Mutations in SARS-CoV 2

Aim: The sole purpose of this report is to stress the incidence of mutations in coronavirus and thereby emphasizing the use of ***Gujarat Biotechnology Research Centre(GBRC)*** decoded sequence of Coronavirus mutates in India for further experiments

Just like any other organism evolves to survive, coronavirus too has mutated itself.

Just look at a snippet of the bat virus RNA nucleotide sequence the human virus was derived from...

AAAATCAAAGCTTGTGTTGAAGAAGTTACAACAACTCTGGAAGAAACTAAGTT

...and a snippet from the human COVID-19's RNA nucleotide sequence...

AAAATTAAGGCTTGCATTGATGAGGTTACCACAACACTGGAAGAAACTAAGTT

...clearly, the coronavirus has changed its internal structure to adapt to the new species of their host (to be more precise, about 20% of the internal structure of the coronavirus was mutated), but maintained enough such that it is still true to its origin species.

In fact, research has shown COVID-19 has mutated repeatedly in ways to boost its survival. In our fight to defeat the coronavirus, we need to find not just how the virus can be destroyed, but how the virus mutates and how those mutations can be addressed.

In this report, the following concepts are covered

- Surface-level explanation of what RNA nucleotide sequences are
- Use of K-Means to create genome information clusters
- Use of PCA to visualize the clusters

Genome sequencing, commonly compared to "decoding," is the process of analyzing deoxyribonucleic acid (DNA) taken from a sample. Within every normal cell are 23 pairs of chromosomes, structures that house DNA.

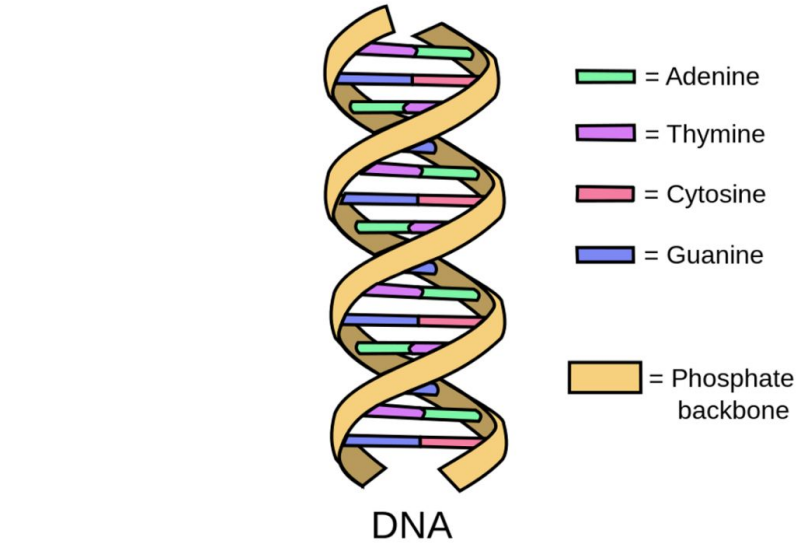


Figure 1 DNA and its nucleotides

The curled double helix structure of DNA allows it to unwind into a ladder shape. This ladder is made out of paired chemical letters called bases. There are only four of these present in DNA: adenine, thymine, guanine, and cytosine. Adenine joins only with thymine, and guanine joins only with cytosine. These bases are represented with A, T, G, and C, respectively.

These bases form a code of sorts that instructs the organism how to construct proteins — it is the DNA that essentially controls how the virus acts.

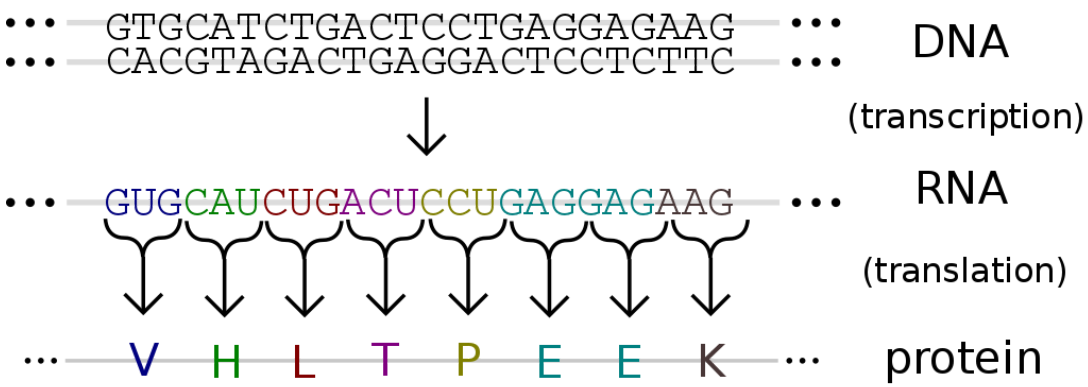


Figure 2 Process of Transcription

Using specialized equipment, including sequencing instruments and specialized tags, the DNA sequences of specific fragments are revealed. Information obtained from this undergoes further analysis and comparison to allow researchers to identify changes in genes, associations with diseases and phenotypes, and identify potential drug targets.

