Gene 909 (2024) 148291

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Geometric analysis of SARS-CoV-2 variants

Mengcen Guan [a,*], Nan Sun [a,*], Stephen S.-T. Yau [a,b,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing, China
[b] Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, China

ARTICLE INFO

Edited by: Sk. Sarif Hassan

Keywords:
SARS-CoV-2
Natural vector
Genome Space
Convex Hull Classification
The Nearest Neighbor Classification
Phylogenetic Analysis

ABSTRACT

SARS-CoV-2 as a severe respiratory disease has been prevalent around the world since its first discovery in 2019. As a single-stranded RNA virus, its high mutation rate makes its variants manifold and enables some of them to have high pathogenicity, such as Omicron variant, the most prevalent virus now. Research on the relationship of these SARS-CoV-2 variants, especially exploring their difference is a hot issue. In this study, we constructed a geometric space to represent all SARS-CoV-2 sequences of different variants. An alignment-free method: natural vector method was utilized to establish genome space. The genome space of SARS-CoV-2 was constructed based on the 24-dimensional natural vector and the appropriate metric was determined through performing phylogenetic analysises. Phylogenetic trees of different lineages constructed under the selected natural vector and metric coincided with the lineage naming standards, which means lineages with same alphabetical prefix cluster in phylogenetic trees. Furthermore, the relationships between the various GISAID clades as depicted by the natural graph primarily matched the description provided in the GISAID clade naming.The validity of our geometric space was demonstrated by these phylogenetic analysis results. So in this research, we constructed a geometry space for the genomes of the novel coronavirus SARS-CoV-2, which allows us to compare the different variants. Our geometric space is valuable for resolving the issues insides the virus.

1. Introduction

SARS-CoV-2, the novel coronavirus that broke out at the end of 2019 has been continuously spreading globally Asselah et al. (2021), and has not completely disappeared for more than three years, with a continuous increase in confirmed cases and deaths Adil et al. (2021). Some vaccines have been successfully developed, but the constantly mutating virus poses great challenges to vaccine application Vasireddy et al. (2021). The current epidemic has had a serious impact on global economic development and people's lives in various countries Zhang et al. (2020), and it is still unpredictable when the epidemic will end. Up to now, new variants continue to occur, such as Omicron variant. Thus, researches on SARS-CoV-2, including genomic analysis, classification of virus variants, and analysis of its characteristics, are crucial for understanding the pathophysiology of viruses as well as for preventing and controlling infections.

According to the World Health Organization, there are 13 variants of SARS-CoV-2: Omicron, Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Eta, Theta, Iota, Kappa, Lambda, Mu; *https*://*www.who.int*/*en*/*activities*/*tracking*−*SARSCoV*−2*variants*/. According to the classification standard of GISAID, SARS-CoV-2 can be divided into 12 evolutionary branches: G, GH, GK, GKA, GR, GR, GR, GV, L, O, S, V; In addition, there are also Pango lineages categories *https*://*cov*−*lineages.org*/*lineage_list.html*. The division and naming of these variants of SARS-CoV-2 are based on sequence alignment algorithms to study their differences. Moreover, due to the increasing speed of virus mutations, the emergence of new variants is becoming more and more rapid. A method for quickly studying the relationship between different variants and accurately classifying them is very important. For example, when a new virus sequence appears, its relationship with existing variants is needed to be quickly determined. Our article proposes a geometric space based on non-alignment algorithms to achieve this.

In revealing the relationship of different variants, traditional alignment methods (BLAST Altschul et al. (1997), MAFFT Katoh et al. (2009)) are time-consuming, also the evolutionary relationships of different variables cannot be intuitively represented. Through alignment-free

Abbreviations List: SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; 1NN, 1-nearest neighbor algorithm; BME, TaxAdd_BME; NJ, Neighbor-Joining; KNN, K-nearest neighbor algorithm; WHO, World Health Organization; MSA, Multiple Sequence Alignment.

* Corresponding authors at: Department of Mathematical Sciences, Tsinghua University, Beijing, China (S.S.T. Yau).
E-mail addresses: gmc20@mails.tsinghua.edu.cn (M. Guan), sunn19@mails.tsinghua.edu.cn (N. Sun), yau@uic.edu (S.S.-T. Yau).

https://doi.org/10.1016/j.gene.2024.148291
Received 15 November 2023; Received in revised form 23 January 2024; Accepted 14 February 2024
Available online 28 February 2024

natural vector method Deng et al. (2011), which was proposed by Yau and his colleagues to represent one protein/gene sequence by one mathematical vector, we can establish a space belonging to SARS-CoV-2, where sequences intuitively correspond to points in the Euclidean space. Our research focuses on the construction of the geometric space of the SARS-CoV-2 sequence in Euclidean space. The similarity of the original sequences can be reflected through the vectors' similarity in the geometric space, then the study on the sequences can be transformed into the study on vectors in Euclidean space. All of these are based on a fundamental premise that natural vectors and sequences are one to one correspondence, which has been proven in the previous study Deng et al. (2011). In this paper, through convex hull analysis and the nearest neighbor classification, we found sequences belonging to the same

$$(n_A, n_G, n_C, n_T, \mu_A, \mu_G, \mu_C, \mu_T, D_2^A, D_2^G, D_2^C, D_2^T, D_3^A, D_3^G, D_3^C, D_3^T, D_4^A, D_4^G, D_4^C, D_4^T, D_5^A, D_5^G, D_5^C, D_5^T).$$

variants naturally cluster together. Furthermore, we can study evolutionary relationships by studying the distance relations of the points from different variants. For points that are closer in space, their biological relationships are closer. This is the purpose of constructing this geometric space.

Our data set was downloaded from GISAID and included 115390 whole-genome sequences covering all variants of SARS-CoV-2. Convex hulls constructed by natural vectors with high orders showed that sequences from different variants had different features and their convex hulls were naturally disjoint. Based on k-mer natural vector method and KNN(K-NearestNeighbor) classification analysis Guo et al. (2003), we chose 7-mer natural vector for its highest accuracy of classification, which could reach 0.9824. Additionally, we analyzed the phylogenetic relationships between SARS-CoV-2 variants from the perspectives of two different classification standards (GISAID and Pango lineages). Our results have demonstrated the effectiveness of the 7-mer natural vector in solving COVID-19 cases. This vector has been successfully applied in many research fields Zhao et al. (2018, 2019, 2021).

2. Materials and methods

2.1. Sequence coding and feature generation

2.1.1. Natural vector

The original natural vector is a 12-dimensional numerical vector used to encode a DNA sequence and describe the distribution of the nucleotides A, G, C, and T.Deng et al. (2011). The definition is as follows. Given a sequence with a length of n: $S = s_1 s_2 \ldots s_n, s_i \in \{A, G, C, T\}$, the indicator functions for A, C, G, T ($w_k(.), k \in \{A, G, C, T\}$) are defined as:

$$w_k(s_i) = \begin{cases} 1, if\ s_i = k \\ 0, otherwise \end{cases}$$

The indication functions describe the position information of the four nucleotides in sequences. The components of the natural vector can be calculated based on these functions:

- The counts of nucleotides, denoted as n_k, are given by $n_k = \sum_{i=1}^{n} w_k(s_i)$;
- The average locations, denoted as μ_k, are given by $\mu_k = \sum_{i=1}^{n} i \frac{w_k(s_i)}{n_k}$;
- The second central moment of positions, denoted as D_2^k, are determined by $= \sum_{i=1}^{n} \frac{(i-\mu_k)^2 w_k(s_i)}{n_k n}$.

The 12-dimensional natural vector is composed by these three components:

$$(n_A, n_G, n_C, n_T, \mu_A, \mu_G, \mu_C, \mu_T, D_2^A, D_2^G, D_2^C, D_2^T).$$

The counts, average locations and the central moments of four nucleotides A, G, C, T are natural parameters associated to a DNA sequence. And these combined numerical parameters are sufficient to characterize each DNA sequence, so These parameters give a complete understanding of four nucleotides A, G, C and T.

