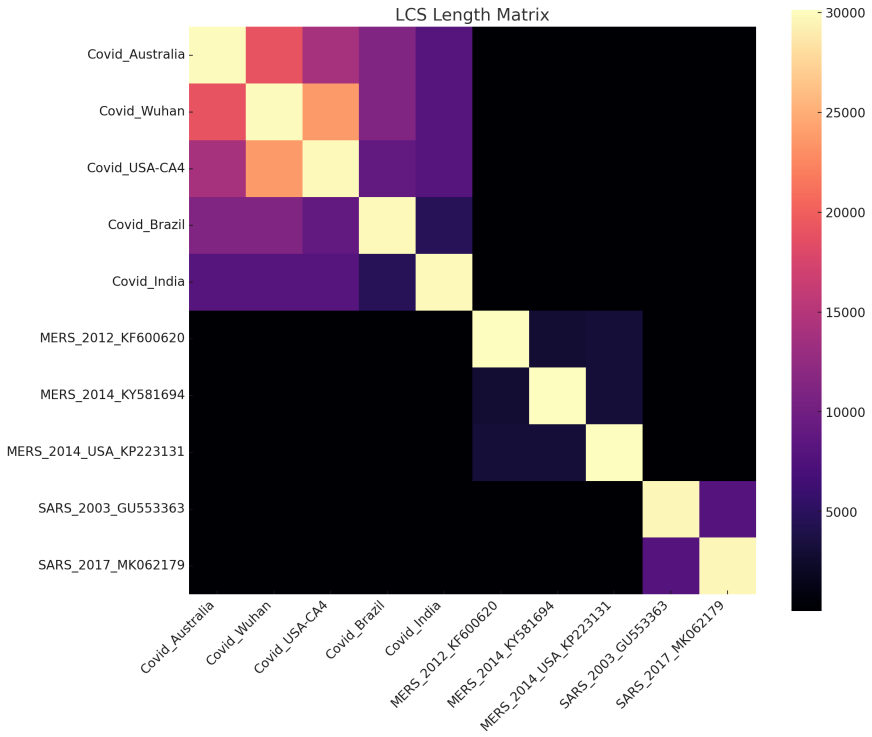

Figure 2: Color-coded heatmap visualization of the similarity matrix D (Table 1).

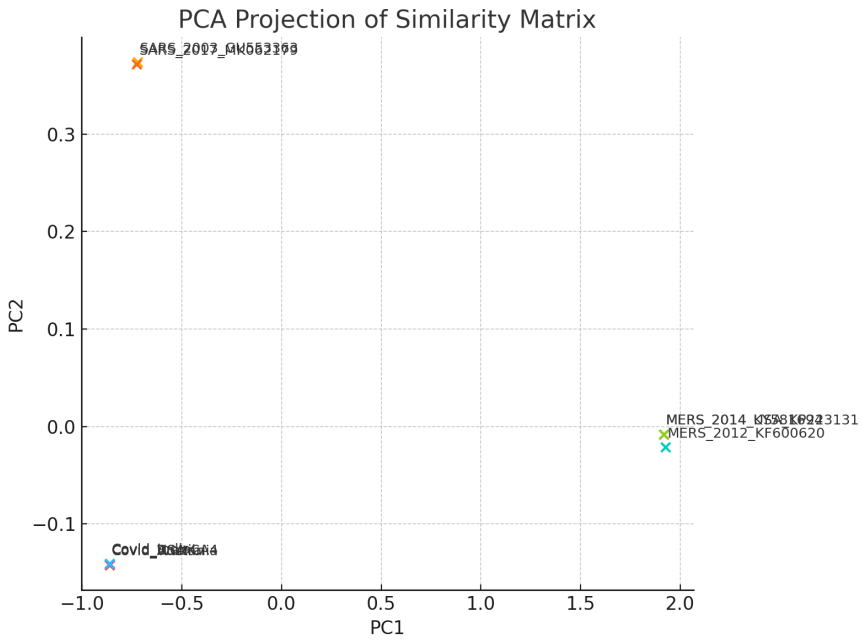

Figure 3: Example: Principal Component Analysis (PCA) based on sequence similarity.

