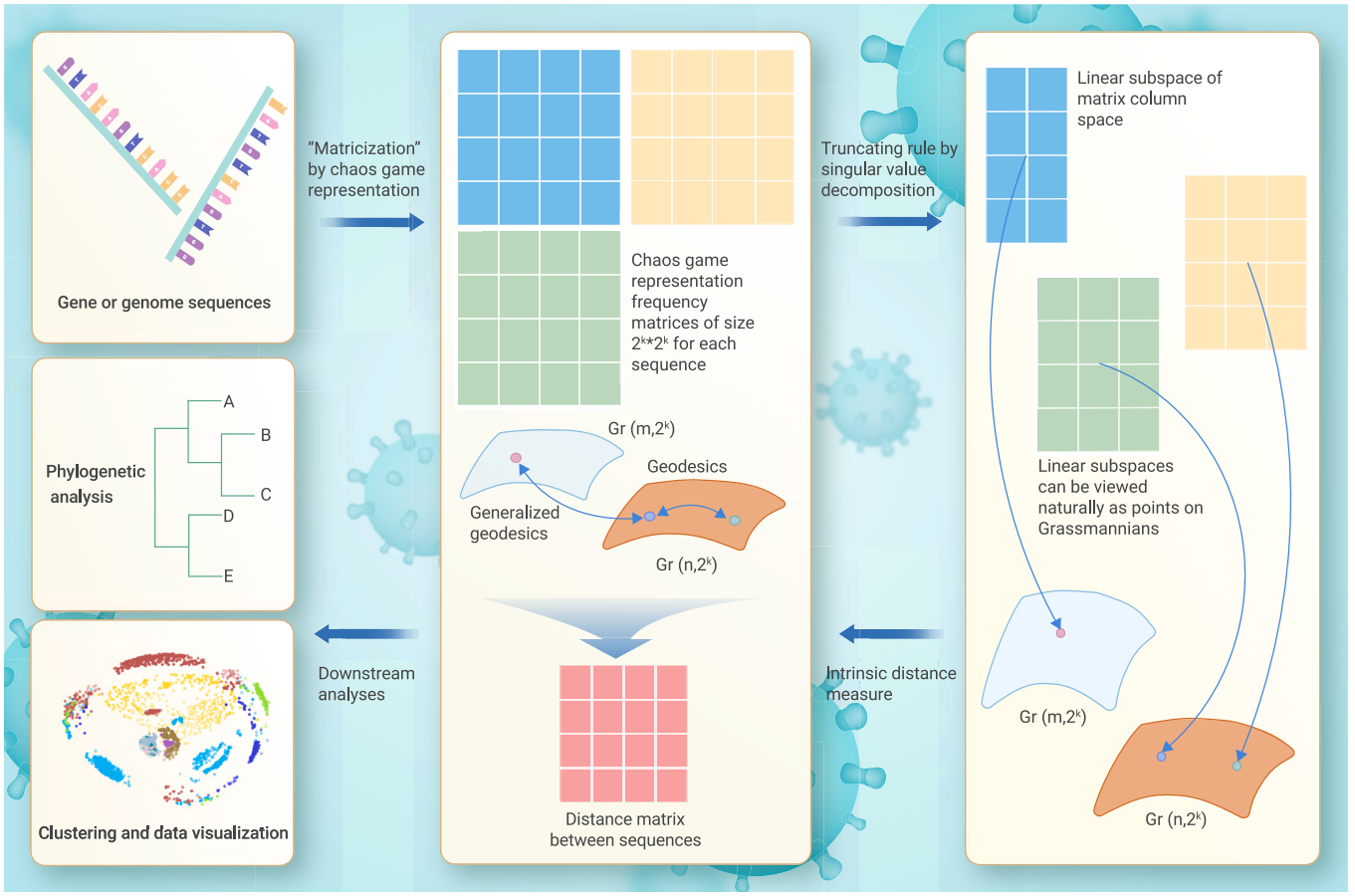
REPORT

Exploring geometry of genome space via Grassmann manifolds

Xiaoguang Li,[1,4] Tao Zhou,[2,4] Xingdong Feng,[1,*] Shing-Tung Yau,[2,3,*] and Stephen S.-T. Yau[2,3,*]
*Correspondence: feng.xingdong@msg.sufe.edu.cn (X.F.); styau@tsinghua.edu.cn (S.-T.Y.); yau@tsinghua.edu.cn (S.S.-T.Y.)
Received: February 6, 2024; Accepted: July 18, 2024; Published Online: July 22, 2024; https://doi.org/10.1016/j.xinn.2024.100677

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- An alignment-free method that involves embedding genome sequences into Grassmann manifolds of different dimensions.
- Geodesic distance is used to capture the relation between chaos game representation frequency matrices of genomes.
- The distribution of genomes on the manifolds is demonstrated in 3D space using multi-dimensional scaling.
- Our method provides a new perspective on manifold-based embedding for alignment-free approaches.

REPORT

Exploring geometry of genome space via Grassmann manifolds

Xiaoguang Li,[1,4] Tao Zhou,[2,4] Xingdong Feng,[1,*] Shing-Tung Yau,[2,3,*] and Stephen S.-T. Yau[2,3,*]

[1]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

[2]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

[3]Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, China

[4]These authors contributed equally

*Correspondence: feng.xingdong@msg.sufe.edu.cn (X.F.); styau@tsinghua.edu.cn (S.-T.Y.); yau@tsinghua.edu.cn (S.S.-T.Y.)

Received: February 6, 2024; Accepted: July 18, 2024; Published Online: July 22, 2024; https://doi.org/10.1016/j.xinn.2024.100677

Citation: Li X., Zhou T., Feng X., et al., (2024). Exploring geometry of genome space via Grassmann manifolds. The Innovation **5(5)**, 100677.

It is important to understand the geometry of genome space in biology. After transforming genome sequences into frequency matrices of the chaos game representation (FCGR), we regard a genome sequence as a point in a suitable Grassmann manifold by analyzing the column space of the corresponding FCGR. To assess the sequence similarity, we employ the generalized Grassmannian distance, an intrinsic geometric distance that differs from the traditional Euclidean distance used in the classical k-mer frequency-based methods. With this method, we constructed phylogenetic trees for various genome datasets, including influenza A virus hemagglutinin gene, Orthocoronavirinae genome, and SARS-CoV-2 complete genome sequences. Our comparative analysis with multiple sequence alignment and alignment-free methods for large-scale sequences revealed that our method, which employs the subspace distance between the column spaces of different FCGRs (FCGR-SD), outperformed its competitors in terms of both speed and accuracy. In addition, we used low-dimensional visualization of the SARS-CoV-2 genome sequences and spike protein nucleotide sequences with our methods, resulting in some intriguing findings. We not only propose a novel and efficient algorithm for comparing genome sequences but also demonstrate that genome data have some intrinsic manifold structures, providing a new geometric perspective for molecular biology studies.