Natural vector with high order moments can also be considered: $D_m^k = \sum_{i=1}^{n} \frac{(i-\mu_k)^m w_k(s_i)}{n_k^{m-1} n^{m-1}}$, which contain more information of nucleotide distribution. For example, 24 dimensional natural vector with 2-th to 5-th order moments is:

2.1.2. K-mer Natural Vector

The k-mer natural vector further takes into account the distribution of k-mers across the entire sequence. K-mer is a contiguous subsequence of length k that serves as a tool for estimating genomic characteristics Liu (2013). For each given k, the number of k-mer is fixed. There consists of 4^k k-mers for a DNA sequence. The specific definition of k-mer natural vector is similar to the original natural vector.

For the same sequence: $S = s_1 s_2 \ldots s_n, s_i \in \{A, G, C, T\}, Str_i (1 \leqslant i \leqslant n-k+1)$is the i-th continuous sub-sequence of length k: $Str_1 = s_1 s_2 \ldots s_k, Str_2 = s_2 s_3 \ldots s_{k+1}, \ldots, Str_{n-k+1} = s_{n-k+1} s_{n-k+2} \ldots s_n$. The indicator functions for each k-mer *kStr* is defined firstly:

$$w_{kStr}(Str_i) = \begin{cases} 1, if\ Str_i = kStr \\ 0, otherwise \end{cases}$$

Is describe the position information of the k-mers in sequences. Then the components of the k-mer natural vector are:

- The counts of k-mers *kStr* are given by $n_{kStr} = \sum_{i=1}^{n-k+1} w_{kStr}(Str_i)$;
- The average positions of k-mers *kStr* are given by $\mu_{kStr} = \sum_{i=1}^{n-k+1} i \frac{w_{kStr}(Str_i)}{n_{kStr}}$;
- The second central moment of positions for k-mers *kStr* are given by $D_2^{kStr} = \sum_{i=1}^{n-k+1} \frac{(i-\mu_{kStr})^2 w_{kStr}(Str_i)}{n_{kStr} * n}$.

Then a DNA sequence can be mapped into $3 * 4^k$-dimensional Euclidean space:

$$(n_{kStr_1}, n_{kStr_2}, \ldots, n_{kStr_{4^k}}, \mu_{kStr_1}, \mu_{kStr_2}, \ldots, \mu_{kStr_{4^k}}, D_2^{kStr_1}, D_2^{kStr_2}, \ldots, D_2^{kStr_{4^k}}).$$

2.2. Convex hull construction and convex hull principle

2.2.1. Convex hull definition

Convex hull of a point set $C = \{x_1, x_2, \ldots, x_n\}$ is the minimal convex set that contains all these points. The convex hull can be represented as: $convC = \{\theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n | \theta_1 + \theta_2 + \ldots + \theta_n = 1, 0 \leqslant \theta_i \leq 1, i \in \{1, 2 \ldots, n\}\}$, which is the linear combination of n points. Here, x_i is the natural vector.

2.2.2. Convex hull principle

Based on natural vectors with high order moments, we can convert nucleotide sequences to points in Euclidean space. Points from the same biological group can then form a convex hull. Convex Hull Principle states that convex hulls constructed from sequences belonging to distinct

Table 1
The number of sequences for the 12 Clades based on the GISAID classification criteria.

	GISAID Clade	Number of Sequences	Number of Filtered Sequences
1	G	9495	4839
2	GH	18952	9722
3	GK	115669	50444
4	GKA	4	3
5	GR	38378	25081
6	GRA	11389	3027
7	GRY	27820	16917
8	GV	2739	1724
9	L	594	460
10	O	2812	1596
11	S	1866	1483
12	V	134	94
	Total	229852	115390

Table 2
The number of sequences for the 12 WHO labels based on the WHO classification criteria.

	WHO Label	Main Pango Lineages	Number of Filtered Sequences
1	Alpha	B.1.1.7 + Q.*	20168
2	Beta	B.1.351 + B.1.351.2 + B.1.351.3	675
3	Delta	B.1.617.2 + AY.*	50894
4	Epsilon	B.1.427 + B.1.429	605
5	Eta	B.1.525	75
6	Gamma	P.1 + P.1.*	4406
7	Iota	B.1.526	215
8	Kappa	B.1.617.1	85
9	Lambda	C.37 + C.37.1	75
10	Mu	B.1.621 + B.1.621.1	397
11	Omicron	B.1.1.529 + BA.*	3001
12	Zeta	P.2	437
	Total		81033

COVID-19 clades or lineages are mutually disjoint.

2.2.3. Convex hull intersection detection

Linear programming method can be applied to check convex hull intersection Sun et al. (2021). Given two finite point sets $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ $(a_i, b_j \in R^N, i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, m\})$, their corresponding linear combinations meet the formula $\sum_{i=1}^{n} \alpha_i a_i = \sum_{j=1}^{m} \beta_j b_j$ if two convex hulls intersect, α_i and β_j are non-negative coefficients that suit $\sum_{i=1}^{n} \alpha_i = 1, \sum_{j=1}^{m} \beta_j = 1$. So we can check the convex hull intersection through solving the following linear programming:

$$\min \quad 0 \tag{1}$$

$$s.t. \quad \sum_{i=1}^{n} \alpha_i a_i = \sum_{j=1}^{m} \beta_j b_j$$
$$\sum_{i=1}^{n} \alpha_i = 1, \qquad \alpha_i \geqslant 0, j = 1, 2, 3 \ldots, n \tag{2}$$
$$\sum_{j=1}^{m} \beta_j = 1 \qquad \beta_j \geqslant 0, j = 1, 2, 3 \ldots, m$$

If the linear programming problem has feasible solution, the two convex hulls, *convA*, *convB*, intersect with each other, otherwise they are disjoint.

3. Dataset

We retrieved all complete genome sequences of SARS-CoV-2 from GISAID https://gisaid.org/ as of June 4, 2022, all originating from human hosts. To ensure the accuracy of the analysis, low-quality sequences were removed from the dataset, retaining a total of 115,390 sequences belonging to 12 GISAID clades. Details for each clade are presented in Table 1. Additionally, only 81,033 sequences from 115,390 have WHO labels, and the information of each label is provided in Table 2.

The total of 115,390 sequences spans six continents, with 27,665 genomes from Asia, 50,643 from Europe, 2,801 from Africa, 6,454 from South America, 27,775 from North America, and only 40 genomes from

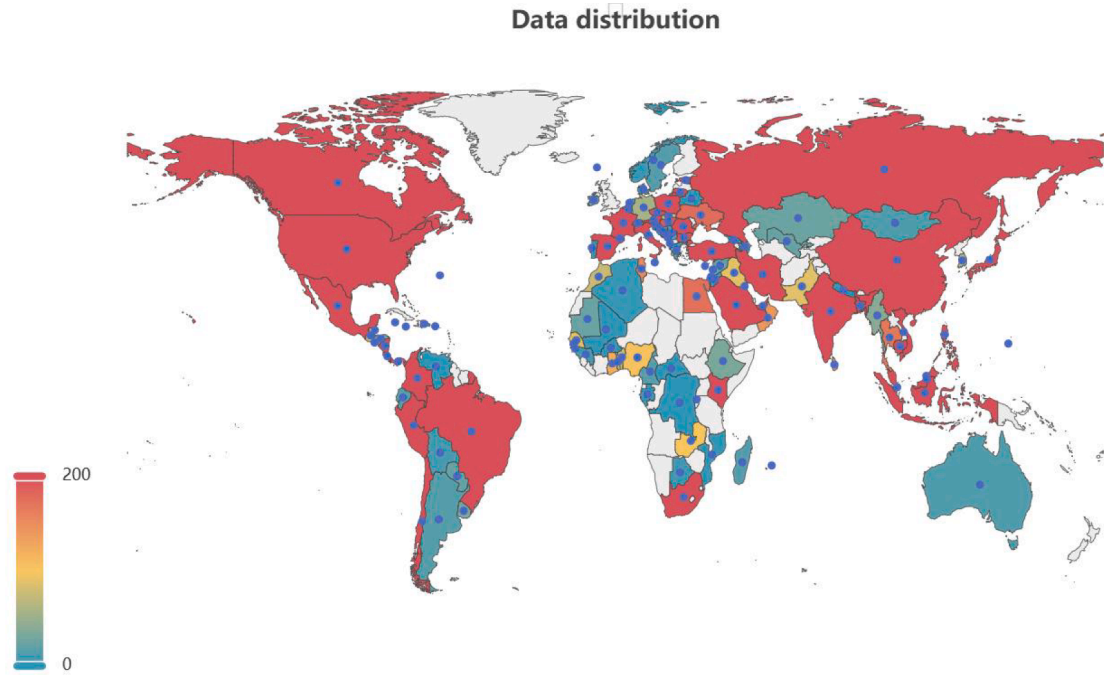

Fig. 1. Global Distribution of Data: regions marked in red indicate higher distribution, while those in blue indicate lower distribution.

