Machine learning-based approach KEVOLVE efficiently identifies SARS-CoV-2 variant-specific genomic signatures

Dylan Lebatteux[1,2], Hugo Soudeyns[2], Isabelle Boucoiran[2], Soren Gantt[2], Abdoulaye Baniré Diallo[1,3]*,

1 Department of Computer Science, Université du Québec à Montréal, Montreal, QC H2X 3Y7, Canada.
2 CHU Sainte-Justine Research Centre, Montréal, QC H3T 1C5, Canada
3 Centre CERMO-FC, Pavillon des Sciences biologiques, Montreal, QC, H2X 3Y7, Canada

* diallo.abdoulaye@uqam.ca

Abstract

Machine learning has proven to be a powerful tool for the identification of distinctive genomic signatures among viral sequences. Such signatures are motifs present in the viral genome that differentiate species or variants. In the context of SARS-CoV-2, the identification of such signatures can contribute to taxonomic and phylogenetic studies, help in recognizing and defining distinct emerging variants, and focus the characterization of functional properties of polymorphic gene products. Here, we study KEVOLVE, an approach based on a genetic algorithm with a machine learning kernel, to identify several genomic signatures based on minimal sets of k-mers. In a comparative study, in which we analyzed large SARS-CoV-2 genome dataset, KEVOLVE performed better in identifying variant-discriminative signatures than several gold-standard reference statistical tools. Subsequently, these signatures were characterized to highlight potential biological functions. The majority were associated with known mutations among the different variants, with respect to functional and pathological impact based on available literature. Notably, we found show evidence of new motifs, specifically in the Omicron variant, some of which include silent mutations, indicating potentially novel, variant-specific virulence determinants. The source code of the method and additional resources are available at: `https://github.com/bioinfoUQAM/KEVOLVE`.

Author summary

Advances in cloning and sequencing technologies have yielded a vast repository of viral genomic sequence data. To analyze this complex and massive data, Machine learning, which refers to the development and application of computer algorithms that improve with experience, has proven to be efficient. Although many methods have been developed to classify viruses into different characteristic groups, it is often difficult to explain the predictions of these methods. To overcome this, we are working in our laboratory on the design of machine learning based methods for discriminative signatures identification within viral genomic sequences. These signatures which are a specific motifs to groups of viruses known to be pervasive in their genome, are used to 1) build accurate and explainable prediction tools for pathogens and 2) highlight mutations potentially associated with functional changes. In this paper we present the potential of our latest approach KEVOLVE. We first compare it to three discriminating motif identification tools with data sets covering several SARS-CoV-2 variant genomes. We then focus on the identified motifs by KEVOLVE to analyze the mutations associated with the different variants and the potential changes in biological functions that they may involve.

Introduction

SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) is the etiological agent of COVID-19. This highly pathogenic coronavirus was discovered in December 2019 in the city of Wuhan, China. It belongs to the betacoronavirus genus, which includes SARS-CoV-1 and MERS-CoV. The genome of SARS-CoV-2 consists of a single-stranded RNA of 29,903 nucleotides (Fig 1 from [1]). Its sequence identity with SARS-CoV and MERS-CoV is 79.5% and 50% at the nucleotide sequence level, respectively [2, 3]. The SARS-CoV-2 genome contains 11 genes encoding 15 Open Reading Frames (ORF), which result in between 29 and 33 viral protein products [1]. SARS-CoV- 2 is associated with a very high mutation rate ranging from 5.2 to 8.1 $\times 10^{-3}$ substitutions/site/year [4, 5], higher than human immunodeficiency virus (HIV), which has a mutation rate of 3 to 8 $\times 10^{-3}$ substitutions/site/year [6]. Many of these mutations, principally in the spike gene, are associated with increased SARS-CoV-2 transmission rates [7], and the development of new variants associated with reduced efficacity of current COVID-19 vaccines and antibody-based treatments [8].

Given this rapid rate of evolution, it is important to be able to efficiently identify genomic signatures that discriminate between the different variants of SARS-CoV-2 and highlight potential functional changes. These signatures are defined as species or variant-specific motifs that are pervasive throughout the viral genome [9]. In the context of SARS-CoV-2, the identification of this type of signature can contribute to taxonomic [10] and phylogenetic [11] studies to differentiate distinct groups of variants, provide an explanation for their evolutionary history [9], as well as to facilitate mechanistic studies to elucidate the functional basis of variant-specific differences in virulence [12].

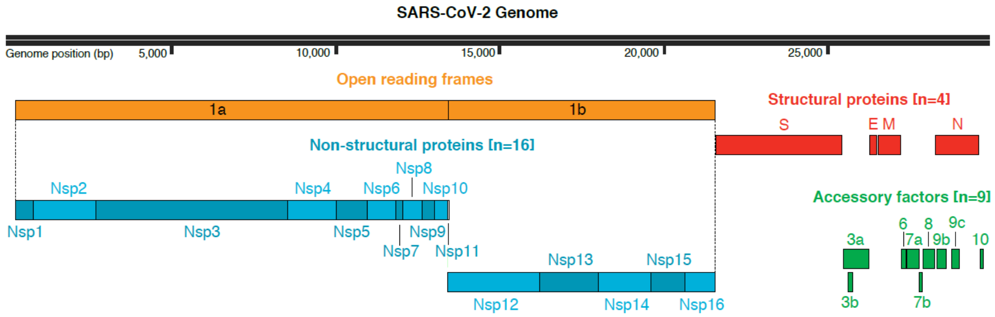


Fig 1. SARS-CoV-2 genome organization

To identify discriminating motifs that constitute genomic signatures among different groups of biological sequences, the traditional approach is to first compute multiple sequence alignment [13] with tools as: MUSCLE [14], Clustal W/X [15] or MAFFT [16]. These alignments can then be analyzed to identify the divergent genomic regions that constitute the discriminating motifs. However, the use of multiple alignment approaches has significant limitations, particularly when applied to viral genomes [12].

First, alignment-based approaches are generally computationally- and time-intensive and are therefore less well suited to dealing with very large viral sequence datasets that are increasingly available [17]. Indeed, computing an accurate multi-sequence alignment is an NP-hard problem with $(2N)!/(N!)^2$ possible alignments for two sequences of length N [18], which means that in some case, the alignment cannot be solved within a realistic time frame [19]. Even with dynamic programming, the time requirement is on the order of the product of the lengths of the input sequences [20].

Second, alignment algorithms assume that homologous sequences consist of a series of more or less conserved linearly arranged sequence segments. However, this assumption, named collinearity, is often questionable, especially for RNA viruses [19]. This is because RNA viruses show extensive genetic variation due to high mutation rates, as well as high frequencies of genetic recombination, horizontal gene transfer, and gene duplication, leading to the gain or the loss of genetic material [21].

Finally, performing multiple alignments often requires adjusting several parameters (e.g., substitution matrices, deviation penalties, thresholds for statistical parameters) that are dependent on prior knowledge about the evolution of the compared sequences [19]. The adjustment of these parameters is therefore sometimes arbitrary and requires a trial-and-error approach. Many experiments have shown that minor variations in these parameters can significantly affect the quality of alignments [22].

To overcome the limitations of discriminative motif identification among different groups of biological sequences using multiple sequence alignment, specialized statistical-based tools have been developed. The most popular of these method is MEME [23, 24], which is dedicated to motif identification. MEME has a discriminative mode [25] that considers two sets of sequences and identifies the enriched motifs that discriminate the first set (primary) from the second (control). A suite of other MEME tools has been developed, of which STREME [26] is the latest and most powerful for motif discovery in sequence datasets. The STREME algorithm is based on a generalized suffix tree and evaluates motifs using a statistical test of the enrichment of matches to the motif in a primary set of sequences compared to a set of control sequences [26].

