RESEARCH ARTICLE

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

Dylan Lebatteux[1], Hugo Soudeyns[2,3,4], Isabelle Boucoiran[5], Soren Gantt[2,3], Abdoulaye Baniré Diallo[1]*

1 Department of Computer Science, Université du Québec à Montréal, Montréal, Québec, Canada, **2** CHU Sainte-Justine Research Centre, Montréal, Québec, Canada, **3** Department of Microbiology, Infectious Diseases and Immunology, Faculty of Medicine, Université de Montréal, Montréal, Québec, Canada, **4** Department of Pediatrics, Faculty of Medicine, Université du Québec à Montréal, Montréal, Québec, Canada, **5** Department of Obstetrics and Gynecology, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada

* diallo.abdoulaye@uqam.ca

Citation: Lebatteux D, Soudeyns H, Boucoiran I, Gantt S, Diallo AB (2024) Machine learning-based approach KEVOLVE efficiently identifies SARS-CoV-2 variant-specific genomic signatures. PLoS ONE 19(1): e0296627. https://doi.org/10.1371/journal.pone.0296627

Editor: Nagarajan Raju, Emory University, UNITED STATES

Received: February 21, 2022

Accepted: December 7, 2023

Published: January 19, 2024

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files located on GitHub (https://github.com/bioinfoUQAM/Kevolve).

Funding: This work is funded by the Canadian Institute of Health Research. Abdoulaye Banire Diallo is supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). Hugo Soudeyns and Isabelle Boucoiran are supported by an infrastructure grant from Réseau SIDA et MI of Fonds de la recherche

Abstract

Machine learning was shown to be effective at identifying distinctive genomic signatures among viral sequences. These signatures are defined as pervasive motifs in the viral genome that allow discrimination between species or variants. In the context of SARS-CoV-2, the identification of these signatures can assist in taxonomic and phylogenetic studies, improve in the recognition and definition of emerging variants, and aid in the characterization of functional properties of polymorphic gene products. In this paper, we assess KEVOLVE, an approach based on a genetic algorithm with a machine-learning kernel, to identify multiple genomic signatures based on minimal sets of *k*-mers. In a comparative study, in which we analyzed large SARS-CoV-2 genome dataset, KEVOLVE was more effective at identifying variant-discriminative signatures than several gold-standard statistical tools. Subsequently, these signatures were characterized using a new extension of KEVOLVE (KANALYZER) to highlight variations of the discriminative signatures among different classes of variants, their genomic location, and the mutations involved. The majority of identified signatures were associated with known mutations among the different variants, in terms of functional and pathological impact based on available literature. Here we showed that KEVOLVE is a robust machine learning approach to identify discriminative signatures among SARS-CoV-2 variants, which are frequently also biologically relevant, while bypassing multiple sequence alignments. The source code of the method and additional resources are available at: https://github.com/bioinfoUQAM/KEVOLVE.

Introduction

Severe acute respiratory syndrome coronavirus (SARS-CoV-2) is the etiological agent of coronavirus disease 2019 (COVID-19). This highly infectious coronavirus was first identified in December 2019 in Wuhan, China [1]. It belongs to the betacoronavirus genus, which includes

du Québec-santé (FRQS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: Soren Gantt receive research funds from Moderna for a COVID-19 vaccine trial, as well as research and consulting fees from Moderna and Merck related to CMV vaccine development, and consulting fees from GSK related to vaccine safety.; Abdoulaye Banire Diallo is co-founder of My Intelligent Machines. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

SARS-CoV-1 and Middle East respiratory syndrome-related coronavirus (MERS-CoV) [2]. The genome of SARS-CoV-2 is a single-stranded RNA molecule composed of approximately 30,000 nucleotides. The nucleotide sequence identity of SARS-CoV-2 with SARS-CoV-1 and MERS-CoV is 79.5% and 50%, respectively [3, 4]. The SARS-CoV-2 genome encodes 29 different proteins, including 16 nonstructural proteins, 4 structural proteins, and 9 accessory proteins (see Fig 1 adapted from [5]). The N (nucleocapsid) protein contains the viral RNA genome, while the S (spike), E (envelope), and M (membrane) proteins together form the viral envelope [6]. SARS-CoV-2 exhibits a notably high mutation rate, with numerous mutations—particularly in the spike gene—correlated to increased SARS-CoV-2 transmission rates [7], augmented fusogenic and pathogenic properties of the virus [8], as well as the emergence of new variants that could diminish the efficacy of existing COVID-19 vaccines and antibody-based therapies [9].

Given its rapid rate of evolution, it is important to be able to efficiently identify genomic signatures that can distinguish between different variants of SARS-CoV-2 and highlight potential functional changes. These signatures, also known as species- or variant-specific motifs that are prevalent throughout the viral genome [10], can contribute to taxonomic [11] and phylogenetic [12] studies to differentiate distinct groups of variants, provide insight into their evolutionary history [10], help to understand the structure of the viral quasispecies [13], and facilitate mechanistic studies to determine the functional basis of variant-specific differences in virulence [14]. To identify discriminative motifs, or genomic signatures, among different groups of biological sequences, the traditional approach is to compute multiple sequence alignments using tools such as MUSCLE [15], Clustal W/X [16], or MAFFT [17]. These alignments are then analyzed to identify divergent genomic regions that constitute the discriminative motifs. However, multiple alignment approaches have significant limitations when applied to viral genomes [14].

First, alignment-based approaches are generally computationally and time-intensive, making them less well suited for dealing with large viral sequence datasets that are increasingly available [18]. In fact, computing an accurate multi-sequence alignment is an NP-hard problem with $(2N)!/(N!)^2$ possible alignments for two sequences of length N [19], which means that in some cases, the alignment cannot be solved within a realistic time frame or involves significant compromise in accuracy [17]. 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

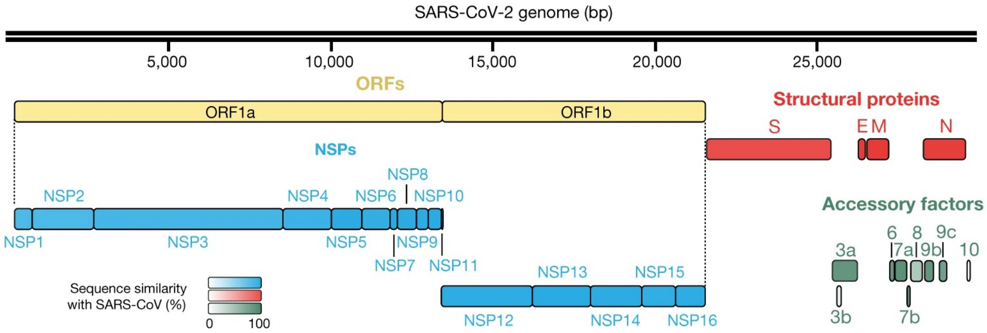


Fig 1. SARS-CoV-2 genome organization. Four structural proteins (red), 16 non-structural proteins (NSPs; blue), and 9 accessory factors (green) are shown. ORFs (open reading frames; yellow) 1a and 1b encode polyproteins. The protein sequence similarity with SARS-CoV homologues (when homologues exist) is depicted by the color intensity.

https://doi.org/10.1371/journal.pone.0296627.g001

questionable, especially for RNA viruses [21]. 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 [22]. Finally, performing multiple alignments often requires adjusting several parameters, such as substitution matrices, deviation penalties, and thresholds for statistical parameters, which are dependent on prior knowledge about the evolution of the compared sequences [21]. However, the adjustment of these parameters is sometimes arbitrary and requires a trial-and-error approach, and research has shown that small variations in these parameters can significantly impact the quality of alignments [23].

To address the limitations of discriminative motif identification using multiple sequence alignment, specialized statistical-based tools were developed, such as MEME [24, 25]. MEME has a discriminative mode [26] that identifies enriched motifs that distinguish a primary set of sequences from a control set. Other MEME tools were also developed, including STREME [27], the most powerful tool for discovering motifs in sequence datasets. STREME uses a generalized suffix tree and evaluates motifs using a statistical test that compares the enrichment of matches to the motif in the primary set of sequences to the control set [27]. In recent years, a series of machine-learning techniques were developed and widely used in the field of genomics, and were proven to be highly effective for solving complex and large-scale data analysis problems [28]. For example, the CASTOR study [29] demonstrated the usefulness of machine learning models coupled with restriction fragment length polymorphism (RFLP) signatures for classifying viral genomic sequences, achieving f1-scores ≥ 0.99 for predicting hepatitis B virus and human papillomavirus genomes. However, these signatures were found to have limitations in predicting human immunodeficiency viruses (HIV) sequences, resulting in an f1-score ≤ 0.90. To address this issue, the KAMERIS study [30] used k-mers (nucleotide subsequences of length k) to characterize the sequences provided to the learning model. To reduce the exponential number of features (4^k) associated with k-mers, KAMERIS applied truncated singular value decomposition for dimensionality reduction, but this transformation affected the ability to identify and analyze relevant features identified by the machine-learning model for discriminating between groups of sequences.

In response to this challenge, CASTOR-KRFE [31] was developed as a method for identifying minimal sets of genomic signatures based on minimal sets of k-mers to discriminate among multiple groups of genomic sequences. During cross-validation evaluations covering a wide range of viruses, CASTOR-KRFE successfully identified minimal sets of motifs, which when combined with supervised learning algorithms, resulted in average f1-scores ≥ 0.96 [31]. However, this study was limited to identifying the optimal set of motifs, rather than exploring suboptimal sets in the feature space, which can be a major limitation when dealing with viral sequences with high genomic diversity or when attempting to infer biological functions based on the identified motifs. To overcome this limitation, KEVOLVE [32] was developed as a new method that uses a genetic algorithm incorporating a machine-learning kernel to identify multiple minimal subsets of discriminative motifs. A preliminary comparative study on HIV nucleotide sequences showed that the KEVOLVE-identified motifs allowed for the construction of models that outperformed specialized HIV prediction tools [32]. In the context of the COVID-19 pandemic, this paper assessed the performance of KEVOLVE in a comparative study with several reference tools (MEME, STREME, and CASTOR-KRFE) for identifying discriminative motifs in the genomes of SARS-CoV-2 variants. The identified motifs were then analyzed using the new KEVOLVE extension (KANALYZER) to extract the associated information, and this information, which is discussed in light of the available literature to highlight the potential biological functions of the sequences/motifs in question.

Materials and methods

To assess the accuracy of KEVOLVE in identifying discriminative motifs, we conducted a comparative study with specialized tools. This involved using 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 variant. In a second step, we used these signatures and a supervised learning algorithm to fit a prediction model on the training sequences. Then, we evaluated the quality of the signatures by predicting the trained models on a large test set of unknown sequences. Finally, we used KANALYZER, the latest extension of KEVOLVE, to analyze the variant-discriminative motifs identified by KEVOLVE and assess their potential functional impact based on their location in the genome, as previously described in the literature.