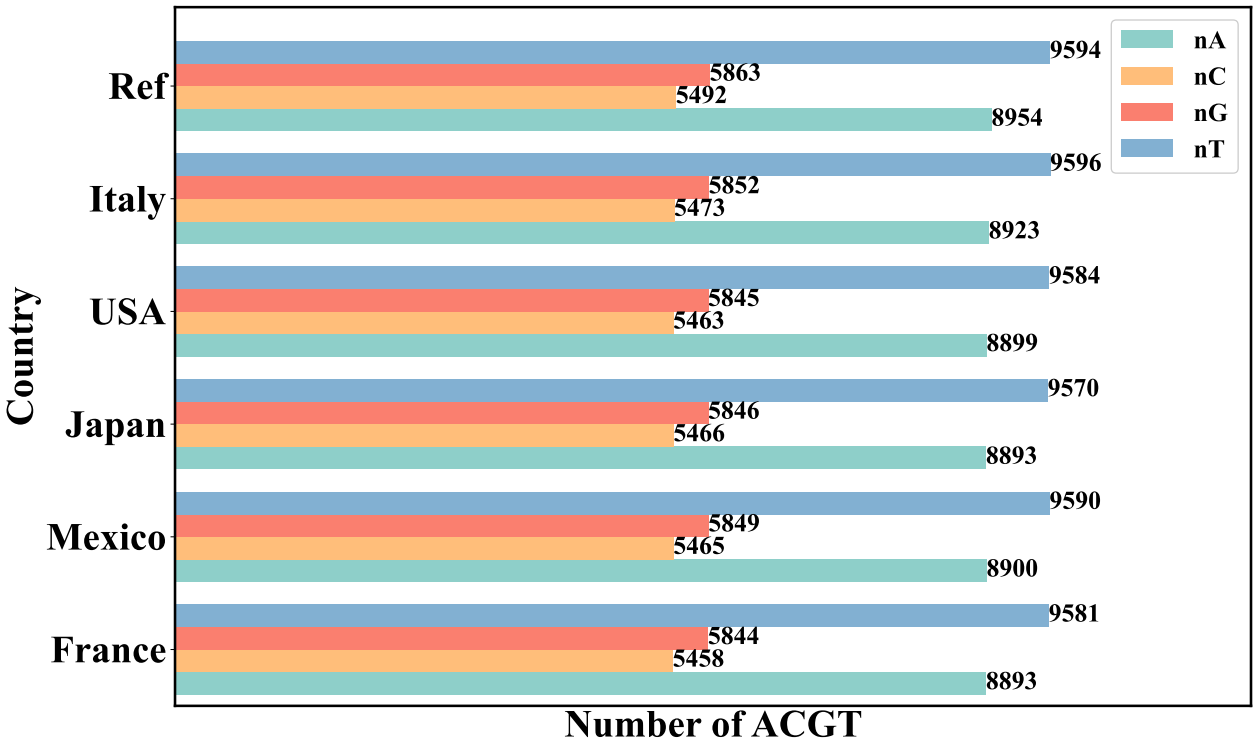

Fig. 2. Comparative analysis of ACGT counts between the five countries with the highest number of genomes and the Reference Sequence. The ACGT count for each country represents the average count of nucleotides.

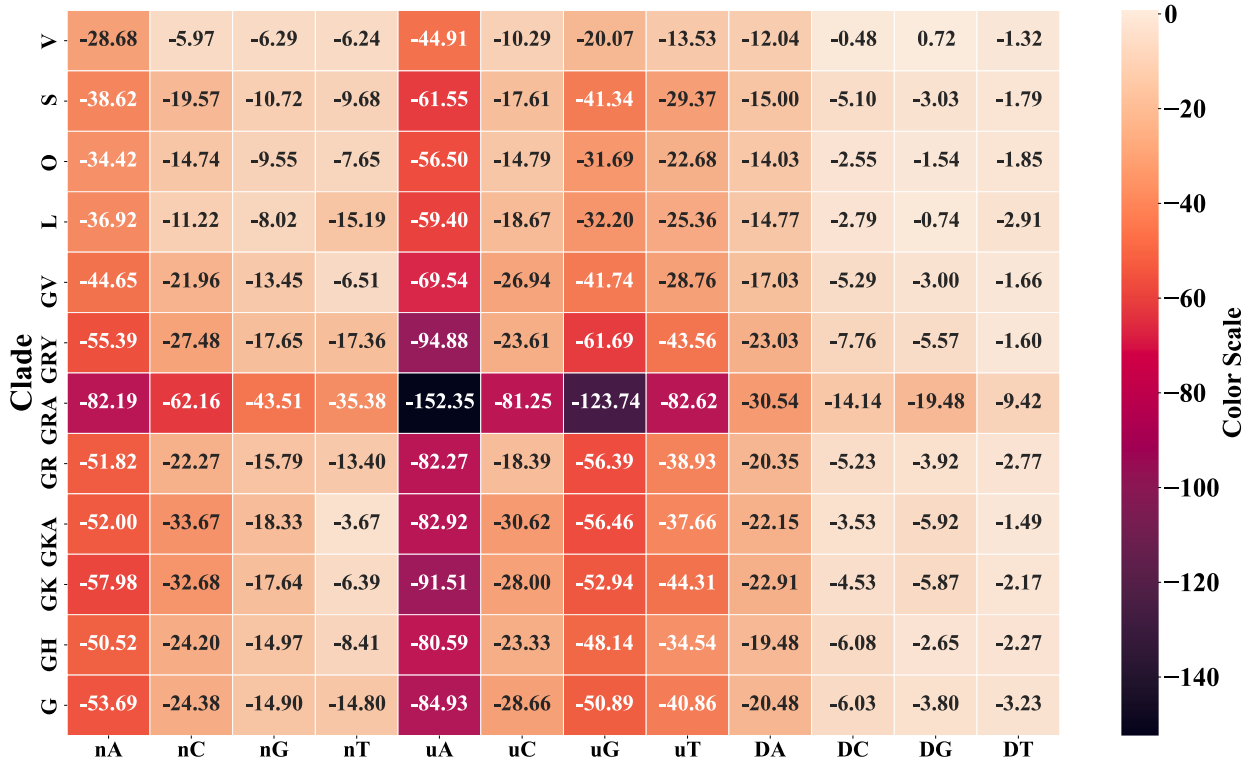

Fig. 3. Heatmap illustrating the difference in the ACGT's distribution characteristics for the 12 GISAID clades compared to the reference sequence. The horizontal axis includes 12 distribution statistics: the counts of ACGT (nA, nC, nG, nT), the average location of ACGT ($\mu A, \mu C, \mu G, \mu T$), and the second central moment of ACGT (DA, DC, DG, DT). The vertical axis represents the 12 GISAID clades.