INTRODUCTION

The genome space is the entire set of genomes of all living organisms, which can be used to gain a comprehensive understanding of the dynamic genomic evolutionary process.[1] Mathematically, the genome space can be regarded as a moduli space of genomes, where each point in the space represents a unique genome, and the distance between two points corresponds to the biological distance between those corresponding genomes.[2] Therefore, in biology, it is an important issue to understand the geometry of genome space.[3]

The study of genome space geometry is still at an early stage. One well-established approach is using principal component analysis (PCA) methods based on Euclidean distance to establish the mapping between the distributions of certain species such as humans and their genomes. These methods usually apply PCA to downscale the genome-type data of loci in genome sequences to two-dimensional indices and then visualize the relationship between the distribution of the genome in Euclidean space with the actual geospatial distribution of some species.[4-7] Other studies have employed multidimensional scaling (MDS) methods to visualize the relationship among the genomes of different species, such as humans,[8] fruits,[9] bacteria,[10] vertebrates,[11] and horses.[12] These studies demonstrate that the landscape of various species can be well presented at the genetic level by combining distance matrices derived from genomic data with downscaling visualization methods. In addition, since the SARS-CoV-2 pandemic, numerous studies have focused on visualizing and analyzing the landscape of coronaviruses using similar methods for both complete genome sequences[13] and genomic fragments, such as the spike protein coding sequence.[14-16]

A good understanding of the geometry of the genome space enables scientists to make genome comparisons with biological significance by measuring the distance between points in the genome space. Genome comparison is an indispensable task of bioinformatics, allowing researchers to study evolutionary relationships among various species, and construct phylogenetic trees by comparing sequence similarity, which is also an important research topic in computational biology. Genome comparison also plays a vital role in identifying related genes, inferring gene functions, and tracing the origin of genes.[17]

The traditional genome comparison method is referred to as "sequence alignment," which employs a specific mathematical model or algorithm to identify the maximum number of matching bases or residues between two or more sequences. When aligning more than two sequences, this process is known as multiple sequence alignment (MSA). The most commonly used MSA methods include CLUSTALW,[18] MAFFT,[19] and MUSCLE.[20] In the realm of phylogenetic reconstruction, state-of-the-art methods, such as maximum likelihood or Bayesian inference, are primarily MSA-based approaches. These methods generally yield reliable results when the sequences under study are closely related and can be reliably aligned. Recently, there has been considerable attention from researchers toward applying deep learning to phylogenetic reconstruction. According to the review,[21] three main research paths have emerged: a quartet-based method that starts from inferring simple quartet topology and then amalgamating them[22]; a distance-based method that infers sequence distance matrices through neural networks[23]; and the construction of phylogenetic trees directly from the results of MSA using generative adversarial networks.[24] However, it is crucial to note that despite these advancements, most work in this field relies on the results of MSA, which means that phylogenetic reconstruction of large-scale sequences is still not free from the expensive MSAs that are usually time consuming, memory intensive, and impractical for large-scale genome sequences. Therefore, it has become an active research area to develop efficient and accurate alignment-free methods to compare whole genome sequences, given the exponential growth in genome sequences facilitated by modern sequencing technologies.

An alignment-free alternative involves transforming molecular sequences into objects suitable for analysis using established mathematical tools in linear algebra, probability statistics, and information theory, etc. These methods can then redefine and calculate the similarity or distance between sequences. In contrast to traditional approaches relying on MSA, alignment-free methods can efficiently extract information to facilitate rapid computation.[25] There are numerous alignment-free methods for sequence analysis, leveraging various approaches such as word frequencies, the length of matching words, informational content between sequences, chaos game representation, and graphical representation of DNA sequences.[26] For instance, Blaisdell[27] employed the k-mer model based on the classic string representation for genome sequence comparison. Qi et al.[28] introduced composition vector tree (CVTree) using a composition vector approach. Li et al.[29] proposed a normalized compression distance (NCD) based on information theory. Kantorovitz[30] utilized k-mer counts for comparing regulatory sequences. Sims et al.[31] utilized feature (or k-mer) frequency profiles (FFP) of whole genomes for genome comparison. Deng et al.[32] developed the alignment-free natural vector method. For a comprehensive overview of alignment-free methods, we refer to the review paper.[26]

The chaos game representation (CGR) is a highly effective tool for transforming genome sequences into CGR frequency matrices.[33] It holds significant utility across various bioinformatics domains, including alignment-free sequence comparison, phylogenetic analysis, and as an encoding method for machine learning tasks.[34] Several alignment-free methods for sequence comparison and phylogenetic analysis leverage CGR.[35-41] For instance, Hatje and Kollmar[38] used the Euclidean distance and the Pearson distance between the CGR frequency matrices to produce an alignment-free method (we refer to it as FCGR). Pei et al.[41] introduced the extended natural vector combining the CGR method. For a thorough introduction to CGR and its applications in bioinformatics, we refer to the review paper.[34] Previous studies have also utilized CGR in combination with other methods to reconstruct coronavirus phylogeny, as evidenced by Sengupta et al.[42] and Paul et al.[43] However, it is important to acknowledge that these

studies focused solely on a limited subset of members within the Orthocoronavirinae subfamily. Consequently, they cannot offer a comprehensive understanding of the entire genome landscape.

Since CGR transforms a genome sequence into a matrix, our goal is to extract information and analyze features of the original sequence from this matrix. To achieve this, we focus specifically on the column space of each FCGR, considering the pairing between purines and pyrimidines. In mathematics, the moduli space of subspaces—the Grassmann manifold—serves as a fundamental geometric object that parameterizes the set of all linear subspaces of a certain dimension in a fixed ambient vector space. The geometric properties of the Grassmann manifold are well understood mathematically, and many quantities, such as the geodesic distance between two points in a Grassmann manifold, can be efficiently computed. These characteristics render the Grassmann manifold an appropriate tool with the potential to be applied to our genome comparison problem. For a more detailed introduction to the Grassmann manifold, please refer to the supplemental information and the references therein.

In this paper, we propose a novel approach for studying the geometry of the genome space by considering a genome sequence as a point in a suitable Grassmann manifold based on the column space of the corresponding FCGR. To explore the intrinsic geometry of the genome space with the proposed method, we define the distance between genome sequences as the generalized Grassmannian distance[44] between their corresponding points in the genome space. Specifically, when the points are in the same Grassmann manifold, the distance is the usual geodesic distance, whereas, when they are in different Grassmann manifolds, it is the distance between a point in one of the Grassmann manifolds and a specific Schubert variety within the Grassmannian.[44] With this distance, we constructed phylogenetic trees for various datasets, which numerically demonstrates that our method is efficient and accurate in comparison to both MSA and alignment-free methods. In addition, we employed MDS to visualize the genome sequences of all members of the Orthocoronavirinae subfamily, along with the genome sequences of the major SARS-CoV-2 variants and the spike protein coding sequences. This visualization was represented in three-dimensional Euclidean space, presented as a point cloud. Notably, the visualization process did not use any clustering technique, but instead, it directly utilized the generalized Grassmann distance matrix as an input to MDS. The results of this visualization offer a novel and more intuitive perspective for understanding the relationships within the Orthocoronavirinae subfamily and the relationships between the Omicron variant of SARS-CoV-2 and other variants.

RESULTS

Evaluation of FCGR-SD as an alignment-free method

To evaluate the effectiveness of FCGR-SD as an alignment-free sequence comparison method, we conducted assessments using various datasets and compared it against competing methods. Since the choice of the parameter n (see Equation 2.1 in the supplemental information) in FCGR-SD directly depends on the sequence length, we constructed two datasets with different orders of magnitude in sequence length. The first dataset consisted of 30 influenza A virus segment 4 hemagglutinin (HA) gene sequences with an average length of 1,718.8, while the second dataset contained 44 coronavirus complete genome sequences with an average length of 29,759. For each dataset, we assess the performance of six methods, namely FCGR-SD, CVTree, NCD, FFP, FCGR, and MAFFT. We evaluated their performance by establishing phylogenetic relationships, comparing the topological structures of phylogenetic trees, and measuring the running time of each method.

To numerically evaluate the performance of different methods and identify satisfactory (though not necessarily optimal) parameters in our approach, we calculated the Robinson–Foulds (RF) distance between the phylogenetic tree constructed by each method and the reference tree constructed by MAFFT (see section benchmark for comparison: the multiple sequence alignment based on fast fourier transform in the supplemental information for the reason of choosing MAFFT as the benchmark). The RF distance quantifies the topological disagreement between the inferred method and the reference trees, where a small value indicates close similarity in tree topology, while a large value suggests limited overlap in bipartitions between the two trees.[45] We employed the R package *TreeDist*[46] to compute the RF distance.