In parallel, machine learning methods have been widely used in the field of genomics over recent years and have proved to be highly effective for solving complex and massive data analysis problems [27]. For viral genomic sequence classification CASTOR [28] has shown the relevance of RFLP (Restriction fragment length polymorphism) signatures coupled with machine learning models. These models obtained in cross-validation evaluations performance in terms of F1-score $> 99\%$ for the prediction of viral genomes of hepatitis B and human papillomavirus. However, these signatures showed some limitations for HIV sequence prediction where the F1-score dropped below 0.90. Subsequently, KAMERIS [29] addressed this problem by using k-mers (nucleotide subsequences of length k) to characterize the sequences given to the learning model. To tackle the problem of the number of exponential number of features (4^k) associated with k-mers, KAMERIS performs a dimensionality reduction using truncated singular value decomposition. However, this transformation significantly affects the ability to explain the predictions of the model.

For this reason, CASTOR-KRFE [30] is a method that focuses on the identification of minimal sets genomic signatures based on minimal sets of k-mers to discriminate among several groups of genomic sequences. During cross-validation evaluations covering a wide range of viruses, CASTOR-KRFE successfully identified minimal sets of motifs. Subsequently, these motifs, coupled with supervised learning algorithms, have allowed to build prediction models resulting in average F1-score > 0.96 [30]. However, this study is limited to identifying an optimal set of motifs, instead of exploring the suboptimal sets of the feature space. This may have major consequences when dealing with in sets of viral sequences with high genomic diversity or when attempting to infer biological functions based on the identified motifs. To overcome the lack of flexibility of CASTOR-KRFE, KEVOLVE [31] a new method based on a genetic algorithm including a machine learning kernel was designed to identify multiple minimal subsets of discriminative motifs. A preliminary comparative study of HIV nucleotide sequences showed that KEVOLVE-identified motifs allowed the construction of models that out-performed specialized HIV prediction tools.

Here, we evaluate the KEVOLVE, whose search function has been improved in order to identify smaller sets of motifs while trying to respect the same discriminative performance criteria. We compared several reference tools (MEME, STREME and CASTOR-KRFE) to identify discriminating motifs among SARS-CoV-2 genome sequences. The motifs were first identified in a restricted set of nucleotide sequences associated with different variants of SARS-CoV-2. Second, the motifs were used to build prediction models that were assessed through the classification of a large set of SARS-CoV-2 sequences. Third, the motifs identified by KEVOLVE were analyzed in order to highlight the potential biological functions of the sequences/motifs in questions. Finally, a specific analysis was dedicated to the new variant of concern, Omicron, that was recognized on 24 November 2021 in South Africa (`https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern`).

Materials and methods

To assess the relative accuracy of KEVOLVE to identify discriminating motifs, we performed a comparative study with specialized tools. This involved for each tool to identify a subset of discriminating motifs in a set of training sequences of SARS-CoV-2 variants. These sets of motifs were designed to provide genomic signatures specific to each SARS-CoV-2 variant. In a second step these signatures combined with a supervised learning algorithm and the training sequences to fit a prediction model. Then, the quality of the signatures was assessed through the prediction of trained models on a large test set of unknown sequences. Finally, we analyzed in line with the literature, the variant-discrimination motifs identified by KEVOLVE according to their location in the genome, to assess the potential functional impact of these mutations.

Discriminative motif identification tools

The first tool that was evaluated was KEVOLVE [31]. KEVOLVE, is a new method based on a genetic algorithm including a machine learning kernel. KEVOLVE implementation is based on two main units: 1) an identification unit that provides subsets of features that are minimal and likely to provide the best performance metrics; and 2) a prediction unit that applies an ensemble classifier using the subsets of features.

The second tool that was evaluated was CASTOR-KRFE [30]. It is an alignment-free machine learning approach for identifying a set of genomic signatures based on k-mers to discriminate between groups of nucleic acid sequences. The core of CASTOR-KRFE is based on feature elimination using SVM (SVM-RFE). CASTOR-KRFE identifies an optimal length of k to maximize classification performance and minimize the number of features. This method also provides a solution to the problem of identifying the optimal length of k-mers for genomic sequence classification [32].

The third tool that was evaluated was MEME (discriminative mode) [25], a tool from the MEME suite [24] specialized in motif identification. MEME inputs two sets of sequences and identifies enriched motifs that discriminate the primary set from the control set. In discriminative mode, the algorithm first calculates a position-specific prior from the two sets of sequences. It then searches the first set of sequences for motifs using the position-specific prior to inform the search based on the discriminative prior D [33]. In addition, MEME considers as a parameter a potential motif distribution type to be identified to improve the sensitivity and quality of the motif search. In discriminative mode, the two available options are: 1) zero or one occurrence per sequence (zoops), where MEME assumes that each sequence may contain at most one occurrence of each motif; and 2) one occurrence per sequence (oops), where MEME assumes that each sequence in the dataset contains exactly one occurrence of each motif.

The last tool evaluated was STREME [26], which during a recent comparative study of motif identification was found to be more accurate, sensitive and thorough than several widely used algorithms [26]. STREME algorithm makes use of a data structure called a generalized suffix tree and evaluates motifs using a one-sided statistical test of the enrichment of matches to the motif in a primary set of sequences compared to a set of control sequences STREME assumes that each primary sequence may contain zoops but the motif discovery will not be negatively affected if a primary sequence contains more than one occurrence of a motif.

Dataset

To set up the most comprehensive evaluation framework possible, we built a dataset of 226,532 complete SARS-CoV-2 genomes. The sequences of this initial dataset covering variants Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Epsilon (B.1.427/9), Zeta (P.2), Eta (B.1.525), Iota (B.1.526) and Kappa (B.1.617.1) were downloaded on August 1, 2021 from the GISAID database [34]. In addition, in the context of the emergence of the new Omicron variant (B.1.1.529) and its interest for global public health, 72 sequences of this variant were also included for our studies on December 15, 2021. Due to small sample size of Omicron during the study, the sequences were not included in our comparative study. However, a specific analysis section was dedicated to it. We specify that only complete genomes with high coverage were included in our data set (Table 1) and the list of accession ids of the sequences used in our different dataset is available on our GitHub repository.

Table 1. Genomic sequence dataset of SARS-CoV-2 variants.

WHO label	SARS-CoV-2 lineage	Country of origin	Number of sequences
Alpha	B.1.1.7	United Kingdom	50000
Beta	B.1.351	South Africa	17126
Gamma	P.1	Brazil	36929
Delta	B.1.617.2	India	50000
Epsilon	B.1.427/9	USA (California)	34118
Eta	B.1.525	United Kingdom / Nigeria	2334
Iota	B.1.526	USA (New York)	28572
Zeta	P.2	Brazil	3745
Kappa	B.1.617.1	India	3708
Omicron	B.1.1.529	South Africa	72
Total number of sequences			**226604**

Subsequently, this initial dataset was partitioned into two independent subsets. The first (training subset) was composed of 2,250 randomly selected sequences (250 sequences for the 9 types of variants). The second (testing subset) was composed of the remaining sequences (224,282 sequences).

