EPiC Series in Computing
Volume 83, 2022, Pages 66–75
Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

SARS-CoV-2 variants classification and characterization

Sofia Borgato[1]*, Marco Bottino[1]*, Marta Lovino[2]†, and Elisa Ficarra[2]

[1] Politecnico di Torino
DAUIN, Corso Duca Degli Abruzzi 24, Torino, Italy, s274110@studenti.polito.it;
sofia.borgato@studenti.polito.it; 0000-0001-7124-8319
[2] Università degli Studi di Modena e Reggio Emilia
Via Vivarelli 10/1 Modena, Italy, marta.lovino@unimore.it; elisa.ficarra@unimore.it;
0000-0002-8061-2124

Abstract

As of late 2019, the SARS-CoV-2 virus has spread globally, giving several variants over time. These variants, unfortunately, differ from the original sequence identified in Wuhan, thus risking compromising the efficacy of the vaccines developed. Some software has been released to recognize currently known and newly spread variants. However, some of these tools are not entirely automatic. Some others, instead, do not return a detailed characterization of all the mutations in the samples. Indeed, such characterization can be helpful for biologists to understand the variability between samples. This paper presents a Machine Learning (ML) approach to identifying existing and new variants completely automatically. In addition, a detailed table showing all the alterations and mutations found in the samples is provided in output to the user. SARS-CoV-2 sequences are obtained from the GISAID database, and a list of features is custom designed (e.g., number of mutations in each gene of the virus) to train the algorithm. The recognition of existing variants is performed through a Random Forest classifier while identifying newly spread variants is accomplished by the DBSCAN algorithm. Both Random Forest and DBSCAN techniques demonstrated high precision on a new variant that arose during the drafting of this paper (used only in the testing phase of the algorithm). Therefore, researchers will significantly benefit from the proposed algorithm and the detailed output with the main alterations of the samples.

Data availability: the tool is freely available at https://github.com/sofiaborgato/-SARS-CoV-2-variants-classification-and-characterization.

1 Scientific Background

At the end of 2019, a new virus of SARS-CoV was spotted in the Chinese region of Wuhan. The virus causes a severe respiratory illness later called COVID-19, which led to a global pandemic. As a result, in early December 2020 FDA authorized the first vaccine for emergency use[10].

*These authors contributed equally
†Corresponding author

Likewise, in late December 2020, The European Commission authorized the first vaccine to prevent COVID-19 in the EU, following evaluation by EMA [17]. However, at the same time, due to some critical mutations in the virus's genome, new lineages of the viruses, commonly known as variants, began to spread, with the risk of making the vaccines less effective. A sample isolated from pneumonia patients who were some of the workers in the Wuhan seafood market found that strains of SARS-CoV-2 had a length of 29.9 kb. Structurally, SARS-CoV-2 has four main structural proteins, including spike (S) glycoprotein, small envelope (E) glycoprotein, membrane (M) glycoprotein, and nucleocapsid (N) protein. The most important variants spreading at the moment this paper is being written are the following ones[9]:

- **VOC Alpha 2012012/01 GRY, lineage B.1.1.7 (English variant)**[4], first detected in October 2020, it is correlated with a significant increase in the rate of COVID-19 infection in the United Kingdom, associated partly with the N501Y mutation.

- **VOC Beta GH/501Y.V2, lineage B.1.351 (South-African variant)**[20], was first detected in South Africa and reported by the country's health department. The South African health department indicated that the variant may have driven the second wave of the COVID-19 epidemic in the country due to the variant spreading faster than other earlier variants of the virus.

- **VOC Gamma GR/501Y.V3, lineage P.1 (Brazilian variant)**[21] was detected in Tokyo on January 2021. A study found that P.1 infections can produce nearly ten times more viral load than persons infected by one of the other Brazilian lineages (B.1.1.28 or B.1.195).

- **VOI Epsilon, lineage B.1.427/B.1.429 (Californian variant)**[5],[11] was first detected in Fall 2020 in Northern California. CDC has listed B.1.429 and the related B.1.427 as "variants of concern" and cites a preprint for saying that they exhibit a 20% increase in viral transmissibility and moderately reduce neutralization by plasma collected by people who have previously been infected by the virus or who have received a vaccine against the virus.

