Cluster Analysis of Visual Differences on Pairs of SARS-CoV-2 Genomes

Minghan Zhu
Yunnan University
Jeffrey Zheng (conjugatelogic@yahoo.com)
Yunnan University

Research Article

Keywords: SARS-CoV-2, clustering, COVID-19, average entropy, combined entropy, integrated entropy, genetic variation

Posted Date: January 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-72027/v2

Cluster Analysis of Visual Differences on Pairs of SARS-CoV-2 Genomes

Minghan Zhu, Jeffrey Zheng

Abstract This paper represents sample results for the C2 and C3 modules of the MAS. Genomic index maps were generated for SARS-CoV-2 genomes from different samples worldwide, and eleven other viral genomes were selected for comparisons. Each 2D genomic index map acts as a pair of X-Y coordinate scatter points located in a specific geometric region restricted. Supported by the principle of entropy invariance and visual distributions, it is convenient to identify the variations of SARS-CoV-2 genomes in the global scope. In view of this powerful transformation structure, future exploration of the detailed systematic expansion of theory and medical applications for COVID-19 patients is required.
Keyword SARS-CoV-2, clustering, COVID-19, average entropy, combined entropy, integrated entropy, genetic variation

Minghan Zhu
School of Software, Yunnan University
e-mail: crystalj000@163.com

Jeffrey Zheng
Key Laboratory of Quantum Information of Yunnnan, School of Software, Yunnan University
e-mail: conjugatelogic@yahoo.com

Funding Supported by the NSFC (62041213), the Key Project of Quantum Communication Technology (2018ZI002)

Introduction

From the end of 2019 to the present, unexplained pneumonia has erupted around the world. Infected persons have developed fever, blood clotting symptoms, and whitening of the lungs, and severe cases have died. Currently, there are more than 3.1 million confirmed cases worldwide. The following picture is a drawing of the team of the key laboratory of machine perception and intelligence of the Ministry of Education, School of Information Science and Technology, Peking University in Fig. 1 (see vis.pku.edu.cn for details).

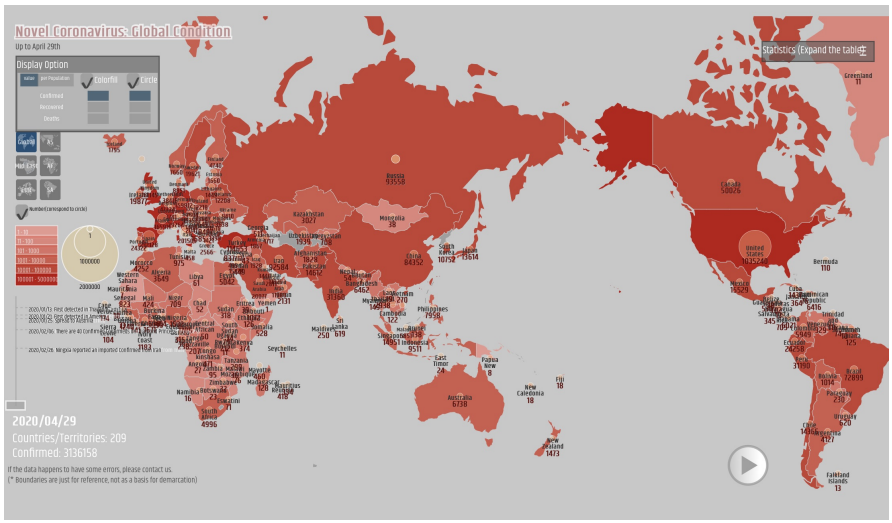


Fig. 1 Global outbreak analysis maps

After testing by the expert group, the analysis revealed that the unexplained pathogen was a new coronavirus. Coronaviruses are a large family of viruses, and new coronaviruses are new strains of coronavirus that have never been found in humans before. This discovery is a new coronavirus after the outbreak of SARS (severe acute respiratory syndrome) in 2003 and the first human discovery of MERS-CoV (Middle East respiratory syndrome coronavirus) in 2012. Subsequently, the International Committee for Classification of Viruses declared that the new coronavirus was named ”SARS-CoV-2” (Severe Acute Respiratory Syndrome Coronavirus 2). New coronavirus pneumonia is still in a stage of rapid change, with fever, fatigue, and dry cough as the main manifestations. It spreads rapidly and threatens people’s

lives and health. The new coronavirus is highly infectious and has a low lethality rate, but the possibility of genetic variation is not ruled out.