Influenza A virus HA gene. The first dataset consists of 30 linear cRNA sequences of the HA gene of influenza A viruses from six different subtypes, namely H1N1, H2N2, H3N2, H5N1, H7N3, and H7N9. For each subtype, we collected five sequences from the NCBI nucleotide database. The lengths of these sequences range from 1,695 to 1,773, with an average length of 1,718.8. Therefore, we set $n = 6 \approx \log_4(1{,}718.8)$, which corresponds to 64 × 64 FCGR. We chose the optimal value of the parameter p as 0.71, which leads to the lowest RF distance to the reference tree with our method.

The phylogenetic trees are generated by the methods MAFFT (combining the maximum likelihood (ML) method), FCGR- SD, CVTree, FFP, NCD, and FCGR, shown in Figure 1, which implies that all the methods accurately classify the influenza A viruses into six subtypes. However, our method correctly identifies the major cluster H3-H7, as well as two secondary clusters, H7N3-H7N9 and H2N2-H5N1, consistent with the reference method MAFFT. In contrast, no competing alignment-free methods perform similarly.

In order to assess our method more precisely, we calculated the RF distance between the resulting trees of those five alignment-free methods and the reference tree provided by MAFFT. The results summarized in Table 1 demonstrate that our method's phylogenetic tree exhibits the smallest RF distance compared to the remaining four alignment-free methods. This suggests that the tree topology of our method in this dataset is the closest among five alignment-free methods to that of the alignment-based method MAFFT.

SARS-CoV-2 complete genome. To determine the optimal truncating parameter for our study on coronavirus, we downloaded all genome sequences from the Global Initiative on Sharing All Influenza Data (GISAID) and randomly selected 44 sequences from 8 of the major variants (Alpha, Beta, Gamma, Delta, Omicron, Lambda, Mu, and GH/490R), so that each group contains 4–6 sequences that are relatively evenly distributed geographically. This yielded the dataset consisting of 44 genome sequences with the length ranging from 29,378 to 29,866 with an average length of 29,759. We then set $n = 8 \approx \log_4(29{,}759)$, which corresponds to 256 × 256 FCGR, and we chose the optimal value of the parameter $p = 0.75$ to obtain the lowest RF distance with our method.

Figure 2 displays the phylogenetic trees generated by those six methods considered in the previous subsection, which demonstrates that all methods accurately divide the SARS- CoV-2 genomes into eight clusters, with Omicron being a single major cluster. This indicates that Omicron is markedly distinct from other variants of SARS-CoV-2, which is consistent with the conclusions given by Xia et al.[47] The evolutionary tree produced by our proposed method effectively identifies the closer relationship between Alpha and Gamma, as well as Mu, GH/490R, Delta, and Lambda, but it fails to recognize the relationship between Beta and other variants. In contrast, the competing methods perform even worse in discerning the phylogenetic relationships between these clusters. As shown in Table 1, the phylogenetic tree generated by our method again exhibits the smallest RF distance to MAFFT compared to the remaining four alignment-free methods in this dataset.

Time statistics. To compare the computational efficiency of different methods, we conducted all calculations on the same computer (Intel i5-12500H with 8G + 32G DDR5 4800 MHz RAM) and cleared the memory before each calculation. Table 2 summarizes the time costs of FCGR-SD, CVTree, NCD, FFP, FCGR, and MAFFT to obtain pairwise distance matrix. As shown in Table 2, due to the complexity of the singular value decomposition, the computational efficiency of FCGR- SD is slightly lower than that of FFP. However, it outperforms all the rest of the methods in terms of computing speed. At the same time, FCGR-SD is able to produce results closest to the phylogenetic tree constructed by MAFFT.

Toward survival manifold

In evaluating the performance of FCGR-SD, we observed that similar organisms have comparable truncation dimensions in the first-stage singular value decomposition. As the sequence length affects the truncation dimension, it was necessary to construct a dataset with similar sequence lengths and the same values for the parameter n to investigate whether closely related organisms tend to occupy similar truncating dimensions.

Therefore, we selected various viral sequences of the Orthocoronavirinae subfamily as our dataset. We downloaded 2,669 genome sequences from Orthocoronavirinae subfamily except for SARS-CoV-2 from the NCBI database, and then, we added 614 SARS-CoV-2 genome sequences from GISAID database, resulting in a dataset of 3,283 sequences with the length ranging from 26,592 to 31,526, all corresponding to $n = 8$.

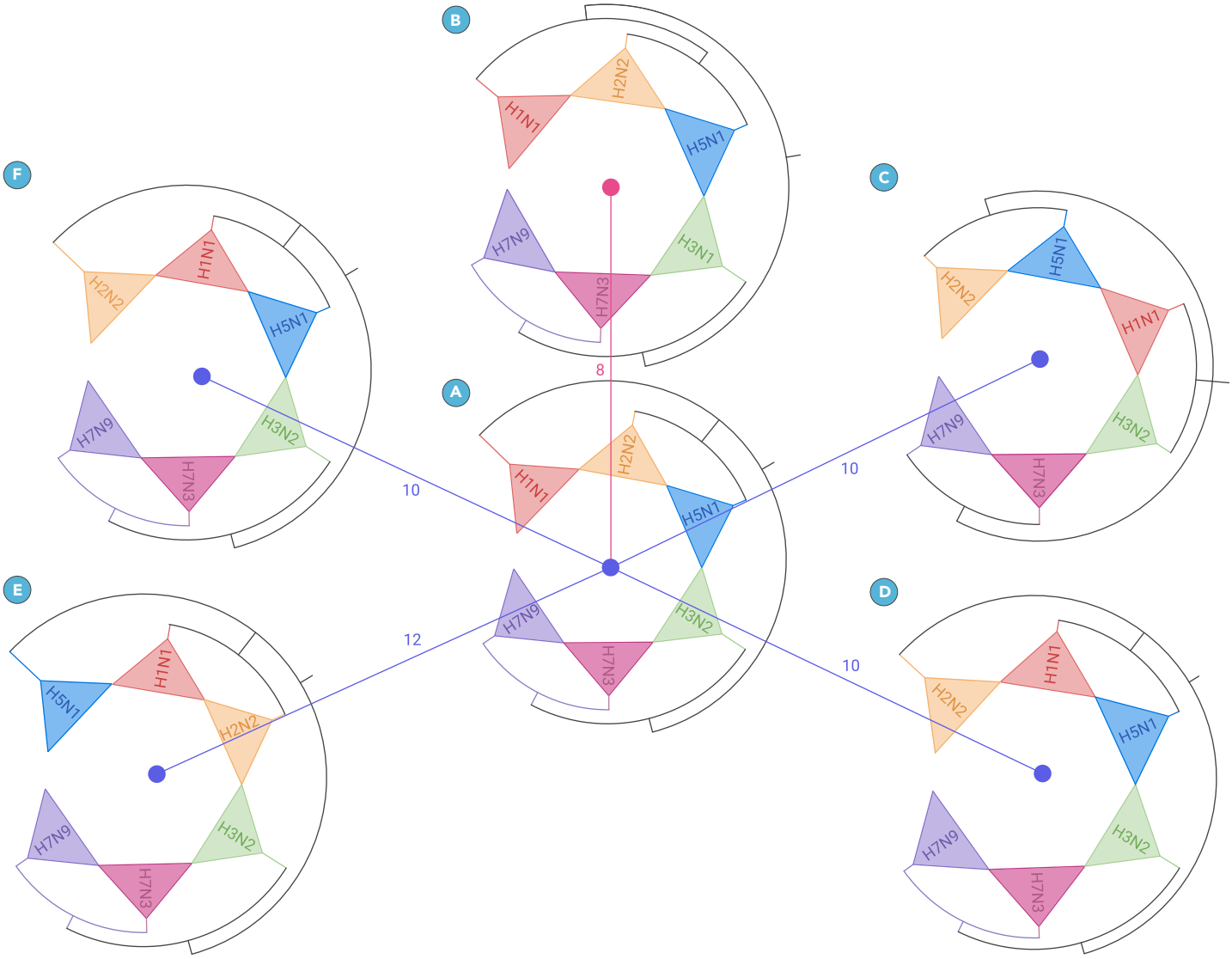


Figure 1. Phylogenetic trees of 30 HA gene cRNA sequences of influenza A virus Each subtype is represented by a specific color. The methods employed for tree construction are as follows: (A) MAFFT alignment + maximum likelihood tree building; (B) FCGR-SD with parameters $k = 6$, $p = 0.71$; (C) CVTree; (D) FFP; (E) NCD; and (F) FCGR. The numbers displayed indicate the RF distance.