- **VOI Eta G/484K.V3, lineage B.1.525 (Nigerian variant)**[13] The first cases were detected in December 2020 in the UK and Nigeria. B.1.525 appeared to have significant mutations already seen in some of the other newer variants, which is partly reassuring as their likely effect is, to some extent, more predictable.

- **VOC Delta G/478K.V1, lineage B.1.617 (Indian variant)**[12], was first identified in Maharashtra, India, in October 2020, but it reached a global spread in Spring 2021. Emerging research suggests the variant may be more transmissible than previously evolved ones.

At this pandemic stage, keeping the spread of new variants under control becomes a key issue. In this context, inspired by a multitude of applications in bioinformatics[16, 15, 14, 18, 7], several methods of variants classification have been proposed exploiting Machine Learning (ML) and Deep Learning (DL) techniques[8, 6, 22]. These methods provide efficient tools for the classification and clustering of SARS-CoV-2 samples. However, these tools can identify new variants without providing the user with a detailed characterization and synthesis of the newly identified variant, which is crucial in the field. Therefore, this paper aims to provide a general pipeline to classify and cluster SARS-CoV-2 samples and verify if a new variant is detected. In addition, the proposed pipeline provides an in-depth characterization in terms of critical mutations of the new variant. This characterization of the known and new variants can help the experts in clinical settings in the recognition process, by describing the most common mutations for each cluster and their location, which would help understanding more quickly its danger.

2 Materials and Methods

This work aims to provide a tool that can automatize the process of analysis and description of the SARS-CoV-2: the tool describes the key mutations of the group of samples and classifies them according to the variant of each sample, giving as input the FASTA/FASTQ files of genome samples of the virus. This section will cover the dataset creation and the proposed algorithm.

2.1 Dataset

The samples used to train the tool were downloaded from the global science initiative GISAID[2], which provides open access to whole-genome sequences of SARS-CoV-2. In addition, we downloaded 7 FASTA files containing the genome from the original Wuhan cases and 6 variants divided into *Variants of Interest* (VOI) and *Variants of Concern* (VOC) [1]. The dataset structure is described in Table 1.

Name	First detected in	# samples	Submission Period
Original (Wuhan-Hu-1)	China	1000	01/01/2020 - 24/03/2021
VOC Alpha 2012012/01 GRY (B.1.1.7)	UK	1000	08/04/2021 - 09/04/2021
VOC Beta GH/501Y.V2 (B.1.351)	South Africa	1000	04/04/2021 - 09/04/2021
VOC Gamma GR/501Y.V3 (P.1)	Brazil	1000	17/03/2021 - 09/04/2021
VOI Eta G/484K.V3 (B.1.525)	UK/Nigeria	1000	04/01/2021 - 09/04/2021
VOI Epsilon (B.1.427/B.1.429)	California	1000	04/04/2021 - 09/04/2021
VOC Delta G/478K.V1 (B.1.617)	India	500	01/04/2021 - 22/04/2021

Table 1: Dataset composition

Each sample was associated with a numeric label for the corresponding variant. Every sample was *complete* ($> 29kb$) and *high coverage* (only entries with $< 1\%$ undefined bases, $< 0.05\%$ of unique amino acid mutations and verified insertions/deletions), according to the GISAID notation [2]. We used the genome NC_045512.2 provided by National Center for Biology Information (NCBI)[3] as a reference to be compared with the samples in order to highlight their mutations. In order to prepare the dataset, some preprocessing is needed, including alignment and tables creation.

Alignment of the samples

First, the pipeline reads the FASTA files obtained from GISAID by creating a dataframe with a row for each genome sample. The genomes are sequences of nucleotides represented by a string of letters. When the nucleotide is known the letter can be one between **A** for adenine, **C** for cytosine, **G** for guanine, **T** for thymine. Different letters can be used if the nucleotide is unknown, according to the probabilities of being one of the previous bases. We decided to change all of these letters with **X**s to symbolize the unknown bases in the sequence. The second step of the preprocessing phase consists in the alignment of the sequences to the reference genome.