Setting the length of k

A preliminary step of this comparative study consists in setting the parameter k for the length of the motifs to be discovered for the respective identification tools. For this purpose, we used CASTOR-KRFE, giving it as input the training sequence set. For the associated parameters, we set the performance threshold to be maintained by reducing the number of features to $T = 0.99$ and the minimum/maximum length of k-mers to be explored to $k\text{-}min = 1$ and $k\text{-}max = 20$ respectively. As output, CASTOR-KRFE identified the following subset which is composed of 9 motifs of length $k = 8$: [AACTAAAA, ATATCTGG, AATTTCTC, ATAGAATG, CCGGTATA, CATAGCGC, TAGTGAAT, TCTTGCAT, CAAAGTAG]. During the CASTOR-KRFE identification process, this subset of motifs, coupled with a supervised prediction model based on a linear SVM, was evaluated by 5-fold cross-validation on the training set and obtained a weighted F1-score > 0.99. In addition, the length of k-mers that was identified was consistent with other studies using k-mers for viral sequence classification [9, 30, 32].

Benchmarking

To assess the relevance of the discriminating motifs identified by each tool, we were inspired by the evaluation conducted in [30]. For CASTOR-KRFE, the previously identified subset of motifs coupled with the set of training sequences and an SVM were used to fit a prediction model. From this model the testing sequence set was predicted. Regarding KEVOLVE, the identification unit was used using as input the training set of sequences as well as the following parameters: $n\_iterations = 1000$, $n\_solutions = 100$, $n\_chromosomes = 2500$, $n\_genes = 1$, $objective\_score = 0.99$, $crossover\_rate = 0.2$, $mutation\_rate = 0.1$ and $variance\_threshold = 0.01$. Initially, KEVOLVE was designed to identify multiple discriminating subsets to build a single ensemble prediction model. However, in this evaluation, for each identified subset, a model was trained and evaluated by predicting the test set.

Table 2. Summary of the evaluated motif identification tools and their associated parameters.

Tools	Number of motifs	Motifs width	Site distribution	Discovery mode	Performance threshold
STREME	1 2 3	8	None	None	Number of motifs
MEME	1 2 3 1 2 3	8	Zero or One Occurrence Per Sequence (zoops) One Occurrence Per Sequence (oops)	Discriminative	Number of motifs
CASTOR KRFE	Auto	8	None	None	0.99
KEVOLVE	Auto	8	None	None	0.99

The Tools column provides information about the different tools evaluated in this study. The column Number of motifs indicates the number of discriminating motifs that have been asked to identify for each tool. For CASTOR-KRFE and KEVOLVE no parameter is filled in because these tools automatically try to minimize the number of motifs. The column Motifs width corresponds to the length of the discriminating motifs to be identified. The column Site distribution refers for MEME to how the discriminating motifs are supposed to be present in the sequences to improve the sensitivity and quality of the search. The column Discovery mode also indicates for MEME the type of search to perform. In this context we have selected the mode for identifying motifs that discriminates between groups of sequences given as input. The Performance Threshold column refers to a quality criterion that the identified motif must satisfy. For MEME and STREME the n best motifs in terms of $p\text{-}value$ are selected where n corresponds to the number in the Number of motifs column. For CASTOR-KRFE and KEVOLVE, the algorithms will search until they obtain a set of motifs that satisfy an F1-score > 0.99 during their internal evaluation.

To evaluate MEME, considering its limitation to take as input a binary set (primary set and control set), we set up the following process: for each variant v present in the training set V, we select all sequences belonging to v to form the primary set. All other sequences in V were used to form the control set. Then, we applied MEME to discover the motifs that discriminated the primary set from the control set. This process was repeated for each v belonging to V in order to build a set of motifs that can discriminate each variant from the others. Then, this set was used to train a model and predict the testing set in the same configuration as CASTOR-KRFE and KEVOLVE. For the associated distribution site parameters, both zoops and oops options were evaluated. In addition, for the motifs to be identified, we ran the experiments to discover 1, 2 and 3 motifs of width 8 for each variant. This involved for MEME the training and evaluation of six prediction models.

Finally, for STREME, we applied the same iterative process to identify the motifs as for MEME. As mentioned before, for the motif distribution type, STREME does not require an input parameter and handles this automatically. Finally, as for MEME, we have identified and evaluated subsets of motifs of variable size (1 motif per variant, 2 motifs per variant and 3 motifs per variant). The tools and their overall configuration are summarized and compared in the (Table 2).

Results and Discussion

Number of identified motifs

First, we focused on the number of motifs identified by each tool, For KEVOLVE, its identification unit gave as output a total of 21 subsets composed of 8 motifs of length $k = 8$. This was the smallest subsets of motifs capable of discriminating between the 9 groups of SARS-CoV-2 variants among all tools. Even though each subset of KEVOLVE was composed of unique motifs, some motifs overlapped between subsets. CASTOR-KRFE identified a set composed of 9 discriminating motifs, which is slightly better than KEVOLVE. To describe the results of MEME and STREME we use the name of the tool followed by the distribution type and the number of motifs to identify. If this is not specified we discuss the overall tool. With the option of 1 motif per variant, MEME zoops, MEME oops and STREME were each able to constitute a subset of 9 discriminative motifs, which is similar to CASTOR-KRFE. For the 2 motifs per variant option, MEME zoops, MEME oops and STREME identified 14, 17 and 18 discriminative motifs respectively. Finally for the 3 motif per variant configuration, the identified subsets reached the size of 18, 24 and 26 respectively for MEME zoops, MEME oops and STREME. These results show that by increasing the number of motifs to be discovered, MEME zoops tends to identify more motifs that are redundant unlike MEME oops, STREME, CASTOR-KRFE and KEVOLVE.

Prediction performances

Regarding the predictive results associated with the model based on the motifs identified by each tool, Table 3, 4 and 5 illustrate respectively the predictive performance of the testing set according to each variant in terms of precision, recall and F1-score. Specifically, for KEVOLVE the scores shown represent the average results followed by the standard deviation obtained by 21 predictive models trained from the sets identified by the algorithm.

The best results of this comparative study are obtained by CASTOR-KRFE and KEVOLVE. For all the variants, they have scores > 0.9 in terms of Precision, Recall and F1-score. Their average score for all performance metrics is above 0.97. We note that the performance of CASTOR-KRFE is slightly better, however, KEVOLVE

Table 3. Precision of the models associated with each tool for the prediction of the testing set.

	Alpha (B.1.1.7)	**Beta (B.1.351)**	**Delta (B.1.617.2)**	**Epsilon (B.1.427/9)**	**Eta (B.1.525)**	**Gamma (P.1)**	**Iota (B.1.526)**	**Kappa (1.617.1)**	**Zeta (P.2)**	**Mean**
MEME_ZOOPS_3	**0.999**	0.710	0.442	0.361	**0.997**	**1.000**	**0.999**	**0.993**	**0.980**	0.831
MEME_ZOOPS_2	**0.999**	0.695	0.538	0.723	**0.998**	**1.000**	**0.999**	**0.993**	**0.980**	0.881
MEME_ZOOPS_1	**0.998**	0.716	0.500	0.351	0.000	**1.000**	**0.999**	**0.993**	**0.979**	0.726
MEME_OOPS_3	**0.999**	0.615	0.329	0.276	**0.997**	**1.000**	0.341	0.878	**0.976**	0.712
MEME_OOPS_2	**0.999**	0.599	0.225	0.277	**0.997**	**1.000**	0.209	0.877	**0.976**	0.684
MEME_OOPS_1	**0.998**	0.652	0.425	0.271	0.000	**0.999**	0.000	0.876	**0.975**	0.577
STREME_3	**0.999**	0.140	**0.998**	**0.981**	0.851	**1.000**	0.606	0.896	**0.998**	0.830
STREME_2	**0.909**	0.070	**0.998**	**0.981**	0.851	**1.000**	0.293	0.896	**0.994**	0.776
STREME_1	**0.914**	0.000	**0.999**	**0.980**	0.851	**1.000**	0.290	**0.991**	**0.994**	0.780
CASTOR_KRFE	**0.997**	**0.991**	**0.999**	**0.997**	**0.992**	**1.000**	**0.943**	**0.969**	**0.978**	**0.985**
KEVOLVE	**0.999** ± 0.001	**0.994** ± 0.003	**0.998** ± 0.003	**0.996** ± 0.003	**0.983** ± 0.011	**0.997** ± 0.006	**0.986** ± 0.006	**0.905** ± 0.092	**0.909** ± 0.068	**0.974** ± 0.038

Table 4. Recall of the models associated with each tool for the prediction of the testing set.