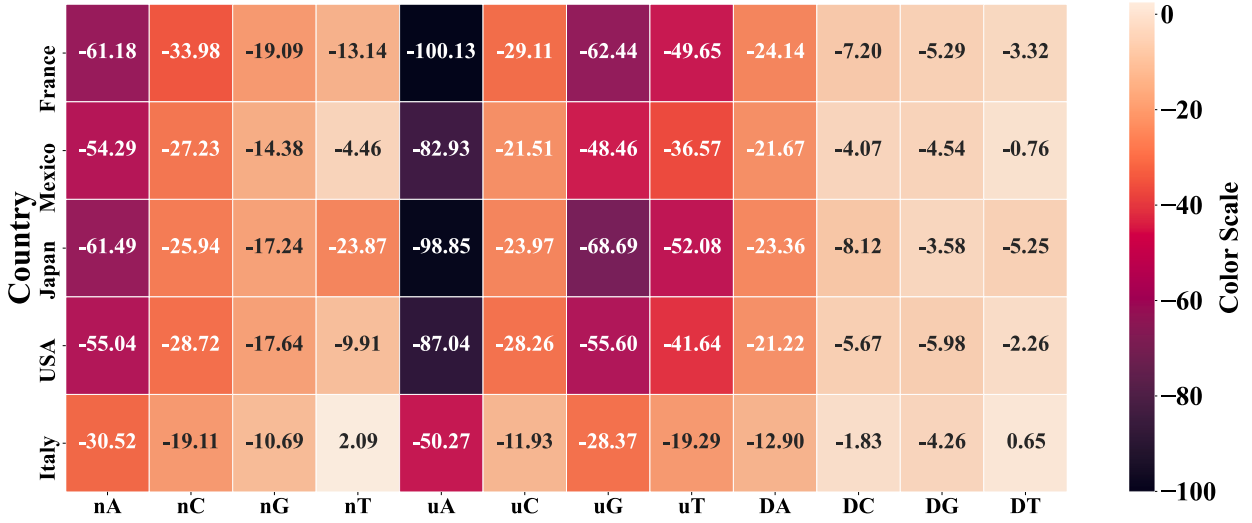

Fig. 4. Heatmap of the difference in the ACGT's distribution characteristics of top five countries compared to the reference sequence. The horizontal axis includes 12 statistics of distribution: the number of ACGT(nA, nC, nG, nT), the average location of ACGT($\mu A, \mu C, \mu G, \mu T$), the second central moment of ACGT(DA, DC, DG, DT). The vertical axis represents the five countries.

Table 3
Subtype composition of top five countries.

	France	Mexico	Japan	USA	Italy
V	0.026	0.000	0.021	0.013	0.120
S	0.518	0.204	1.698	0.333	0.060
O	0.303	0.163	1.223	0.373	0.622
L	0.007	0.007	1.329	0.067	0.000
GV	1.025	0.034	0.000	0.040	10.442
GRY	20.807	4.206	1.223	10.215	34.478
GRA	0.884	2.627	1.729	2.200	1.426
GR	2.734	26.323	87.801	8.975	11.486
GKA	0.011	0.000	0.000	0.000	0.000
GK	66.808	57.425	2.6360	39.712	35.402
GH	5.0011	3.219	0.928	33.751	1.345
G	1.875	5.791	1.413	4.321	4.618

Oceania. The SARS-CoV-2 dataset is widely distributed globally as shown in Fig. 1. France has the highest count with 27,034 sequences, followed by Mexico with 14,694 genomes, Japan with 9,484 genomes, the USA with 7,499 genomes, Italy with 4,980 genomes, Brazil with 4,369 genomes, Malaysia with 3,482 genomes, Austria with 3,329 genomes, Spain with 2,905 genomes, and Indonesia with 2,470 genomes.

4. Results

4.1. Statistical analysis

To discover the differences between our data set and reference sequence of SARS-CoV-2, we conducted a comparative analysis of the distributions of nucleotides AGCT from various perspectives, including countries, GISAID clades, and genders.

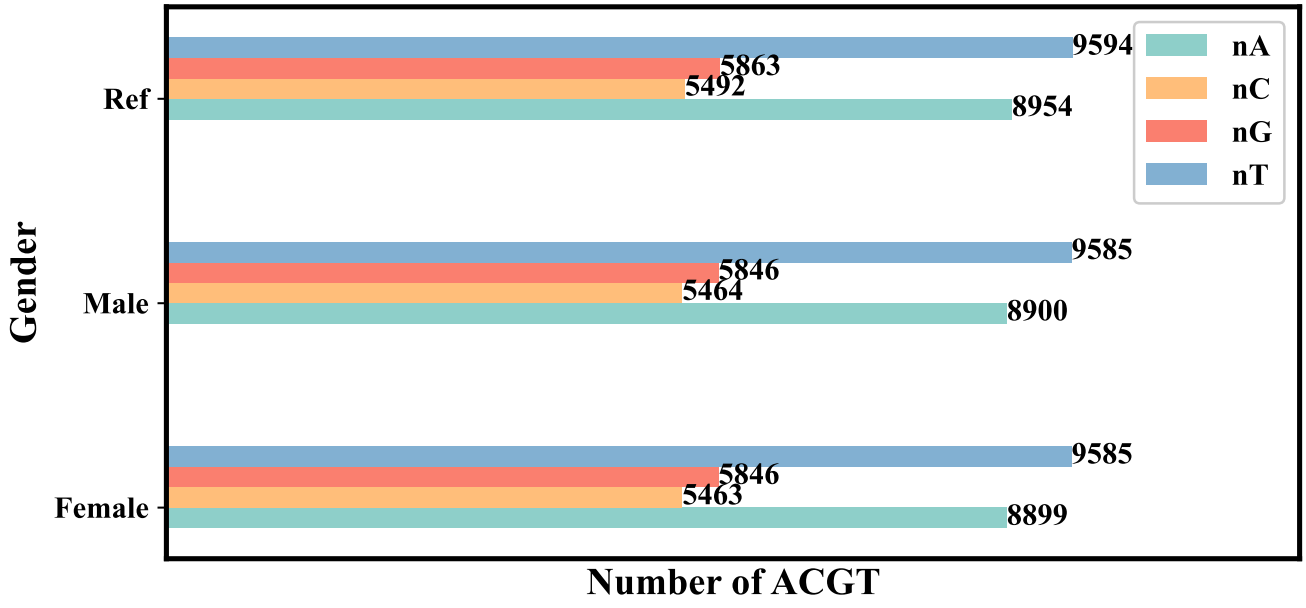

Fig. 5. Comparison of the number of nucleotides ACGT between Famale/Male and reference sequence: the number of nucleotides ACGT of each gender is the average number of nucleotides.