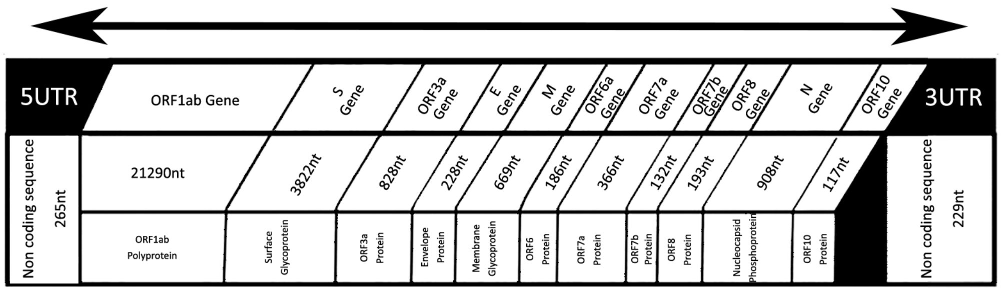


Figure 1: Description of the genome structure in SARS-CoV-2

This step is required to highlight and describe the mutations characterizing each genome. In particular, the local alignment technique has proven to be less sensitive to unknown bases in the samples. This technique, though, can be computationally expensive. Therefore, the tool allows dividing the genome in a subsequence of a fixed length, then separately aligned and concatenated. In order to ensure the stability of the process, consecutive sub-string have an intersecting section.

The performance of the aligner can vary according to the choice of the different scores: **1) correct match**: score assigned when the basis of the sequence matches the one of the reference. Set to 2 in this paper; **2) mismatch**: score assigned when the basis of the sequence does not match the one of the reference. It is set to -0.1; **3) gap**: score assigned when a first gap is inserted in the sequence or in the reference. It is set to -2; **4) repeated gap**: score assigned when there is more than one consecutive gap in the sequence or in the reference. It is set to -0.2;

These parameters have been empirically optimized for the alignment of SARS-CoV-2 sequences. This combination, indeed, is stable to the presence of sequences of **X**s, which have to be considered as mismatches.

Tables construction

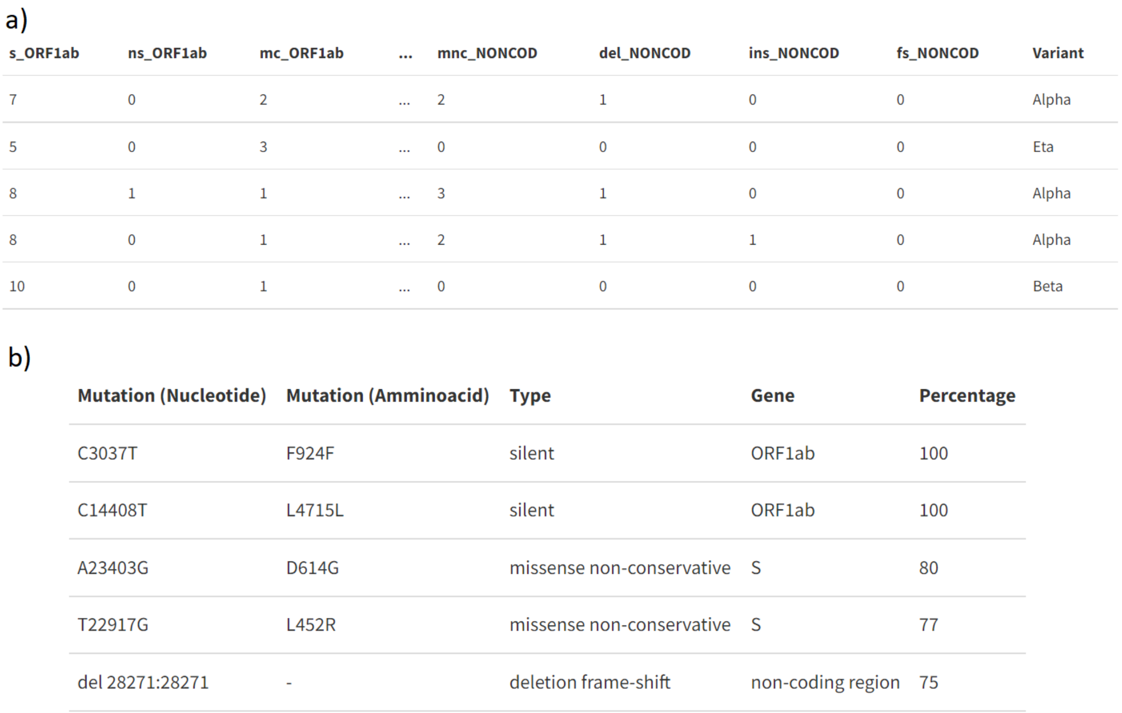
The final step of the preprocessing phase compares the aligned sequences and the reference genome.