Discriminative motif identification tools

We first evaluated KEVOLVE [32], a machine learning method based on a genetic algorithm for identifying multiple minimal sets of *k*-mers to discriminate nucleotide sequences. KEVOLVE takes as input a set of labeled nucleotide sequences and a parameter *k*, which corresponds to the length of the *k*-mers used to represent the sequences in an occurrence matrix. KEVOLVE starts by using a meta-transformer to remove *k*-mers with low discriminative contribution based on importance weights assigned by a linear Support Vector Machine (SVM). Then, the genetic algorithm begins its search by initializing several subsets (chromosomes) composed of a reduced set of *k*-mers (genes). Each chromosome is evaluated in a cross-validation process where prediction models are trained and tested on nucleotide sequences represented by the genes in the chromosome. The chromosomes with the best scores are then subjected to mutation/crossover processes. The mutation process involves randomly substituting a gene with another within a chromosome, and the crossover process involves exchanging genes between different chromosomes. In addition, the genes in the best chromosome have an increased probability of being selected in the next iteration. The next generation is then composed of the best current chromosomes and new chromosomes, which are generated based on the updated probability of selection. This process is repeated and coupled with a progressive increase in chromosome size until a stopping criterion is met (number of iterations or performance score of the solutions). The detailed KEVOLVE pseudo code is available in the original article [32], and the algorithm code can be accessed in the GitHub repository.

The second tool we evaluated was CASTOR-KRFE [31], 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). It identifies the optimal length of *k* to maximize classification performance and minimize the number of features, providing a solution to the problem of identifying the optimal length of *k*-mers for genomic sequence classification [33]. The third tool we evaluated was MEME (discriminative mode) [26], a tool from the MEME suite [25] specialized in motif identification. MEME takes two sets of sequences as input and identifies enriched motifs that discriminate the primary set from the control set. By default, MEME assumes that all positions in the sequences have an equal chance of being a motif site. However, in discriminative mode, the algorithm uses additional information such as sequence conservation, nucleosome positioning, and negative examples to compute a measure of the probability that a discriminative motif starts at each position in each sequence [26]. This measure, called "position specific prior" (PSP), is then used to guide the sequence motif discovery algorithm in the primary set, resulting in motifs that are more likely to discriminate it from the control set [34]. MEME also allows for the specification of a potential motif distribution type to improve the sensitivity and quality of the motif search. There are two available options in discriminative mode: zero or one occurrence per sequence (ZOOPS), where MEME assumes that each sequence may contain at most one occurrence of each motif, and one occurrence per sequence (OOPS), where MEME assumes that each sequence in the dataset contains exactly one occurrence of each motif. The last tool we evaluated was STREME [27], which was found to be more accurate, sensitive, and thorough than several widely used algorithms in a recent comparative study [27]. STREME's algorithm uses 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 control set. STREME assumes that each primary sequence may contain ZOOPS of the motif, but the discovery of the motif will not be negatively affected if a primary sequence contains more than one occurrence.

Dataset

To set up the most comprehensive evaluation framework possible, we built a dataset of 334,956 SARS-CoV-2 genomes representing the different variants defined by the World Health Organization (WHO) with at least 100 available sequences. The sequences for this dataset, covering variants Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Kappa (B.1.617.1), Epsilon (B.1.427/B.1.427), Iota (B.1.526), Eta (B.1.525), Lambda (C.37), and Omicron (B.1.1.529/BA.x), were downloaded on November 1, 2022 from the NCBI database [35] using their command line data download tool (https://www.ncbi.nlm.nih.gov/datasets/docs/v2/how-tos/virus/get-sars2-genomes/). We only included complete genomes with high coverage (less than 1% missing nucleotides) in our dataset (Table 1), and the list of accession ids for the sequences used in our different datasets is available on our GitHub repository.

Benchmarking. We assessed the performance of the different tools to identify discriminative motifs using an established approach [31]. We performed a repeated *K*-fold evaluation 100 times with a different randomization at each repetition. For each iteration, 2,500 sequences were used to form a training set and the rest (332,456) were used as a testing set. In the training set, the variants were represented by 250 sequences, with the exception of Kappa, which was represented by 100 sequences due to the low number of available sequences. Alpha and Omicron were each represented by 350 and 300 sequences, respectively, due to the large number of available sequences. At each iteration, the training sets were given as input to each tool to identify the motifs that discriminate the sequences of the variants. The identified motifs, along with the training sequences, were used to train a machine-learning algorithm (linear-SVM). Indeed, linear SVMs are one of the most commonly used approaches in the classification of viral

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

WHO Label	Pango Lineage	Number of sequences
Alpha	B.1.1.7	175,212
Beta	B.1.351	695
Gamma	P.1	8,129
Delta	B.1.617.2	9,408
Kappa	B.1.617.1	127
Epsilon	B.1.427/B.1.429	14,674
Iota	B.1.526	19,274
Eta	B.1.525	716
Lambda	C.37	428
Omicron	B.1.1.529/BA.x	106,293
Total number of sequences		334,956

https://doi.org/10.1371/journal.pone.0296627.t001 genomes, including SARS-CoV-2 [36]. They have also shown robustness when combined with *k*-mer occurrence vectors to represent sequences [32]. The ability to exploit the weights assigned to characteristics (based on *k*-mers in our case) makes them particularly interesting for highlighting regions of interest in viral genomes. This model was then used to predict the test set, and different performance metrics were calculated. For each iteration, we computed the unweighted average of precision, recall, and f1-score. By computing each metric as an unweighted average, we avoided the dominance effect of prevalent variants, as demonstrated in Eqs 1, 2 and 3.

$$\text{precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Positives}_i} \tag{1}$$

$$\text{recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Negatives}_i} \tag{2}$$

$$\text{f1} - \text{score} = \frac{1}{N}\sum_{i=1}^{N}2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{3}$$

The distributions of the different performance metrics for each tool are illustrated through violin plots in Fig 2A–2C. In addition, to visualize the prediction by class more specifically, we computed the average confusion matrix with its standard deviation for each tool (Fig 3A–3E). Finally, Fig 2D illustrates the average number of unique motifs identified by each tool during the hundred iterations to train the prediction models.

Identification of discriminating motifs and tool settings. In the identification phase of the discriminative motifs, we set the length of the motifs to *k* = 9 for two reasons. First, this length is consistent with other studies that have used *k*-mers for viral sequence classification [10, 31, 33]. Second, the selection of a multiple of 3 is consistent with the codon size, and as we use sliding windows with a step of 1 to calculate the number of *k*-mers, encompassing all reading frames, we believe this method facilitates the capture of potential amino acid-level mutations. For KEVOLVE, we set the following search parameters: *n_chromosomes* = 100 (the number of chromosomes generated at each iteration), and *n_genes* = 1 (the number of genes composing the chromosome in the first generation). Initiating with a unitary instance allows KEVOLVE to ascertain the optimal size during its search process since this is unknown, and the training sets vary throughout the evaluation. The stopping criterion parameters were set at *n_iterations* = 1000 and *n_solutions* = 10. We utilized the default crossover and mutation rates from a previous study [32] for these parameters. For CASTOR-KRFE, we set the performance threshold to be maintained while reducing the number of features to *T* = 0.99.

To evaluate MEME, considering its limitation to take as input a binary set, we implemented the following process: for each variant *v* in the training set *V*, we selected all sequences belonging to *v* to form the primary set and used the remaining sequences in *V* to form the control set. We then applied MEME to discover motifs that discriminated the primary set from the control set. This process was repeated for each variant *v* in order to build a set of motifs that could discriminate each variant from the others. This set of motifs was used to train a model and predict the testing set in the same configuration as CASTOR-KRFE and KEVOLVE. Both the ZOOPS and OOPS options were evaluated for the associated distribution site parameters. Additionally, to strongly characterize the different groups of sequences, we performed experiments to discover 10 motifs of width 9 for each variant. This choice allows us to theoretically characterize each training set with 100 motifs, assuming there are no duplicates. We applied the same