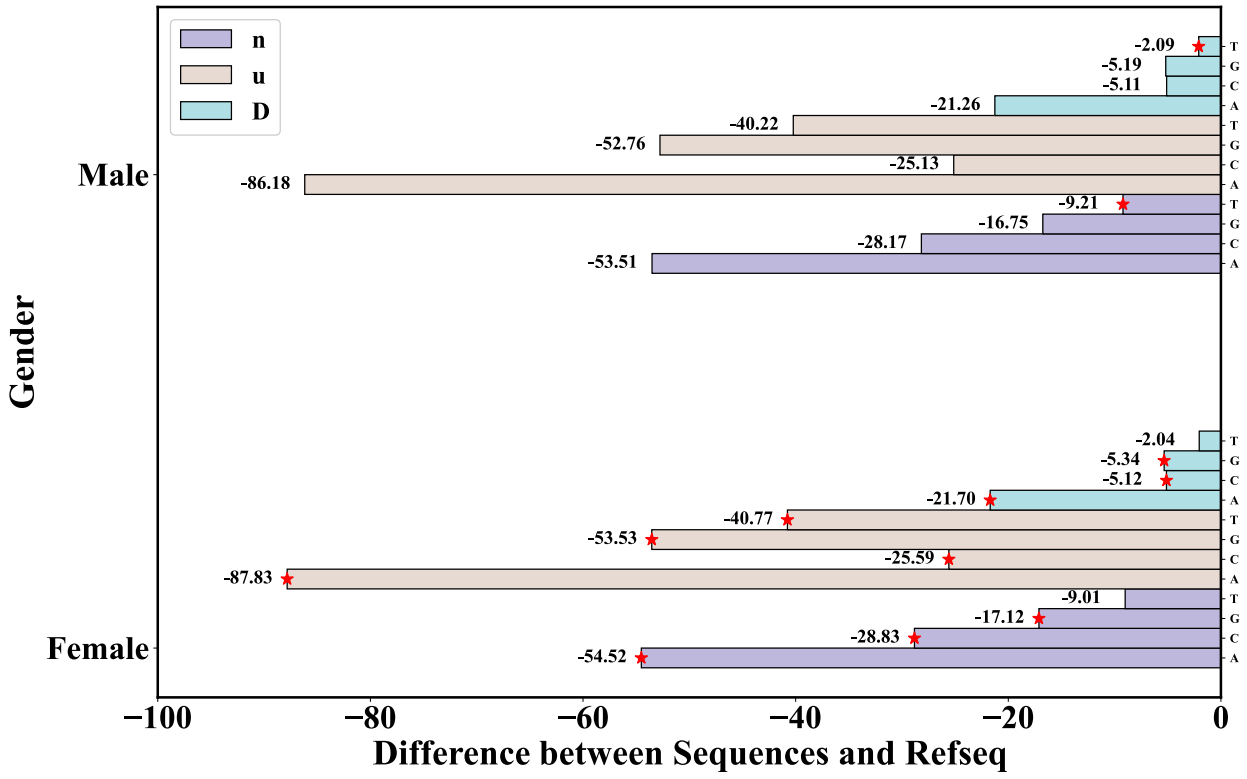

Fig. 6. Difference in the ACGT's distribution characteristics of Famale/Male and the reference sequence: red star represents the greater difference between two genders compared to reference sequence.

There was no statistically significant difference in the counts of ACGT between the five countries and the reference sequence, as depicted in Fig. 2. Except for Italy, the number of ACGT decreased slightly in other four countries. There was an increase in nucleotide T compared to the reference sequence for Italy.

Considering the average location and second-order central moments of the four nucleotides, we compared these statistics of the twelve GISAID clades with those of reference sequence. The heatmap in Fig. 3 showes that in vertical comparison, the difference between clades and reference sequence mainly resides n_A, μ_A, μ_G, and μ_T. The distribution of adenine A in all clades significantly differs from that of reference sequence. The color of n_A is darker than the number of T,C,G for all clades, as is the case with uA,DA, which implies that adenine A of these clades is worthy of our attention. In horizontal comparison, the most substantial difference is observed between clade GRA and reference sequence, which agrees with the fact that GRA, being an Omicron strain, is highly infective and pathogenic due to its numerous mutation sites in Spike protein.

Fig. 4 illustrates the differences of ACGT's distribution characteristics between the top five countries or regions with the maximum data volume and reference sequence. In vertical comparison, the gaps between these countries and reference sequence mainly lie in nA, uA, uG, and uT, especially in the n_A, the gap between these countries and reference sequence is greatest, which agrees with the result of heatmap 3. Horizontal comparison reveals that there is a significant difference between France, Japan and reference sequence. By analyzing the proportion of clades in these countries, as presented in Table 3: we discover that with regard to France, the clades with the highest proportion are GK (Delta dominated) and GRY (Alpha), which differ significantly from reference sequence and the differences are visible on the heatmap of twelve GISAID clades. For Japan, the clade with the highest proportion is GR (Gamma, Alpha), which is also significantly different from reference sequence. Thus, the percentages of their clade composition can account for the significant difference between France, Japan, and reference sequence.

Table 4
Intersection results of different spaces with different order central moments(j is the highest order in natural vector).

Euclidean space	j = 2	j = 3	j = 4	j = 5	j = 6	j = 7	j = 8
	R^{12}	R^{16}	R^{20}	R^{24}	R^{28}	R^{32}	R^{36}
No.of disjoint convex hull pairs	14	18	18	66	66	66	66
No.of intersectant convex hull pairs	52	48	48	0	0	0	0

From the perspective of genders, 47,465 sequences of the total data set were males, 47,778 sequences were females and genders of the rest were unknown. Firstly, we created a histogram of the ACGT counts Fig. 5 as well as a histogram of the differences in the ACGT distributions (counts, average positions, and center moments) between the males, females, and reference sequence groups Fig. 6. Both genders have fewer nucleotide counts than reference sequence, according to the ACGT distribution histogram. Taking first-order and second-order moments into consideration, it is observed that, while there is no significant difference between males and females and reference sequence overall, the difference between females and reference sequence is greater than that of males. The specific difference is that, except for the number and second-order moment of thymine T, the difference in females is greater than that in males and reference sequence..

4.2. Convex hull analysis

For each SARS-CoV-2 genome sequence, we calculated its natural vectors with different order central moments, and constructed the

Table 5
The classification results of four metrics: k values range from 1 to 7, two weights ($1/2^n, 1/n^2$), and two norms (L-1 norm and L-2 norm)

L1-norm	Weight	D_1	D_2	D_3	D_4	D_5	D_6	D_7
	$1/2^n$	0.9144	0.9575	0.9681	0.9700	0.9704	0.9710	0.9810
	$1/n^2$	0.9144	0.9558	0.9681	0.9699	0.9705	0.9709	0.9781
L2-norm	Weight	D_1	D_2	D_3	D_4	D_5	D_6	D_7
	$1/2^n$	0.9100	0.9558	0.9633	0.9669	0.9673	0.9709	0.9824
	$1/n^2$	0.9100	0.9479	0.9622	0.9671	0.9674	0.9686	0.9788