We computed the truncated dimension of the first-stage singular value decomposition for each organism in the dataset. As illustrated in Figure 3, the truncation dimension exhibits diversity among virus types.

Firstly, we observe that the distribution of truncation dimensions is relatively concentrated for most viruses when the truncation parameter p is specified (e.g., SARS-CoV-2 is entirely concentrated at dimension 42). The most loosely distributed viruses in the figure are the "D5 | Unc-DeltaCoVs," which are Deltacoronaviruses not yet categorized at the subgenus level in the NCBI database. This suggests that these unclassified sequences likely belong to more than one subgenus.

Secondly, different sequence types of the same species may share similar truncation dimensions. For instance, bovine coronavirus, human coronavirus OC43, and porcine hemagglutinating encephalomyelitis virus, all classified as betacoronavirus 1, exhibit similar truncating dimensions.

Table 1. RF distance between MAFFT reference tree and resulting trees

Dataset	FCGR-SD	CVTree	FFP	NCD	FCGR
30 INF	8	10	10	12	10
44 SARS2	42	48	48	56	52

Furthermore, we note that this diversity in truncating dimensions reflects the variability in the degree of concentration of singular values in the FCGR. For the same truncation parameter p, sequences with lower truncation dimensions tend to have a more concentrated distribution of singular values, suggesting that their theoretically "optimal truncation dimensions" will be lower and vice versa. Consequently, assuming we can determine the optimal truncation dimension for each sequence, this dimension also delineates the Grassmann manifold that best fits the sequence, which we refer to as the "survival manifold" of the sequence.

This introduces challenges in measuring distances between sequences, as employing intrinsic geodesic distances directly within the same Grassmann manifold may not be applicable. With the introduction of survival manifolds, it becomes necessary to consider how distances are defined between manifolds of different dimensions. We address this by enabling the measurement of distances between different survival manifolds through the application of geometric results on cross-dimensional generalizations of geodesic distances.

A geometric dive into SARS-CoV-2 genome sequences

We evaluated the accuracy and efficiency of our method for assessing sequence similarity on larger datasets by conducting nearest-neighbor classification on two extensive datasets. From the dataset composed of 3,283

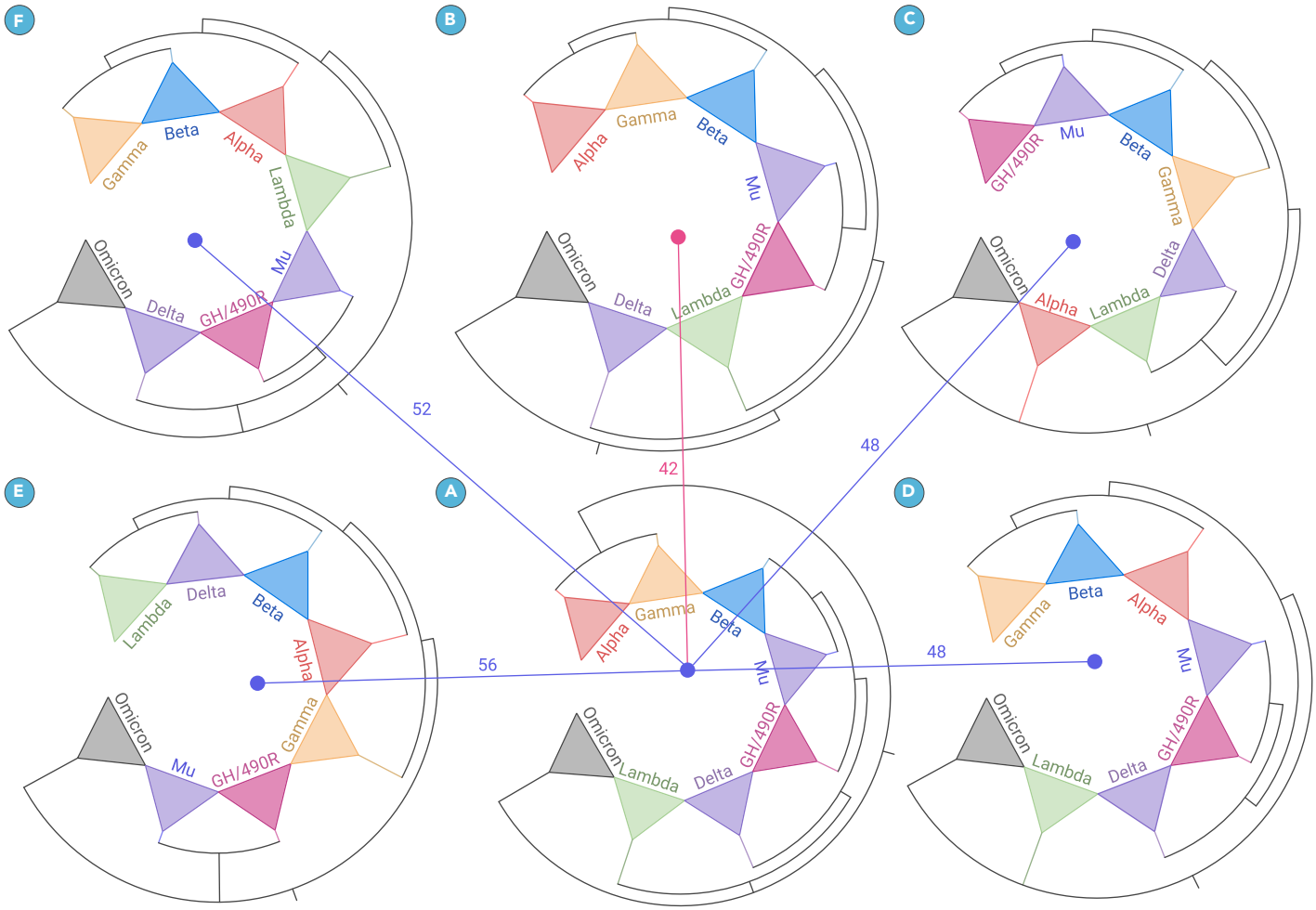


Figure 2. Phylogenetic trees of 44 genome sequences of SARS-CoV-2 Each variant is represented by a specific color. The methods employed for tree construction are as follows: (A) MAFFT alignment + maximum likelihood tree building; (B) FCGR-SD with parameters $k = 8$, $p = 0.75$; (C) CVTree; (D) FFP; (E) NCD; and (F) FCGR. The numbers displayed indicate the RF distance.

Orthocoronavirinae sequences, we filtered out 28 sequences missing subgenus tags, left with 3,255 sequences distributed across 4 genera and 18 subgenera. We calculated a distance matrix of size 3,255 × 3,255 using FCGR-SD and then evaluated nearest-neighbor classification accuracy using genus and subgenus classifications as labels, respectively. For each sequence, we identified the nearest sequence label and verified whether they matched. We achieved classification accuracies of 98.99% at the genus level and 98.53% at the subgenus level.

