Visualizations of Combinatorial Entropy Index on Whole SARS-CoV-2 Genomes

Yang Zhou, Jeffrey Zheng

Abstract In this paper, a set of SARS-CoV-2 genomes from four countries are selected for visualizations under the C1 modules of the metagenomic analysis system MAS. Based on the variant construction and the theory of information entropy, the module makes statistics on the number of bases in SARS-CoV-2 sequences to calculate the base probability measures in segments to generate the combinatorial entropy index data from the base probability measures. Under visualization technology, the combinatorial entropy index is projected on 2D clustering genomic index maps and 1D histogram maps to provide projection results. The visual results provide intuitive and easy properties to analyze complicated clustering among genomes to support clustering analysis of SARS-CoV-2 genomes in batches, showing the distribution characteristics of SARS-CoV-2 genomes in different countries or regions conveniently.
Keyword metagenomic analysis aystem MAS, combinatorial entropy, variant construction, genomic index, geometric visualization

Yang Zhou[1] e-mail: zhouyang_se@163.com Jeffrey Zheng[1,2,3] e-mail: conjugatelogic@yahoo.com
[1]Yunnan University, Kunming
[2]Key Laboratory of Software Engineering of Yunnan
[3]Key Laboratory of Quantum Information of Yunnan,

This work was supported by the NSFC (62041213), the Key Project on Electric Information and Next Generation IT Technology of Yunnan(2018ZI002).

Introduction

The COVID-19 outbreak is showing a global trend [1]. Researchers have launched joint research on SARS-CoV-2 worldwide. At present, with the increase in SARS-CoV-2 sequences submitted by various countries, the analysis of SARS-CoV-2 sequences is one of the focuses. There are a large number of gene sequences sampled from thousands of COVID-19 patients worldwide. As of April 30, 2020, there are 15819 SARS-CoV-2 sequences counted by CNCB[2], which are respectively from 266 regions. For these sequences, we need to perform a preliminary analysis. The metagenomic analysis system MAS [3, 4, 5] is an available and effective method that focuses on exploring general information from collections of whole genomic sequences intrinsically included in virus RNA genomes on SARS-CoV-2 samples. This paper adopts the C1 module in the MAS. This module uses the theory of information entropy and the variant construction [6, 7] to generate a genomic index, which can be used to show the population differences of SARS-CoV-2 sequences between different regions.

According to the concept of information entropy in information theory, the more orderly a system is, the lower its information entropy is. The more chaotic a system is, the higher its information entropy is. Information entropy is a measure of the degree of system ordering, so it can describe the randomness of gene sequences to a certain extent. RNA is a kind of natural sequence that has natural randomness, so information entropy can be the most important criterion [8]. Based on the theory of information entropy, combinatorial entropy is used in the MAS to support the cluster analysis of the SARS-CoV-2 genome in batches. Combinatorial entropy can quickly, stably and conveniently describe the distribution characteristics of RNA sequences [9], select different bases as the elements of calculating combinatorial entropy, and index RNA sequences. Although the occurrence of a phenomenon is accidental, through the same conditions, a large number of experiments, an obvious non-accidental law can appear. Therefore, through experiments on a large number of sequences, we can conclude the distribution law of sequence combinatorial entropy under genes. By analyzing the distribution law of combinatorial entropy, we can obtain the correlation between some features of the SARS-CoV-2 sequences and the distribution of combinatorial entropy. Based on this idea, 2D clustering genomic index maps and 1D histogram maps are used to describe the relationship between the regional distribution of SARS-CoV-2 sequences and combinatorial entropy.

Aim of The Study

This paper is devoted to exploring information from the whole gene sequence of the RNA genome of the SARS-CoV-2 sample, quantitatively measuring the whole gene sequence [10], and representing the gene sequence as a genome index. The genome index is mapped to the geometric measurement area to facilitate visualization of the collection of genomes. The diagram contains two distribution modes, 2D clustering genomic index maps and 1D histogram maps. This module can solve the classification and content-based indexing problems of large quantities of gene sequences, and can provide a solution for managing gene sequences globally.