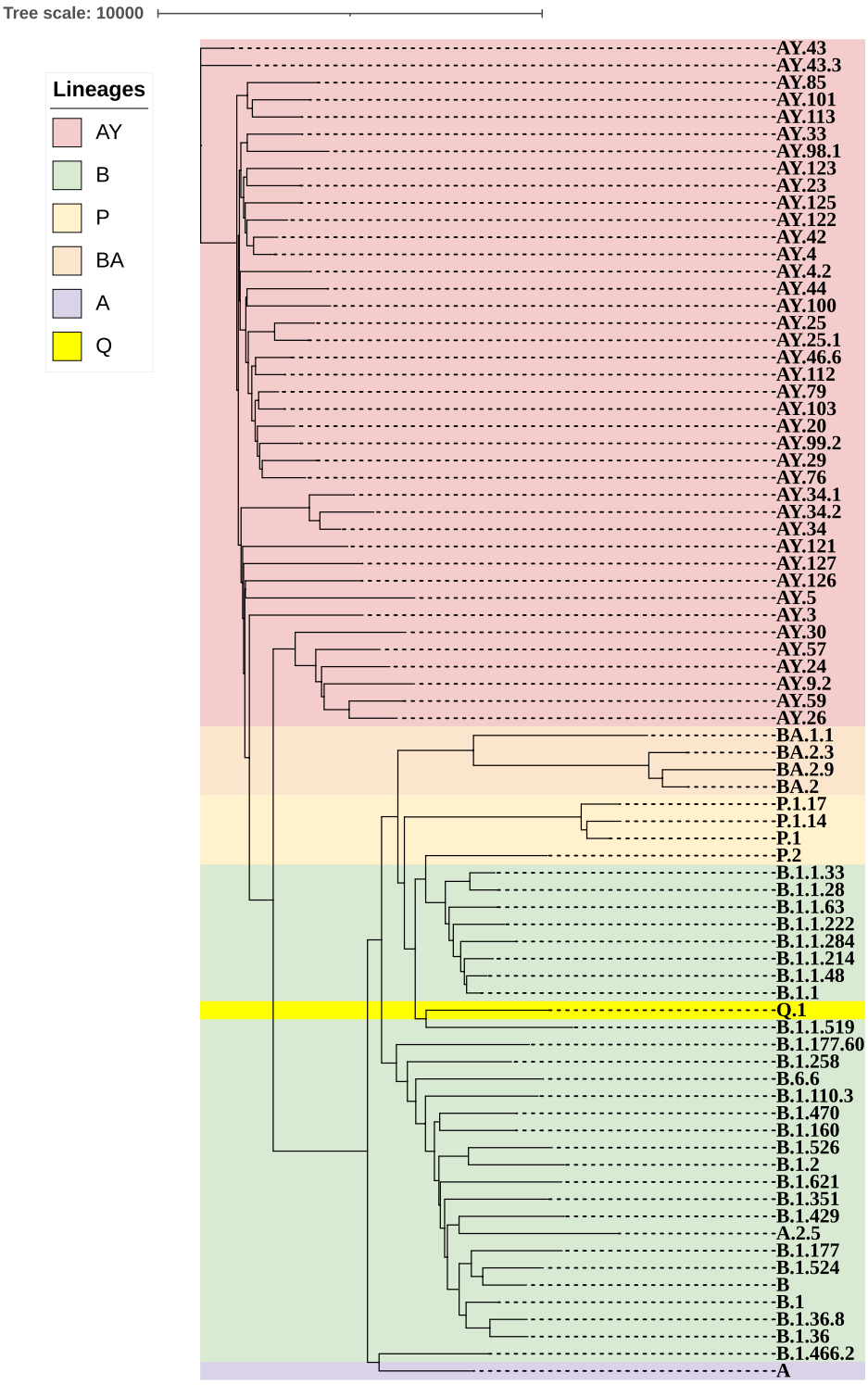

Fig. 7. The phylogenetic BME tree using FastME(by Tree refinement with Subtree Pruning and Regrafting) on 78 lineages shown in Table S1.

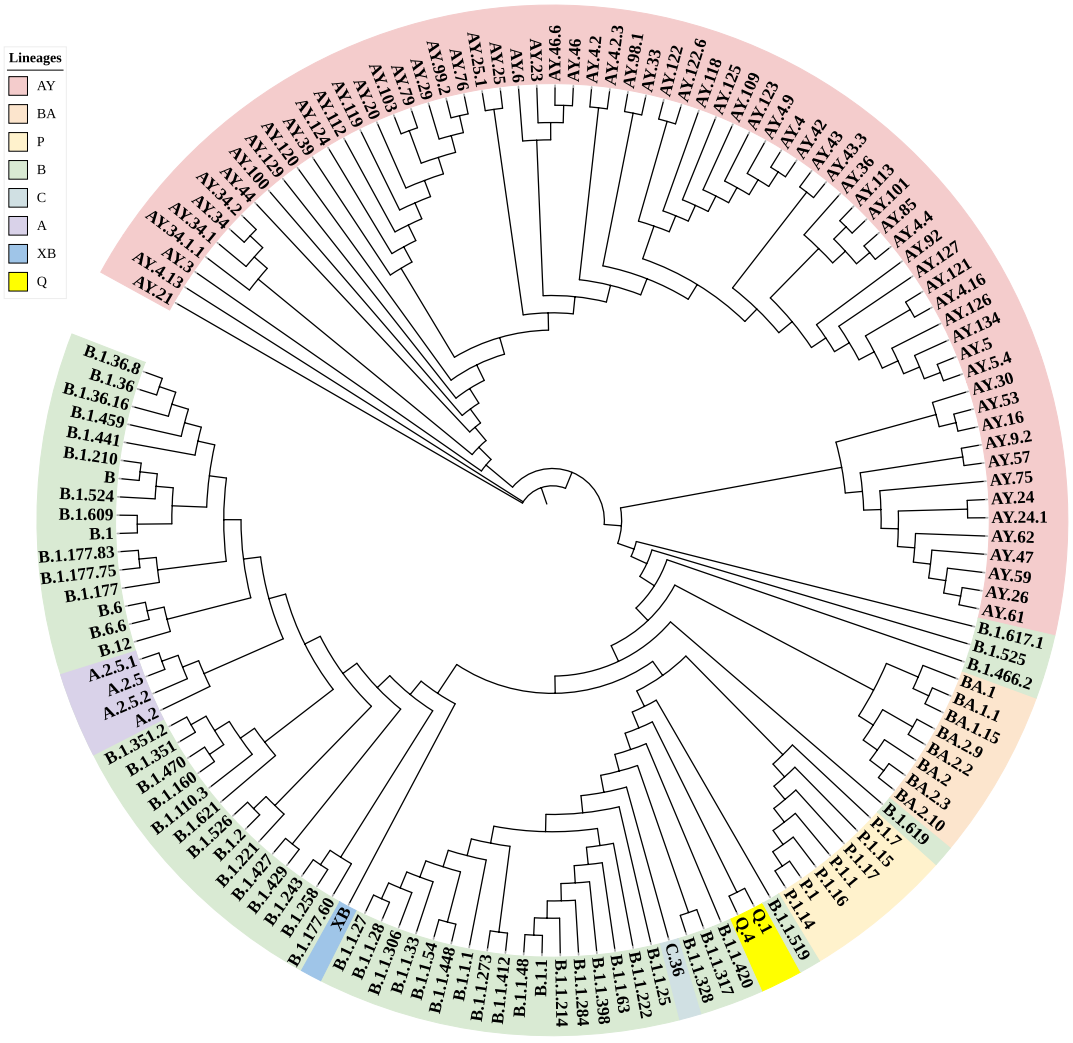

Fig. 8. The phylogenetic NJ tree using FastME(by Tree refinement with Subtree Pruning and Regrafting) on 146 lineages shown in Table S2.

convex hull of each GISAID clade in corresponding euclidean space. In the original 12-dimensional space, 14 pairs out of 66 pairs of convex hulls intersect with each other. As the higher order central moments are added to natural vector, the number of intersectant convex hull pairs decline and all pairs are disjoint in 24-dimensional space, where natural vector covers second to fifth order central moments. Table 4 displays the intersection results of several spaces in detail. The findings demonstrate that in order to define sequences, the fifth order central moment is required.

4.3. 1-Nearest neighbor analysis

To examine the classification performance of the SARS-CoV-2 sequences, both natural vectors and k-mer natural vectors were applied. The 1-nearest neighbor algorithm(1NN) was used to classify the sequences to different GISAID clades. In an attempt to determine which metric had the best classification accuracy, we experimented with several k values, weights ($1/2^n$,$1/n^2$), and two norms (L-1 norm and L-2 norm). We employed four metrics, which are mixtures of several k-mer distances:

$$\begin{aligned}
&L1-norm\,with\,1/2^n\,weight: D_n = d_1 + 1/2 * d_2 + \ldots + 1/2^k * d_k,\\
&d_i = \|vec_1 - vec_2\|_1\\
&L1-norm\,with\,1/n^2\,weight: D_n = d_1 + 1/2^2 * d_2 + \ldots + 1/k^2 * d_k,\\
&d_i = \|vec_1 - vec_2\|_1\\
&L2-norm\,with\,1/2^n\,weight: D_n = d_1 + 1/2 * d_2 + \ldots + 1/2^k * d_k,\\
&d_i = \|vec_1 - vec_2\|_2\\
&L2-norm\,with\,1/n^2\,weight: D_n = d_1 + 1/2^2 * d_2 + \ldots + 1/k^2 * d_k,\\
&d_i = \|vec_1 - vec_2\|_2,
\end{aligned} \tag{3}$$

d_i is the distance between two i-mer natural vectors. The classification results are shown in Table 5.

The highest accuracy is 0.9824 under the L2-norm metric using $1/2^n$ weighted 1 to 7 mer natural vectors, which means 98.24% of sequences are correctly classified into their respective GISAID clades. The classification accuracy of natural vectors alone is not high, only 0.9127 for 12-dimensional natural vectors.

4.4. Phylogenetic tree and natural graph

The good performance of L2-norm metric using $1/2^n$ weighted 1 to 7 mer natural vectors provides a fancy tool to carry out phylogenetic analysis. Similar sequences with the same label are expected to cluster together, and established classification standards like the GISAID and WHO labels can serve as normative references for sequence clustering.