Genetic variation is a sudden, heritable variation of a genomic DNA molecule. At the molecular level, gene variation refers to the change in base pair composition or sequence in the structure of a gene. While genes are remarkably stable, replicating themselves precisely as cells divide, this stability is relative. Under certain conditions, a gene can suddenly change from its original existence form to a new existence form, that is, a new gene suddenly appears at a site to replace the original gene. In biology, it refers to changes in the genetic code of a cell (usually DNA in the nucleus). It includes point mutations caused by changes in a single base or the deletion, duplication, and insertion of multiple bases. The cause can be errors in the replication of genetic genes when cells divide or exposure to chemicals, radiation or viruses. Mutations often cause cells to malfunction or die and can even cause cancer in higher organisms. However, mutations are also seen as the "driving force" of evolution: undesirable mutations are weeded out by natural selection, and beneficial mutations accumulate. Neutral mutations accumulate without affecting the species, resulting in a discontinuous equilibrium.

A disordered world is impossible to produce life, and a living world is bound to be orderly. Biological evolution is from single cell to multicellular, from simple to complex, from low to high, that is to say, to a more orderly and precise direction, which is a direction of entropy subtraction, which is exactly the opposite of the direction of isolated system to increase entropy. However, life is a "dissipative structure", which is thought to be an open system far from the equilibrium state. By exchanging material and energy with the outside world, it may change from the original disordered state to a state ordered in time, space or function under certain conditions. The ordered structure is maintained by the constant dissipation of matter and energy. Through the continuous exchange of matter, energy and information with the outside world, the total entropy value of the life system can be reduced so that the order degree is constantly improved, and the life system can develop dynamically.

Based on vector logic, modern matrix theory, geometric measure theory, combinatorial algebra and discrete mathematics, variant construction starts from n 0-1 variables to form 2^n states and 2^{2^n} functions via vector permutation and complement operations on state space to establish a variant logic framework to contain $2^n! \times 2^{2^n}$ configurations as a variation space. Variant measurement acts as a core of quantitative measurement, starting from m 0-1 variables to explore relevant clustering conditions on 2^m states. Many sample applications have been developed for 40 years using variant construction [1]-[7], such as content-based image retrieval, medical image processing, bat echo identifications, DNA maps, hierarchical organization, phase space classification, feature extraction, filtering, combinations, projections and conjugate transformations [8].

In this paper, we analyze the sequence of genome sequencing, segments, shifts and combinations of viruses and quantify the four kinds of base arrangements to form a differentiated quantization matrix. Combined with the invariance principle

of entropy, the mean entropy and superposition entropy are calculated, and the clustering distribution is illustrated. Showing different gene sequence structures.

Data Sources

The genomic sequence data used in this article are downloaded from the open source databases NCBI (National Center for Biotechnology Information) and GISAID (Global Shared Influenza Data Initiative) [11]-[12]. The description of the data used is shown in Fig. 2:

Distribution Characteristics and Method Description

Download representative new coronavirus SARS-CoV-2 and various influenza viruses from NCBI and GISAID. First, the genome sequence was screened, cleaned and segmented. The segmented data were shifted according to its segmented length m, one base was moved each time, and a total of m times was moved. The number of A, T, C and G bases in the corresponding segments was calculated. Second, the calculated results of the same genome sequence were replaced and combined, and the numbers were counted according to the same counting information contained in different segments. At the same time, the probability distribution of four ATCG bases was calculated. Combined with the entropy calculation formula of the second law of thermodynamics, the superposition entropy and average entropy of different genome sequences of each virus were recorded. The superposition entropy was used as the X-axis and the average entropy as the Y-axis to form a quantitative matrix of $16 \times (m+1)^2$. Finally, combining the visualization analysis method in the variant logic system from the macroscopic Angle display genome sequence could mutate the different characteristics of base pairs, using the invariance principle of entropy, change the related parameters, the new coronavirus and different flu virus genomes for clustering analysis, distinguish looking for optimization of the clustering method, and discover SARS-CoV-2 as a possible source of development. For the specific formula derivation process, please refer to the paper [10]. The flow of the method used is shown in Fig. 3.