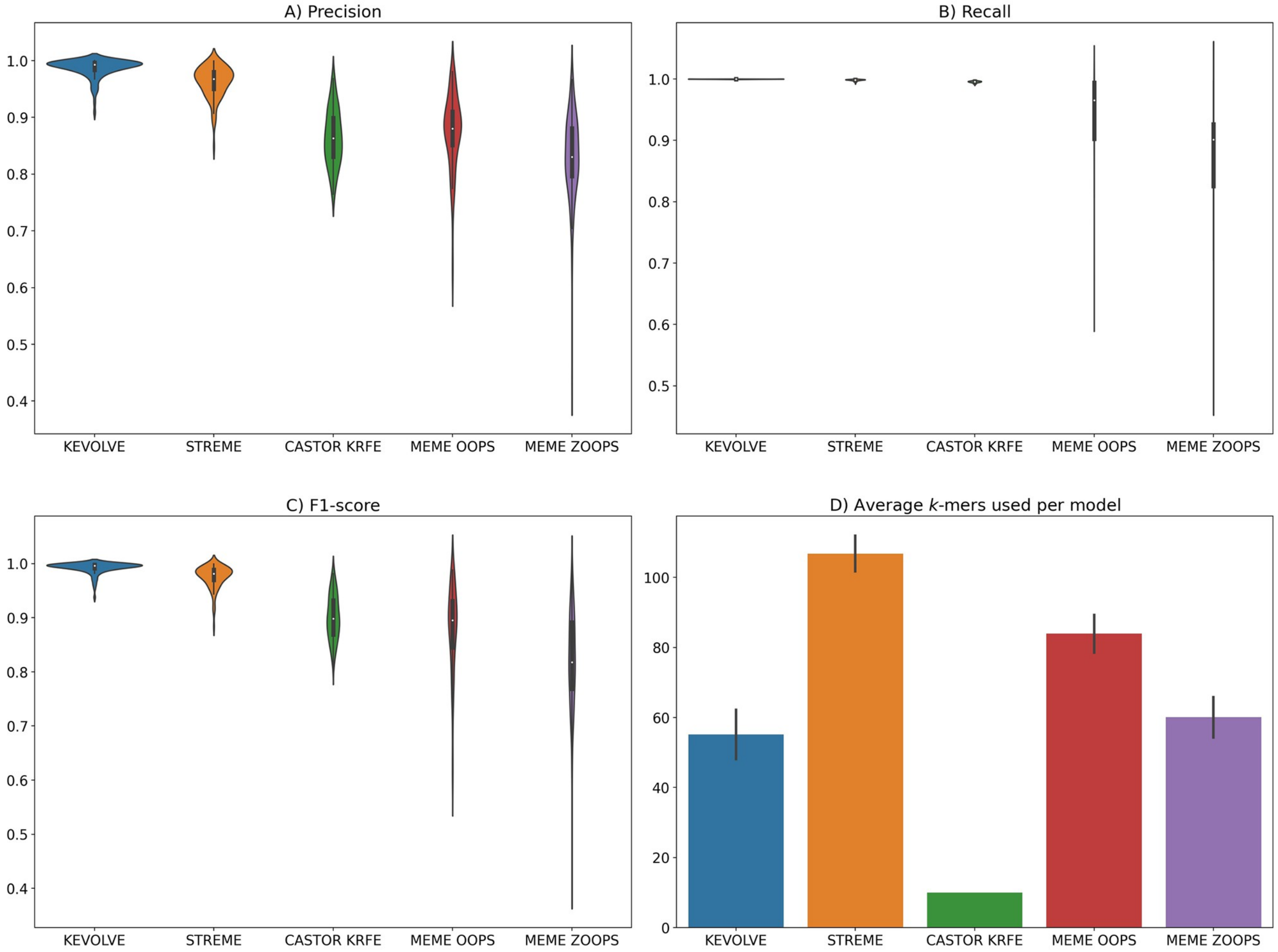
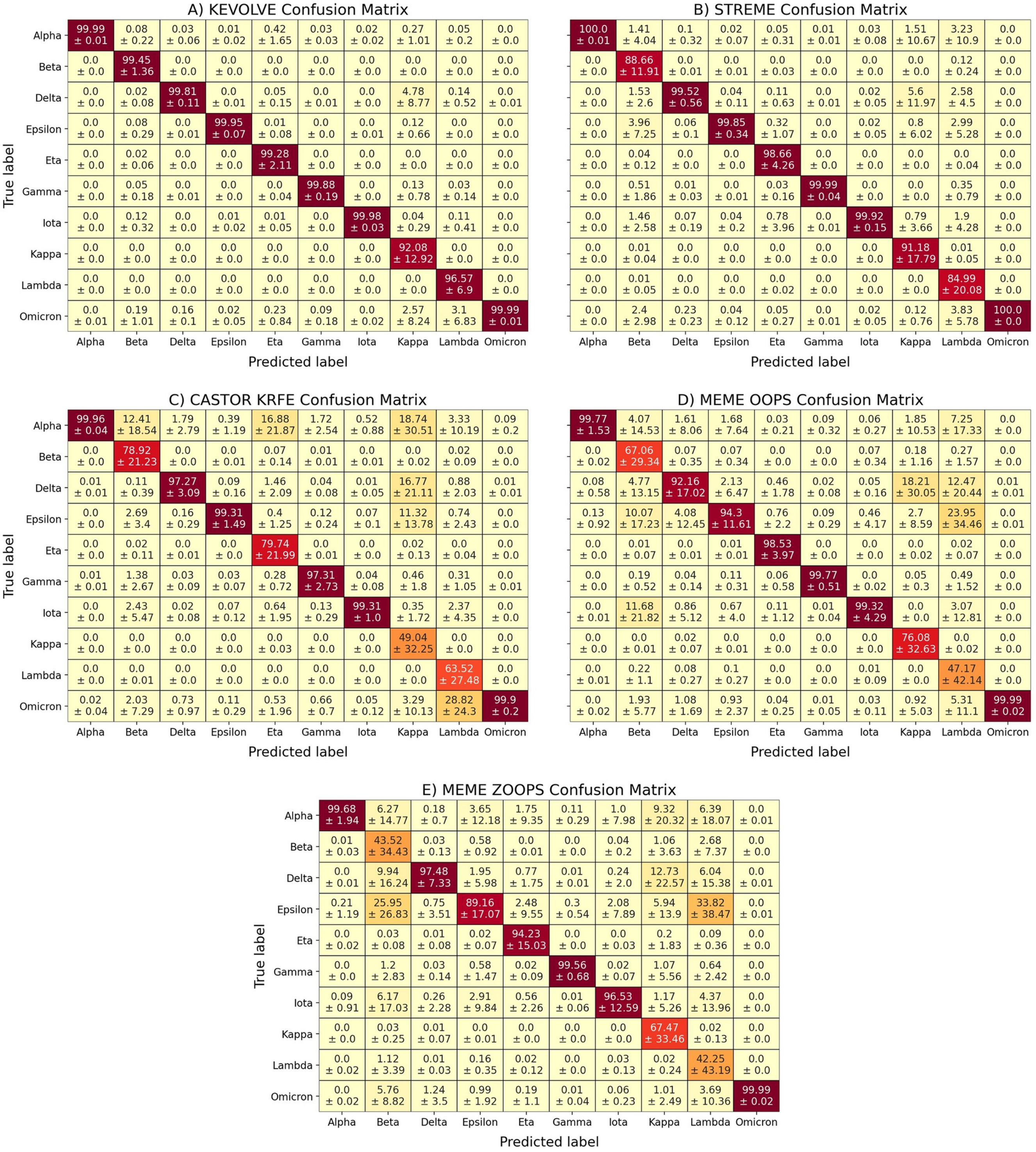


Fig 2. Results of the comparative study. A-C) The violin plots illustrate the distributions of the performance metrics, including Precision, Recall, and F1-score, obtained for the test set predictions during the cross-validation evaluation of 100 iterations. D) The bar plot depicts the average number of motifs identified by each approach to build their prediction model. The black vertical bar indicates the standard deviation.

https://doi.org/10.1371/journal.pone.0296627.g002

iterative process for identifying motifs to STREME. As mentioned previously, STREME does not require an input parameter for the motif distribution type and handles this automatically. Moreover, considering the number of experiments involved in evaluating the tools of the MEME suite because of their limitation to not handle multi-class sequences, it was not feasible to perform it on their web platform. To handle this, we set up virtual Linux environments where we installed the MEME suite version 5.5.0 with all the necessary dependencies for its functioning. Then several Shell/Python scripts were developed to run the different experiments and process the output files to extract the identified motifs. Finally, we specified that for the tools that identify multiple sets of motifs (KEVOLVE, MEME and STREME), the union of the motifs is used to represent the sequences through the feature matrix at each iteration.

Analysis of the biological significance of the motifs identified by KEVOLVE. To broaden the utility of KEVOLVE beyond identifying discriminative motifs and building prediction models for nucleotide sequences, we developed KANALYZER [37]. KANALYZER is an extension of KEVOLVE that uses pairwise alignment and parallel computing. It takes as input a reference sequence in GenBank format, a list of nucleotide sequences labeled by their



A) KEVOLVE Confusion Matrix

	Alpha	Beta	Delta	Epsilon	Eta	Gamma	Iota	Kappa	Lambda	Omicron
Alpha	99.99 ± 0.01	0.08 ± 0.22	0.03 ± 0.06	0.01 ± 0.02	0.42 ± 1.65	0.03 ± 0.03	0.02 ± 0.02	0.27 ± 1.01	0.05 ± 0.2	0.0 ± 0.0
Beta	0.0 ± 0.0	99.45 ± 1.36	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Delta	0.0 ± 0.0	0.02 ± 0.08	99.81 ± 0.11	0.0 ± 0.01	0.05 ± 0.15	0.0 ± 0.01	0.0 ± 0.0	4.78 ± 8.77	0.14 ± 0.52	0.0 ± 0.01
Epsilon	0.0 ± 0.0	0.08 ± 0.29	0.0 ± 0.01	99.95 ± 0.07	0.01 ± 0.08	0.0 ± 0.0	0.0 ± 0.01	0.12 ± 0.66	0.0 ± 0.0	0.0 ± 0.0
Eta	0.0 ± 0.0	0.02 ± 0.06	0.0 ± 0.0	0.0 ± 0.0	99.28 ± 2.11	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Gamma	0.0 ± 0.0	0.05 ± 0.18	0.0 ± 0.01	0.0 ± 0.0	0.0 ± 0.04	99.88 ± 0.19	0.0 ± 0.0	0.13 ± 0.78	0.03 ± 0.14	0.0 ± 0.0
Iota	0.0 ± 0.0	0.12 ± 0.32	0.0 ± 0.0	0.01 ± 0.02	0.01 ± 0.05	0.0 ± 0.0	99.98 ± 0.03	0.04 ± 0.29	0.11 ± 0.41	0.0 ± 0.0
Kappa	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	92.08 ± 12.92	0.0 ± 0.0	0.0 ± 0.0
Lambda	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	96.57 ± 6.9	0.0 ± 0.0
Omicron	0.0 ± 0.01	0.19 ± 1.01	0.16 ± 0.1	0.02 ± 0.05	0.23 ± 0.84	0.09 ± 0.18	0.0 ± 0.02	2.57 ± 8.24	3.1 ± 6.83	99.99 ± 0.01

B) STREME Confusion Matrix

	Alpha	Beta	Delta	Epsilon	Eta	Gamma	Iota	Kappa	Lambda	Omicron
	100.0 ± 0.01	1.41 ± 4.04	0.1 ± 0.32	0.02 ± 0.07	0.05 ± 0.31	0.01 ± 0.01	0.03 ± 0.08	1.51 ± 10.67	3.23 ± 10.9	0.0 ± 0.0
	0.0 ± 0.0	88.66 ± 11.91	0.0 ± 0.01	0.0 ± 0.01	0.0 ± 0.03	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.12 ± 0.24	0.0 ± 0.0
	0.0 ± 0.0	1.53 ± 2.6	99.52 ± 0.56	0.04 ± 0.11	0.11 ± 0.63	0.0 ± 0.01	0.02 ± 0.05	5.6 ± 11.97	2.58 ± 4.5	0.0 ± 0.0
	0.0 ± 0.0	3.96 ± 7.25	0.06 ± 0.1	99.85 ± 0.34	0.32 ± 1.07	0.0 ± 0.0	0.02 ± 0.05	0.8 ± 6.02	2.99 ± 5.28	0.0 ± 0.0
	0.0 ± 0.0	0.04 ± 0.12	0.0 ± 0.0	0.0 ± 0.0	98.66 ± 4.26	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.04	0.0 ± 0.0
	0.0 ± 0.0	0.51 ± 1.86	0.01 ± 0.03	0.0 ± 0.01	0.03 ± 0.16	99.99 ± 0.04	0.0 ± 0.0	0.0 ± 0.0	0.35 ± 0.79	0.0 ± 0.0
	0.0 ± 0.0	1.46 ± 2.58	0.07 ± 0.19	0.04 ± 0.2	0.78 ± 3.96	0.0 ± 0.01	99.92 ± 0.15	0.79 ± 3.66	1.9 ± 4.28	0.0 ± 0.0
	0.0 ± 0.0	0.01 ± 0.04	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	91.18 ± 17.79	0.01 ± 0.05	0.0 ± 0.0
	0.0 ± 0.0	0.01 ± 0.05	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.02	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	84.99 ± 20.08	0.0 ± 0.0
	0.0 ± 0.0	2.4 ± 2.98	0.23 ± 0.23	0.04 ± 0.12	0.05 ± 0.27	0.0 ± 0.01	0.02 ± 0.05	0.12 ± 0.76	3.83 ± 5.78	100.0 ± 0.0