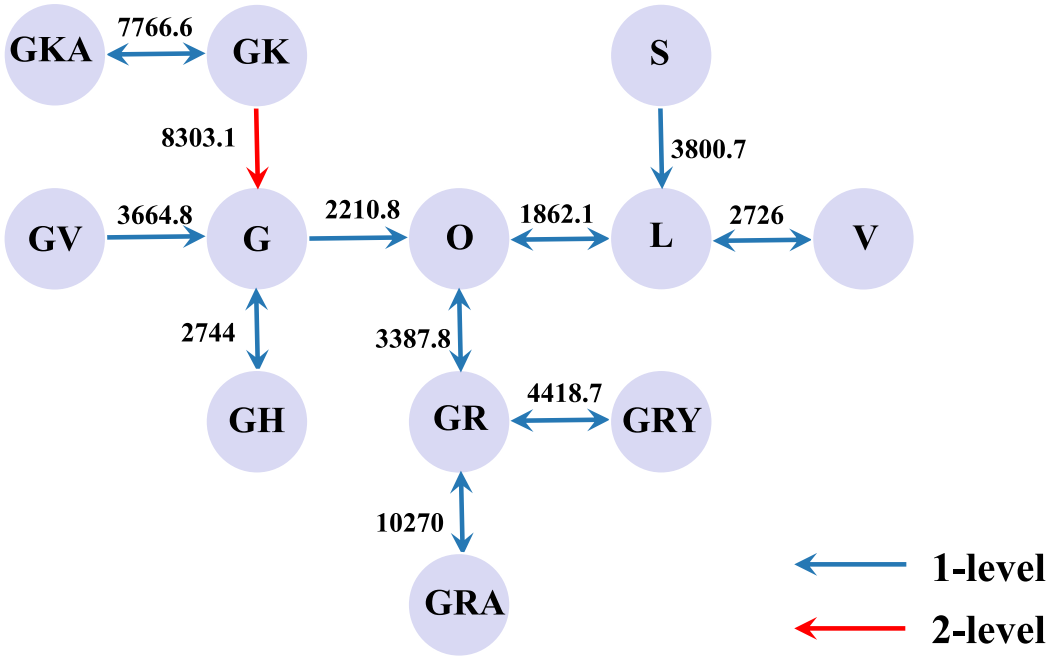

Fig. 9. The natural graph of twelve GISAID clades of total data set, the blue lines represent the 1-level connected components and the red ones 2-level.

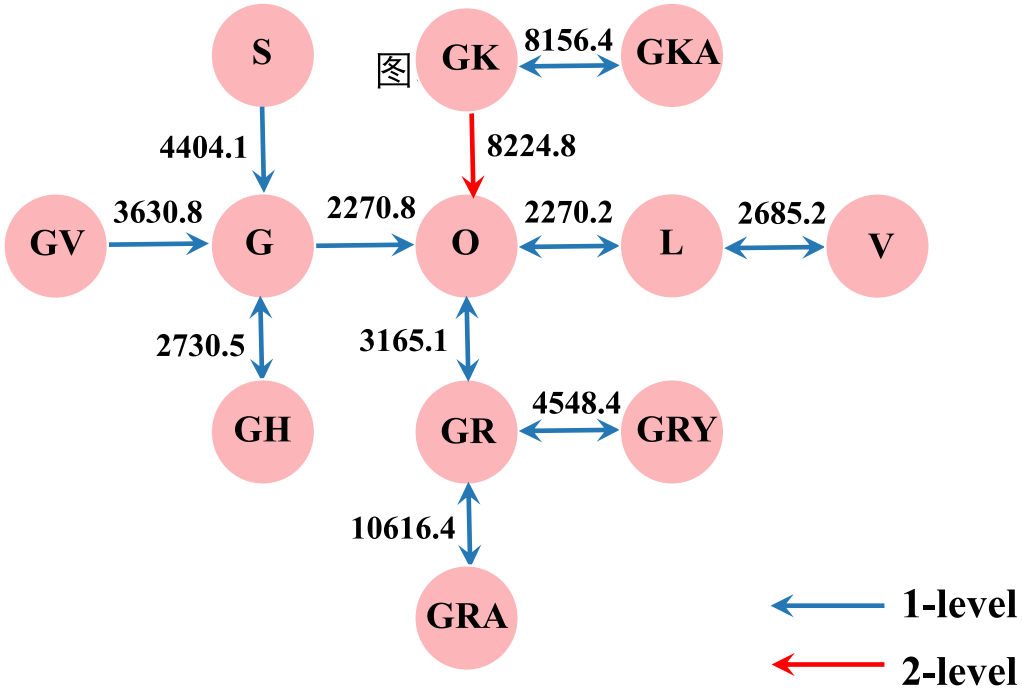

Fig. 10. The natural graph of twelve GISAID clades of females, the blue lines represent the 1-level connected components and the red ones 2-level.

Lineages are named using an alphabetical prefix (such as B or BA) and numerical suffix (such as ".1" or ".1.1.5")lineages (2021). When a new lineage is defined, the Pango system assigns an additional number to the name of its parent lineage (e.g.. BA.2.75 is a sublineage of BA.2). As the virus continues to change, the Pango lineage names can become very long. Lineages with longer names may be given alphabetic aliases and numbering continues (e.g.. "BA" stands for "B.1.1.529," thus BA.2 is the same as B.1.1.529.2). In this section, we hope that lineages with the same alphabetical prefix cluster in phylogenetic trees generated by our alignment-free approach and selected metric, which will intuitively depict the relationships between Pango lineages. To do this, we represented each Pango lineage using its average vector.

Certain lineages in the SARS-CoV-2 data set contain limited data, so in order to create phylogenetic trees, we used lineages with considerable data volume. Two lineage sets were chosen for the phylogenetic tree construction, the specific lineage information is shown in

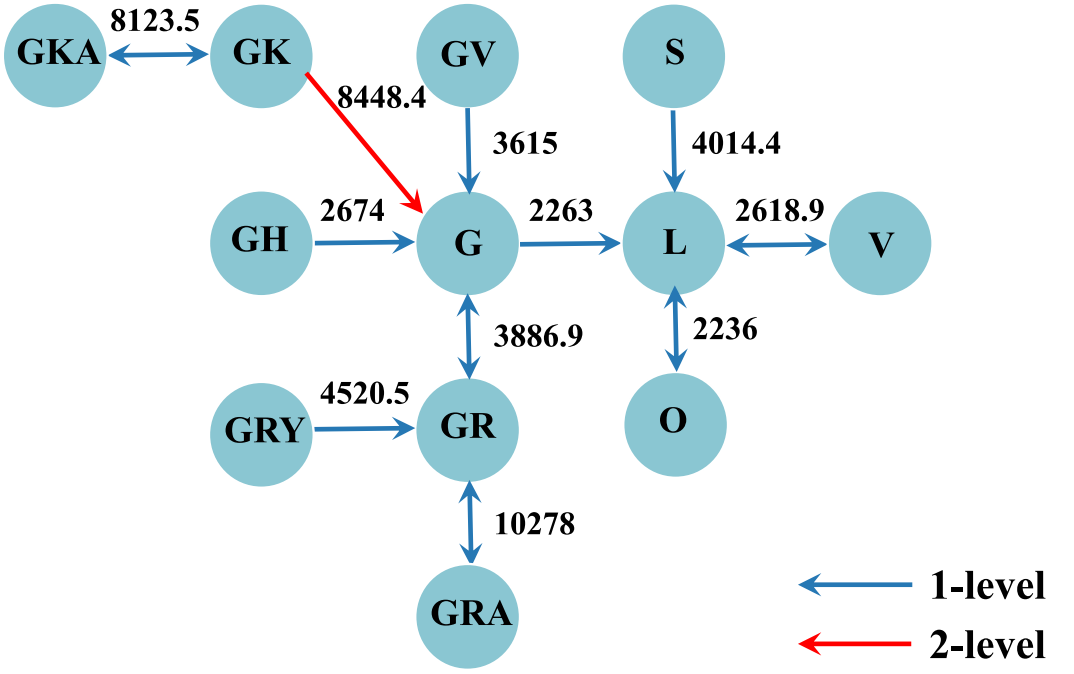

Fig. 11. The natural graph of twelve GISAID clades of males, the blue lines represent the 1-level connected components and the red ones 2-level.