The genome sequence, a long string of ‘A’s, ‘T’s, ‘G’s, and ‘C’s, represents how the organism reacts to its environment. Mutations to an organism are created by altering the DNA. Looking at the genome sequence is a strong way to analyze coronavirus mutations.

Get to know the data.

The data, which can be found on Kaggle here, looks like this:

	query acc.ver	subject acc.ver	% identity	alignment length	mismatches	gap opens	q. start	q. end	s. start	29882.2	evalue	bit score
0	MN997409.1	MT020881.1	99.990	29882	3	0	1	29882	1	29882	0.0	55166
1	MN997409.1	MT020880.1	99.990	29882	3	0	1	29882	1	29882	0.0	55166
2	MN997409.1	MN985325.1	99.990	29882	3	0	1	29882	1	29882	0.0	55166
3	MN997409.1	MN975262.1	99.990	29882	3	0	1	29882	1	29882	0.0	55166
4	MN997409.1	LC522974.1	99.993	29878	2	0	4	29881	1	29878	0.0	55164
...	...	...	...	...	...	...	...	...	...	...	...	...
257	MN997409.1	AY283796.1	79.325	1925	357	35	19	1923	3	1906	0.0	1312
258	MN997409.1	AY282752.2	82.304	17716	2948	169	3956	21577	3868	21490	0.0	15175
259	MN997409.1	AY282752.2	80.063	5417	988	68	22539	27910	22414	27783	0.0	3936
260	MN997409.1	AY282752.2	90.189	1641	142	12	28257	29882	28088	29724	0.0	2121
261	MN997409.1	AY282752.2	79.305	1928	358	35	16	1923	1	1907	0.0	1312

262 rows × 12 columns

Each one of the rows represents one mutation of the bat virus. The coronavirus has already created 262 mutations (and counting) of itself to increase survival rates.

Some important columns:

query acc.ver represents the original virus identifier.

subject acc.ver is the identifier for a virus mutation.

% identity represents what percent of the sequence is the same as the original virus.

alignment length represents how many items in the sequence are the same, or aligned.

mismatches represent the number of items that the mutation and the original differ on.

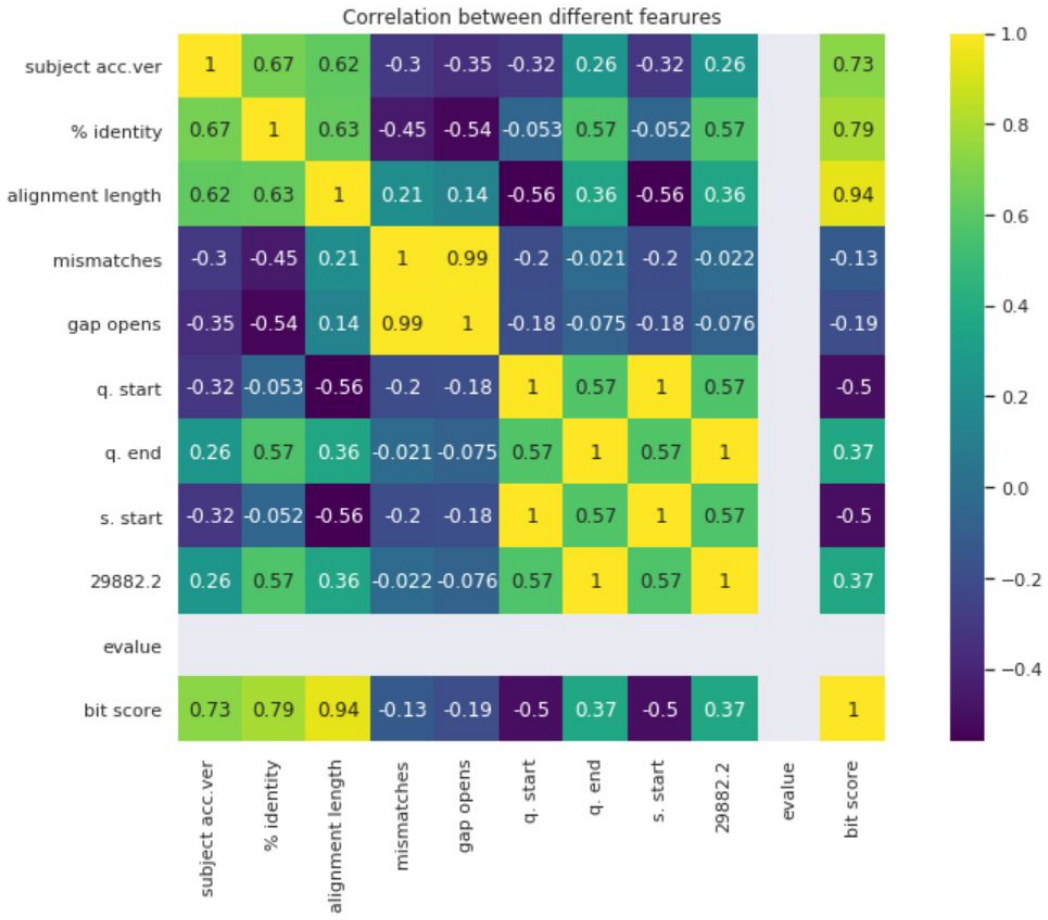
bit score represents a measure to represent how good an alignment is; the higher the score, the better the alignment.