C) CASTOR KRFE Confusion Matrix

	Alpha	Beta	Delta	Epsilon	Eta	Gamma	Iota	Kappa	Lambda	Omicron
Alpha	99.96 ± 0.04	12.41 ± 18.54	1.79 ± 2.79	0.39 ± 1.19	16.88 ± 21.87	1.72 ± 2.54	0.52 ± 0.88	18.74 ± 30.51	3.33 ± 10.19	0.09 ± 0.2
Beta	0.0 ± 0.0	78.92 ± 21.23	0.0 ± 0.0	0.0 ± 0.01	0.07 ± 0.14	0.01 ± 0.01	0.0 ± 0.01	0.0 ± 0.02	0.02 ± 0.09	0.0 ± 0.0
Delta	0.01 ± 0.01	0.11 ± 0.39	97.27 ± 3.09	0.09 ± 0.16	1.46 ± 2.09	0.04 ± 0.08	0.01 ± 0.05	16.77 ± 21.11	0.88 ± 2.03	0.01 ± 0.01
Epsilon	0.0 ± 0.0	2.69 ± 3.4	0.16 ± 0.29	99.31 ± 1.49	0.4 ± 1.25	0.12 ± 0.24	0.07 ± 0.1	11.32 ± 13.78	0.74 ± 2.43	0.0 ± 0.0
Eta	0.0 ± 0.0	0.02 ± 0.11	0.0 ± 0.01	0.0 ± 0.0	79.74 ± 21.99	0.0 ± 0.01	0.0 ± 0.0	0.02 ± 0.13	0.0 ± 0.04	0.0 ± 0.0
Gamma	0.01 ± 0.01	1.38 ± 2.67	0.03 ± 0.09	0.03 ± 0.07	0.28 ± 0.72	97.31 ± 2.73	0.04 ± 0.08	0.46 ± 1.8	0.31 ± 1.05	0.0 ± 0.01
Iota	0.0 ± 0.0	2.43 ± 5.47	0.02 ± 0.08	0.07 ± 0.12	0.64 ± 1.95	0.13 ± 0.29	99.31 ± 1.0	0.35 ± 1.72	2.37 ± 4.35	0.0 ± 0.0
Kappa	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.03	0.0 ± 0.0	0.0 ± 0.0	49.04 ± 32.25	0.0 ± 0.0	0.0 ± 0.0
Lambda	0.0 ± 0.0	0.0 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	63.52 ± 27.48	0.0 ± 0.0
Omicron	0.02 ± 0.04	2.03 ± 7.29	0.73 ± 0.97	0.11 ± 0.29	0.53 ± 1.96	0.66 ± 0.7	0.05 ± 0.12	3.29 ± 10.13	28.82 ± 24.3	99.9 ± 0.2

D) MEME OOPS Confusion Matrix

	Alpha	Beta	Delta	Epsilon	Eta	Gamma	Iota	Kappa	Lambda	Omicron
	99.77 ± 1.53	4.07 ± 14.53	1.61 ± 8.06	1.68 ± 7.64	0.03 ± 0.21	0.09 ± 0.32	0.06 ± 0.27	1.85 ± 10.53	7.25 ± 17.33	0.0 ± 0.0
	0.0 ± 0.02	67.06 ± 29.34	0.07 ± 0.35	0.07 ± 0.34	0.0 ± 0.0	0.0 ± 0.0	0.07 ± 0.34	0.18 ± 1.16	0.27 ± 1.57	0.0 ± 0.0
	0.08 ± 0.58	4.77 ± 13.15	92.16 ± 17.02	2.13 ± 6.47	0.46 ± 1.78	0.02 ± 0.08	0.05 ± 0.16	18.21 ± 30.05	12.47 ± 20.44	0.01 ± 0.01
	0.13 ± 0.92	10.07 ± 17.23	4.08 ± 12.45	94.3 ± 11.61	0.76 ± 2.2	0.09 ± 0.29	0.46 ± 4.17	2.7 ± 8.59	23.95 ± 34.46	0.0 ± 0.01
	0.0 ± 0.0	0.01 ± 0.07	0.0 ± 0.01	0.01 ± 0.01	98.53 ± 3.97	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.02	0.02 ± 0.07	0.0 ± 0.0
	0.0 ± 0.0	0.19 ± 0.52	0.04 ± 0.14	0.11 ± 0.31	0.06 ± 0.58	99.77 ± 0.51	0.0 ± 0.02	0.05 ± 0.3	0.49 ± 1.52	0.0 ± 0.0
	0.0 ± 0.01	11.68 ± 21.82	0.86 ± 5.12	0.67 ± 4.0	0.11 ± 1.12	0.01 ± 0.04	99.32 ± 4.29	0.0 ± 0.0	3.07 ± 12.81	0.0 ± 0.0
	0.0 ± 0.0	0.0 ± 0.01	0.02 ± 0.07	0.0 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	76.08 ± 32.63	0.0 ± 0.02	0.0 ± 0.0
	0.0 ± 0.01	0.22 ± 1.1	0.08 ± 0.27	0.1 ± 0.27	0.0 ± 0.0	0.0 ± 0.0	0.01 ± 0.09	0.0 ± 0.0	47.17 ± 42.14	0.0 ± 0.0
	0.0 ± 0.02	1.93 ± 5.77	1.08 ± 1.69	0.93 ± 2.37	0.04 ± 0.25	0.01 ± 0.05	0.03 ± 0.11	0.92 ± 5.03	5.31 ± 11.1	99.99 ± 0.02

E) MEME ZOOPS Confusion Matrix

	Alpha	Beta	Delta	Epsilon	Eta	Gamma	Iota	Kappa	Lambda	Omicron
Alpha	99.68 ± 1.94	6.27 ± 14.77	0.18 ± 0.7	3.65 ± 12.18	1.75 ± 9.35	0.11 ± 0.29	1.0 ± 7.98	9.32 ± 20.32	6.39 ± 18.07	0.0 ± 0.01
Beta	0.01 ± 0.03	43.52 ± 34.43	0.03 ± 0.13	0.58 ± 0.92	0.0 ± 0.01	0.0 ± 0.0	0.04 ± 0.2	1.06 ± 3.63	2.68 ± 7.37	0.0 ± 0.0
Delta	0.0 ± 0.01	9.94 ± 16.24	97.48 ± 7.33	1.95 ± 5.98	0.77 ± 1.75	0.01 ± 0.01	0.24 ± 2.0	12.73 ± 22.57	6.04 ± 15.38	0.0 ± 0.01
Epsilon	0.21 ± 1.19	25.95 ± 26.83	0.75 ± 3.51	89.16 ± 17.07	2.48 ± 9.55	0.3 ± 0.54	2.08 ± 7.89	5.94 ± 13.9	33.82 ± 38.47	0.0 ± 0.01
Eta	0.0 ± 0.02	0.03 ± 0.08	0.01 ± 0.08	0.02 ± 0.07	94.23 ± 15.03	0.0 ± 0.0	0.0 ± 0.03	0.2 ± 1.83	0.09 ± 0.36	0.0 ± 0.0
Gamma	0.0 ± 0.0	1.2 ± 2.83	0.03 ± 0.14	0.58 ± 1.47	0.02 ± 0.09	99.56 ± 0.68	0.02 ± 0.07	1.07 ± 5.56	0.64 ± 2.42	0.0 ± 0.0
Iota	0.09 ± 0.91	6.17 ± 17.03	0.26 ± 2.28	2.91 ± 9.84	0.56 ± 2.26	0.01 ± 0.06	96.53 ± 12.59	1.17 ± 5.26	4.37 ± 13.96	0.0 ± 0.0
Kappa	0.0 ± 0.0	0.03 ± 0.25	0.01 ± 0.07	0.0 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	67.47 ± 33.46	0.02 ± 0.13	0.0 ± 0.0
Lambda	0.0 ± 0.02	1.12 ± 3.39	0.01 ± 0.03	0.16 ± 0.35	0.02 ± 0.12	0.0 ± 0.0	0.03 ± 0.13	0.02 ± 0.24	42.25 ± 43.19	0.0 ± 0.0
Omicron	0.0 ± 0.02	5.76 ± 8.82	1.24 ± 3.5	0.99 ± 1.92	0.19 ± 1.1	0.01 ± 0.04	0.06 ± 0.23	1.01 ± 2.49	3.69 ± 10.36	99.99 ± 0.02

Fig 3. Results of the comparative study. A-E) The confusion matrices represent the average prediction performance as a function of the different variants for each tool over the 100 iterations. Each cell shows the average percentage of the assigned instance in the top value, and the standard deviation in the bottom value.

https://doi.org/10.1371/journal.pone.0296627.g003

classes related to the organism of the reference sequence, and a list of discriminative motifs associated with the studied sequences. KANALYZER aims to understand the reasons behind a motif's discriminatory potential by identifying the variations associated with it within different groups of variants. A variation is defined as a nucleotide sequence derived from an initial *k*-mer that has undergone one or more nucleotide changes. KANALYZER generates a report for each motif as output, containing information on their variations that occur in the different nucleotide sequences, their genomic localization, their frequencies of appearance according to the different types of variants, and the resulting mutations at the amino acid level in the case of coding regions. In this study, we used the KANALYZER extension to extract information associated with the discriminative motifs identified by KEVOLVE. The information was derived from the 334,956 sequences we collected and used the SARS-CoV-2 reference sequence NC_045512.2 (Wuhan-Hu-1 isolate, complete genome) for analysis.

Results and discussion

Prediction performances

Initially, we examined the number of discriminative motifs identified by each tool, as summarized in Fig 2D. CASTOR-KRFE identified the lowest average number of motifs at 10 per iteration, which is minimally constrained by the number of classes in the input dataset. KEVOLVE, MEME ZOOPS, and MEME OOPS identified an average of 55, 60, and 84 motifs, respectively. Finally, STREME identified the highest average number of motifs at 107 per iteration, including several degenerate motifs that were converted into classical motifs. The predictive performance of the models based on the motifs identified by each tool is shown in Fig 2A–2C in terms of precision, recall, and f1-score, respectively.

KEVOLVE performed the best, with an average score of 0.99 across all metrics. The associated confusion matrix (Fig 3A) for KEVOLVE indicates that misclassifications sometimes occur, with Kappa sequences being incorrectly predicted as Delta and Omicron in 4.8% and 2.6% of cases, on average. For Lambda variants, approximately 3.1% of the sequences were incorrectly predicted as Omicron. STREME models, which are based on approximately twice as many motifs as KEVOLVE, yielded the second-best predictions with an average performance of 0.96, 1.00, and 0.98 for precision, recall, and f1-score, respectively. The associated confusion matrix for STREME (Fig 3B) revealed some limitations for Lambda sequences, with more than 12.5% of them being incorrectly predicted as Alpha, Delta, Epsilon, or Omicron, on average. There was also an average of 9% of Beta sequences that were misclassified in a similar manner as Lambda sequences. Like KEVOLVE, STREME models had difficulty predicting certain Kappa variant sequences (≈ 9% on average), with many of them being incorrectly assigned as Delta.

The CASTOR-KRFE method had an average precision of 0.86, a recall of 1.00, and an f1-score of 0.90. The confusion matrix for the CASTOR-KRFE method (Fig 3C) indicates that it shares the same challenges as the KEVOLVE and STREME methods in inaccurately classifying some Kappa variant sequences, with 19% of these sequences being incorrectly predicted as Alpha, 17% as Delta, and 11% as Epsilon, on average. There were also limitations in the classification of Lambda variants, with nearly 29% of the sequences being incorrectly assigned to Omicron. In addition, more than 12% of the Beta variant sequences were incorrectly assigned to Alpha, on average, and 17% of the Eta variant sequences were incorrectly assigned to Alpha. The MEME OOPS and MEME ZOOPS models showed the poorest prediction performance, with average precisions of 0.97 and 0.83, average recalls of 0.94 and 0.88, and average f1-scores of 0.88 and 0.82, respectively. Both models frequently made classification errors with Lambda variants, which were often incorrectly predicted to be Epsilon. Beta variants were sometimes incorrectly predicted to be Epsilon, Delta, or Iota, and Kappa variants were often incorrectly predicted to be Delta. More detailed results can be seen in the confusion matrices shown in Fig 3D and 3E.