The uptake of working with aligned sequences allows splitting them into different genes according to the division of the reference and evaluating the mutations separately for each gene. The results of this evaluation are then automatically summarized into three output tables (gene sequences, mutation statistics, key mutation), which supply a thorough description of the sample given as input.

Gene sequences table. This table contains in each row the sequence split according to the gene division in Figure 1 and the label of the variant. variants.

Mutations statistics table. This table describes numerically the kind and the number of mutations divided by region. The mutations can be divided according to the following:

1. Silent substitutions: single-base substitutions which code for the same amino acid and do not affect the functioning of the protein; **2. Nonsense substitutions**: single-base substitutions that result in a premature termination codon which signals the end of translation. This interruption causes the protein to be abnormally shortened; **3. Missense substitutions**: single-base substitutions, which results in the generation of a codon that specifies a different amino acid and hence leads to a different polypeptide sequence. Depending on the type of



a)

s_ORF1ab	ns_ORF1ab	mc_ORF1ab	...	mnc_NONCOD	del_NONCOD	ins_NONCOD	fs_NONCOD	Variant
7	0	2	...	2	1	0	0	Alpha
5	0	3	...	0	0	0	0	Eta
8	1	1	...	3	1	0	0	Alpha
8	0	1	...	2	1	1	0	Alpha
10	0	1	...	0	0	0	0	Beta

b)

Mutation (Nucleotide)	Mutation (Amminoacid)	Type	Gene	Percentage
C3037T	F924F	silent	ORF1ab	100
C14408T	L4715L	silent	ORF1ab	100
A23403G	D614G	missense non-conservative	S	80
T22917G	L452R	missense non-conservative	S	77
del 28271:28271	-	deletion frame-shift	non-coding region	75

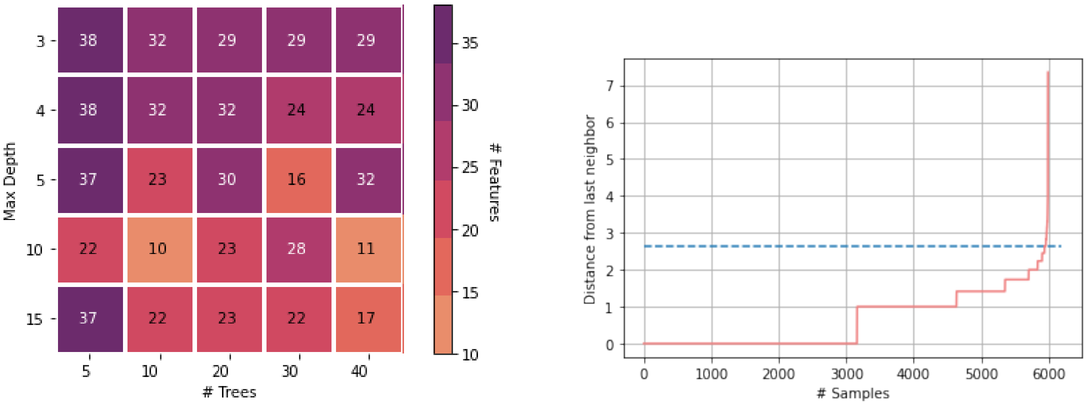
Figure 2: **a)** Example of **mutation statistics table**. Every column has the format type_region and counts how many mutations of a specific type happened in a region of the virus genome, for every sample. The last column contains the common name of the variant classified. **b)** Example of **key mutation description table**, describing the most common mutations in the dataset given as input in descending order according to the frequency.