Materials and Methods

A total of 381 sequences selected in this paper are from the GISAID [11]. The collection location for these gene sequences included the following four countries: China, the United States, Australia, and Belgium. To ensure the accuracy of the information entropy calculation, the selected sequences are of high quality and high coverage and do not contain missing bases.

For a given sequence, its total length is N, which is divided into $M = N/m$ segments according to the segment length m. Within each segment, the frequency c_i of each type of base in A, T, C

and G can be calculated. Dividing the frequency c_i by the segment length m, the frequency p_i of the base in the segment can be obtained.The p_i calculated from each segment can calculate the entropy of a base in the corresponding sequence according to the following formula.

$$entropy = -\sum_{i=0}^{M-1} p_i \log p_i$$

For the four meta bases A, T, C and G, we select two of them and calculate their entropy separately. Then, each sequence can obtain two entropies, which are projected on the 2D plane as a binary group. After projecting several sequences, their distribution can be observed on the 2D plane. For the 2D distribution of multiple sequences, its projection in the horizontal direction and vertical direction can be drawn. With the number of sequences within a certain range of entropy as the Y value, the number distribution of multiple sequences in a certain direction can be observed.

In this paper, the segment length $m = 16$ is selected, and different base combinations are tried. The measurement results are divided into five parts, namely, the overall projection of four countries and the decomposition projection of four countries.

Result

The results are presented in five parts, with the overall projection of four countries as the first part and the remaining four parts as the projections of four countries. For each part of the projection, 2D clustering genomic index maps are included, along with two sets of 1D histogram maps.

Four Countries

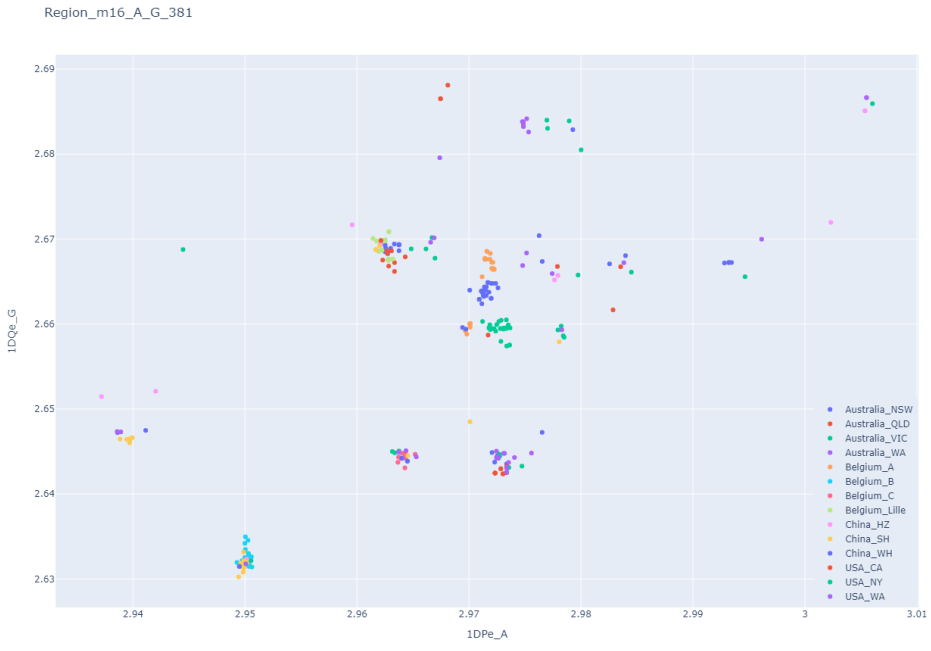

Fig. 1 SARS-CoV-2 of 2D genomic indices on combinatorial entropy maps in four countries