Development and Significance

After the SARS-CoV-2 genome sequence was publicly released, researchers at home and abroad used different analysis methods to guess the source host of the new coronavirus. A joint study found that the average genetic sequences of SARS-CoV-2, SARS and MERS viruses are more than 70% and 40% similar, respectively,

Samples	NO.	*Locality*
SARS-COV-2	$(2019-nCoV)$	
	$EPI-ISL-412978$	China
	$EPI-ISL-417310$	England
		Turkey
		South Africa
		Singapore
	$EPI-ISL-416521$	SaudiArabia
		Russia
		America
		Mexico
	$LC-528233$	Japan
	$EPI-ISL-412974$	Italy
	$MT-012098$	India
	$EPI-ISL-410720$	France
		Chile
	$EPI-ISL-413014$	Canada
	$EPI-ISL-413016$	Brazil
	$EPI-ISL-417426$	Belgium
	$EPI-ISL-407193$	South Korea
	$EPI-ISL-408977$	Australia
	$A3-EPI-ISL-406862$	Germany
Human Coronavirus		
	$NC-002645$	HCOV-229E
	$NC-006577$	HCOV-HKU1
	$NC-006213$	HCOV-OC43
	$NC-005831$	HCOV-NL63
Deadly Coronavirus		
	$AY-508724$	SARS
	$JX-869059$	MERS
	$NC-002549$	EBOLA
Animals Coronavirus		
	$KX-022602$	PDCOV
	$SL-CovZC45$	Bat
	$MT-084071$	Pangolin
Other Virus		
		H1N1
		H3N2