amino acid substitution the missense mutation is either *conservative* or *nonconservative*. **4. Deletions**: mutations where one or more bases are lost from the reference genome. According to the number of lost bases, the deletion can be in-frame or cause a frameshift, resulting in a garbled message or a non-functional product. **5. Insertions**: mutations where one or more bases are added concerning the reference genome. Once again, they can be in-frame or frame-shift. An example of the mutations statistics table is reported in 2 a).

We wrote Python code to compare every sample with the reference and identify every mutation. This table counts the number of mutations for each gene divided according to the specific type as described above.

Mutations description table. This table describes the most frequent mutations in the input sample. Every time a mutation is encountered when comparing the sequences and the reference genome, different features are saved to describe it:

1) the nucleotide position, 2) the amino acid position 3) the gene 4) the type of mutation as described in the previous paragraph. An example of the mutations description table is reported in Figure 2 b). Afterward, the most frequent mutations are saved in the table in descending order according to the frequency.



(a) The heat-map describes the relationship between hyper-parameters of the Random Forest and the optimal number of features selected by RFE. The columns represent different number of trees, while the rows represent their possible depths.

(b) The y axis represents the distance to the k-th nearest neighbor, x axis represents the number of points that have the k-th nearest neighbor within that distance.

Figure 3: Random forest(a) and clustering(b) hyper-parameters.

2.2 The algorithm

The proposed tool performs the supervised classification of new samples (test samples) from FASTA/FASTQ files of SARS-CoV-2 sequences. In order to do so, the algorithm is trained on 6000 sequences, coming from the original Wuhan samples and 5 variants known until April 2021 (alpha, beta, gamma, epsilon, and eta). First, all details about the training samples are reported in Table[1]. Then, it automatically creates the mutation statistics table to train the ML algorithm. With proper hyper-parameter tuning, many ML algorithms have been explored for the classification task (e.g., Support Vector Machine - SVM, Multi-Layer perceptron - MLP). The test set is made up of 150 samples, equally divided into the 5 SARS-CoV-2 variants.

Since all classifiers returned an accuracy higher than 99%, the Random Forest (RF) classifier has been chosen because it is one of the fastest and best performing. In addition, the Recursive Features Elimination (RFE) was chosen as a feature selection method to optimize the results of the RF. RF hyper-parameters (the best number of estimators of the model (trees) and their maximum depth) were selected by performing a grid search for the highest accuracy. Since all models scored $> 99\%$ accuracy on stratifying 5-fold cross-validation, we only looked for the minimum number of features. This choice resulted in a number of estimators and a maximum depth equal to 10. Figure [3a] represents the result of the RF hyper-parameters optimization. The columns represent a different number of trees, while the rows represent their possible depths.

In addition to the classification, the tool allows the user to cluster new samples providing the FASTA/FASTQ files of SARS-Cov-2 sequences. The method automatically builds the mutation statistics table, and the new samples are grouped into clusters corresponding to known variants (alpha, beta, gamma, epsilon, and eta) or eventually new ones. In order to do so, the tool concatenates the new samples to a control database containing 1000 samples for each known variant (see Table[1]). Eventual new clusters should represent a group of similar observations that are sufficiently different from the known variants present in the control database at the execution. Therefore, among many ML clustering algorithms tested, DBSCAN has been selected for its density-based properties. However, since the number of clusters is not known in advance, we implemented an automatized version of DBSCAN [19]. The parameters to be tuned in a DBSCAN algorithm are the minimum number μ of points to define a cluster and the maximum distance ϵ between two samples in the same cluster in the features space. The best performing distance for this task is the euclidean distance and the optimal value of μ is 20. This value was found empirically as the best-performing given the structure of the input table, but other values, as long as lesser than the number of features (N = 39), led to comparable results. Since the value of ϵ can vary according to the dataset given as input, we automatized the algorithm proposed for the ϵ best choice as in [19]. A suitable value for ϵ can be found by calculating the distance to the nearest μ points for each point, sorting and plotting the results (Figure 4b).