	Alpha (B.1.1.7)	**Beta (B.1.351)**	**Delta (B.1.617.2)**	**Epsilon (B.1.427/9)**	**Eta (B.1.525)**	**Gamma (P.1)**	**Iota (B.1.526)**	**Kappa (1.617.1)**	**Zeta (P.2)**	**Mean**
MEME_ZOOPS_3	**0.999**	0.496	0.004	**0.951**	**0.984**	**0.999**	**0.993**	0.779	**0.962**	0.796
MEME_ZOOPS_2	**0.999**	0.498	**0.964**	0.020	**0.984**	**0.984**	**0.993**	0.779	**0.961**	0.798
MEME_ZOOPS_1	**0.999**	0.492	0.001	**0.962**	0	**0.974**	**0.993**	0.784	**0.960**	0.685
MEME_OOPS_3	**0.999**	0.534	0.002	**0.914**	**0.984**	**0.999**	0.027	0.783	**0.961**	0.689
MEME_OOPS_2	**0.999**	0.513	0.001	**0.937**	**0.984**	**0.984**	0.005	0.783	**0.961**	0.685
MEME_OOPS_1	**0.999**	0.474	0.001	**0.961**	0	**0.974**	0	0.790	**0.960**	0.573
STREME_3	**0.993**	0.001	**0.998**	**0.981**	**0.991**	**0.980**	**0.978**	**0.982**	**0.910**	0.868
STREME_2	0.008	0.001	**0.998**	**0.981**	**0.991**	**0.980**	**0.979**	**0.982**	**0.912**	0.759
STREME_1	0.008	0	**0.986**	**0.983**	**0.991**	**0.983**	**0.983**	**0.936**	**0.912**	0.753
CASTOR_KRFE	**0.985**	**0.983**	**0.998**	**0.988**	**0.996**	**0.985**	**0.994**	**0.983**	**0.996**	**0.990**
KEVOLVE	**0.989** ± 0.007	**0.987** ± 0.005	**0.996** ± 0.003	**0.991** ± 0.003	**0.990** ± 0.005	**0.996** ± 0.003	**0.994** ± 0.002	**0.987** ± 0.071	**0.991** ± 0.004	**0.991** ± 0.003

Table 5. F1-score of the models associated with each tool for the prediction of the testing set.

	Alpha (B.1.1.7)	**Beta (B.1.351)**	**Delta (B.1.617.2)**	**Epsilon (B.1.427/9)**	**Eta (B.1.525)**	**Gamma (P.1)**	**Iota (B.1.526)**	**Kappa (1.617.1)**	**Zeta (P.2)**	**Mean**
MEME_ZOOPS_3	**0.999**	0.496	0.004	**0.951**	**0.984**	**0.999**	**0.993**	0.779	**0.962**	0.796
MEME_ZOOPS_2	**0.999**	0.498	**0.964**	0.020	**0.984**	**0.984**	**0.993**	0.779	**0.961**	0.798
MEME_ZOOPS_1	**0.999**	0.492	0.001	**0.962**	0	**0.974**	**0.993**	0.784	**0.960**	0.685
MEME_OOPS_3	**0.999**	0.534	0.002	**0.914**	**0.984**	**0.999**	0.027	0.783	**0.961**	0.689
MEME_OOPS_2	**0.999**	0.513	0.001	**0.937**	**0.984**	**0.984**	0.005	0.783	**0.961**	0.685
MEME_OOPS_1	**0.999**	0.474	0.001	**0.961**	0	**0.974**	0	0.790	**0.960**	0.573
STREME_3	**0.993**	0.001	**0.998**	**0.981**	**0.991**	**0.980**	**0.978**	**0.982**	**0.910**	0.868
STREME_2	0.008	0.001	**0.998**	**0.981**	**0.991**	**0.980**	**0.979**	**0.982**	**0.912**	0.759
STREME_1	0.008	0	**0.986**	**0.983**	**0.991**	**0.983**	**0.983**	**0.936**	**0.912**	0.753
CASTOR_KRFE	**0.985**	**0.983**	**0.998**	**0.988**	**0.996**	**0.985**	**0.994**	**0.983**	**0.996**	**0.990**
KEVOLVE	**0.989** ± 0.007	**0.987** ± 0.005	**0.996** ± 0.003	**0.991** ± 0.003	**0.990** ± 0.005	**0.996** ± 0.003	**0.994** ± 0.002	**0.987** ± 0.071	**0.991** ± 0.004	**0.991** ± 0.003

identified 21 subsets of discriminating motifs of size 8 while CASTOR identified only one subset of size 9.

The next best performer is STREME with STREME 3 obtaining average scores between 0.83 and 0.86 for Presicion, Recall and F1-Score. The drop in performance compared to CASTOR-KRFE and KEVOLVE is because it was not able to identify motifs that could characterize the Beta variant and showed scores close to 0. The performance of STREME 1 and 2 then drops to average scores between 0.75 and 0.78 for the different performance metrics. These results are explained because the majority of the Alpha sequences were predicted as Iota variants.

Then MEME ZOOPS performed worse than STREME. MEME ZOOPS 2 and 3 achieved scores between 0.79 and 0.86 for Precision, Recall and F1-score. Like STREME, MEME ZOOPS had difficulties in predicting the Beta variant. For both models many Kappa sequences were incorrectly predicted as Beta or Epsilon. In addition, MEME ZOOPS 3 for Delta and MEME ZOOPS 2 for Epsilon, respectively obtained an F1-score close to 0. For MEME ZOOPS 3, many Beta and Epsilon sequences were predicted Delta, and for MEME ZOOPS 3 the majority of Epsilon sequences were assigned to other types of variants. Finally, MEME ZOOP 1 presents the same errors as MEME ZOOP 3. Moreover, it was not able to identify any discriminative motifs related to Eta leading to scores of 0 for this variant.

Finally, the least performing models are those based on the motifs identified by MEME OOPS. MEME OOPS 3 and 2 obtained average scores between 0.68 and 0.72 for Precision, Recall and F1-score. These models have had similar problems as MEME ZOOPS in predicting the sequences of Beta, Delta and Kappa. In addition, MEME OOPS was not able to identify discriminative motifs associated with the Iota variant involving performance metric scores close to 0 for Iota sequence prediction. Finally, MEME OOPS 1 also encountered the same difficulties as MEME ZOOP 1 for the prediction of the Eta variant, which dropped its average scores to about 0.58 for Precision, Recall and F1-score.

Analysis of identified motifs by KEVOLVE

KEVOLVE was able to identify 74 unique motifs that discriminated between different sequences of SARS-CoV-2 variants.Fig 2 is a cluster map of 34 motifs that allows to visualize the discriminating potential by their percentage of presence/absence according to the different families of variants. The selected motifs are those that allow to discriminate one or several groups of variants by their presence or absence at more than 97%. Note that these motifs were identified by KEVOLVE during the comparative study from the training set of 2,250 SARS-CoV-2 genomes (250 sequences for the 9 types of variants). However, the results shown in Fig 2 are based on the 226,532 variant sequences (training and testing sets).