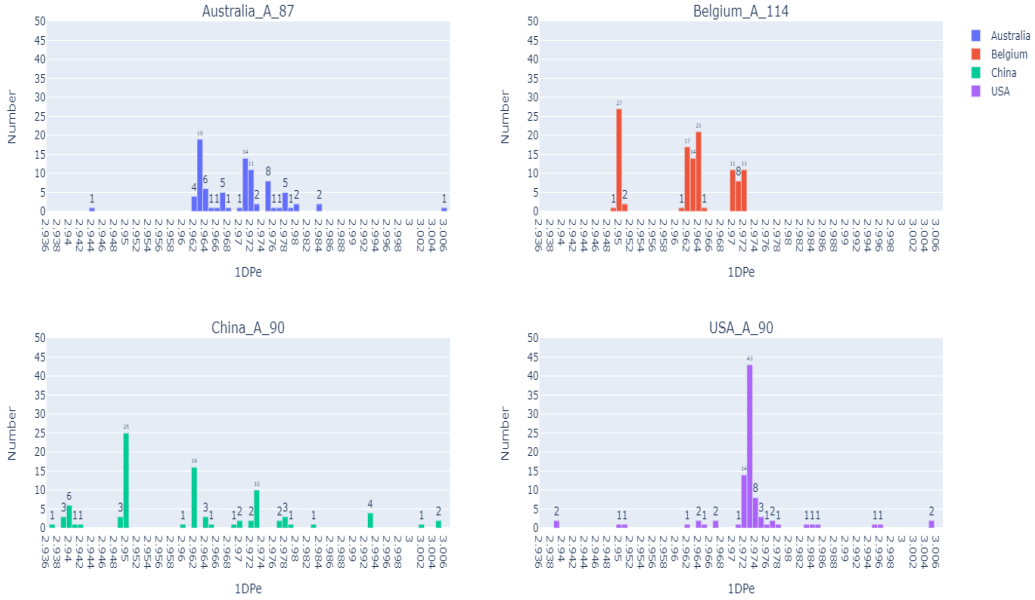

Fig. 2 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in four countries:A

Fig. 3 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in four countries:T

China

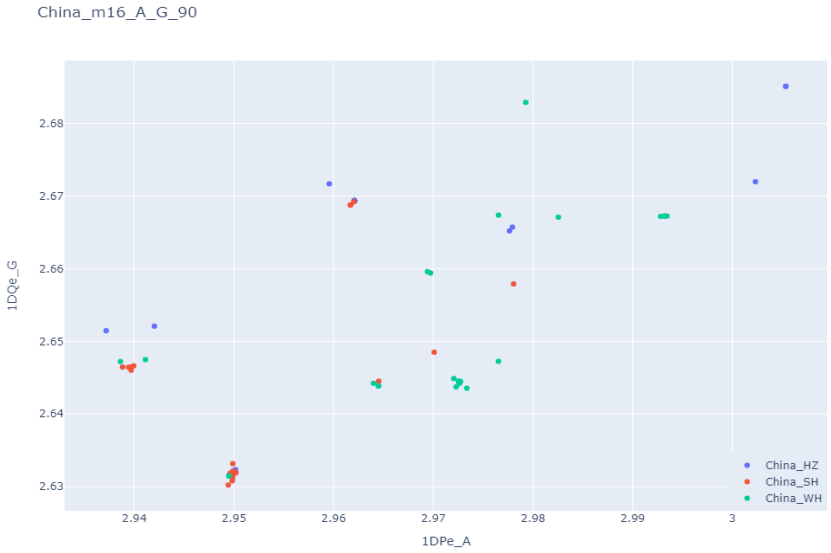

Fig. 4 SARS-CoV-2 of 2D genomic indices on combinatorial entropy maps in China