Supplementary materials. The phylogenetic tree 7 indicates that two primary branches are formed by lineages with the alphabetical prefixes 'AY' and 'B'. Furthermore, 'BA' lineage clusters, 'P' lineage with the exception of 'P.2' gathers together, and 'P.2' lineage is not distant from 'P' branch. A clustering pattern can also be observed that all lineages with prefix 'B.1.1.' distribute in a single branch. The Supplementary materials contain a list of the sequences covered by phylogenetic tree 7. (see Fig. 8).

There are additional lineages with other alphabetical prefixes (e.g.. XB,C) among the 146 lineages. In the phylogenetic tree containing 146 lineages, 'AY' lineage forms one branch with shorter branch length compared to the others. Lineage 'A', 'Q', 'P', 'BA' respectively gather together and they are closer to lineage 'B'. Therefore, both phylogenetic trees support the usefulness of our weighted metric and k-mer natural vector.

Another graphical tool for describing the distance relationships between sets is the natural graph, which was initially shown by Yu et al. (2013). Drawing a two-level natural graph requires two steps, which correspond to two levels of the graph. Level-1: This is the result of the first step to find the closest elements to all the points. Then the drawn lines which directs to their nearest points are called level-1. The direction in the graph can show the closest elements of each element based on their biological distances. Level-2:After the first turn to link the nearest points in the graph, we get many sub-graphs. We compute the distance matrix for these graphs. The distance between two graphs is defined as the minimum of all distance between any element in one graph and any element in the other graph. Then the lines linked different graphs are called level-2. In this section, we drew one natural graph of all sequences and separate natural graphs for both genders. The natural graphs can roughly reflect the spatial distribution of twelve GISAID clades.

The natural graph 9 shows that: In level-1, GK and GKA are closest to each other. GRA,GRY are closest to GR. GV, GH are closest to G in level-1, GK is linked to G in level-2, which means it was further away from G. The natural graph of females 10 shows little change compared to the overall one, with the following differences: Clade S is connected to G. S is closest to G in level-1. 1-level structure GK-GKA is connected to clade O. In level-2, the GK-GKA sub- graph is connected to clade O, which means it is closest to O in level-2. It is different from the overall graph, in which the level-2 closest clade is G. The 1-level structures of GR, GRY, and GRA, as well as GK and GKA, are still maintained locally, and there haven't been any notable modifications to the overall structure. The differences between natural graphs of all sequences and males 11 lie in: Clade L, GR are directly connected to G, while clade O is directly connected to L. The 1-level structures of GR, GRY and GRA; GK and GKA are still maintained locally. L,GR are closest to G in level-1. O is closest to L and not the center point of the natural graph.

Overall, the natural graphs we obtained are consistent with description that given in GISAID clade naming. Starting from S and L Tang et al. (2020), S continues to be at a moderate level, L splits into initial versions of G and V, G further splits into GR and GH, and later GV. According to observations, GR splits into GRY. Later, a new branch splits out from the basic branch G, forming branch GK. More information of these lineage naming standards are shown in the supplementary material.

5. Discussion

Overall, our research determined the SARS-CoV-2 geometric space using the natural vector approach. As a subspace in the Euclidean space, the distances between points corresponding to different SARS-CoV-2 sequences reflect their biological distances, the efficiency and dependability of this geometric space and the distances have been proven by phylogenetic and 1-NN(1-nearest neighbor) classification analyses. This space also enables mathematical techniques like linear programming and convex hulls to be applied in the further research. Moreover, our findings imply that the L2-norm metric with $1/2^n$ weighted 1 to 7 mer natural vectors has the best performance. This can handle the complex calculations when MSA(Multiple Sequence Alignment) is used to deal with long sequences or large data volume. When a new mutation sequence appears, quick comparison with all other SARS-CoV-2 sequences can be finished based on our geometric space, just by calculating their Euclidean distances. Then an initial prediction can be made

and targeted treatment methods can be considered. Our research have shown some significant discoveries, but they have also had certain drawbacks. First, the data scale has an impact on classification performance; for instance, small set size will result in low accuracy because one misclassified point will drastically decrease the accuracy.For our paper, we have downloaded as many reliable sequences as possible. Perhaps we can calculate classification accuracy by removing 5%, 10%, 20% of the sequences from our data set to improve the performance. Second, the next nearest neighbor would be overlooked because our natural graph is limited to showing the distance relationships between the two closest sets. Perhaps future research will lead to the creation of a better natural graph model.

6. Conclusion

In this work, we established convex hulls for various clades by analyzing the SARS-CoV-2 data set using natural vectors with introduced high order central moments. We found that convex hulls of different clades did not intersect in 24-dimensional euclidean space (including 5th order moments), meeting convex hull principle. This suggests that the information covered by our natural vectors varies amongst sequences belonging to distinct clades. Furthermore, we performed 1NN classification at the clade level using k-mer natural vectors. Using $1/2^n$ weighted 1 to 7 mer natural vectors, the L2-norm metric is selected to achieve the highest classification accuracy. Under this metric, the classification accuracy reaches 0.9824. In order to demonstrate the relationships between different clades, we established phylogenetic trees and natural graphs under selected metrics, and discovered the relationships between clades, for example, Fig. 9 shows that GRA and GRY, are clades derived from GR and are closest to GR. Consequently, this validates the efficacy of our chosen metric and natural vectors. The geometric space is an effective tool in SARS-CoV-2 sequence analysis.

Author Contributions

SS-TY conceived the project and designed the study. NS collected data and MC carried out the data analysis including figures drawing and wrote the preliminary version of the paper. All authors have read and agreed to the published version of the manuscript.

Funding

This work is supported by National Natural Science Foundation of China (NSFC) grant (12171275) and Tsinghua University Education Foundation fund (042202008).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

SS-TY is grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gene.2024.148291.

References

Adil, M.T., Rahman, R., Whitelaw, D., Jain, V., Al-Taan, O., Rashid, F., Munasinghe, A., Jambulingam, P., 2021. Sars-cov-2 and the pandemic of covid-19. Postgraduate Med. J. 97 (1144), 110–116.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucl. Acids Res. 25 (17), 3389–3402.

Asselah, T., Durantel, D., Pasmant, E., Lau, G., Schinazi, R.F., 2021. Covid-19: Discovery, diagnostics and drug development. J. Hepatol. 74 (1), 168–184.

Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.-T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PloS one 6 (3), e17293.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. Knn model-based approach in classification. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3–7, 2003. Proceedings. Springer, pp. 986–996.

Katoh, K., Asimenos, G., Toh, H., 2009. Multiple alignment of dna sequences with mafft. Bioinform. DNA Sequence Anal. 39–64.

lineages, P., 2021. Global report investigating novel coronavirus haplotypes. Website, URL https://cov-lineages.org/global_report.html.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., Fan, W., 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv preprint arXiv:1308.2012.

Sun, N., Pei, S., He, L., Yin, C., He, R.L., Yau, S.S.-T., 2021. Geometric construction of viral genome space and its applications. Comput. Struct. Biotechnol. J. 19, 4226–4234.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al., 2020. On the origin and continuing evolution of sars-cov-2. Natl. Sci. Rev. 7 (6), 1012–1023.

Vasireddy, D., Vanaparthy, R., Mohan, G., Malayala, S.V., Atluri, P., 2021. Review of covid-19 variants and covid-19 vaccine efficacy: what the clinician should know? J. Clin. Med. Res. 13 (6), 317.

Yu, C., Deng, M., Cheng, S.-Y., Yau, S.-C., He, R.L., Yau, S.S.-T., 2013. Protein space: A natural method for realizing the nature of protein universe. J. Theor. Biol. 318, 197–204. URL https://www.sciencedirect.com/science/article/pii/S0022519312005802.

Zhang, D., Hu, M., Ji, Q., 2020. Financial markets under the global pandemic of covid-19. Finance Res. Lett. 36, 101528.

Zhao, X., Tian, K., He, R.L., Yau, S.S.-T., 2019. Convex hull principle for classification and phylogeny of eukaryotic proteins. Genomics 111 (6), 1777–1784.

Zhao, X., Tian, K., Yau, S.S.-T., 2018. A new efficient method for analyzing fungi species using correlations between nucleotides. BMC Evol. Biol. 18, 1–13.