Some statistical measures of each of the columns

	subject acc.ver	% identity	alignment length	mismatches	gap opens	q. start	q. end	s. start	29882.2	evalue	bit score
count	263.000000	263.000000	263.000000	263.000000	263.000000	263.000000	263.000000	263.000000	263.000000	263.0	263.00000
mean	35.790875	86.064958	10711.114068	919.235741	57.821673	11295.684411	21970.517490	11212.646768	21888.837262	0.0	14240.34981
std	24.874763	7.609654	10530.955700	1085.784789	60.654416	12022.056513	10652.171126	11963.759134	10619.704332	0.0	19226.72192
min	0.000000	77.559000	1603.000000	0.000000	0.000000	1.000000	1923.000000	1.000000	1672.000000	0.0	1011.00000
25%	16.000000	80.048000	1925.000000	142.000000	12.000000	16.000000	21577.000000	1.050000	21489.000000	0.0	2101.00000
50%	32.000000	82.304000	5417.000000	359.000000	35.000000	3956.000000	27910.000000	3875.000000	27783.000000	0.0	3936.00000
75%	49.000000	90.189000	17716.000000	989.000000	68.000000	22539.000000	29875.500000	22429.000000	29729.000000	0.0	15175.00000
max	99.000000	100.000000	29882.000000	2952.000000	172.000000	28257.000000	29882.100000	28137.000000	30256.000000	0.0	55182.00000

Looking at the % identity column, it is interesting to see the minimum alignment percent a mutation has with the original virus — about 77.6 percent. The rather large standard deviation of 7 percent for % identity means that there is a wide range of mutation. This is supported by a massive standard deviation in bit score — the standard deviation is larger than the mean!

A good way to visualize data is through a correlation heatmap. Each cell represents how correlated one feature is with another.



A lot of the data is highly correlated with each other. This makes sense since most of the measures are variations of each other. One thing to take note of is the alignment length's high correlation with bit score.

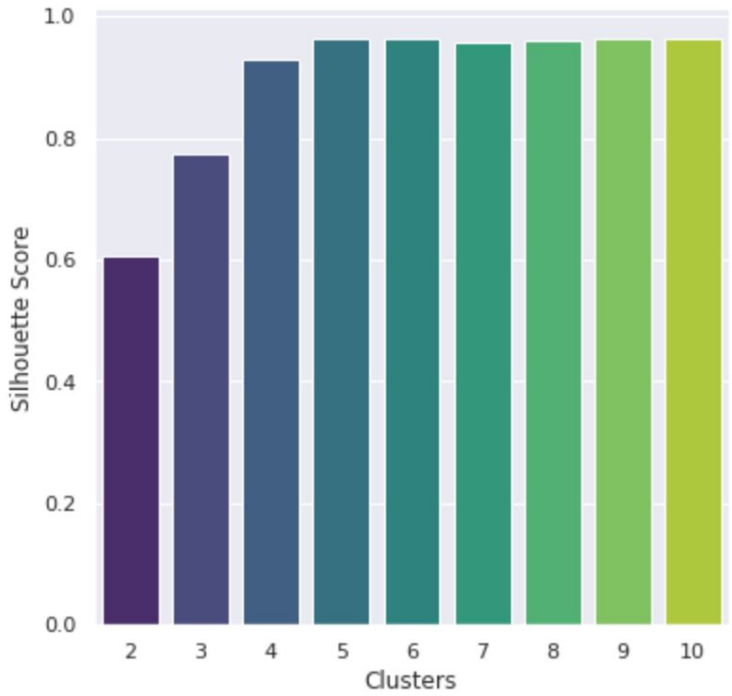
Using K-Means to Create Mutation Clusters