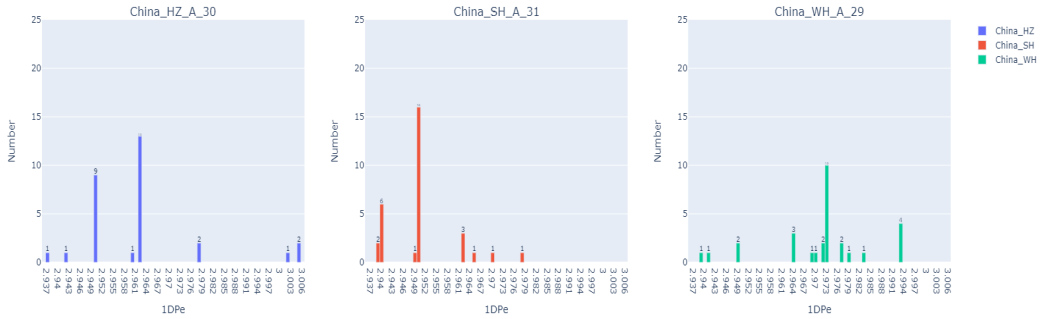

Fig. 5 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in China:A

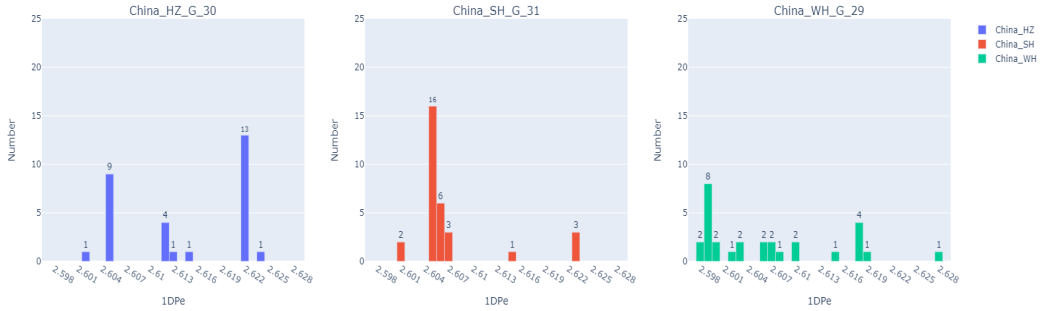

Fig. 6 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in China:G

Australia

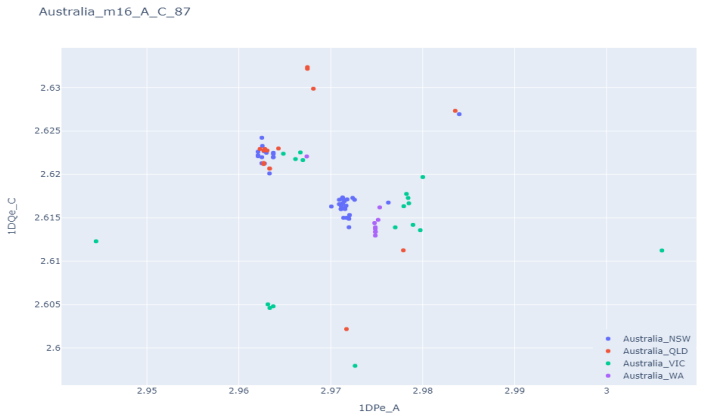

Fig. 7 SARS-CoV-2 of 2D genomic indices on combinatorial entropy maps in Australia