Moving to a finer level of classification, we focused on classifying the Pango lineage within SARS-CoV-2. Previous work in this area is documented in Ali et al.,[48] where the authors compared classification outcomes using a dataset of 7,000 amino acid sequences of SARS-CoV-2 spike proteins from 22 randomly selected variants from the GISAID database. State-of-the-art classification results ranged from 84% to 85% in the k-nearest-neighbor-based classification results. For comparison, we extracted sequences belonging to the 22 variants from a pool of 10,348 DNA sequences encoding the SARS-CoV-2 spike protein, yielding a dataset of 2,610 sequences containing 21 PANGOLIN variants (excluding AY.4, which is absent from our dataset). Employing this dataset, we performed nearest-neighbor classification using PANGOLIN variants as labels and achieved an accuracy of 85.13%.

Table 2. Time consumed of methods compared

Dataset	FCGR-SD	CVTree	NCD	FFP	FCGR	MAFFT+ML
30 INF	0.07 s	0.64 s	0.14 s	0.06 s	0.23 s	3 s + 8 min 16 s
44 SARS2	1.86 s	7.91 s	10.43 s	1.22 s	5.83 s	16 min + 18 h 12 min

These findings demonstrate that by utilizing the FCGR-SD method, we can achieve relatively high nearest-neighbor classification accuracies ranging from the genus level to the variant level, indicating the validity of embedding coronavirus sequences into Grassmann manifolds.

To visualize the distribution of the Orthocoronavirinae dataset and different variants of SARS-CoV-2 on Grassmann manifolds, we require a distance matrix-based visualization method such as MDS, t-SNE, LLE, UMAP, or ISOMAP. A number of previous studies have attempted to visualize the SARS-CoV-2 genome sequence or the S protein coding sequence within two-dimensional Euclidean space, primarily utilizing methods such as t-SNE.[14-16,49,50] However, our goal is to preserve the original manifold structure as accurately as possible in a low-dimensional space, and t-SNE or another popular method may lead to distortions in the original geometry. Since the distance matrix inherently contains the structure information, we use a non-metric MDS method that conserves the distance matrix.

MDS[51] is a suitable method for visualizing data distribution since it constructs a low-dimensional space using the similarity between sample pairs to ensure that the distances and similarities in the high-dimensional space are consistent. The main difference between metric and non-metric MDS lies in whether the input matrix is Euclidean. We employ the non-metric MDS method in our work to allow for non-Euclidean distances. Our visualization approach is similar to ISOMAP[52] but is more intrinsic, because we use the generalized Grassmannian distance instead of constructing a neighborhood graph and using the geodesic distance (nearest distance between each pair of data point) on the graph.

We calculated distance matrices for 3,283 coronavirus sequences, 10,348 SARS-CoV-2 genome sequences, and 10,348 spike protein coding sequences

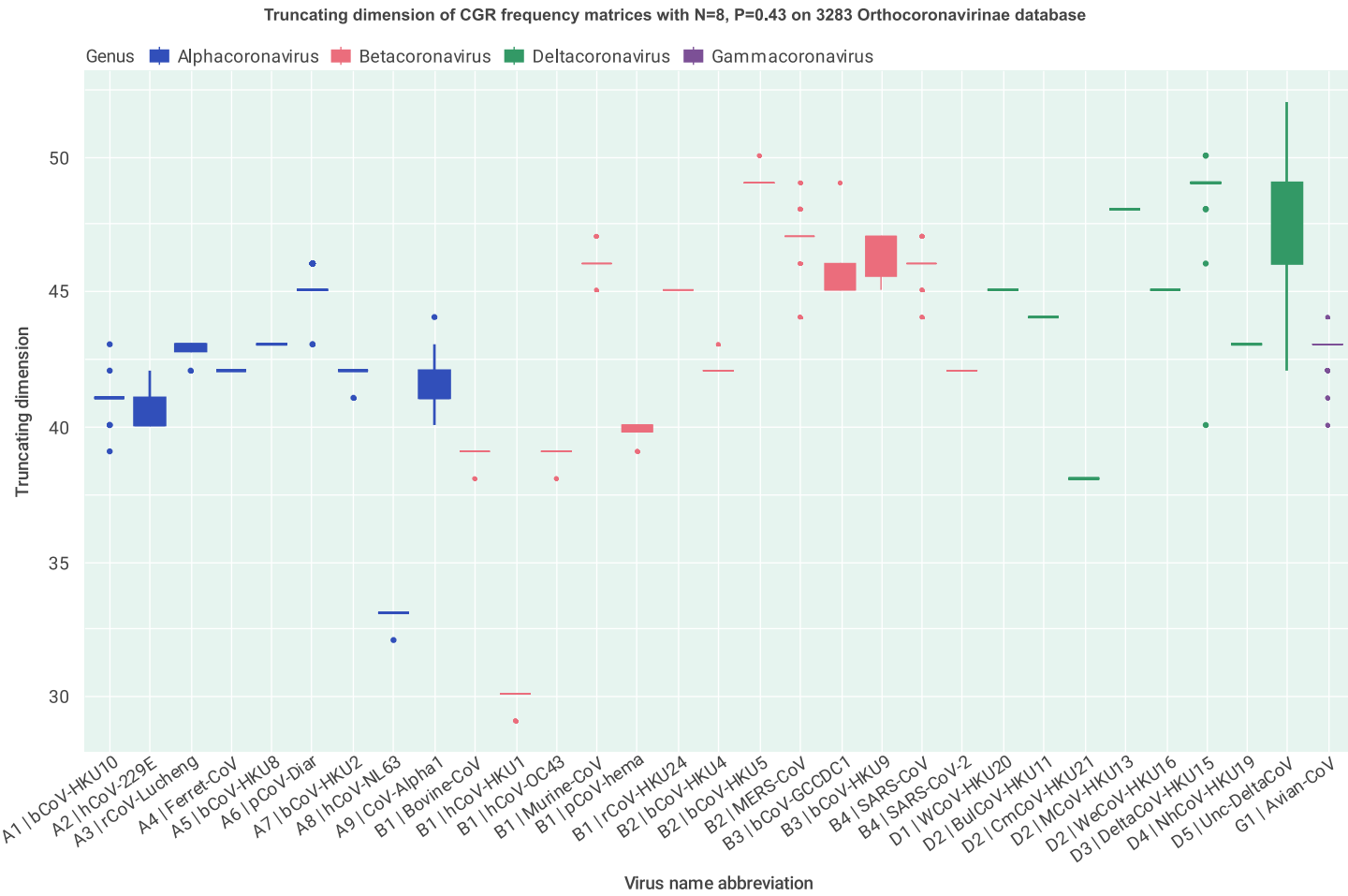


Figure 3. Boxplot of truncated dimensions of 3,283 Orthocoronavirinae sequences under optimal parameter selection (p = 0.43) Labels on the x axis follow the format "$x|y$," where the y part is an abbreviation of the virus name. In the x part, the initial letter denotes the genus of the virus (A for alphacoronavirus, B for betacoronavirus, D for deltacoronavirus, and G for gammacoronavirus); the number following the initial letter indicates different subgenera (e.g., A1 represents the Decacovirus subgenus under alphacoronavirus).

of SARS-CoV-2, respectively. We then used MDS for dimensionality reduction visualization, and the results are shown in Figures 4 and 5.