Biological significance of KEVOLVE-identified motifs

To extract biological information related to the motifs identified by KEVOLVE, we first combined all the motifs identified during different iterations. We used these motifs to represent all 334,956 sequences in our dataset and trained a SVM model. We ranked the motifs based on their discriminant contribution, as determined by the importance weights assigned by the model. We subsequently used KANALYZER to analyze the top 50 non-overlapping, non-redundant motifs (with regards to highlighted mutations), which encompass at least the first third of the most discriminating motif set according to the SVM-assigned weights. This analysis was conducted alongside the full set of sequences and the reference sequence NC_045512.2. The results are summarized in Table 2. In cases where KANALYZER did not produce results for a specific motif, we assumed that it was located in a genomic region with high nucleotide variability (e.g., near residues 203-205 of the nucleocapsid protein [38]) or involved numerous successive deletions (e.g., the large 9-base SGF deletion in OR1ab [39]). To improve the signal for these motifs, we extended them to 30 nucleotides based on a consensus sub-sequence from the genomes where they were initially present. These extended motifs (ID 3, 8, 12, 18, and 38 in Table 2) were then analyzed using KANALYZER like the others. As shown on Table 2, the majority of the identified motifs were located in the coding regions of structural proteins, particularly the S protein. These motifs tended to involve missense mutations, which can have significant impacts on the infectivity, tropism, and pathogenesis of the virus even when few changes are involved [40].

Motif 1, located in the S glycoprotein, is an interesting example. It has a variation present in Beta variants and in 90% of Omicron variants that involves the K417N mutation. A second variation of motif 1, found in Gamma variants, involves the K417T mutation. Both mutations occur in the receptor binding domain (RBD) of S protein, which plays a crucial role in viral infection by interacting with the host ACE2 cell surface receptor. According to published reports, these mutations may potentially decrease binding ACE2 [41] and facilitate immune escape [42]. In contrast to the K417N/T mutations, the N501Y substitution found in the RBD-ACE2 interface was shown to result in one of the largest increases in ACE2 affinity conferred by a single RBD mutation [41]. This substitution, which is associated with the variation of motif 4, is present in several different variants, including Alpha, Beta, Gamma, and Omicron. According to Nelson et *al*. [43], the additional presence of the E484K mutation can further enhance virus binding to ACE2, while the presence of the K417N substitution can stabilize this binding. The combination of these mutations may result in the emergence of a mutant, whith the potential to evade host immune responses [43]. In addition, tests in individuals who received the Moderna or Pfizer-BioNTech SARS-CoV-2 vaccines suggest that the presence of the K417N, N501Y, and E484K mutations may result in a small but significant reduction in viral neutralization, potentially impacting the effectiveness of these vaccines against certain variants [44].

KEVOLVE highlighted several other notable mutations in the S protein, including the P681H and P681R substitutions. P681H is present in the sequences of both Alpha and Omicron variants, and its proximity to the furin protease cleavage site is thought to increase the cleavage of the S protein, potentially contributing to the rapid transmission of these variants [45]. This mutation was suggested to enhance SARS-CoV-2 infectivity [46]. The P681R substitution, which is highly conserved in the Delta and Kappa variants, appears to be associated

Table 2. Mutational landscape of the motifs identified by KEVOLVE.

ID	REFERENCE *K*-MERS	LOCATIONS	VARIATIONS	AMINO ACID CHANGE	VARIANTS
1	CTGGAAAGA	S	CTGGAAATA	K417N	Beta (98%) / Omicron (90%)
			CTGGAACGA	K417T	Gamma (99%)
2	AATTGCTAT	M	AATTGCTAC	I82T	Delta (98%) / Eta (99%)
			AATTGCTAG	I82S	Kappa (96%)
3	AGTTGGATGGAAAGTGAGTTCAGAGTTTAT	S	AGTTGGATGGAAAGTG——GAGTTTAT	Del156-157 / R158G	Delta (92%)
			AGTTGTATGGAAAGTGAGTTCAGAGTTTAT	W152C	Epsilon (98%)
			AGTTGGATGAAAAGTGAGTTCAGAGTTTAT	E154K	Kappa (79%)
4	ACCCACTAA	S	ACCCACTTA	N501Y	Alpha (99%) / Beta (98%) / Gamma (99%) / Omicron (97%)
5	GCTAGAAAA	ORF8	GCTATAAAA	R52I	Alpha (99%)
6	CAAACTAAA	None	CAAACTATA	No CDS	Epsilon (99%)
			CAAACTTAA		Lambda (99%) / Omicron (99%)
7	CCTCGGCGG	S	CATCGGCGG	P681H	Alpha (99%) / Omicron (99%)
			CGTCGGCGG	P681R	Delta (99%) / Kappa (97%)
8	CCAGGCAGCAGTAGGGGAACTTCTCCTGCT	N	CCAGGCAGCAGTAAACGAACTTCTCCTGCT	R203K / G204R	Alpha (94%) / Lambda (96%) / Omicron (98%)
			CCAGGCAGCAGTAGGGGAATTTCTCCTGCT	T205I	Beta (98%) / Epsilon (99%) / Eta (98%)
			CCAGGCAGCAGTATGGGAACTTCTCCTGCT	R203M	Delta (97%) / Kappa (92%)
			CCAGGCAGCTCTAAACGAACTTCTCCTGCT	R203K / G204R	Gamma (96%)
			CTAGGCAGCAGTAGGGGAACTTCTCCTGCT	P199L	Iota (70%)
			CCAGGCAGCAGGAGGGGAACTTCTCCTGCT	S202R	Iota (27%)
9	CAACCAGAA	S	TAACCAGAA	T19I	Omicron (71%)
			GAACCAGAA	T19R	Delta (96%)
			CAAACAGAA	T20N	Gamma (98%)
10	TTCAGAGCG	ORF3a	TTCATAGCG	Q57H	Beta (98%) / Epsilon (99%) / Iota (98%)
11	CTTGGTGCA	S	TTTGGTGCA	L699F	Beta (100%) / Iota (71%)
12	TTGGTTCCATGCTATACATGTCTCTGGGAC	S	TTGGTTCCATGCTA——TCTCTGGGAC	Del69 / Del70	Alpha (97%) / Omicron (6%)
			TTGGTTCCATGTTA——TCTCTGGGAC	A67V / Del69-70	Eta (98%) / Omicron (6%)
13	AAATGCACC	N	AAATGGACC	A12G	Eta (99%)
			AAATGCACT	P13L	Iota (28%) / Lambda (97%) / Omicron (99%)
14	TTACGCAAT	ORF1ab	CTACGCAAT	L3201P	Iota (99%) / Lambda (99%)
15	TGTATAGAT	S	GGTATAGAT	L452R	Delta (98%) / Epsilon (99%) / Kappa (100%) / Omicron (5%)
			AGTATAGAT	L452Q	Lambda (99%)

(*Continued*)

Table 2. (Continued)

ID	REFERENCE *K*-MERS	LOCATIONS	VARIATIONS	AMINO ACID CHANGE	VARIANTS
16	GTTGCAGCC	5' UTR	TTTGCAGCC	No CDS	Beta (6%) / Delta (98%) / Kappa (100%)
17	CCACTGAGA	S	CCATTGAGA	T95I	Delta (20%) / Kappa (88%) / Iota (99%) / Omicron (27%)
18	CATTTTTGGGTGTTTATTACCACAAAAACA	S	CATTTTTGGGTGT—TTACCACAAAAACA	Del144	Alpha (98%) / Eta (99%)
			CATTTTTGGATGTTTATTACCACAAAAACA	G142D	Delta (62%) / Kappa (69%) / Omicron (70%)
			CATTTTTGG———ACCACAAAAACA	Del142-144 / Y145D	Omicron (27%)
19	AGATCAGTT	ORF7a	AGATCAGCT	V82A	Delta (94%) / Kappa (100%)
20	CTAAGAGGT	S	CTACGAGGT	K77T	Delta (52%)
			TTAAGAGGT	T76I	Lambda (98%)
21	AGGAATCAC	ORF1ab	GGGAATCAC	K6711R	Delta (53%)
			GGGAAGCAC	K6711R / S6713A	Kappa (94%)
22	TTAATCTTA	S	TTAATTTTA	L18F	Beta (33%) / Gamma (99%)
23	ATATCCTTT	S	ATATCCTTG	S982A	Alpha (99%)
			ATATCTTTT	L981F	Omicron (27%)
24	GACTCAGAC	S	GACTCACAC	Q677H	Eta (98%)
			TACTCAGAC	Q675H	Lambda (8%)
25	AACTTCAAG	S	AACTTCAAA	D950N	Delta (96%)
			AACTCCAAG	Silent	Kappa (19%)
26	AATGATCCA	S	AATTATCCA	D138Y	Gamma (98%)
			AATCATCCA	D138H	Lambda (5%)
27	TACACCAAA	N	TACACCGAA	Silent	Eta (99%)
28	CACAACTGT	ORF8	CATAACTGT	T11I	Iota (99%)
29	CTAATTCTC	S	CTAAGTCTC	N679K	Omicron (99%)
30	AGAGTTCCT	E	AGAGTTCTT	P71L	Beta (99%)
31	CAATGGAAC	M	GAATGGAAC	Q19E	Omicron (96%)
32	GCTCCAATT	S	GCTCCAAAT	N969K	Omicron (99%)
33	AGACATTGC	S	AGACATTGA	A570D	Alpha (99%)
34	AAAGTGGAA	ORF1ab	AAATTGGAA	K1655N	Beta (99%)
35	GTTGGACCT	S	GTTGGACCC	F888L	Eta (99%)
36	TGTTTTTCT	S	TGTTTTTTT	L5F	Iota (99%)
37	AAAATATCT	ORF1ab	ACAATATCT	K1795Q	Gamma (99%)
38	ACTAGTTTGTCTGGTTTTAAGCTAAAAGAC	ORF1ab	ACTAGTTTG———AAGCTAAAAGAC	Del3675–3677	Alpha (99%) / Beta (95%) / Gamma (99%) / Eta (99%) / Iota (99%) Lambda (99%) / Omicron (71%)
			ACTAG———TTTTAAGCTAAAAGAC	Del3674–3676	Omicron (28%)
39	GTCAACCAA	S	GTCAACCAT	Q954H	Omicron (99%)
40	CTTACTGTT	S	CTTAATGTT	T859N	Lambda (99%)
41	GTACATCGA	ORF8	GTGCATCGA	Y73C	Alpha (99%)

(*Continued*)

Table 2. (Continued)