The goal of our K-Means is to find clusters of mutations, so we can derive insights on the nature of the mutations and how to address them.

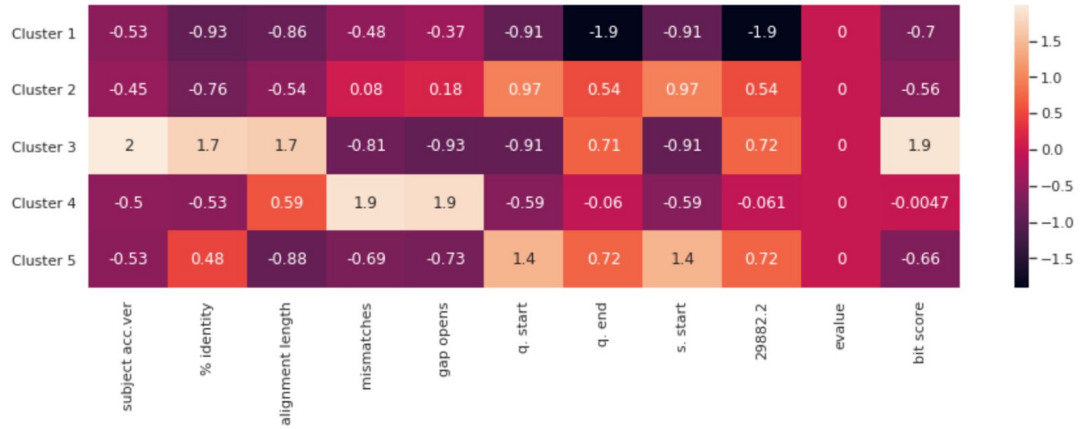
However, we still need to choose the number of clusters k. While this is as simple as plotting out the points in two dimensions, this is unachievable in higher dimensions (if we want to retain the most information). Methods like the elbow method to choose k are subjective and inaccurate, so instead, we will use the silhouette method.

The silhouette method is a score given to k clusters on how well the clusters suit the data.

```
For n_clusters = 2 The average silhouette_score is : 0.6051456304424345
For n_clusters = 3 The average silhouette_score is : 0.7737574709178652
For n_clusters = 4 The average silhouette_score is : 0.9304299438199628
For n_clusters = 5 The average silhouette_score is : 0.9626586534510194
For n_clusters = 6 The average silhouette_score is : 0.9625756535242548
For n_clusters = 7 The average silhouette_score is : 0.9585239650538647
For n_clusters = 8 The average silhouette_score is : 0.9614207899227608
For n_clusters = 9 The average silhouette_score is : 0.9626169262248865
For n_clusters = 10 The average silhouette_score is : 0.9644636830559539
```



It seems that 5 clusters seems to be the best for the data. Now, we can determine the cluster centers. These are the points in which each cluster is centered around, and represent a numerical evaluation of (in this case) the 5 main types of mutations.



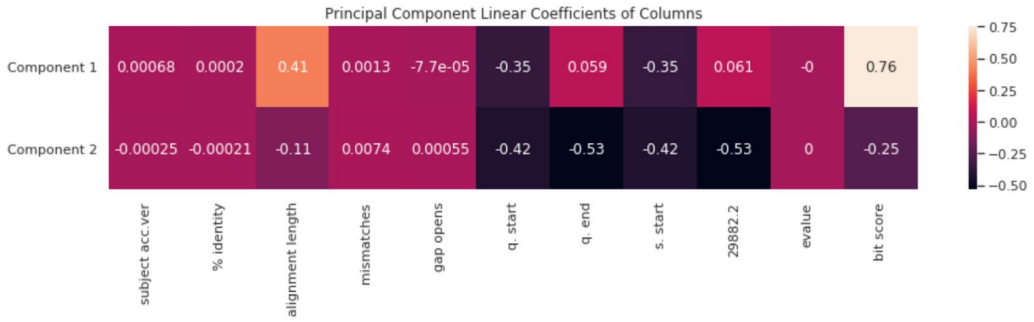
This heatmap represents each cluster's attributes, by column. Because the points were scaled, the actual annotated values do not quantitatively mean anything. However, scaled values in each column can be compared. You can get a visual sense for the relative attributes of each of the mutation clusters. If scientists were to develop a vaccine, it should address these main clusters of virii.