In Figure 4, we observed that coronaviruses are distributed on an approximately spherical surface with different genera occupying different local areas. SARS-CoV-2 is located on an island-shaped area surrounded by SARS, which resembles Antarctica and the 60° south latitude circle. However, the "island" of SARS-CoV-2 is divided into two areas: one part consists of Omicron BA.2, BA.4, and BA.5, while the other part comprises other mutant strains. In Figure 4B, we noticed that within the Omicron variant, BA.2, BA.4, and BA.5 form one cluster, while BA.1 and other variants form another. This finding suggests that BA.1 is closer to other types of variants than the rest of Omicron variants, which may be characterized in practice by lower infectivity, etc. The results in Figure 5A are consistent with the previous coronavirus visualization results, showing that BA.1 is closer to other types of mutant strains than BA.2, BA.4, and BA.5. Finally, in Figure 5B, we observed that the spike protein coding sequences of the Omicron variant are more disparate than those of the remaining variants, which is consistent with the conclusions given by Simon-Loriere and Schwartz.[53]

From the visualization results, we concluded that Omicron BA.1 already showed characteristics closer to the non-Omicron variants than the other Omicron variants. This is in agreement with previous virologists that some divergence has occurred within Omicron and that the later variants such as BA.4 and BA.5 have shown significant genomic-level changes compared to BA.1.[54,55] In addition, based on the visualization of the genome of the SARS-CoV-2 and the S protein coding sequence, we further concluded that the differentiation within Omicron is not significantly reflected at the S protein level. Therefore, it implies that the differentiation differences within Omicron may be caused by a combination of individual genes rather than by differences in S proteins alone.

DISCUSSION

In this paper, we propose a novel approach on exploring the geometry of genome space by considering a genome sequence as a point in an appropriate Grassmann manifold via the column space of the corresponding frequency matrices of CGR. With this geometric explanation, we propose a new alignment-free genome comparison method, named FCGR-SD, which uses the generalized Grassmannian distance as a geometric distance measure.

Compared to traditional MSA methods, FCGR-SD occupies several advantages. Neither similar lengths of genome sequences nor normalization of the frequency matrix are necessary with FCGR-SD. Moreover, the proposed method has a lower computational complexity since it only uses singular value decomposition. Numerical results based on several datasets have also verified the superior performance of the proposed method FCGR-SD in large genome data analysis.

Genomes are considered as points on different Grassmann manifolds based on the molecular structure similarity between bases with the proposed method. The phylogenetic tree constructed with the proposed method has a more accurate topological structure, indicating that genetic data usually have some intrinsic manifold structures with biological significance. This topic is worth further in-depth study in genome sequence analysis. In our future work, we will explore more detailed and larger-scale studies of the survival manifold phenomena to gain deeper insights into the underlying biological mechanisms.

MATERIALS AND METHODS

See the supplemental information for details.

REFERENCES

1. Bellgard, M.I., Itoh, T., Watanabe, H., et al. (1999). Dynamic evolution of genomes and the concept of genome space. Ann. N. Y. Acad. Sci. ***870***(1): 293–300. https://doi.org/10.1111/j.1749-6632.1999.tb08891.x.

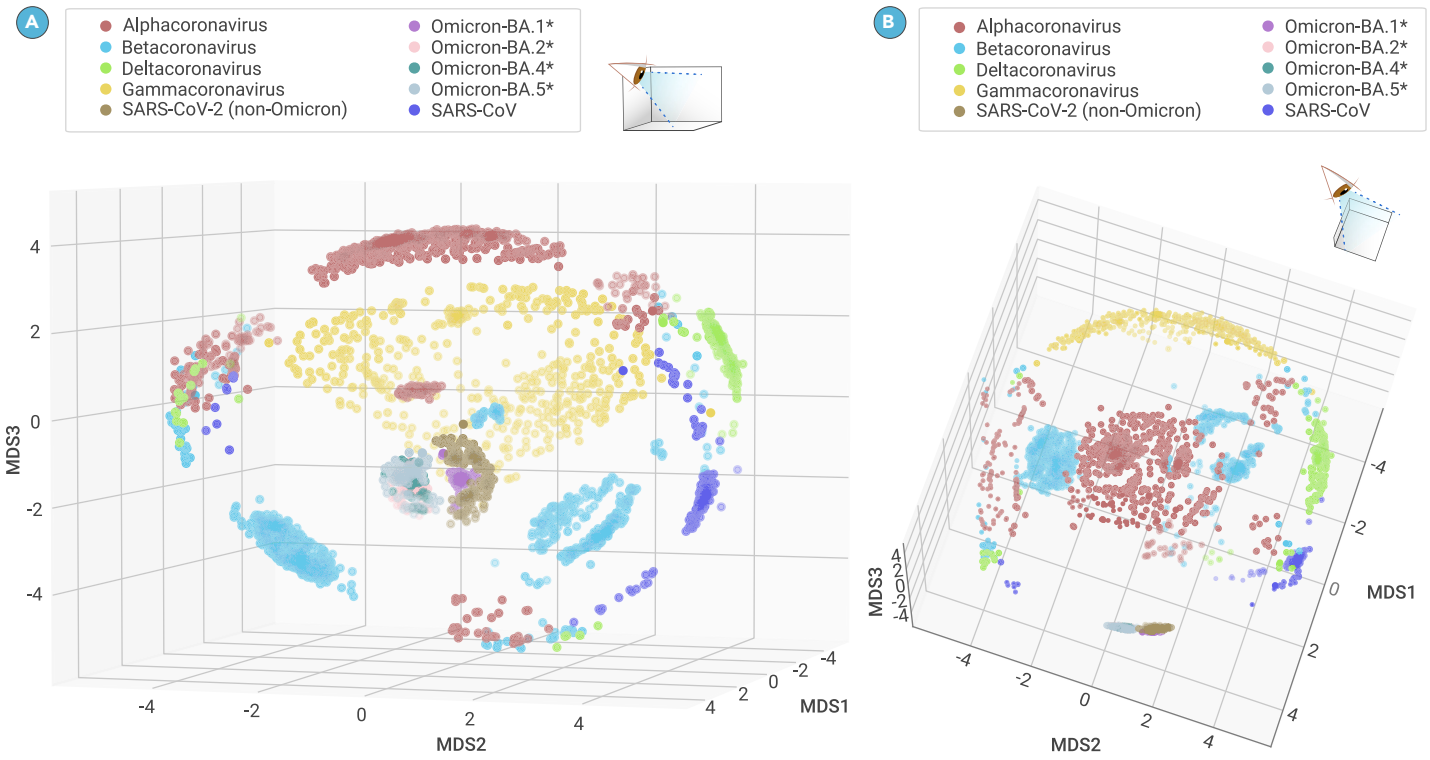


Figure 4. 3D visualization of Orthocoronavirinae dataset (A) 3D visualization of 3,283 coronavirus sequences. (B) 3D visualization of 3,283 coronavirus sequences, focusing on the area constituted by SARS-CoV-2.