ID	REFERENCE *K*-MERS	LOCATIONS	VARIATIONS	AMINO ACID CHANGE	VARIANTS
42	AGAAAAGTA	ORF1ab	AGAAAAATA	Silent	Eta (99%)
43	GTCTCTAGT	S	GTCTCTATT	S13I	Epsilon (98%)
44	ATCATAACC	ORF3a	ATCATAACT	Silent	Omicron (99%)
45	ATCTCAGAT	ORF1ab	ATCTCATAT	D5584Y	Epsilon (98%)
46	GGTTCATCC	ORF3a	GGTTCACCC	S253P	Gamma (98%)
47	AACTCGTCT	5' UTR	AACTCTTCT	No CDS	Beta (99%)
48	CCAACCCAC	S	CCGACCCAC	Q498R	Omicron (97%)
49	CCTTTCTGC	ORF7b	CCTTTCTGT	Silent	Omicron (99%)
50	AAGGAAGAC	N	AAGGAAGGC	D63G	Delta (96%)

The ID column is used to reference motifs and their associated information within the text. The REFERENCE *K*-MERS column comprises the motifs in their original form as seen in the reference sequence NC_045512.2. The LOCATIONS column pinpoints the genomic region where the motifs reside. The VARIATIONS column illustrates the changes stemming from the initial motifs that transpire across different sequences. The AMINO ACID CHANGE column details the distinct amino acid level mutations induced by the variations. The VARIANTS column represents the percentage of variations' occurrence within different groups of variants. The nucleotides subject to mutations are highlighted by underlining. All data is sourced from the comprehensive SARS-CoV-2 dataset (334,956 sequences).

https://doi.org/10.1371/journal.pone.0296627.t002

with enhanced fusogenicity and pathogenicity [8]. The Omicron variant is distinguished by the N679K substitution, which is associated with motif 28 and also located near the furin cleavage site [47]. When combined with P681H, both substitutions allow for the inclusion of basic amino acids near the furin cleavage site, facilitating the partition of the S protein into S1 and S2 subunits and enhancing virus fusion and infection [48]. Among other notable mutations in the S protein, KEVOLVE identified the double Del156-157 and R158G substitution (highlighted by motif 3), which are located in the N-terminal domain (NTD) of the protein and are unique to the Delta variant. These mutations, known as vaccine breakthrough mutations [49], may potentially contribute to enhanced transmissibility or reduced sensitivity to pre-existing neutralizing antibodies [50].

Motif 3 also allowed the identification of the W152C and E154K mutations, which are present in more than 98% of Epsilon variants and ≈ 80% of Kappa variants. The W152C mutation, in particular, is correlated with the S13I mutation associated with motif 43, which together have important biological consequences that may allow immune evasion [51]. According to [51], mass spectrometry and structural studies showed that the S13I and W152C mutations resulted in a complete loss of neutralization for 10 of 10 NTD-specific monoclonal antibodies, due to the remodeling of the NTD antigenic supersite by the shift of the signal peptide cleavage site and the formation of a new disulfide bond. Other examples of mutations that affect the ability of SARS-CoV-2 to bind to specific antibody molecules (antigenicity) include the L18F, T19R/I, and T20N substitutions, which are highlighted by motifs 9 and 22. L18F is found in the Gamma variant and in ≈ 35% of Beta genomes. T19R and T19I are present in 96% of Delta variants and 71% of Omicron variants, respectively, while T20N is a Gamma-specific mutation. Epitope binding of 41 NTD-specific monoclonal neutralizing antibodies (mAbs) identified six antigenic sites, one of which, termed the "NTD supersite", is recognized by all known NTD-specific mAbs and consists of residues 14-20, 140-158, and 245-264 [52]. The mutations associated with motifs 9 and 22 therefore include substitutions close to these antigenic regions of the NTD, including L18F, which is known to reduce neutralization by some antibodies [53]. A last example of motif located in the S protein identified by KEVOLVE that involves major impacts on the characteristics of SARS-CoV-2 is motif 14. A first variation of this motif, present in Delta, Epsilon, Kappa, and a minority of Omicron variants (5%), involves the L452R substitution. Located in the spike RBD which interacts directly with ACE2, this mutation was shown to increase spike stability, viral infectivity, viral fusogenicity, and viral replication [54]. The L452Q substitution, which is present in the Lambda variant, appears to be correlated with the T76I mutation associated with motif 20. These specific mutations are major contributors to the increased infectivity of the Lambda variant compared to other variants [55].

Regarding the mutations of interest associated with the motifs identified by KEVOLVE outside the S protein are I82T and I82S which are located in the M protein. M protein is highly conserved with low mutation rates and is a key element in virion morphogenesis and assembly, facilitating the release of viral particles from host cells and enhancing glucose transport during replication [56]. The I82T mutation, found in Delta and Eta variants, was suggested to enhance viral replicative fitness by altering cellular glucose uptake [57]. The I82S mutation, which is currently unique to Kappa, has not yet been well studied for its effects on SARS-CoV-2 [58]. Motif 8, located in the highly immunogenic and abundantly expressed N protein, is a last relevant example of a motif associated with mutations of interest. KANALYZER's analysis of this motif has identified variations in that region that involve P199L, S202R, R203K/M, G204R, and T205I, at least one of which is found in every major natural variant [59]. The R203K/G204R mutation, which is present in the majority of Alpha, Gamma, Lambda, and Omicron variants, was shown to confer replication advantages likely related to ribonucleocapsid (RNP) assembly, and to be associated with increased infectivity, adaptability, and virulence of SARS--CoV-2 [60]. The R203M mutation, present in Delta and Kappa, as well as the S202R mutation present in ≈ 27% of Iota variants, were shown to increase viral infectivity by ≈ 50-fold [59]. Addition of the P199L mutation (present in ≈ 70% of Iota variants) to S202R and R203K/M increases transmissibility by four to seven times and enhances luciferase activity, which is positively correlated with the more efficient assembly of virus-like particles and more effective mRNA delivery [59]. Overall, the highly variable region of residues 203-205 in the N protein of SARS-CoV-2, which includes the T205I substitution specific to Beta, Epsilon, and Eta, was associated with increased replication and pathogenicity [38]. The motif analysis reports generated by KANALYZER and the accession numbers of the sequences used in our study are available on our GitHub directory (https://github.com/bioinfoUQAM/KEVOLVE). In addition, all identified mutations were manually confirmed using resources found at https://covdb.stanford.edu/variants/ and https://covariants.org/.

Motifs identified by KEVOLVE/KANALYZER as genomic signature of SARS-CoV-2 variants

In the comparative study, we used KEVOLVE to identify motifs that discriminate between different classes of SARS-CoV-2 variants. We then selected the top 50 non-overlapping and non-redundant motifs determined by the importance weights assigned by the model.

These 50 motifs were subsequently input into KANALYZER to characterize and identify their variations within the different SARS-CoV-2 variant groups (Column "VARIATIONS" of Table 2)." In total, we obtained 125 motifs and their associated variations, which are represented in the form of a cluster map (Fig 4). This map illustrates the frequency of absence/presence of each motif across different SARS-CoV-2 variants. Although these motifs were identified by KEVOLVE from a training subset of 2,500 sequences, the frequencies shown in Fig 4 are computed from the entire dataset of 334,956 sequences. By examining the columns, it is possible to identify different profiles and clusters of absence/presence of motifs specific to various variants. For example, Omicron has a cluster of 7 motifs that are unique to this variant

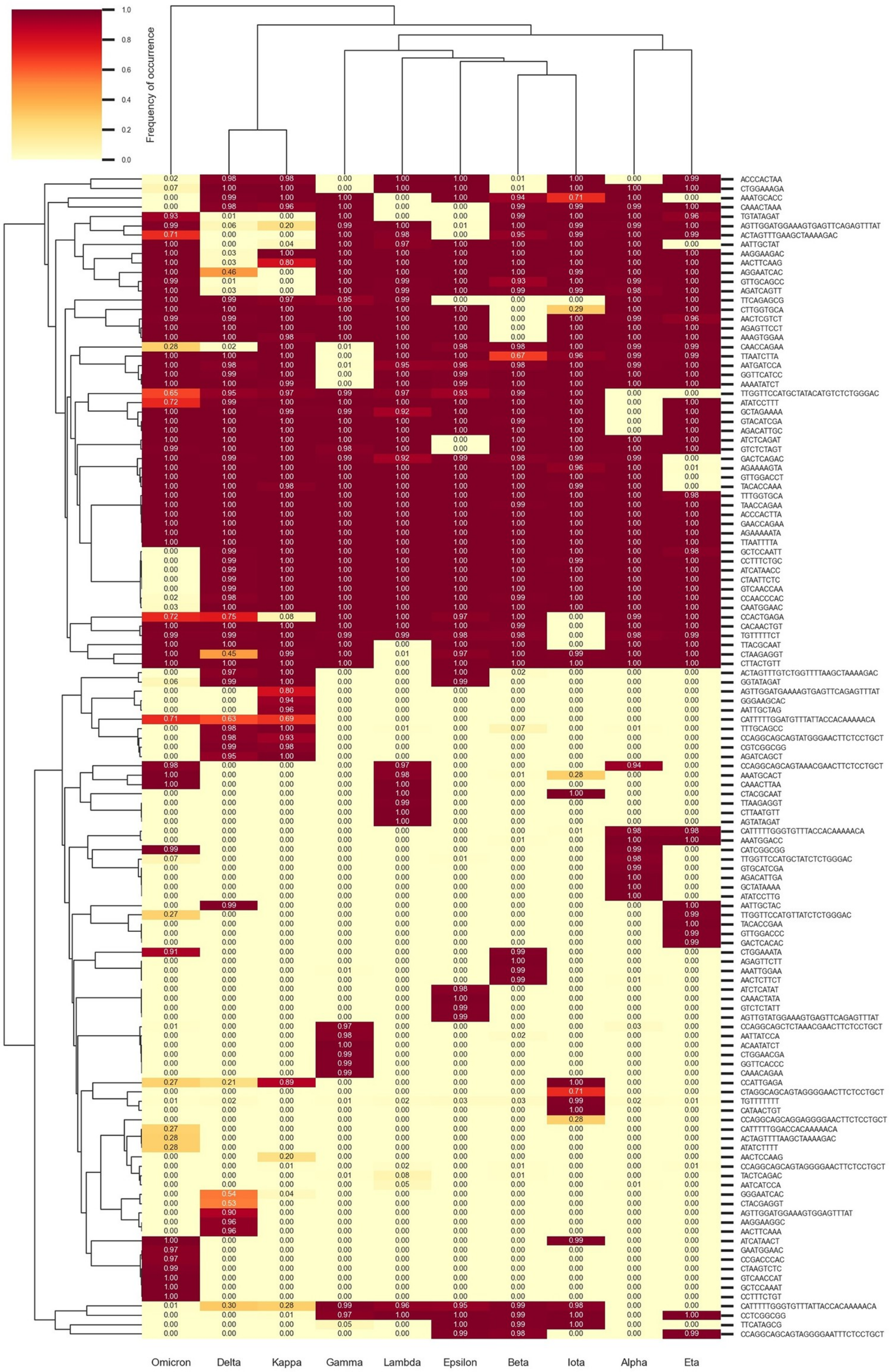


Fig 4. Cluster map of motif occurrence frequency according to SARS-CoV-2 variants.

https://doi.org/10.1371/journal.pone.0296627.g004

(located in the lower left of the cluster map), with the exception of the ATCATAACT motif, which is also present in Iota. Towards the middle of the cluster map, we can see a second cluster of 7 motifs that appear in all variants except Omicron. These two Omicron-specific clusters contribute to its distance from the other SARS-CoV-2 variants. In summary, this figure illustrates KEVOLVE's ability to identify motifs in temporally conserved regions starting with a limited set of sequences and to generalize to a larger dataset of sequences collected since the start of the COVID-19 pandemic. The identified motifs provide genomic signatures that can be used to generate peptide or oligonucleotide libraries for rapid and accurate detection of listed pathogens with tools such as VirScan [61] or to design specific primer sets for the classification of SARS-CoV-2 variants with artificial intelligence [62]. These approaches, which use models built from a restricted number of motifs and sequences, can efficiently classify large sets of sequences, which is crucial during major viral outbreaks where swift identification of the virus' taxonomic classification and genomic sequence origin is necessary for effective strategic planning, containment, and treatment [10]. In addition, the identified genomic signatures, along with the reports generated by KANALYZER, provide valuable insights that can help understand the viral evolution and transmission, the mechanisms through which the virus causes disease, and the development of treatments and vaccines. These approaches, which use models built from a restricted number of motifs and sequences, can efficiently classify large sets of sequences, which is crucial during major viral outbreaks where swift identification of the virus' taxonomic classification and genomic sequence origin is necessary for effective strategic planning, containment, and treatment [10].