In the figure we can see motifs that allow to discriminate the different variants from the others. For the Alpha variant the ACTACAGA motif for example is absent from their genome while it is present in all the other variants. For the Beta variant, the AAAGTGGA motif is absent in about 97% of the cases in contrast to the other variants. Considering the Delta variant, we observe that the GACCTTAA and CGGTTCAC motifs are absent from their genome while they are present in the other variants. Focusing on the Epsilon variant, several motifs such as ATAGCGCT, CCTGTATA and TTACCTTA by their absence and presence can be used to discriminate it from other variants. The CCGCAATG motif, which is present only in Eta genomes, provides a genomic signature for this variant. The absence of the AAATATCT and GGGAATTT motifs in Gamma's genome allow discrimination from other variants where they are present. For Iota and Kappa it is for example the presence of the ATAACTGT and TATCTTAA motifs respectively that allow them to be distinguished from other variants. Finally, the Zeta variant can be characterized by the absence of the TGTATCAA motif, which is in contrast present in the genome of all other variants.

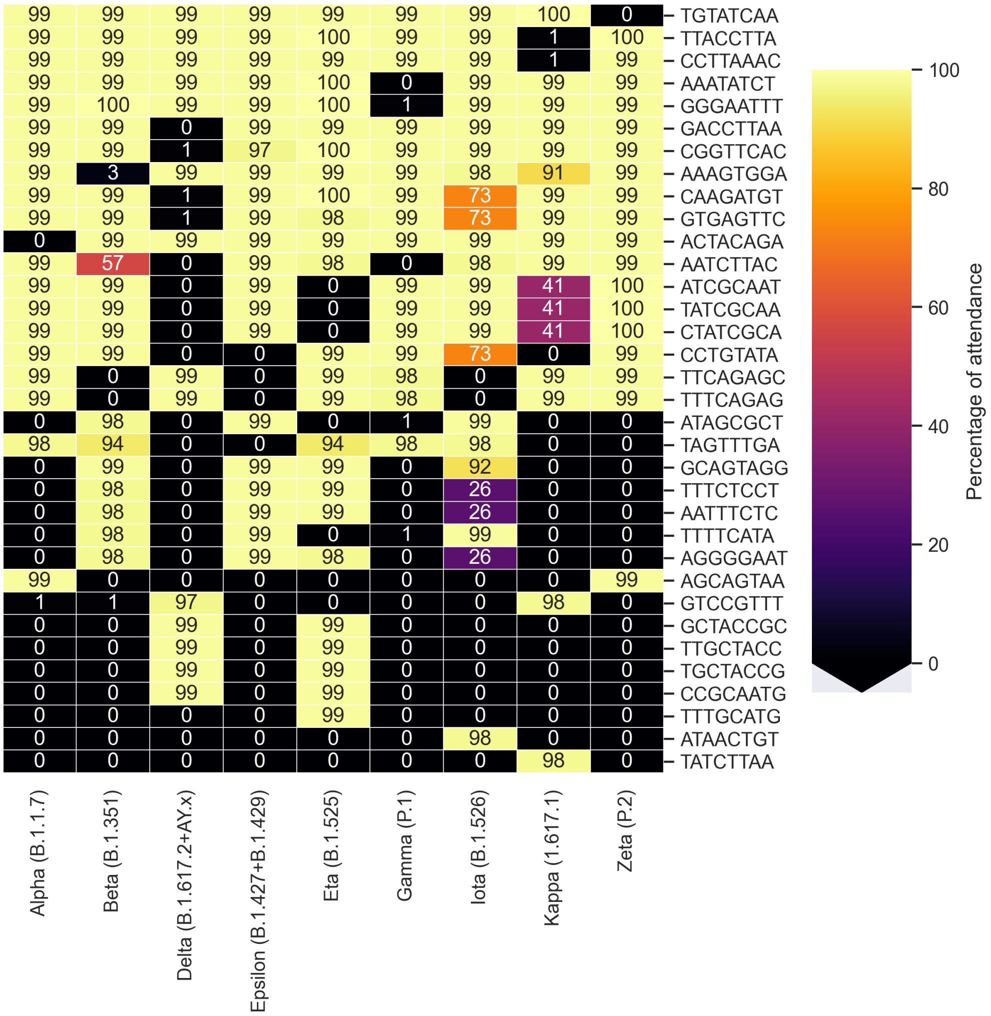
In summary, KEVOLVE identified from a small portion of the dataset multiple motifs that can discriminate by their absence or presence between different groups of SARS-CoV-2 variants. The discriminative potential of these motifs can be generalized to larger data sets as well as to constitute genomic signatures associated with SARS-CoV-2 variants.

Biological interests of the identified motifs

We analyzed the variant-discrimination motifs identified by KEVOLVE according to their location in the genome, to assess their potential functional impact of these mutations.

Preliminary sequence analyses

To study the motifs, we first used UGENE bioinformatics software [35] to perform multiple alignment by selecting 50 genomes per variant family from our training sequence set. In addition, 50 Omicron variant genomes were included to form a set of

