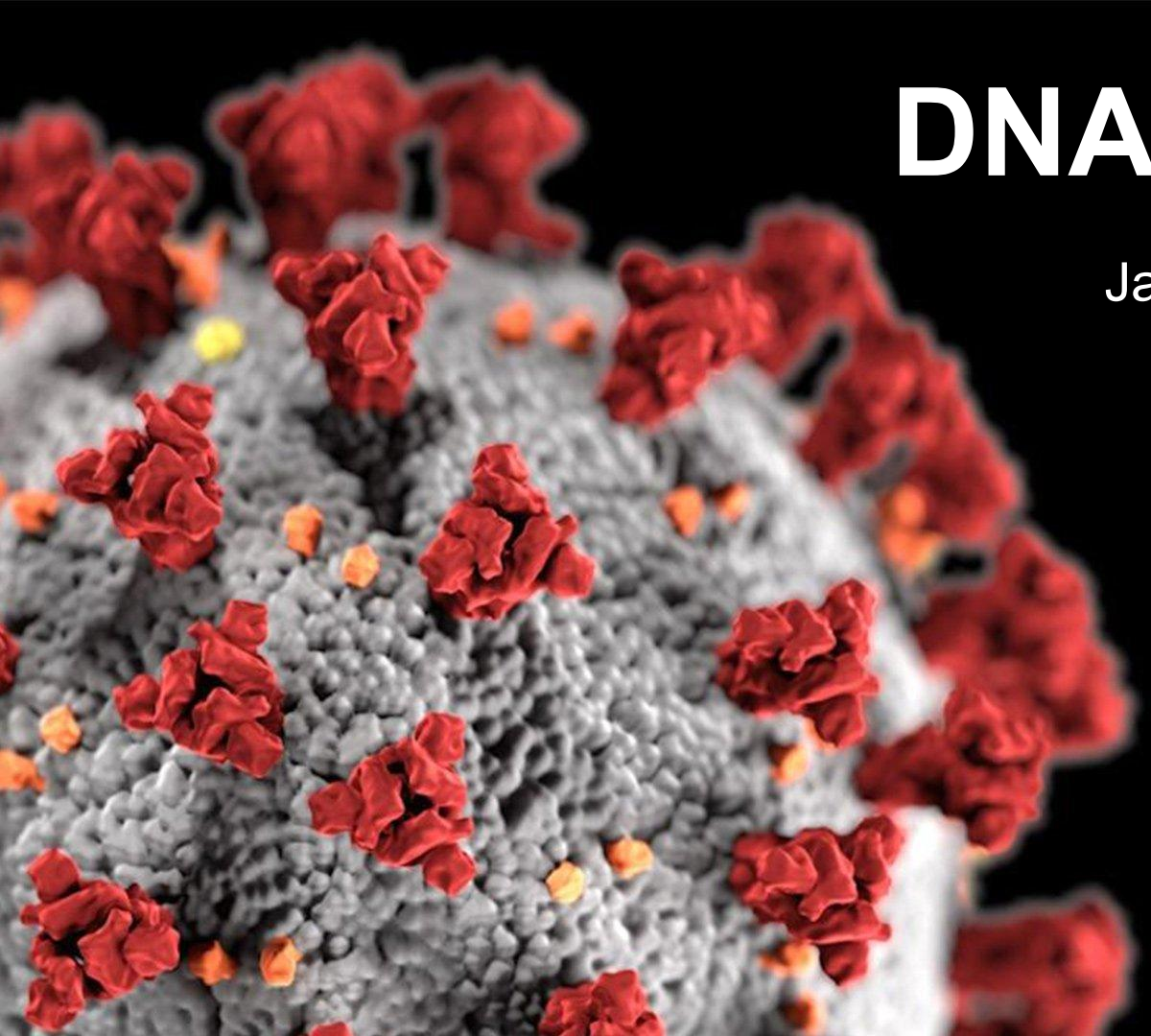
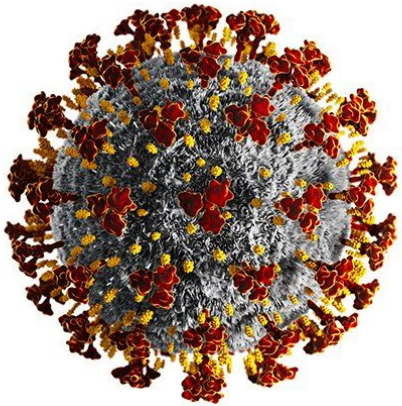