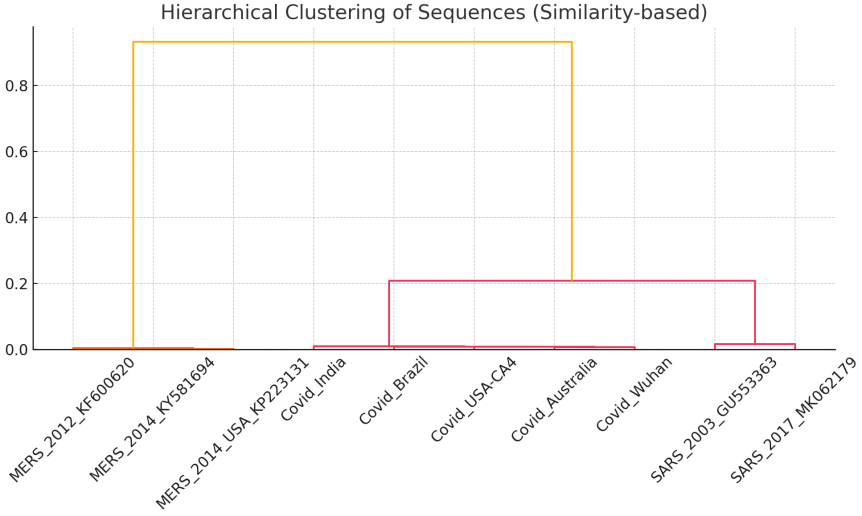

Figure 4: Dendrogram showing hierarchical clustering based on the similarity matrix D.

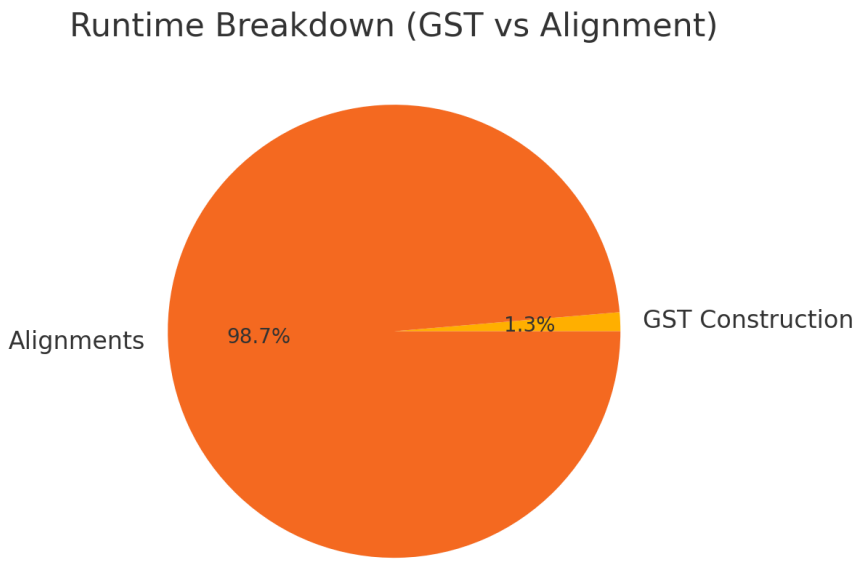

Figure 5: Pie chart illustrating the relative time spent on GST construction vs. Alignment tasks.

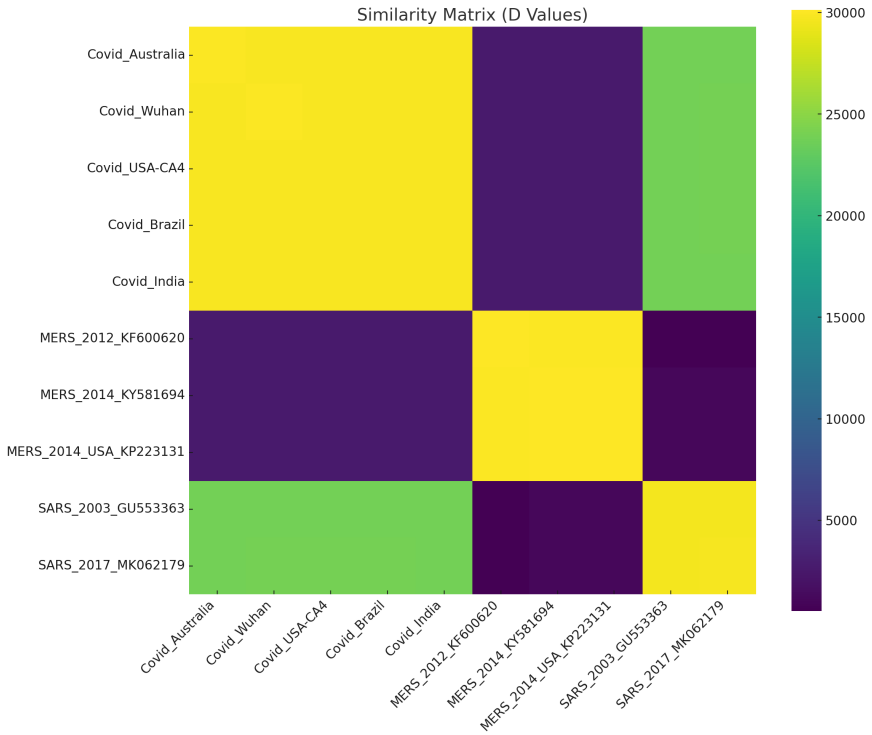

Figure 6: Similarity matrix color coded on value thresholds.