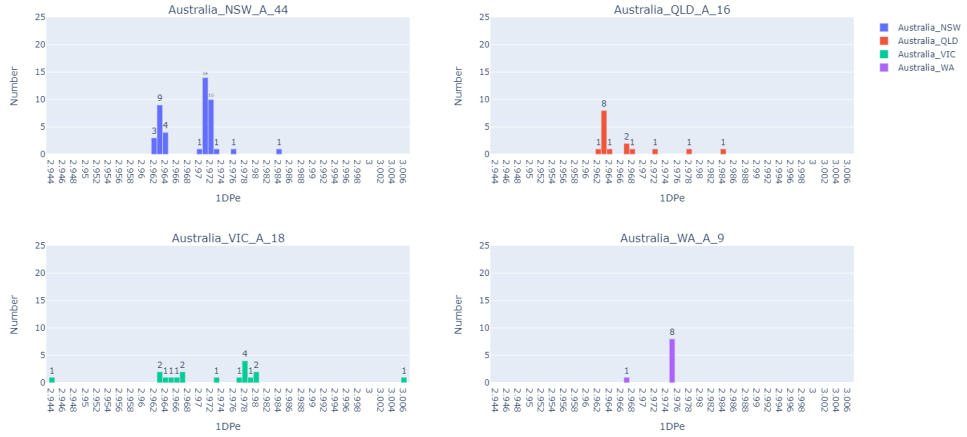

Fig. 8 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in Australia:A

Fig. 9 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in Australia:T

USA

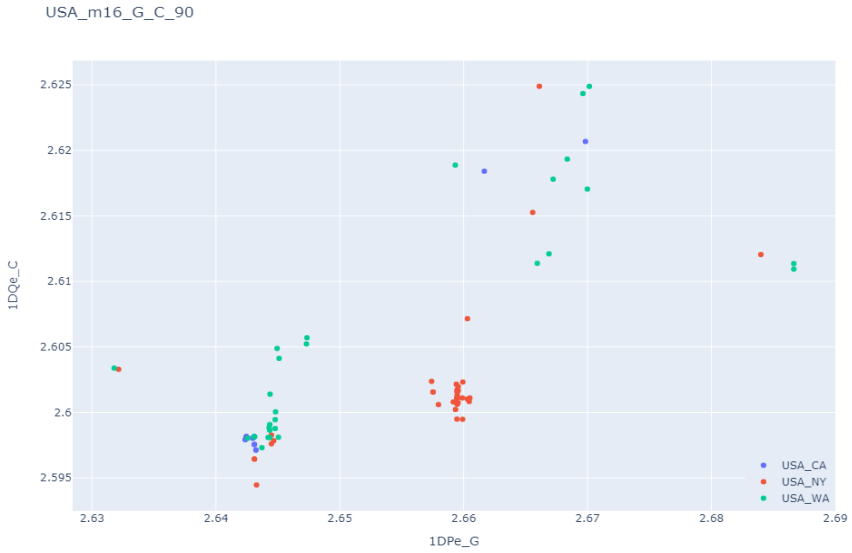

Fig. 10 SARS-CoV-2 of 2D genomic indices on combinatorial entropy maps in USA

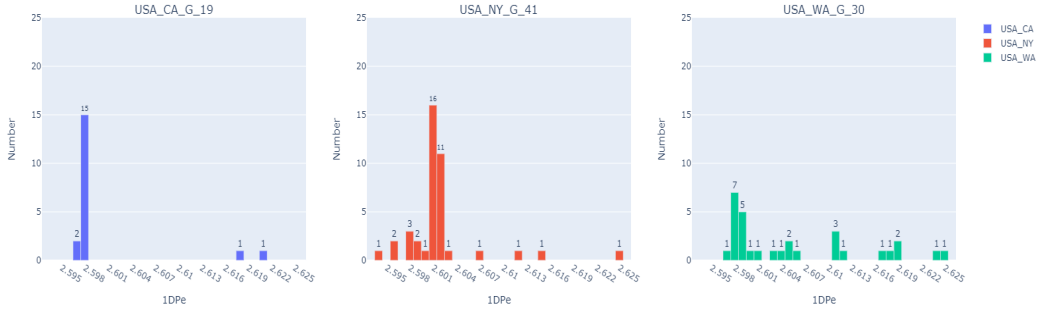

Fig. 11 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in USA:C

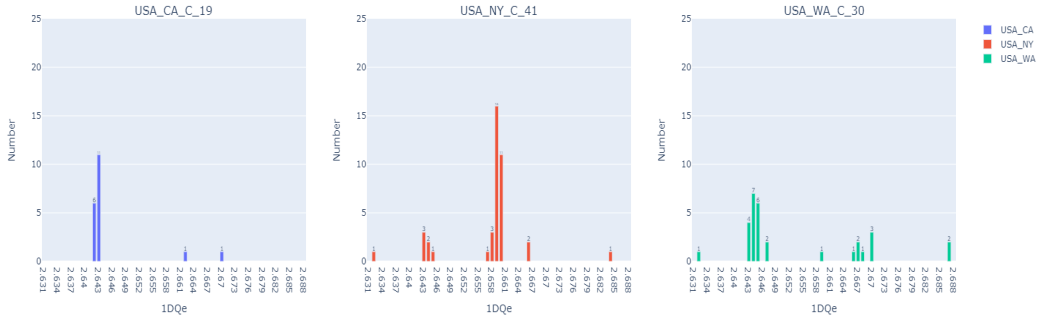

Fig. 12 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in USA:G

Belgium

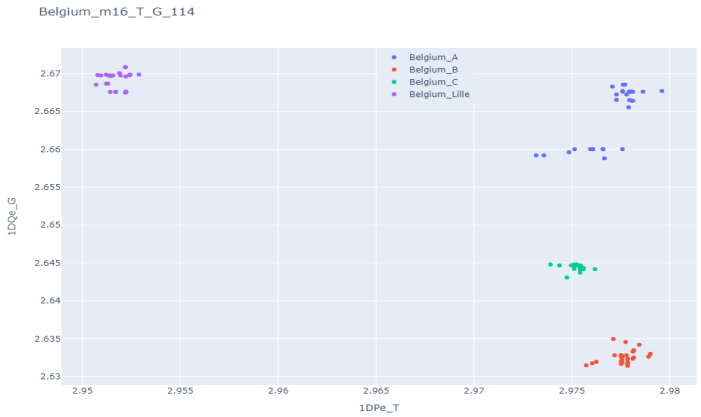

Fig. 13 SARS-CoV-2 of genomic indices on combinatorial entropy maps in Belgium

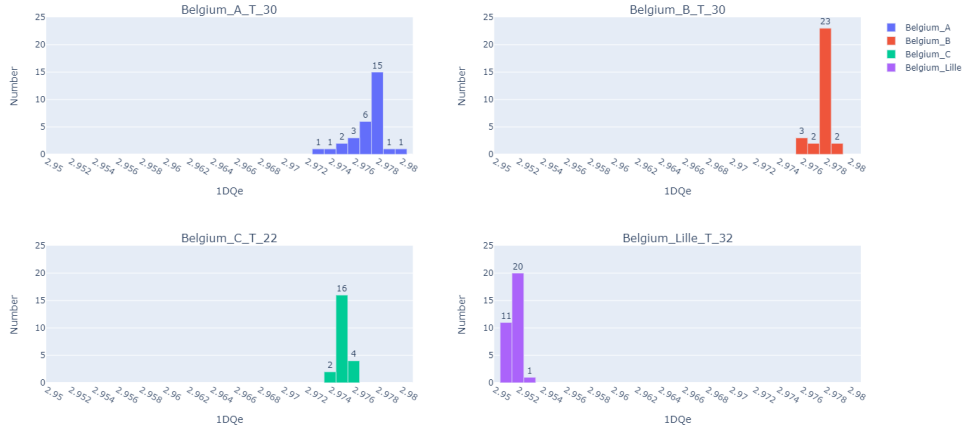

Fig. 14 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in Belgium:T

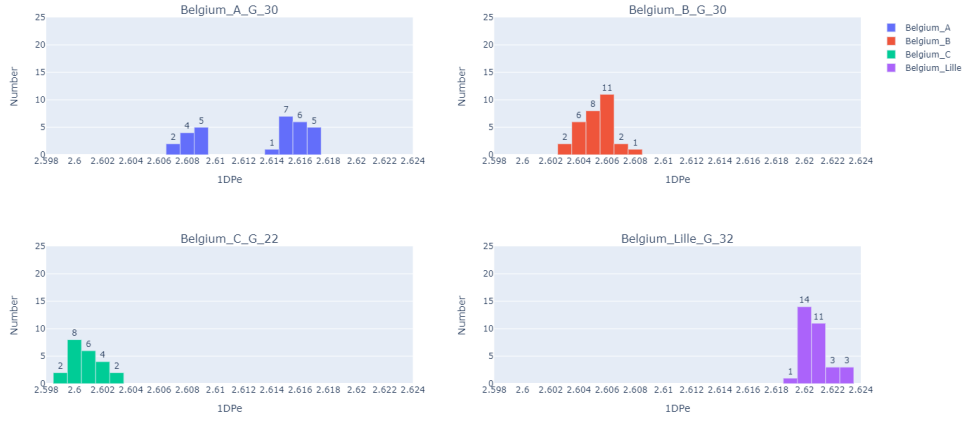

Fig. 15 SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in Belgium:C

Discussions

The 2D scatter plots in the above charts represent the distribution of the genomic index composed of base A and base G for a given number of SARS-CoV-2 sequences. Fig. 1 shows the distribution of 381 SARS-CoV-2 sequences in four countries. It can be seen from the figure that some gene sequences will be clustered at certain locations. These clustered points show that they are associated with the location of sequence acquisition. At the same time, there are also some regional sequences that do not meet the clustering situation. To observe the distribution more clearly, we projected the 2D scatter plot and counted the number of sequences near a certain coordinate