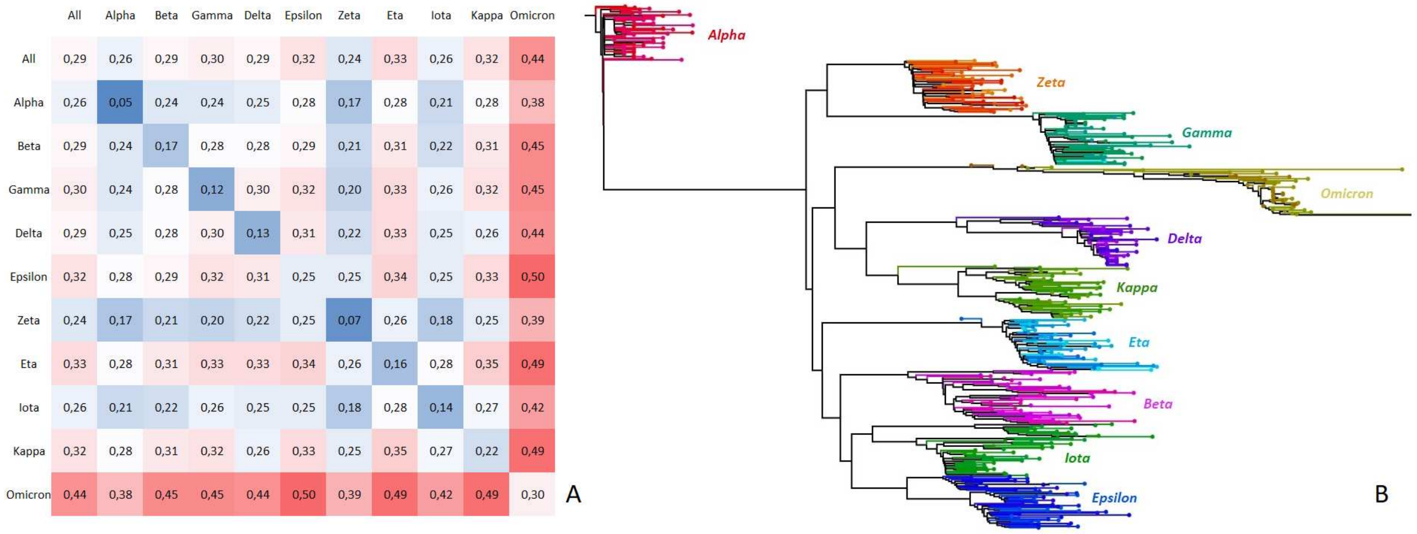


	Alpha (B.1.1.7)	Beta (B.1.351)	Delta (B.1.617.2+AY.x)	Epsilon (B.1.427+B.1.429)	Eta (B.1.525)	Gamma (P.1)	Iota (B.1.526)	Kappa (1.617.1)	Zeta (P.2)
TGTATCAA	99	99	99	99	99	99	99	100	0
TTACCTTA	99	99	99	99	100	99	99	1	100
CCTTAAAC	99	99	99	99	99	99	99	1	99
AAATATCT	99	99	99	99	100	0	99	99	99
GGGAATTT	99	100	99	99	100	1	99	99	99
GACCTTAA	99	99	0	99	99	99	99	99	99
CGGTTCAC	99	99	1	97	100	99	99	99	99
AAAGTGGA	99	3	99	99	99	99	98	91	99
CAAGATGT	99	99	1	99	100	99	73	99	99
GTGAGTTC	99	99	1	99	98	99	73	99	99
ACTACAGA	0	99	99	99	99	99	99	99	99
AATCTTAC	99	57	0	99	98	0	98	99	99
ATCGCAAT	99	99	0	99	0	99	99	41	100
TATCGCAA	99	99	0	99	0	99	99	41	100
CTATCGCA	99	99	0	99	0	99	99	41	100
CCTGTATA	99	99	0	0	99	99	73	0	99
TTCAGAGC	99	0	99	0	99	98	0	99	99
TTTCAGAG	99	0	99	0	99	98	0	99	99
ATAGCGCT	0	98	0	99	0	1	99	0	0
TAGTTTGA	98	94	0	0	94	98	98	0	0
GCAGTAGG	0	99	0	99	99	0	92	0	0
TTTCTCCT	0	98	0	99	99	0	26	0	0
AATTTCTC	0	98	0	99	99	0	26	0	0
TTTTCATA	0	98	0	99	0	1	99	0	0
AGGGGAAT	0	98	0	99	98	0	26	0	0
AGCAGTAA	99	0	0	0	0	0	0	0	99
GTCCGTTT	1	1	97	0	0	0	0	98	0
GCTACCGC	0	0	99	0	99	0	0	0	0
TTGCTACC	0	0	99	0	99	0	0	0	0
TGCTACCG	0	0	99	0	99	0	0	0	0
CCGCAATG	0	0	99	0	99	0	0	0	0
TTTGCATG	0	0	0	0	99	0	0	0	0
ATAACTGT	0	0	0	0	0	0	98	0	0
TATCTTAA	0	0	0	0	0	0	0	98	0

Fig 2. Cluster map representing the percentage of presence of motifs identified by KEVOLVE according to the groups of variants of SARS-CoV-2.

500 sequences that were aligned by MUSCLE algorithm [14] in large alignment mode. From this alignment, we calculated the dissimilarity matrix based on Hamming distances. Finally, the matrix representing the dissimilarity percentages of nucleotide between the different groups of SARS-CoV-2 variants as well as a phylogenetic tree (Fig 3) based on the neighbour-joining method [36] was computed.

From this matrix we can observe that the divergence between the genomes of the several clusters of SARS-CoV-2 variants is less than 1% and the mean divergence between all the sequences is 0.29%. Focusing on the phylogentic tree on the right of Fig 3 we observe that this divergence is sufficient to cluster the variant families. Considering the columns related to Omicron as well as the phylogenetic tree, we observe that Omicron is the most divergent. It diverges by 0.44% compared to the other variants and shows an intra-variant divergence of 0.30%. Lastly, the Alpha, Zeta and Iota variants are the least divergent (0.26%, 0.24% and 0.26% respectively compared to the other variants) and (0.05%, 0.007% and 0.14% intra variant divergence).



	All	Alpha	Beta	Gamma	Delta	Epsilon	Zeta	Eta	Iota	Kappa	Omicron
All	0,29	0,26	0,29	0,30	0,29	0,32	0,24	0,33	0,26	0,32	0,44
Alpha	0,26	0,05	0,24	0,24	0,25	0,28	0,17	0,28	0,21	0,28	0,38
Beta	0,29	0,24	0,17	0,28	0,28	0,29	0,21	0,31	0,22	0,31	0,45
Gamma	0,30	0,24	0,28	0,12	0,30	0,32	0,20	0,33	0,26	0,32	0,45
Delta	0,29	0,25	0,28	0,30	0,13	0,31	0,22	0,33	0,25	0,26	0,44
Epsilon	0,32	0,28	0,29	0,32	0,31	0,25	0,25	0,34	0,25	0,33	0,50
Zeta	0,24	0,17	0,21	0,20	0,22	0,25	0,07	0,26	0,18	0,25	0,39
Eta	0,33	0,28	0,31	0,33	0,33	0,34	0,26	0,16	0,28	0,35	0,49
Iota	0,26	0,21	0,22	0,26	0,25	0,25	0,18	0,28	0,14	0,27	0,42
Kappa	0,32	0,28	0,31	0,32	0,26	0,33	0,25	0,35	0,27	0,22	0,49
Omicron	0,44	0,38	0,45	0,45	0,44	0,50	0,39	0,49	0,42	0,49	0,30

Fig 3. Nucleotide rate dissimilarity matrix and phylogenetic tree of SARS-CoV-2 variant families.

Mutations and potential impact associated with motifs

The motif analyses were supported by the computed multiple alignment and the SnapGene software (from Insightful Science; available at snapgene.com) using the sequence of references NC_045512 (Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome). In addition, we analyzed all 72 Omicron genomes with KEVOLVE to identify motifs that discriminate this variant from others present in the training set. From this analysis, we highlighted in Table 6, the different mutations contained in the identified motifs, where they are located and the associated variants.

Concerning the Alpha variant, the identified motifs highlighted the D1118H mutation located in the Spike glycoprotein, the SGF3675-3677 deletions located in ORF1ab (NSP6), which is also present in the Beta, Gamma, Eta and Iota variants, and the substitutions R203K / G204R, which are shared with the Zeta and Omicron variants. A recent study [37] showed that the 203K/204R mutation located in ORF9 (Protein N) is associated with increased COVID-19 infectivity. Thus, this mutation is potentially a major contributor to the high contagiousness of Omicron.

For the Beta variant, the motifs pointed out the K1655N mutation in ORF1ab (NSP3), the Q57H mutation located in ORF3a, which is present in Epsilon and Iota, as well as the T205I mutation shared with Epsilon and Eta and which is located in ORF9 (Protein N). Regarding the Gamma variant, KEVOLVE identified motifs that contain three characteristic substitutions of this variant [38] which are: K1795Q in ORF1ab (NSP3), R190S and L18F in ORF2 (Spike Protein S1).

For Delta variant-associated motifs, they highlighted D63G (ORF9 (Protein N)), G5063S (ORF1b (NSP12)), D950N (ORF2 (Protein Spike S2)), 156del / 157del (ORF2 (Protein Spike S1)), and T19R (ORF2 (Protein Spike S1)) mutations that are specific to Delta [39, 40].

In the motifs, we also identified the I82T mutation located in ORF5 (membrane protein), which has been proposed to increase replication fitness through alteration of cellular glucose uptake during viral replication [41]. Our analysis also confirms the presence of this mutation in the Eta variant [42]. The L452R mutation located in the spike protein, which increases fusogenicity and promotes viral replication and

Table 6. Mutational landscape associated to the motifs identified by KEVOLVE of the different emerging variants.