Perspective and future directions

For future work, we believe it would be insightful to explore comparisons with approaches that have been developed concurrently and exhibit similarities. One such tool is CouGaR-g, recently published, which introduces an approach using a deep learning model (convolutional neural networks) to classify SARS-CoV-2 sequences represented by frequency chaos game representation [63]. In their study, CouGaR-g demonstrated strong performance with an accuracy exceeding 96% for a test set comprising 19,146 SARS-CoV-2 sequences divided into 11 clades. The authors also utilize saliency maps to highlight relevant k-mers and further demonstrate their association with known marker variants. It could, therefore, be beneficial to conduct experiments to compare the impact of sequence representation, the influence of the choice of machine learning model (especially to investigate performance on GISAID clades such as GR, GRY, or O where CouGaR-g's performance was lower), or to assess the overlap and differences in the *k*-mers identified as significant in correlation with known marker variants.

Another pertinent comparative analysis would consider Nextclade [64] in classifying viral sequences with significant nucleotide divergence. Nextclade conducts pairwise alignments of viral genomes against a reference sequence, discerns mutations, and employs mutational distances to ascertain the nearest match within a phylogenetic framework, thereby designating the query sequence to a closely related clade [64]. While both Nextclade and KEVOLVE demonstrate robustness in SARS-CoV-2 sequence classification, KEVOLVE may offer superior performance for viruses exhibiting substantial nucleotide divergence, such as HIV—with divergence rates between subtypes ranging from 25 to 35% [65] and hepatitis C virus (HCV), where genotypic differences reach 31 to 33% at the nucleotide level [66]. Notably, Nextclade is tailored for rapid alignment of sequences with less than 10% divergence [64], a scenario less applicable to the broad variability seen in HIV or HCV. KEVOLVE employs *k*-mer occurrence vectors for sequence representation and a SVM for prediction, a methodology previously validated for robust viral classification across diverse divergence levels [30–32]. The sensitivity of *k*-mer occurrences, as opposed to mutations at specific positions, is particularly advantageous for sequences with elevated rates of nucleotide divergence [21]. Furthermore, the SVM framework provides a nuanced approach by relating a set of training sequences to their features and assigning weights based on their discriminative value—unlike Nextclade's distance-based classification. This feature weighting proves instrumental in prioritizing mutations for analysis. Nextclade and KEVOLVE each have their place in the genomics toolbox, with specific scenarios where they can distinguish themselves.

Finally, although our current approach and all those mentioned above operate within a closed classification framework, which is limited to the classes defined by the training sequence dataset, we plan to extend it to an open classification context. To achieve this, we propose a strategy to calculate the distance between each new sequence and the existing genomic signature profiles, generated in the cluster map (Fig 4). By using an appropriate distance threshold, we can identify sequences that are significantly distant from known signatures, potentially indicating a new variant. Thresholds can be determined by leveraging the knowledge of distances between genomic signature profiles of different known variants. This method could be based on distance metrics such as Euclidean distance, Manhattan distance, or even a normalized distance based on *k*-mer similarity. Furthermore, to make our approach more flexible and adaptable to new variants, we could also implement an incremental learning mechanism. In this way, each time a new variant is identified above a certain support threshold, the associated sequences could be integrated into the initial training set, and the model would be retrained to account for this new information. This would allow our model to learn and progressively adjust its parameters based on the newly encountered sequences. This approach could facilitate the detection of new variants and enable regular model updates with the integration of new sequences associated with emerging variants.

Conclusion

In this study, we compared the performance of the machine learning-based tools KEVOLVE and CASTOR-KRFE with statistical tools specialized in identifying discriminative motifs in unaligned sequence sets for the classification of SARS-CoV-2 variants. Overall, the models based on the motifs identified by KEVOLVE outperformed the models based on the motifs identified by the statistical tools, while using a lower number of motifs. Models based on STREME motifs achieved the second-best performance (slightly below KEVOLVE), but these models require the use of twice as many motifs. The drop in performance was mainly due to prediction errors for Beta, Kappa, and Lambda variants. CASTOR-KRFE obtained the third-best performance with models based on 10 times fewer motifs than STREME, as the tool only identifies a single subset of motifs, unlike the others. The prediction errors of the CASTOR-KRFE models are associated with the same variants as those of STREME, but they are more pronounced. Finally, the weakest performances were associated with the MEME OOPS/ZOOPS models, with many more errors for the same variants than STREME and CASTOR-KRFE. This study also demonstrated that KEVOLVE and CASTOR-KRFE are able to handle multi-class sets, rather than being limited to binary sets like some other tools. This is an important advantage when analyzing organisms such as SARS-CoV-2, which are constituted of multiple classes of viral variants.

Subsequently, we analyzed the motifs identified by KEVOLVE using KANALYZER, a new extension based on pairwise alignment and parallel computing. This analysis allowed us to identify variations of the discriminative motifs in different classes of SARS-CoV-2 variants, including their frequency, genomic localization, and mutation at the amino acid level. This analysis, performed on all 334,956 sequences belonging to the 10 major variant classes defined by the WHO, showed that the majority of the motifs identified by KEVOLVE were located in structural proteins, with a particular focus on the S protein. The motifs and variations identified were linked to known mutations previously reported in the literature, which are assumed to affect key characteristics of the virus such as infectivity, pathogenicity, tropism, transmission, and evolution. In conclusion, this study demonstrates the utility of KEVOLVE as a robust tool for identifying discriminative motifs of SARS-CoV-2 variants. These motifs provide genomic signatures that can be used to construct oligonucleotide libraries or to build artificial intelligence models for rapid and accurate pathogen detection. Furthermore, KANALYZER allows the analysis of motifs identified by KEVOLVE, providing valuable insights into the biological properties of viruses and viral gene products that serve as targets for the development of vaccines or antiviral therapy.

Author Contributions

Funding acquisition: Abdoulaye Baniré Diallo.

Methodology: Dylan Lebatteux.

Resources: Dylan Lebatteux, Abdoulaye Baniré Diallo.

Software: Dylan Lebatteux, Abdoulaye Baniré Diallo.

Supervision: Abdoulaye Baniré Diallo.

Validation: Dylan Lebatteux.

Writing – original draft: Dylan Lebatteux.

Writing – review & editing: Hugo Soudeyns, Isabelle Boucoiran, Soren Gantt, Abdoulaye Baniré Diallo.

References

1. Gorbalenya A., Baker S., Baric R., De Groot R., Drosten C., Gulyaeva A., et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*. 5, 536–544 (2020) https://doi.org/10.1038/s41564-020-0695-z
2. Zhu N., Zhang D., Wang W., Li X., Yang B., Song J., et al. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal Of Medicine*. (2020) https://doi.org/10.1056/NEJMoa2001017 PMID: 31978945
3. Lee E., Ng M. & Khong P. COVID-19 pneumonia: what has CT taught us?. *The Lancet Infectious Diseases*. 20, 384–385 (2020) https://doi.org/10.1016/S1473-3099(20)30134-1 PMID: 32105641
4. 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*. 395, 565–574 (2020) https://doi.org/10.1016/S0140-6736(20)30251-8 PMID: 32007145
5. Gordon D., Jang G., Bouhaddou M., Xu J., Obernier K., White K., et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. 583, 459–468 (2020) https://doi.org/10.1038/s41586-020-2286-9 PMID: 32353859
6. Kandeel M., Ibrahim A., Fayez M. & Al-Nazawi M. From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. *Journal Of Medical Virology*. 92, 660–666 (2020) https://doi.org/10.1002/jmv.25754 PMID: 32159237
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*. 65, 1075–1082 (2020) https://doi.org/10.1038/s10038-020-0808-9 PMID: 32699345
8. Saito A., Irie T., Suzuki R., Maemura T., Nasser H., Uriu K., et al. Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature*. 602, 300–306 (2022) https://doi.org/10.1038/s41586-021-04266-9 PMID: 34823256

9. Koyama T., Weeraratne D., Snowdon J. & Parida L. Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment. *Pathogens*. 9, 324 (2020)

10. Randhawa G., Soltysiak M., El Roz H., Souza C., Hill K. & Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *Plos One*. 15, e0232391 (2020) https://doi.org/10.1371/journal.pone.0232391 PMID: 32330208

11. Lopez-Rincon A., Tonda A., Mendoza-Maldonado L., Mulders D., Molenkamp R., Perez-Romero C., et al. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. *Scientific Reports*. 11, 1–11 (2021) https://doi.org/10.1038/s41598-020-80363-5 PMID: 33441822

12. Bauer D., Tay A., Wilson L., Reti D., Hosking C., McAuley A., et al. Supporting pandemic response using genomics and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. *Transboundary And Emerging Diseases*. 67, 1453–1462 (2020) https://doi.org/10.1111/tbed.13588 PMID: 32306500

13. Lau B., Pavlichin D., Hooker A., Almeda A., Shin G., Chen J., et al. Profiling SARS-CoV-2 mutation fingerprints that range from the viral pangenome to individual infection quasispecies. *Genome Medicine*. 13, 1–23 (2021) https://doi.org/10.1186/s13073-021-00882-2 PMID: 33875001

14. Slezak T., Hart B. & Jaing C. Design of genomic signatures for pathogen identification and characterization. *Microbial Forensics*. pp. 299–312 (2020) https://doi.org/10.1016/B978-0-12-815379-6.00020-9

15. Edgar R.MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32, 1792–1797 (2004) https://doi.org/10.1093/nar/gkh340 PMID: 15034147

16. Larkin M., Blackshields G., Brown N., Chenna R., McGettigan P., McWilliam H., et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 23, 2947–2948 (2007) https://doi.org/10.1093/bioinformatics/btm404 PMID: 17846036