value according to the region. Fig. 2 and Fig. 3 are the projection statistics of Fig. 1. From Fig. 2 and Fig. 3, we can see that the variation in genomic index distribution in different countries, and most sequences between countries will be approximately 1 to 3 peaks.

Fig. 4 to Fig. 15 show the distribution of the genome sequence index after dividing the regions of the four countries. From the graph, we can observe a unimodal distribution similar to New York, USA, and a bimodal distribution similar to New South Wales, Australia. Most of the regions meet the above two distributions.

Conclusion

The SARS-CoV-2 genome index visualization module in this paper can describe the distribution characteristics of whole genomes stably and rapidly. In 2D clustering genomic index maps, the clustering effect of SARS-CoV-2 genomes can be observed intuitively. In 1D histogram maps, the distribution characteristics of SARS-CoV-2 genomes in different countries or regions are shown conveniently. This module can select different base combinations and parameters to form rich projection combinations. This paper only shows some projection combinations, and interested readers can try other projection combinations. This projection method can also be used to observe the distribution pattern of other virus gene sequences.

Conflict Interest

No conflict of interest has claimed.

Acknowledgements The authors would like to thank GISAID,CNCB for providing invaluable information on the newest dataset collections of SARS-CoV-2 genomes to support this project working smoothly.

References