Variant name	Motifs	Mutation	Mutation position	Amino acid change
Alpha (B.1.1.7)	ACTACAGA ⇒ ACTACACA	D1118H	ORF2 (Protein Spike S2)	Asp1118 ⇒ His
Alpha (B.1.1.7) Beta (B.1.351) Gamma (P.1) Eta (B.1.525) Iota (B.1.526)	TAGTTTGTCTGGTTTTA ⇒ TAGTTTG———A	Del SGF3675-3677	ORF1ab (NSP6)	Deletion mutation
Alpha (B.1.1.7) Zeta (P.2) Omicron (B.1.1.529)	CAGTAGGG ⇒ CAGTAAAC	R203K / G204R	ORF9 (Protein N)	Arg203 ⇒ Lys / Gly204 ⇒ Arg
Beta (B.1.351)	AAAGTGGA ⇒ AAATTGGA	K1655N	ORF1ab (NSP3)	Lys1655 ⇒ Asn
Beta (B.1.351) Epsilon (B.1.427/9) Iota (B.1.526)	TTCAGAGC ⇒ TTCATAGC	Q57H	ORF3a	Gln57 ⇒ His
Beta (B.1.351) Epsilon (B.1.427/9) Eta (B.1.525)	AACTTCTC ⇒ AATTTCTC	T205I	ORF9 (Protein N)	Thr205 ⇒ Ile
Gamma (P.1)	AAATATCT ⇒ CAATATCT	K1795Q	ORF1ab (NSP3)	Lys1795 ⇒ Gln
	GGGAATTT ⇒ GTGAATTT	R190S	ORF2 (Protein Spike S1)	Arg190 ⇒ Ser
	AATCTTAC ⇒ AATTTTAC	L18F	ORF2 (Protein Spike S1)	Leu18 ⇒ Phe
	GCAGTAGG ⇒ GCTCTAAA	Silent Mutation	ORF2 (Protein Spike S2)	N/A
Delta (B.1.617.2)	GACCTTAA ⇒ GGCCTTAA	D63G	ORF9 (Protein N)	Asp63 ⇒ Gly
	CGGTTCAC ⇒ CAGTTCAC	G5063S	ORF1b (NSP12)	Gly5063 ⇒ Ser
	CAAGATGT ⇒ CAAAATGT	D950N	ORF2 (Protein Spike S2)	Asp950 ⇒ Asn
	GTGAGTTC ⇒ GTG——	156del / 157del	ORF2 (Protein Spike S1)	Deletion mutation
	AATCTTAC ⇒ AATCTTAG	T19R	ORF2 (Protein Spike S1)	Thr19 ⇒ Arg
Delta (B.1.617.2) Eta (B.1.525)	ATCGCAAT ⇒ ACCGCAAT	I82T	ORF5 (Protein membranaire)	Ile82 ⇒ Thr
Delta (B.1.617.2) Epsilon (B.1.427/9) Kappa (1.617.1)	CCTGTATA ⇒ CCGGTATA	L452R	ORF2 (Protein Spike S1)	Leu452 ⇒ Arg
Delta (B.1.617.2) Kappa (1.617.1)	GCAGTAGG ⇒ GCAGTATG	R203M	ORF9 (Protein N)	Arg203 ⇒ Met
	GTCCGTGT ⇒ GTCCGTTT	N/A	5' UTR	N/A
Eta (B.1.525)	CTTGCATG ⇒ TTTGCATG	Silent Mutation	ORF2 (Protein Spike S2)	Phe1062 ⇒ Phe
Iota (B.1.526)	ACAACTGT ⇒ ATAACTGT	T11I	ORF8	Thr11 ⇒ Ile
Kappa (B.1.617.1)	TTACCTTA ⇒ TTATCTTA	Silent Mutation	ORF1ab (NSP3)	Tyr1064 ⇒ Tyr
Zeta (P.2)	TGTATCAA ⇒ TGTATTAA	Silent Mutation	ORF1ab (NSP6)	Ile3053 ⇒ Ile
Omicron (B.1.1.529)	GCTGCTAA ⇒ GCGGCTAA	Silent Mutation	ORF1ab (NSP3)	Ala1707 ⇒ Ala
	AGAGGTAT ⇒ AGAGGTGT	I3758V	ORF1ab (NSP6)	Ile3758 ⇒ Val
	ACTAATTC ⇒ ACTAAGTC	N679K	ORF2 (Protein Spike)	Asn679 ⇒ Lys
	TTAAAGAT ⇒ TTAAATAT	D796Y	ORF2 (Protein Spike)	Asp796 ⇒ Tyr
	AATTAGAC ⇒ AATTAGAT	Silent Mutation	ORF2 (Protein Spike)	Asp1146 ⇒ Asp
	CATAACCC ⇒ CATAACTC	Silent Mutation	ORF3a	Thr64 ⇒ Thr
	TATTATGA ⇒ TATTATGC	Silent Mutation	ORF6	Arg20 ⇒ Arg

infectivity [43], was also found in motifs within the Delta, Epsilon and Kappa genomes.

Finally, three substitutions constituting unique features of Omicron were highlighted, by KEVOLVE: I3758V in ORF1ab (NSP6) and N679K and D796Y in ORF2 (Spike protein) [44]. The functional implications of these Omicron variant mutations are unknown, leaving many questions about how they may affect viral fitness and vulnerability to natural and vaccine-mediated immunity [45]. However, the combination of N679K with H655Y and P681H, due to their proximity to the furin cleavage site, could increase the cleavage of spike, enhancing fusion and viral transmission [46].

Conclusion

In this study, we compared the ability of machine learning-based tools to classify SARS-CoV-2 variants compared to statistical tools specialized in discriminative motif identification. We found that the identification of motifs in SARS-CoV-2 genome sequences readily discriminates different groups of variants. However, the machine learning-based approaches, CASTOR-KRFE and KEVOLVE, were generally more efficient. The predictive models based on the motifs (8 for KEVOLVE and 9 for CASTOR-KRFE) identified by these two approaches predict a large set of SARS-CoV-2 variant sequences with an average F1 score greater than 0.98. Furthermore, these two approaches predicted a large set of SARS-CoV-2 variant sequences (over 225,000) with an average F1-score greater than 0.98. In contrast, the model involving the most motifs (26), using STREME, which was the best performing approach after KEVOLVE and CASTOR-KRFE, only obtained an average F1-score of 0.836. In addition, unlike the statistical approaches, KEVOLVE and CASTOR-KRFE, can deal with multi-class sets and are not limited to binary sets. In addition, KEVOLVE is distinguished by its ability to identify multiple discriminative sets unlike other tools that are limited to a single optimal set.

Subsequently, we analyzed the motifs identified by KEVOLVE with respect to their recognized or potential functional importance from the existing literature. Not surprisingly, we found that the majority of SARS-CoV-2 motifs identified by KEVOLVE were associated with known mutations among the different viral variants. However, of interest, several motifs derived from CASTOR-KRFE and KEVOLVE did not correspond to recognized variant-specific mutations. With respect to Omicron, 4 motifs contained what appear to be silent mutations, indicating potentially novel variant-specific virulence determinants [47]. Interestingly, although Omicron displays increased transmissibility and evades vaccine-induced and natural-acquired neutralizing antibodies through its numerous spike mutations, it may also cause less severe disease, perhaps due to altered tissue tropism [48, 49]. As the genetic basis of SARS-CoV-2 virulence remain incompletely understood, variant-discriminating mutations represent valuable targets for understanding differences in viral phenotypes and clinical outcomes.

These results suggest that KEVOLVE is a robust tool for the rapid and accurate determination of SARS-CoV-2 variants. The identified motifs provide genomic signatures that can be used to build peptide or oligonucleotide libraries for rapid and accurate pathogen detection using tools such as VirScan [50]. The identification of motifs by KEVOLVE is automatic and independent of multiple sequence alignments, in contrast to traditional methods by which mutations are associated with variant-discriminating motifs. Indeed, such analyses require manual verification based on annotated reference sequences and multiple sequence alignment, making them impractical for variant discrimination of diverse viruses with large and complex genome structures, such as cytomegalovirus [51]. KEVOLVE and CASTOR-KRFE can also be adapted to allow the automatic analysis of previously-identified motifs, further increasing its efficiency.