PCA for Cluster Visualization

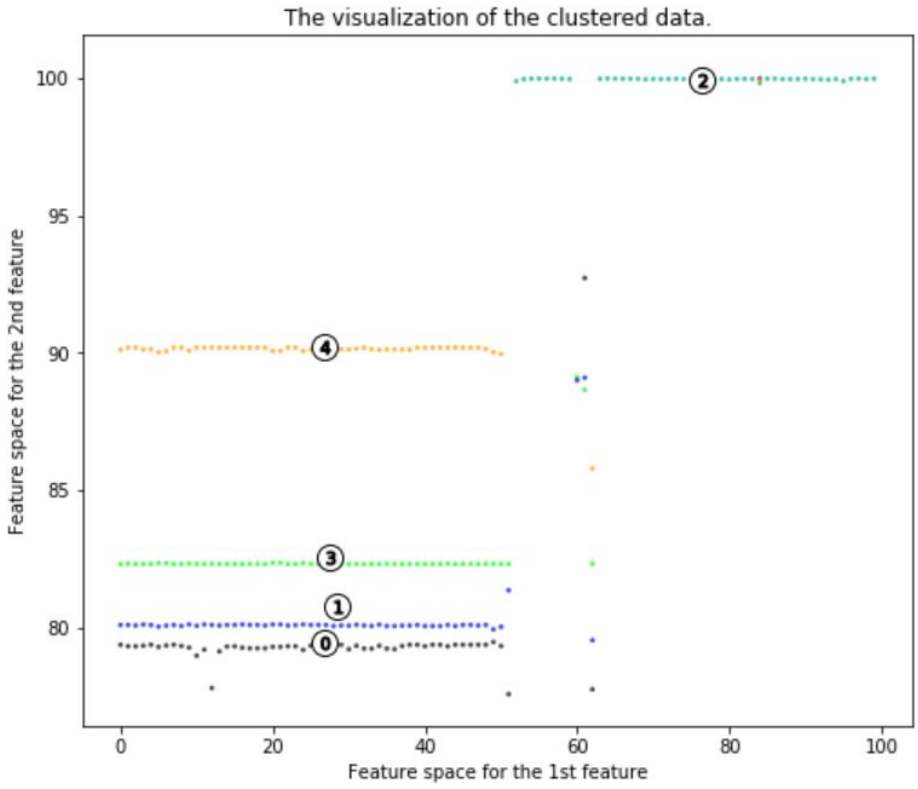
PCA, or Principal Component Analysis, is a method of dimensionality reduction. It selects orthogonal vectors in multidimensional space to represent axes, such that the most information (variance) is retained.

With popular Python library sklearn, implementing PCA can be done in two lines. First, we can check the explained variance ratio. This is the percent of statistical information that is retained from the original dataset. The explained variance ratio, in this case, is 0.9838548580740327, which is astronomically high! We can be assured that whatever analyses we take from PCA will be true to the data.

Each new feature (principal component) is a linear combination of several other columns. We can visualize how important a column is to one of the two principal components with a heatmap.



It is important to understand what having a high value in the first component means — in this case, it is characterized by having a higher alignment length (is closer to the original virus), and component 2 is largely characterized by having a shorter alignment length (mutated farther from the original value). This is also reflected by the larger difference in bit score.



It is clear that there are 5 main strands of the virus mutation. We can take away lots of insights.