The optimal value for ϵ corresponds then to the elbow of this plot. Since it represents a step function, we automatized the choice of the elbow by selecting the index where the distance between two consecutive steps is under a certain threshold. We set the threshold empirically to 30 to be the most stable for the optimal choice of ϵ. The number of clusters created corresponds to the number of known variants, a cluster of the outliers and, if data from new variants are present, one cluster for each new variant.

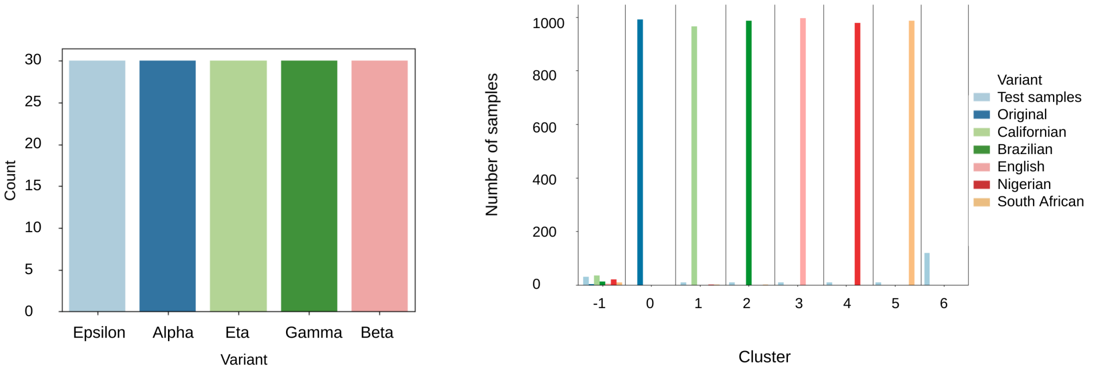
3 Results and discussion

As previously stated, this paper aims to provide a tool to classify and cluster new SARS-CoV-2 samples in searching for possible new variants. In addition, gene sequences, mutation statistics, key mutation tables are reported as an output. The tool has two options.

Classification option: the tool performs the supervised classification and labels the new samples according to the known variants. This option is more accurate and less sensitive to the outliers compared to the clustering option. However, it can be used only and advised for files with $<$ 100 samples. It creates a folder containing: 1) a histogram describing how the samples are divided among the known variants, 2) the three tables -gene sequences, mutations statistics and mutations description, described above-, 3) a descriptive plot of the average number of mutations for each region and variant.

Clustering option: the tool performs the unsupervised clustering of the new samples. It is more sensitive to the outliers but capable of highlighting new lineages of the genome. Therefore it is advised for files with $>$ 100 samples. It creates a folder containing: 1) a histogram describing how the samples are divided among the clusters, 2) the three tables -gene sequences, mutations statistics and mutations description, described above-, 3) a descriptive plot of the average number of mutations for each region and variant, 4) an additional mutation description table for each new variant which describes its key mutations, 5) a text file describing the performance scored by the clustering.

Outputs 1), 2), and 3) are in common between the classification and the clustering options, while 4) and 5) are specific for the clustering one. Figure 4a reports an example of output 1) both for the classification (on the left) and the clustering (on the right). Here, each column represents the number of samples for each variant. In the classification case, we selected 150



(a) Distribution of the test samples among the known variants as output plot of classification option.

(b) Distribution of the original labels among the different clusters as output plot of clustering option. We can see that the clusters from 0 to 5 correspond to the known variants, whereas most of the unknown samples are assigned to a brand new cluster. The cluster -1 collects the outliers.

Figure 4: Examples of output 1) for classification and clustering.

samples (30 for each known variant) as a test set, and the algorithm correctly assigned every sample to its variant.