DNA Detectives

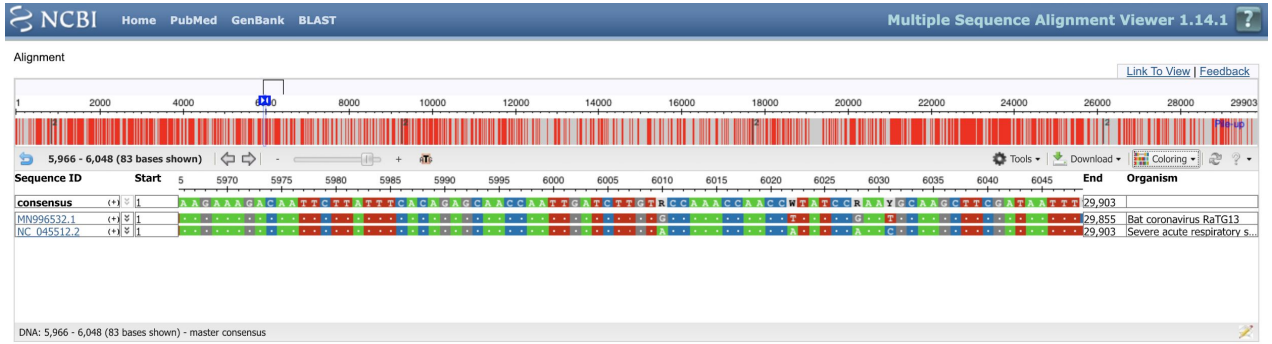
James, Daniel, Irith, Preston

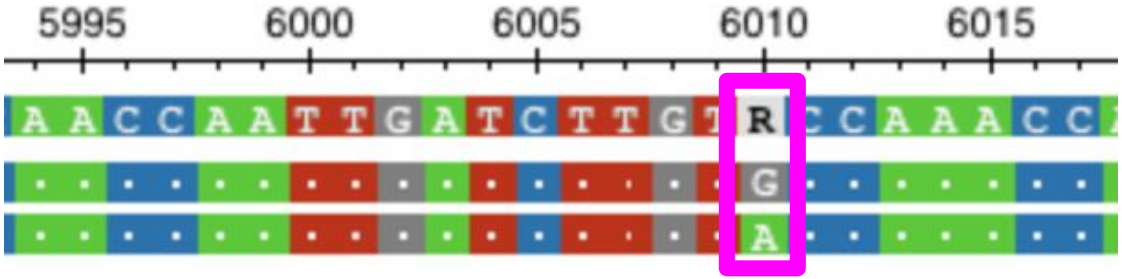
Why?

- COVID-19, the disease caused by the virus SARS-CoV2, has had a widespread economic, political, and social impact throughout the world
- We hope to use our training in AI applications to help predict future hot-spots of this disease, and better understand viral evolution



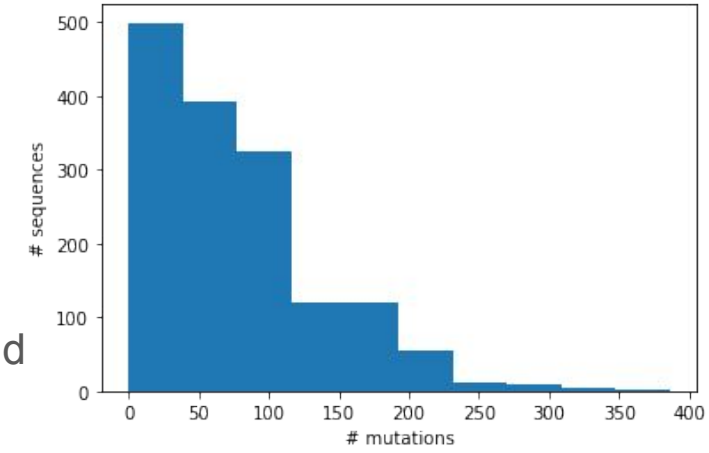
The Dataset





- We used a program called **NCBI** to visualize our data
- Genomic sequences of **1,538** different strains of SARS-CoV-2
- **16** different countries in **3** different regions around the world
- Each sequence was **29,903** bases long

- The **magenta** location represents a **mutation**



Preprocessing

- Categorize countries into 3 regions
  - North America
  - Oceania
  - Asia

```
countries_to_regions_dict = {
        'Australia': 'Oceania',
        'China': 'Asia',
        'Hong Kong': 'Asia',
        'India': 'Asia',
        'Nepal': 'Asia',
        'South Korea': 'Asia',
        'Sri Lanka': 'Asia',
        'Taiwan': 'Asia',
        'Thailand': 'Asia',
        'USA': 'North America' ,
        'Viet Nam': 'Asia'
}
```

- Balance the dataset
  - Our dataset should contain an equal number of genome sequences from North America, Oceania, and Asia

Feature Extraction

- Converted strings of genome sequences to boolean matrix indicating whether there was an A, C, G, T, or missing nucleotide (-) at each location; a 1 (true) means that the sequence has the given base in the given location; a 0 (false) means that it doesn't
- Only 'care about' locations with differences (mutations)

```
AGCCTTGTCATCCGTATC-TTTCAA----
AGCCTTGTCATCCGTATC-TTTCA-----
-GCCTTGTCATCCGTATC-TTTCAACG--
--CCTTGTCATCCGTATC-TTTCAACGTG
--CCTTGTCATCCGTATC-TTTCAAC---
---CTTGTCATCCGTATCTTTTCAAC---
---CTTGTCATCCGTATC-T---------
----TTGTCATCCGTATC-TT--------
------GTCATCCGTATC-TTTCAACGTG
------GTCATCCGTATC-TTTCAACGTG
--------CATCCGTATC-TTTCAA----
--------CATCCGTATC-TTTCA-----
---------ATCCGTATC-TTTCAACGTG
```