In summary, we have shown that machine learning-based tools has numerous advantages over statistical tools and conventional alignment-based methods for efficiently discriminating among SARS-CoV-2 variants. This new approach is independent of multiple sequence alignment and allows users to capture mutations associated with motifs of interest in different groups of viral pathogens. Moreover, these machine learning-based approaches may rapidly identify novel motifs that point toward otherwise unrecognized mutations of functional importance, in new variants such as Omicron. Thus, ML-based/KEVOLVE is a useful adjunct to conventional genomic analyses to classify and understand viral variants.

Acknowledgments

Abdoulaye Baniré Diallo holds both a Natural Sciences and Engineering Research Council of Canada (NSERC) discovery and a Genome Canada grant.

References

1. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020;583(7816):459–468.

2. Lee EY, Ng MY, Khong PL. COVID-19 pneumonia: what has CT taught us? The Lancet Infectious Diseases. 2020;20(4):384–385.

3. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The lancet. 2020;395(10224):565–574.

4. Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019-nCoV in Italy: Where they come from? Journal of medical virology. 2020;92(5):518–521.

5. Sallam M, Ababneh NA, Dababseh D, Bakri FG, Mahafzah A. Temporal increase in D614G mutation of SARS-CoV-2 in the Middle East and North Africa. Heliyon. 2021;7(1):e06035.

6. Klitting R, Fischer C, Drexler JF, Gould EA, Roiz D, Paupy C, et al. What does the future hold for yellow fever virus?(II). Genes. 2018;9(9):425.

7. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. Journal of human genetics. 2020;65(12):1075–1082.

8. Koyama T, Weeraratne D, Snowdon JL, Parida L. Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment. Pathogens. 2020;9(5):324.

9. Randhawa GS, Soltysiak MP, El Roz H, de Souza CP, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. Plos one. 2020;15(4):e0232391.

10. Lopez-Rincon A, Tonda A, Mendoza-Maldonado L, Mulders DG, Molenkamp R, Perez-Romero CA, et al. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Scientific reports. 2021;11(1):1–11.

11. Bauer DC, Tay AP, Wilson LO, Reti D, Hosking C, McAuley AJ, et al. Supporting pandemic response using genomics and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. Transboundary and emerging diseases. 2020;67(4):1453–1462.

12. Slezak T, Hart B, Jaing C. Design of genomic signatures for pathogen identification and characterization. In: Microbial Forensics. Elsevier; 2020. p. 299–312.

13. Zielezinski A, Girgis HZ, Bernard G, Leimeister CA, Tang K, Dencker T, et al. Benchmarking of alignment-free sequence comparison methods. Genome biology. 2019;20(1):1–18.

14. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792–1797.

15. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. bioinformatics. 2007;23(21):2947–2948.

16. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772–780.

17. Bernard G, Chan CX, Chan Yb, Chua XY, Cong Y, Hogan JM, et al. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Briefings in Bioinformatics. 2019;20(2):426–435.

18. Lange K. Mathematical and statistical methods for genetic analysis. vol. 488. Springer; 2002.

19. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome biology. 2017;18(1):1–17.

20. Eddy SR. What is dynamic programming? Nature biotechnology. 2004;22(7):909–910.

21. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics. 2008;9(4):267–276.

22. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008;319(5862):473–476.

23. Bailey TL, Elkan C, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994;.

24. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic acids research. 2015;43(W1):W39–W49.

25. Bailey TL, Bodén M, Whitington T, Machanick P. The value of position-specific priors in motif discovery using MEME. BMC bioinformatics. 2010;11(1):1–14.

26. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37(18):2834–2840.

27. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015;16(6):321–332.

28. Remita MA, Halioui A, Malick Diouara AA, Daigle B, Kiani G, Diallo AB. A machine learning approach for viral genome classification. BMC bioinformatics. 2017;18(1):1–11.

29. Solis-Reyes S, Avino M, Poon A, Kari L. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. PloS one. 2018;13(11):e0206409.

30. Lebatteux D, Remita AM, Diallo AB. Toward an alignment-free method for feature extraction and accurate classification of viral sequences. Journal of Computational Biology. 2019;26(6):519–535.

31. Lebatteux D, Diallo AB. Combining a genetic algorithm and ensemble method to improve the classification of viruses. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 688–693.

32. Zhang Q, Jun SR, Leuze M, Ussery D, Nookaew I. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. Scientific reports. 2017;7(1):1–13.

33. Narlikar L, Gordân R, Hartemink AJ. Nucleosome occupancy information improves de novo motif discovery. In: Annual International Conference on Research in Computational Molecular Biology. Springer; 2007. p. 107–121.

34. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global challenges. 2017;1(1):33–46.

35. Okonechnikov K, Golosova O, Fursov M, Team U. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166–1167.

36. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution. 1987;4(4):406–425.

37. Wu H, Xing N, Meng K, Fu B, Xue W, Dong P, et al. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell host & microbe. 2021;29(12):1788–1801.

38. Slavov SN, Patané JS, Bezerra RdS, Giovanetti M, Fonseca V, Martins AJ, et al. Genomic monitoring unveil the early detection of the SARS-CoV-2 B. 1.351 (beta) variant (20H/501Y. V2) in Brazil. Journal of Medical Virology. 2021;93(12):6782–6787.

39. Gupta N, Kaur H, Yadav P, Mukhopadhyay L, Sahay RR, Kumar A, et al. Clinical characterization and Genomic analysis of COVID-19 breakthrough infections during second wave in different states of India. medRxiv. 2021;.

40. Kannan SR, Spratt AN, Cohen AR, Naqvi SH, Chand HS, Quinn TP, et al. Evolutionary analysis of the Delta and Delta Plus variants of the SARS-CoV-2 viruses. Journal of autoimmunity. 2021;124:102715.

41. Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. Emerging microbes & infections. 2021;10(1):885–893.

42. Chakraborty C, Sharma AR, Bhattacharya M, Agoramoorthy G, Lee SS. Evolution, mode of transmission, and mutational landscape of newly emerging SARS-CoV-2 variants. Mbio. 2021;12(4):e01140–21.

43. Motozono C, Toyoda M, Zahradnik J, Saito A, Nasser H, Tan TS, et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell host & microbe. 2021;29(7):1124–1136.

44. Kannan SR, Spratt AN, Sharma K, Chand HS, Byrareddy SN, Singh K. Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies. Journal of Autoimmunity. 2022;126:102779.

45. Sarkar R, Lo M, Saha R, Dutta S, Chawla-Sarkar M. S glycoprotein diversity of the Omicron variant. MedRxiv. 2021;.

46. He X, Hong W, Pan X, Lu G, Wei X. SARS-CoV-2 Omicron variant: characteristics and prevention. MedComm. 2021;.

47. Berrio A, Gartner V, Wray GA. Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. PeerJ. 2020;8:e10234.

48. Chan MC, Hui KP, Ho J, Cheung Mc, Ng Kc, Ching R, et al. SARS-CoV-2 Omicron variant replication in human respiratory tract ex vivo. 2021;.

49. Diamond M, Halfmann P, Maemura T, Iwatsuki-Horimoto K, Iida S, Kiso M, et al. The SARS-CoV-2 B. 1.1. 529 Omicron virus causes attenuated infection and disease in mice and hamsters. 2021;.

50. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, et al. Comprehensive serological profiling of human populations using a synthetic human virome. Science. 2015;348(6239).

51. Lassalle F, Depledge DP, Reeves MB, Brown AC, Christiansen MT, Tutill HJ, et al. Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus evolution. 2016;2(1).