We selected 140 Delta variant samples (not present in the training dataset) and 10 samples for each known variant as a test set in the clustering case. In this case, the algorithm adds the new 190 test samples to the previous 6000 training ones. As shown in Figure[4b] the tool properly associated the samples from the known variants to the same cluster of the ones from the training set; it also creates a new cluster (cluster 6 in the figure) made up of only new Delta variant samples. In addition, a cluster of outliers samples (cluster -1) is created, which corresponds to 1.6% of the total number of samples.

An example of output 2) -the three tables- can be seen above, while an example of output 3) the descriptive plot of the average number of mutations for each region and variant is reported in 5.

This plot is crucial for biological interpretation of the results since it allows further analyses to compare the different variants. For example, most of the mutations happening on genes ORF1ab and S are samples from the Gamma variant, which on average mutates more in gene S than the other ones. It is also interesting to note that distinction between variants is not sensible to a low number of mutations. For example, it can be seen that even if the samples of the original virus (light blue in Figure[5]) have a non-zero number of mutations, the clustering algorithms still groups them correctly.

Suppose the tool highlights one or more new variants, as in the case reported. In that case, a new mutation description table is created for each new variant, providing information about its key mutations. In terms of format, output 4) is identical to table in Figure 2 b). This table is extremely crucial for a biological interpretation of the variant since it details all encountered mutations.

Output 5) is a text file containing the main clustering evaluation metrics (e.g., in the example described the results are Homogeneity = 0.958, Completeness = 0.941, V-measure = 0.949, Adjusted Rand Index = 0.97, Adjusted Mutual Information = 0.949, Silhouette Coefficient = 0.44).

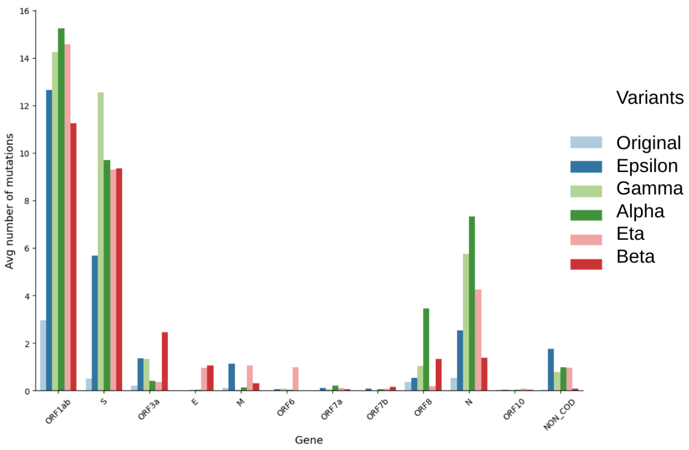


Figure 5: Example of analysis performed with statistics table: each column represents the average number of mutations for each gene divided by color according to the variant of the virus.

4 Conclusion

Summing up, the tool receives as input a FASTA file containing genome sequences from the SARS-CoV-2 virus, aligns them to a reference genome with an algorithm of local alignment, and compares them to the reference to obtain information about the mutations provided by three data structures. This information can be exploited to train supervised classifiers that can assign the samples to the correct variant or train a clustering algorithm to highlight new clusters of genomes, representing new unseen variants.

The main limit of this work is the time used to align the genomes. In addition, this process has a high computational cost compared to the ML part of the tool. Another limit is in the ability of DBSCAN to find new clusters with unseen variants: since this algorithm is density-based, it fails in finding a new variant if the number of samples is too small. We obtained good results if the number of samples from the new variant is > 100. On the other hand, identifying a group of anomalous samples as ”variant” would not make much sense if the number of exemplars is too small.
Further development for this work could make the training dataset updatable whenever new variants are met. By adding the samples from the new variant to the training dataset and labeling them, the supervised classifier would predict new classes for the following observations. This work, in particular, the mutation statistics and mutation description tables, can be useful in clinical applications to automatize the analyses of the isolated samples, which would make much faster the identification and description of new dangerous mutations in clusters of new cases COVID-19.

References

[1] "https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html".

[2] GISAID Dataset. "www.gisaid.org".