1. Zheng J. SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat.IntJBiolSci2020;16(10):1678-1685.doi:10.7150/ijbs.45053.Availablefromhttp://www.ijbs.com/v16p1678.htm.
2. 国家生物信息中心: 2019新型冠状病毒信息库 https://bigd.big.ac.cn/ncov/release_genome.
3. FZ Song, Military Medical Science Press 2011 (Chinese)宋方洲，基因组学，军事医学科学出版社 2011.
4. R. Durbin, S. Eddy, K. Krogh, G. Mitchison, Biological Sequence Analysis, Cambridge Uni-versity Press 2010.
5. R. Durrentt, Probability Models for DNA Sequence Evolution, Springer 2008.
6. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications, SpringerNature 2019 https://www.springer.com/in/book/9789811322815.
7. Jeffrey Zheng, Chris Zheng, Biometrics and Knowledge Management Information Systems,Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Na-ture 2019, 193-202 https://link.springer.com/chapter/10.1007/978-981-13-2282-2_11.
8. Shenkin P S , Erman B , Mastrandrea L D . Information-theoretical entropy as a measure of sequence variability[J]. Proteins Structure Function & Genetics, 1991, 11(4):297-313.

9. Du J F , Wu Y J , Zhang Y X , et al. Large-scale information entropy analysis of important sites in mature and precursor miRNA sequences[J]. Science in China, 2009(08):81-89.
10. Paraskevis D, Kostaki EG, Magiorkinis G, et al. Full-genome evolutionary analysisof the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a re-sult of a recent recombination event. Infect Genet Evol. 2020 Apr;79:104212. doi:10.1016/j.meegid.2020.104212. Epub 2020 Jan 29. PMID: 32004758
11. GISAID: Open access to influenza virus data https://gisaid.org.