2. Yu, C., Liang, Q., Yin, C., et al. (2010). A novel construction of genome space with biological geometry. DNA Res. ***17***(3): 155–168. https://doi.org/10.1093/dnares/dsq008.
3. Vukmirovic, O.G., and Tilghman, S.M. (2000). Exploring genome space. Nature ***405***: 820–822. https://doi.org/10.1038/35015690.
4. Novembre, J., Johnson, T., Bryc, K., et al. (2008). Genes mirror geography within Europe. Nature ***456***(7218): 98–101. https://doi.org/10.1038/nature07331.
5. Kim, K., Baik, H., Jang, C.S., et al. (2019). Genomic GPS: using genetic dis- tance from individuals to public data for genomic analysis without disclosing personal genomes. Genome Biol. ***20***(175): 175. https://doi.org/10.1186/s13059-019-1792-2.
6. Reich, D., Price, A.L., and Patterson, N. (2008). Principal component analysis of genetic data. Nat. Genet. ***40***: 491–492. https://doi.org/10.1038/ng0508-491.
7. Gilbert, E., Shanmugam, A., and Cavalleri, G.L. (2022). Revealing the recent demographic history of Europe viahaplotype sharing in the UK Biobank. Proc. Natl. Acad. Sci. USA ***119***(7): e2119281119. https://doi.org/10.1073/pnas.2119281119.
8. Malaspinas, A.S., Tange, O., Moreno-Mayar, J.V., et al. (2014). bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). Bioinformatics ***30***(20): 2962–2964. https://doi.org/10.1093/bioinformatics/btu410.
9. Biscarini, F., Nazzicari, N., Bink, M., et al. (2017). Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies. BMC Genom. ***18***(1): 432. https://doi.org/10.1186/s12864-017-3781-8.
10. Tsai, M.H., Liu, Y.Y., and Chen, C.C. (2019). OutbreakFinder: a visualization tool for rapid detection of bacterial strain clusters based on optimized multi- dimensional scaling. PeerJ ***7***(7): e7600. https://doi.org/10.7717/peerj.7600.
11. Kari, L., Hill, K.A., Sayem, A.S., et al. (2015). Mapping the space of genomic signatures. PLoS One ***10***(5): e0119815. https://doi.org/10.1371/jour- nal.pone.0119815.
12. McCue, M.E., Bannasch, D.L., Petersen, J.L., et al. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: utility for association map- ping, genetic diversity, and phylogeny studies. PLoS Genet. ***8***(1): e1002451. https://doi.org/10.1371/journal.pgen.1002451.
13. Hie, B., Zhong, E.D., Berger, B., et al. (2021). Learning the language of viral evolution and escape. Science ***371***(6526): 284–288. https://doi.org/10.1126/sci- ence.abd7331.
14. Ali, S., Ali, T.E., Khan, M.A., et al. (2022). Effective and scalable clustering of SARS-CoV-2 sequences. In Proceedings of the 5th International Conference on Big Data Research (ICBDR '21), pp. 42–49. https://doi.org/10.1145/3505745.3505752.
15. Taslim, M., Prakash, C., Ali, S., et al. (2023). Hashing2Vec: Fast Embedding Generation for SARS-CoV-2 Spike Sequence Classification. In Proceedings of the 14th Asian Conference on Machine Learning, p. 189.
16. Ali, S., and Patterson, M. (2021). Spike2Vec: An Efficient and Scal- able Embedding Approach for COVID-19 Spike Sequences. In IEEE in- Ternational Conference on Big Data, pp. 1533–1540. https://doi.org/10.1109/Big- Data52589.2021.9671848.
17. Semple, C., and Steel, M. (2003). Phylogenetics **24** (Oxford University Press on Demand).
18. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. ***22***(22): 4673–4680. https://doi.org/10.1093/nar/22.22.4673.
19. Katoh, K., Misawa, K., Kuma, K.i., et al. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. ***30***(14): 3059–3066. https://doi.org/10.1093/nar/gkf436.
20. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high ac- curacy and high throughput. Nucleic Acids Res. ***32***(5): 1792–1797. https://doi.org/10.1093/nar/gkh340.
21. Mo, Y.K., Hahn, M.W., Smith, M.L., et al. (2024). Applications of machine learning in phylogenetics. Mol. Phylogenet. Evol. ***196***: 108066. https://doi.org/10.1016/j.ympev.2024.108066.
22. Zou, Z., Zhang, H., Guan, Y., et al. (2020). Deep Residual Neural Networks Resolve Quartet Molecular Phylogenies. Mol. Biol. Evol. ***37***(5): 1495–1507. https://doi.org/10.1093/molbev/msz307.
23. Nesterenko, L., Boussau, B., and Jacob, L. (2022). Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.24.496975.
24. Smith, M.L., and Hahn, M.W. (2023). Phylogenetic inference using generative adversarial networks. Bioinformatics ***39***(9): btad543. https://doi.org/10.1093/bioin formatics/btad543.
25. Vinga, S., and Almeida, J. (2003). Alignment-free sequence comparison-a review. Bioinformatics ***19***(4): 513–523. https://doi.org/10.1093/bioinformatics/btg005.
26. Zielezinski, A., Vinga, S., Almeida, J., et al. (2017). Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. ***18***(1): 186. https://doi.org/10.1186/s13059-017-1319-7.
27. Blaisdell, B.E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. USA ***83***(14): 5155–5159. https://doi.org/10.1073/pnas.83.14.5155.
28. Qi, J., Luo, H., and Hao, B. (2004). CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. ***32***(Web Server issue): W45–W47. https://doi.org/10.1093/nar/gkh362.
29. Li, M., Chen, X., Li, X., et al. (2004). The similarity metric. IEEE Trans. Inf. Theor. ***50***(12): 3250–3264. https://doi.org/10.1109/TIT.2004.838101.
30. Kantorovitz, M.R., Robinson, G.E., and Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics ***23***(13): i249–i255. https://doi.org/10.1093/bioinformatics/btm211.
31. Sims, G.E., Jun, S.R., Wu, G.A., et al. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. USA ***106***: 2677–2682. https://doi.org/10.1073/pnas.0813249106.