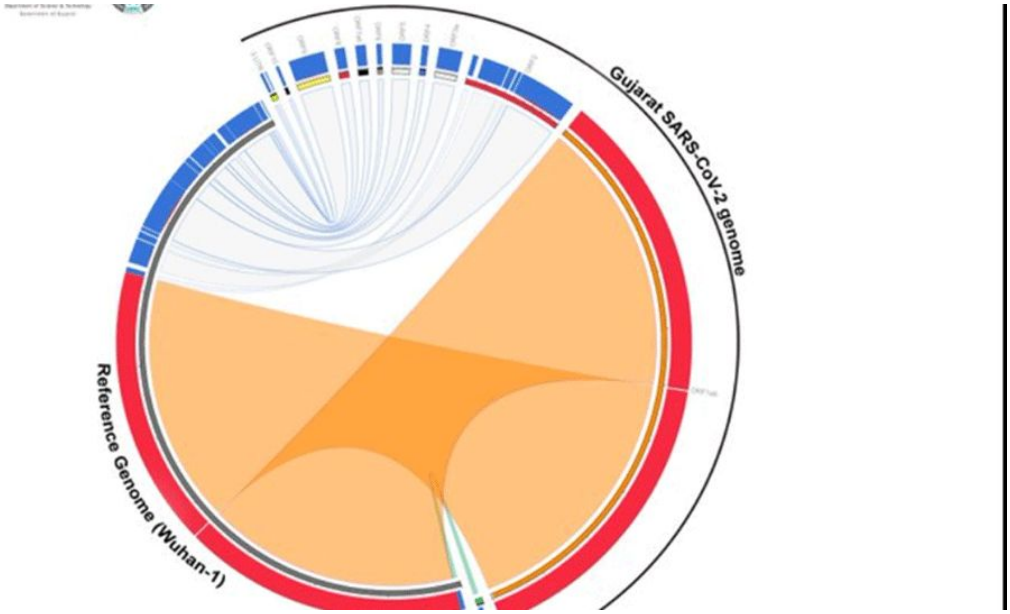
Four of the virus mutations are on the left side of the first principal component, and one on the right side. A signature of the first principal component is a high alignment length. This means that a higher value for a first principal component means a higher alignment length (is closer to the original virus). Lower values of component 1, thus, are farther genetically from the original virus. Most of the virus clusters vary largely from the original virus. Hence, scientists attempting to create a vaccine should be aware that the virus mutates a lot.

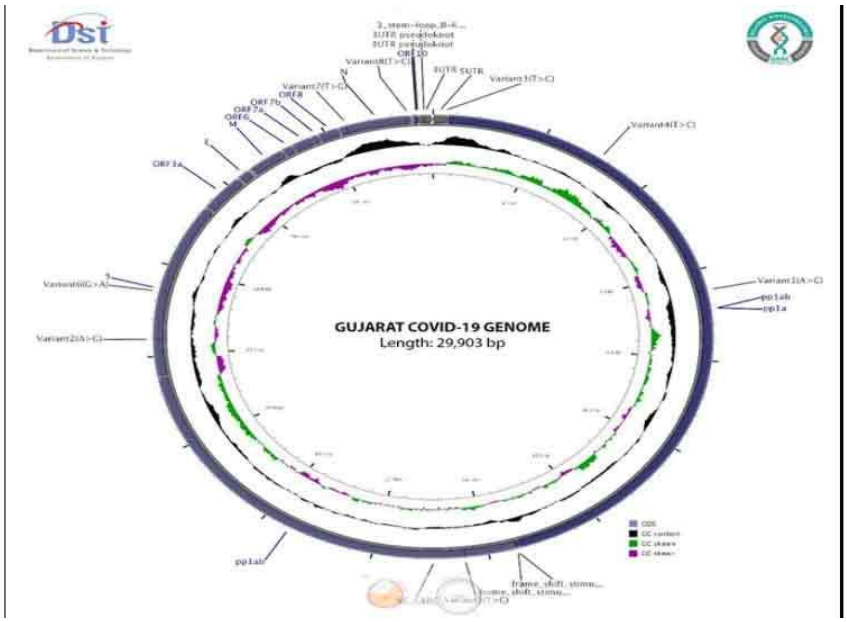
Conclusion

Using K-Means and PCA, were able to identify five main clusters of mutations in the coronavirus. Scientists developing vaccines for the coronavirus can use the cluster centers to gain knowledge about characteristics of each cluster. We were able to visualize the clusters in two dimensions using principal component analysis, and found that the coronavirus has a very high rate of mutation. This may be what makes it so deadly.

Advances in India

Recently, scientists at Gujarat Biotechnology Research Centre (GBRC) Decoded the entire genome of coronavirus





Gujarat Biotechnology Research Centre (GBRC) is the only State Govt laboratory in India that has reported COVID19 whole-genome sequence which will be helpful in tracking origin, drug targets, vaccine & association with

virulence. This is a proud moment for the whole country. When the data is made public, it can be used to find certain compounds which can bind very well with it and further this can be passed as data to a generative neural network which can help in generating new potential compounds from the latent space of reference molecules.