	0_A	0_T	0_G	0_C	0_-	1_A	1_T	1_G	1_C	1_-	2_A	2_T	2_G	2_C	2_-	3_A	3_T	3_G	3_C	3_-
0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
2	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
3	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
4	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0

Raw data: genome sequences

Input matrix: row = sequence number; column = <location>_<base>

	0_A	0_T	0_G	0_C	0_-	...	29902_A	29902_T	29902_G	29902_C	29902_-
4	1	0	0	0	0	...	1	0	0	0	0
5	1	0	0	0	0	...	1	0	0	0	0
6	1	0	0	0	0	...	1	0	0	0	0
7	1	0	0	0	0	...	1	0	0	0	0
9	1	0	0	0	0	...	1	0	0	0	0
10	1	0	0	0	0	...	1	0	0	0	0
11	1	0	0	0	0	...	1	0	0	0	0
14	1	0	0	0	0	...	1	0	0	0	0
15	1	0	0	0	0	...	0	0	0	0	1

Genome sequence #15 does not have an A at position 29902, whereas the other genome sequences do have an A

Results and Accuracy

- We used the **multinomial** class of **logistic regression** model because we have more than two categories (North America, Oceania, and Asia)
- **95%** accuracy
- Our model had trouble with some lineages actually from Asia being predicted as from North America, but lineages from Asia and Oceania were never confused

Accuracy: % 94.92753623188406

	Asia predicted	North America predicted	Oceania predicted
Asia true	23	8	0
North America true	2	210	3
Oceania true	0	1	29

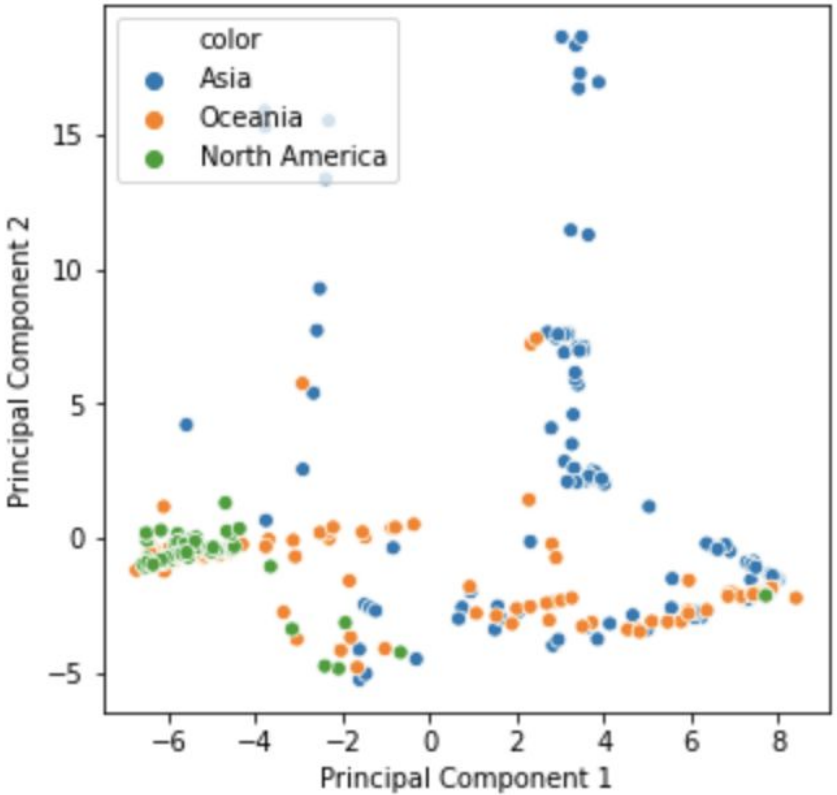
Results and Accuracy, cont.

- Similar to before, except that we predicted based on the **country** rather than **region** of origin. Limited data only to countries with >15 sequences.
- Again, we used the multinomial class of the logistic regression model, as there are more than 2 categories (Australia, China, Hong Kong, India, Taiwan, Thailand, USA)
- **95%** accuracy

Accuracy: 94.52554744525547%

	Australia predicted	China predicted	Hong Kong predicted	India predicted	Taiwan predicted	Thailand predicted	USA predicted
Australia true	32	0	0	0	0	0	4
China true	0	2	1	0	0	0	3
Hong Kong true	0	0	2	0	0	0	1
India true	0	0	0	2	0	0	1
Taiwan true	0	1	0	0	2	0	1
Thailand true	1	0	0	0	0	2	1
USA true	0	0	1	0	0	0	217

Unsupervised Learning: PCA



Next Steps and Future Directions

- Look into mutations that are regional-specific and study which proteins are changing as a result and how this affects the virulence of the particular lineage
- Collect more data (DNA sequences) to determine with more accuracy the origin of the virus (get data from more countries)
- https://nextstrain.org/ncov/global