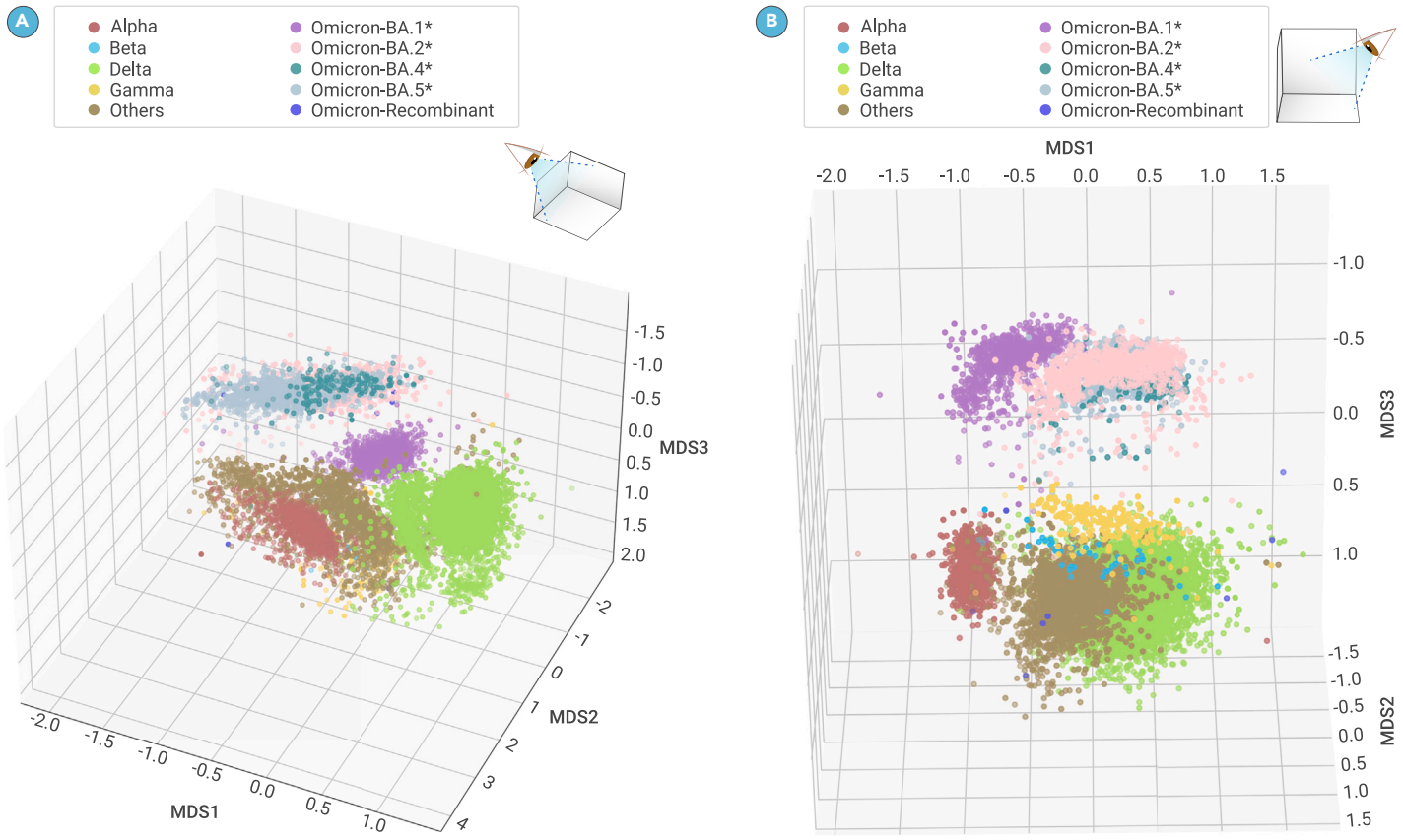


Figure 5. 3D visualization of SARS-CoV-2 genome sequences and the spike protein coding sequences (A) Three-dimensional visualization of 10,348 SARS-CoV-2 genome sequences. (B) Three-dimensional visualization of 10,348 spike protein coding sequences of SARS-CoV-2.

32. Deng, M., Yu, C., Liang, Q., et al. (2011). A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One ***6***(3): e17293. https://doi.org/10.1371/journal.pone.0017293.
33. Jeffrey, H.J. (1990). Chaos game representation of gene structure. Nucleic Acids Res. ***18***(8): 2163–2170. https://doi.org/10.1093/nar/18.8.2163.
34. Löchel, H.F., and Heider, D. (2021). Chaos game representation and its applications in bioinformatics. Comput. Struct. Biotechnol. J. ***19***: 6263–6271. https://doi.org/10.1016/j.csbj.2021.11.008.
35. Deschavanne, P.J., Giron, A., Vilain, J., et al. (1999). Genomic signature: characterization and classification of species assessed by chaos game represen- tation of sequences. Mol. Biol. Evol. ***16***(10): 1391–1399. https://doi.org/10.1093/oxford- journals.molbev.a026048.
36. Almeida, J.S., Carriço, J.A., Maretzek, A., et al. (2001). Analysis of genomic se- quences by Chaos Game Representation. Bioinformatics ***17***(5): 429–437. https://doi.org/10.1093/bioinformatics/17.5.429.
37. Joseph, J., and Sasikumar, R. (2006). Chaos game representation for comparison of whole genomes. BMC Bioinf. ***7***: 243. https://doi.org/10.1186/1471-2105- 7-243.
38. Hatje, K., and Kollmar, M. (2012). A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. Front. Plant Sci. ***3***: 192. https://doi.org/10.3389/fpls.2012.00192.
39. Hoang, T., Yin, C., and Yau, S.S.T. (2016). Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. Genomics ***108***(3-4): 134–142. https://doi.org/10.1016/j.ygeno.2016.08.002.
40. Lichtblau, D. (2019). Alignment-free genomic sequence comparison us- ing FCGR and signal processing. BMC Bioinf. ***20***(1): 742. https://doi.org/10.1186/s12859-019-3330-3.
41. Pei, S., Dong, W., Chen, X., et al. (2019). Fast and accurate genome com- parison using genome images: The Extended Natural Vector Method. Mol. Phylogenet. Evol. ***141***: 106633. https://doi.org/10.1016/j.ympev.2019.106633.
42. Sengupta, D.C., Hill, M.D., Benton, K.R., et al. (2020). Similarity Studies of Corona Viruses through Chaos Game Representation. Comput. Mol. Biosci. ***10***(3): 61–72. https://doi.org/10.4236/cmb.2020.103004.
43. Paul, T., Vainio, S., and Roning, J. (2022). Detection of intra-family coronavirus genome sequences through graphical representation and artificial neural net- work. Expert Syst. Appl. ***194***: 116559. https://doi.org/10.1016/j.eswa.2022.116559.
44. Ye, K., and Lim, L.H. (2016). Schubert Varieties and Distances between Sub- spaces of Different Dimensions. SIAM J. Matrix Anal. Appl. ***37***: 1176–1197. https://doi.org/10.1137/15M1054201.
45. Zielezinski, A., Girgis, H.Z., Bernard, G., et al. (2019). Benchmarking of alignment-free sequence comparison methods. Genome Biol. ***20***: 144. https://doi.org/10.1186/s13059-019-1755-7.
46. Smith, M.R. (2020). TreeDist: Distances between Phylogenetic Trees. R Pack- Age Version 2.4.1 (Comprehensive R Archive Network). https://doi.org/10.5281/zen- odo.3528124.
47. Xia, B., Wang, Y., Pan, X., et al. (2022). Why is the SARS-CoV-2 Omicron variant milder? Innovation ***3***(4): 100251. https://doi.org/10.1016/j.xinn.2022.100251.
48. Ali, S., Sahoo, B., Khan, M.A., et al. (2023). Efficient Approximate Ker- nel Based Spike Sequence Classification. IEEE ACM Trans. Comput. Biol. Bioinf ***20***(6): 3376–3388. https://doi.org/10.1109/TCBB.2022.3206284.
49. Ali, S., Sahoo, B., Zelikovsky, A., et al. (2023). Benchmarking machine learn- ing robustness in Covid-19 genome sequence classification. Sci. Rep. ***13***(1): 4154. https://doi.org/10.1038/s41598-023-31368-3.
50. Zvyagin, M., Brace, A., Hippe, K., et al. (2023). GenSLMs: Genome-scale lan- guage models reveal SARS-CoV-2 evolutionary dynamics. Int. J. High Perform. Comput. Appl. ***37***(6): 683–705. https://doi.org/10.1177/10943420231201154.
51. Borg, I., and Groenen, P.J.F. (2005). Modern Multidimensional Scaling: Theory And Applications, 2nd ed. (Springer Science + Business Media).
52. Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric frame- work for nonlinear dimensionality reduction. Science ***290***(5500): 2319–2323. https://doi.org/10.1126/science.290.5500.2319.
53. Simon-Loriere, E., and Schwartz, O. (2022). Towards SARS-CoV-2 serotypes? Nat. Rev. Microbiol. ***20***(4): 187–188. https://doi.org/10.1038/s41579-022-00708-x.
54. Rössler, A., Netzl, A., Knabl, L., et al. (2022). BA.2 and BA.5 omicron differ immunologically from both BA.1 omicron and pre-omicron variants. Nat. Commun. ***13***(1): 7701. https://doi.org/10.1038/s41467-022-35312-3.
55. Mykytyn, A.Z., Rissmann, M., Kok, A., et al. (2022). Antigenic cartography of SARS-CoV-2 reveals that Omicron BA.1 and BA.2 are antigenically distinct. Sci. Immunol. ***7***(75): eabq4450. https://doi.org/10.1126/sciimmunol.abq4450.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (12171275 and 12371270), the Shanghai Science and Technology Development Funds (23JC1402100), and the Tsinghua University Education Foundation fund (042202008). The authors would like to thank the anonymous referees for their detailed comments and valuable suggestions, which significantly improved the quality of the paper.

AUTHOR CONTRIBUTIONS

X.F., S.S.-T.Y., and S.-T.Y. designed research; X.L. and T.Z. performed research, analyzed data, and wrote the paper with input from X.F., S.S.-T.Y., and S.-T.Y. X.L. and T.Z. contributed equally to this work.

DECLARATION OF INTERESTS

The authors declare no competing interest.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.1016/j.xinn.2024.100677.

LEAD CONTACT WEBSITE

https://math.tsinghua.edu.cn/info/1125/2065.htm.
https://ddgene.github.io/.