[3] Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,completegenome. "https://www.ncbi.nlm.nih.gov/nuccore/NC_045512".

[4] Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations, Dec. 2020.

[5] Detection of the recurrent substitution Q677H in the spike protein of SARS-CoV-2 in cases descended from the lineage B.1.429, Apr. 2021.

[6] S. Ali, Tamkanat-E-Ali, M. A. Khan, I. Khan, and M. Patterson. Effective and scalable clustering of SARS-CoV-2 sequences. *arXiv:2108.08143 [cs, q-bio]*, Oct. 2021. arXiv: 2108.08143.

[7] P. Barbiero, M. Lovino, M. Siviero, G. Ciravegna, V. Randazzo, E. Ficarra, and G. Cirrincione. Unsupervised multi-omic data fusion: The neural graph learning network. In *International Conference on Intelligent Computing*, pages 172–182. Springer, 2020.

[8] S. Basu and R. H. Campbell. Classifying covid-19 variants based on genetic sequences using deep learning models. *bioRxiv*, 2021.

[9] C. Chakraborty, M. Bhattacharya, and A. R. Sharma. Present variants of concern and variants of interest of severe acute respiratory syndrome coronavirus 2: Their significant mutations in S-glycoprotein, infectivity, re-infectivity, immune escape and vaccines activity. *Reviews in Medical Virology*, page e2270, June 2021.

[10] O. o. t. Commissioner. FDA Takes Key Action in Fight Against COVID-19 By Issuing Emergency Use Authorization for First COVID-19 Vaccine, Dec. 2020.

[11] X. Deng, M. A. Garcia-Knight, M. M. Khalid, V. Servellita, C. Wang, M. K. Morris, et al. Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. Technical report, Mar. 2021. Type: article.

[12] M. Hoffmann, H. Hofmann-Winkler, N. Krüger, A. Kempf, I. Nehlmeier, et al. SARS-CoV-2 variant B.1.617 is resistant to Bamlanivimab and evades antibodies induced by infection and vaccination. Technical report, May 2021. Type: article.

[13] Y. Hu, X. Zhao, Z. Li, M. Kang, X. Deng, and B. Li. Two Imported Cases of New Variant COVID-19 First Emerging in Nigeria — Guangdong Province, China, March 12, 2021. *China CDC Weekly*, 3(19):411–413, May 2021.

[14] M. Lovino, G. Bontempo, G. Cirrincione, and E. Ficarra. Multi-omics classification on kidney samples exploiting uncertainty-aware models. In *International Conference on Intelligent Computing*, pages 32–42. Springer, 2020.

[15] M. Lovino, M. S. Ciaburri, G. Urgese, S. Di Cataldo, and E. Ficarra. Deeprior: a deep learning tool for the prioritization of gene fusions. *Bioinformatics*, 36(10):3248–3250, 2020.

[16] M. Lovino, G. Urgese, E. Macii, S. Di Cataldo, and E. Ficarra. A deep learning approach to the screening of oncogenic gene fusions in humans. *International journal of molecular sciences*, 20(7):1645, 2019.

[17] A. C. Pinho. EMA recommends first COVID-19 vaccine for authorisation in the EU, Dec. 2020.

[18] I. Roberti, M. Lovino, S. Di Cataldo, E. Ficarra, and G. Urgese. Exploiting gene expression profiles for the automated prediction of connectivity between brain regions. *International journal of molecular sciences*, 20(8):2035, 2019.

[19] A. Starczewski, P. Goetzen, and J. Er. A new method for automatic determining of the dbscan parameters. *Journal of Artificial Intelligence and Soft Computing Research*, 10:209–221, 07 2020.

[20] H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854):438–443, Apr. 2021.

[21] D. Tian, Y. Sun, J. Zhou, and Q. Ye. The global epidemic of SARS-CoV-2 variants and their mutational immune escape. *Journal of Medical Virology*, n/a(n/a).

[22] H. Torun, B. Bilgin, M. Ilgu, C. Yanik, N. Batur, et al. Machine learning detects sars-cov-2 and variants rapidly on dna aptamer metasurfaces. *medRxiv*, 2021.