17. Katoh K., Rozewicki J. & Yamada K. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings In Bioinformatics*. 20, 1160–1166 (2019) https://doi.org/10.1093/bib/bbx108 PMID: 28968734

18. Bernard G., Chan C., Chan Y., Chua X., Cong Y., Hogan J., et al. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings In Bioinformatics*. 20, 426–435 (2019) https://doi.org/10.1093/bib/bbx067 PMID: 28673025

19. Lange K.Mathematical and statistical methods for genetic analysis. (Springer,2002)

20. Eddy S.What is dynamic programming?. *Nature Biotechnology*. 22, 909–910 (2004) https://doi.org/10.1038/nbt0704-909 PMID: 15229554

21. Zielezinski A., Vinga S., Almeida J. & Karlowski W. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*. 18, 1–17 (2017) https://doi.org/10.1186/s13059-017-1319-7 PMID: 28974235

22. Duffy S., Shackelton L. & Holmes E. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*. 9, 267–276 (2008) https://doi.org/10.1038/nrg2323 PMID: 18319742

23. Wong K., Suchard M. & Huelsenbeck J. Alignment uncertainty and genomic analysis. *Science*. 319, 473–476 (2008) https://doi.org/10.1126/science.1151532 PMID: 18218900

24. Bailey, T., Elkan, C. & Others Fitting a mixture model by expectation maximization to discover motifs in bipolymers. (Department of Computer Science,1994)

25. Bailey T., Johnson J., Grant C. & Noble W. The MEME suite. *Nucleic Acids Research*. 43, W39–W49 (2015) https://doi.org/10.1093/nar/gkv416 PMID: 25953851

26. Bailey T., Bodén M., Whitington T. & Machanick P. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics*. 11, 1–14 (2010) https://doi.org/10.1186/1471-2105-11-179 PMID: 20380693

27. Bailey T.STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. 37, 2834–2840 (2021) https://doi.org/10.1093/bioinformatics/btab203 PMID: 33760053

28. Libbrecht M. & Noble W. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*. 16, 321–332 (2015) https://doi.org/10.1038/nrg3920 PMID: 25948244

29. Remita M., Halioui A., Malick Diouara A., Daigle B., Kiani G. & Diallo A. A machine learning approach for viral genome classification. *BMC Bioinformatics*. 18, 1–11 (2017) https://doi.org/10.1186/s12859-017-1602-3 PMID: 28399797

30. 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*. 13, e0206409 (2018) https://doi.org/10.1371/journal.pone.0206409 PMID: 30427878

31. Lebatteux D., Remita A. & Diallo A. Toward an alignment-free method for feature extraction and accurate classification of viral sequences. *Journal Of Computational Biology*. 26, 519–535 (2019) https://doi.org/10.1089/cmb.2018.0239 PMID: 31050550

32. Lebatteux, D. & Diallo, A. Combining a genetic algorithm and ensemble method to improve the classification of viruses. *2021 IEEE International Conference On Bioinformatics And Biomedicine (BIBM)*. pp. 688-693 (2021)

33. Zhang Q., Jun S., 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*. 7, 1–13 (2017)

34. Narlikar, L., Gordân, R. & Hartemink, A. Nucleosome occupancy information improves de novo motif discovery. *Annual International Conference On Research In Computational Molecular Biology*. pp. 107-121 (2007)

35. Sayers E., Bolton E., Brister J., Canese K., Chan J., Comeau D., et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 50, D20–D26 (2022) https://doi.org/10.1093/nar/gkab1112 PMID: 34850941

36. Ahmed I. & Jeon G. Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. *Interdisciplinary Sciences: Computational Life Sciences*. 14, 504–519 (2022) https://doi.org/10.1007/s12539-021-00465-0 PMID: 34357528

37. Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S. & Diallo, A. KANALYZER: a method to identify variations of discriminative k-mers in genomic sequences. *2022 IEEE International Conference On Bioinformatics And Biomedicine (BIBM)*. pp. 757-762 (2022)

38. Johnson B., Zhou Y., Lokugamage K., Vu M., Bopp N., Crocquet-Valdes P., et al. Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *PLoS Pathogens*. 18, e1010627 (2022) https://doi.org/10.1371/journal.ppat.1010627 PMID: 35728038

39. Tamanaha E., Zhang Y. & Tanner N. Profiling RT-LAMP tolerance of sequence variation for SARS-CoV-2 RNA detection. *PLoS One*. 17, e0259610 (2022) https://doi.org/10.1371/journal.pone.0259610 PMID: 35324900

40. Zhu C., He G., Yin Q., Zeng L., Ye X., Shi Y. et al. Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. *Journal Of Medical Virology*. 93, 5729–5741 (2021) https://doi.org/10.1002/jmv.27132 PMID: 34125455

41. Starr T., Greaney A., Hilton S., Ellis D., Crawford K., Dingens A., et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*. 182, 1295–1310 (2020) https://doi.org/10.1016/j.cell.2020.08.012 PMID: 32841599

42. Barton M., MacGowan S., Kutuzov M., Dushek O., Barton G. & Van Der Merwe P. Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *Elife*. 10 pp. e70658 (2021) https://doi.org/10.7554/eLife.70658 PMID: 34435953

43. Nelson G., Buzko O., Spilman P., Niazi K., Rabizadeh S. & Soon-Shiong P. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y. V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *BioRxiv*. (2021)

44. Wang Z., Schmidt F., Weisblum Y., Muecksch F., Barnes C., Finkin S., et al. mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature*. 592, 616–622 (2021) https://doi.org/10.1038/s41586-021-03324-6 PMID: 33567448

45. Desingu P., Nagarajan K. & Dhama K. Emergence of Omicron third lineage BA. 3 and its importance. *Journal Of Medical Virology*. 94, 1808–1810 (2022) https://doi.org/10.1002/jmv.27601 PMID: 35043399

46. Zuckerman N., Fleishon S., Bucris E., Bar-Ilan D., Linial M., Bar-Or I., et al. A unique SARS-CoV-2 spike protein P681H variant detected in Israel. *Vaccines*. 9, 616 (2021) https://doi.org/10.3390/vaccines9060616 PMID: 34201088

47. Kannan S., Spratt A., Sharma K., Chand H., Byrareddy S. & Singh K. Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies. *Journal Of Autoimmunity*. 126 pp. 102779 (2022) https://doi.org/10.1016/j.jaut.2021.102779 PMID: 34915422

48. He X., Hong W., Pan X., Lu G. & Wei X. SARS-CoV-2 Omicron variant: characteristics and prevention. *MedComm*. 2, 838–845 (2021) https://doi.org/10.1002/mco2.110 PMID: 34957469

49. Muttineni R., Putty K., Marapakala K., KP S., Panyam J., Vemula A., et al. SARS-CoV-2 variants and spike mutations involved in second wave of COVID-19 pandemic in India. *Transboundary And Emerging Diseases*. 69, e1721–e1733 (2022) https://doi.org/10.1111/tbed.14508 PMID: 35266305

50. Fan L., Hu X., Chen Y., Peng X., Fu Y., Zheng Y., et al. Biological significance of the genomic variation and structural dynamics of SARS-CoV-2 B. 1.617. *Frontiers In Microbiology*. 12 pp. 750725 (2021) https://doi.org/10.3389/fmicb.2021.750725 PMID: 34691002

51. Zhang J., Xiao T., Cai Y., Lavine C., Peng H., Zhu H., et al. Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant. *Science*. 374, 1353–1360 (2021) https://doi.org/10.1126/science.abl9463 PMID: 34698504

52. Harvey W., Carabelli A., Jackson B., Gupta R., Thomson E., Harrison E., et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*. 19, 409–424 (2021) https://doi.org/10.1038/s41579-021-00573-0 PMID: 34075212

53. McCallum M., De Marco A., Lempp F., Tortorici M., Pinto D., Walls A., et al. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell*. 184, 2332–2347 (2021) https://doi.org/10.1016/j.cell.2021.03.028 PMID: 33761326

54. Motozono C., Toyoda M., Zahradnik J., Saito A., Nasser H., Tan T., et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. *Cell Host & Microbe*. 29, 1124–1136 (2021) https://doi.org/10.1016/j.chom.2021.06.006 PMID: 34171266

55. Kimura I., Kosugi Y., Wu J., Zahradnik J., Yamasoba D., Butlertanaka E., et al. The SARS-CoV-2 Lambda variant exhibits enhanced infectivity and immune resistance. *Cell Reports*. 38, 110218 (2022) https://doi.org/10.1016/j.celrep.2021.110218 PMID: 34968415

56. Thakur S., Sasi S., Pillai S., Nag A., Shukla D., Singhal R., et al. SARS-CoV-2 Mutations and Their Impact on Diagnostics, Therapeutics and Vaccines. *Frontiers In Medicine*. 9 (2022) https://doi.org/10.3389/fmed.2022.815389 PMID: 35273977

57. Shen L., Bard J., Triche T., Judkins A., Biegel J. & 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*. 10, 885–893 (2021) https://doi.org/10.1080/22221751.2021.1922097 PMID: 33896413

58. Singh P., Sharma K., Singh P., Bhargava A., Negi S., Sharma P., et al. Genomic characterization unravelling the causative role of SARS-CoV-2 Delta variant of lineage B. 1.617. 2 in 2nd wave of COVID-19 pandemic in Chhattisgarh, India. *Microbial Pathogenesis*. 164 pp. 105404 (2022) https://doi.org/10.1016/j.micpath.2022.105404 PMID: 35065253

59. Syed A., Taha T., Tabata T., Chen I., Ciling A., Khalid M., et al. Rapid assessment of SARS-CoV-2–evolved variants using virus-like particles. *Science*. 374, 1626–1632 (2021) https://doi.org/10.1126/science.abl6184 PMID: 34735219

60. 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*. 29, 1788–1801 (2021) https://doi.org/10.1016/j.chom.2021.11.005 PMID: 34822776

61. Xu D. & Tian Y. A comprehensive survey of clustering algorithms. *Annals Of Data Science*. 2, 165–193 (2015) https://doi.org/10.1007/s40745-015-0040-1

62. McMillen T., Jani K., Robilotti E., Kamboj M. & Babady N. The spike gene target failure (SGTF) genomic signature is highly accurate for the identification of Alpha and Omicron SARS-CoV-2 variants. *Scientific Reports*. 12, 1–8 (2022) https://doi.org/10.1038/s41598-022-21564-y PMID: 36347878

63. Avila Cartes J., Anand S., Ciccolella S., Bonizzoni P. & Della Vedova G. Accurate and fast clade assignment via deep learning and frequency chaos game representation. *GigaScience*. 12 pp. giac119 (2023)

64. Aksamentov I., Roemer C., Hodcroft E. & Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal Of Open Source Software*. 6, 3773 (2021) https://doi.org/10.21105/joss.03773

65. Hemelaar J., Gouws E., Ghys P. & Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *Aids*. 20, W13–W23 (2006) https://doi.org/10.1097/01.aids.0000247564.73009.bc PMID: 17053344

66. Simmonds P., Bukh J., Combet C., Deléage G., Enomoto N., Feinstone S., et al. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology*. 42, 962–973 (2005) https://doi.org/10.1002/hep.20819 PMID: 16149085