Fig. 2 Datasets of SARS-CoV-2 and other viruses

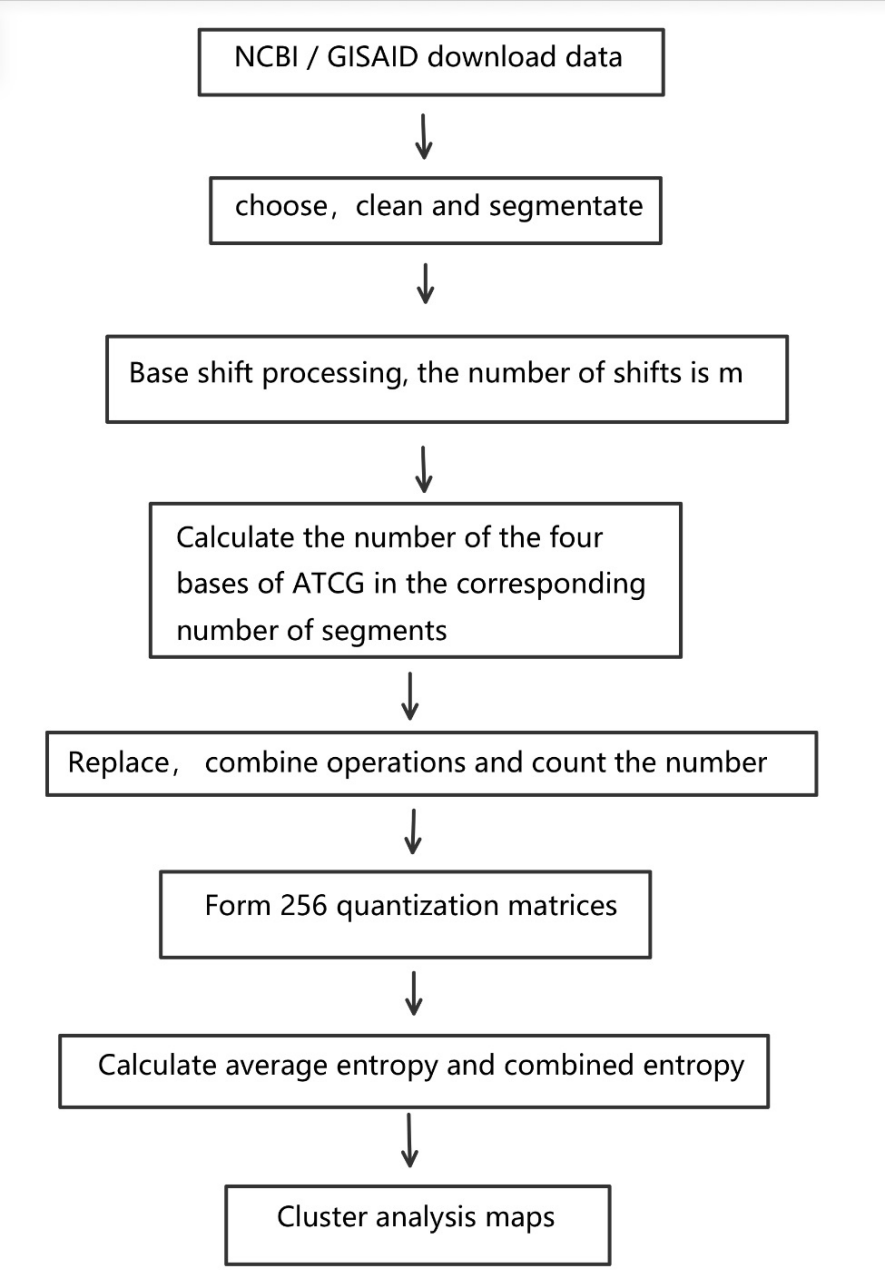


Fig. 3 The flow of the method maps

by the Chinese Academy of Sciences, the Academy of Military Medical Sciences, and the Chinese Academy of Sciences Biological Laboratory[13]. Peking University, GuangXi University of Traditional Chinese Medicine, Ningbo University and Wuhan Institute of Bioengineering have jointly tackled the problem and found that snakes are the most likely wild animal reservoirs that cause the current outbreak of SARS-COV-2 infections [14]. Shi Zhengli's team at the Wuhan Institute of Virology, Chinese Academy of Sciences, presented new challenges to the previous results, believing that bats are the most likely wild animals carrying SARS-CoV-2, with a full gene level of over 90% [15]. The University of Hong Kong and South China Agricultural University have analyzed more than 1,000 metagenomic samples and believe that pangolins may be potential intermediate hosts for SARS-CoV-2. Molecular biology tests revealed that the positive rate of pangolin beta coronavirus is 70%. For isolation and identification of the virus, the typical coronavirus particle structure was observed under an electron microscope. Through genome analysis of the virus, it was found that the sequence similarity of the isolated virus strain and the currently infected human strain is as high as 95% [16]. Xi ' an University of Electronic Technology and Peking University calculated and analyzed the SARS-CoV-2 virus sequence from 21 different countries and the similarity and evolutionary relationship between SARS-CoV-2 virus and its related viruses through 337 genome sequencing sequences. There was an average similarity of up to 99.8%. The analysis of their homologous evolution relationship using MEGA 6.0 shows that the SARS-CoV-2 from different countries has undergone a certain degree of variation [17]. The Xishuangbanna Institute of Botany and Cambridge University uses different clustering algorithms to combine gene base replacement and clustering algorithms to simulate the parallel evolution of viruses and divide the viruses into three types, A, B, and C [18]. The corresponding features of genomic information and sampling information are verified [19]. Nextstrain is an open-source project [20] to harness the scientific and public health potential of pathogen genome data. Applying powerful analytic and visualization tools provides a continually updated view of publicly available data for the community [9]. The series is shown in Fig. 4.

SARS-CoV-2, which is erupting worldwide, is bringing huge disasters to people all over the world, and understanding its source is of great significance for the development of treatment and prevention of future epidemics. At present, the mystery of coronavirus has not been fully uncovered by human beings, and there are no antiviral drugs or vaccines for SARS and MERS. Therefore, the research in this paper based on the whole genomic construction of global invariances by genomic index maps will provide relevant experts in the fields of biology, medicine and health to trace the origin and transmission path of SARS-CoV-2 virus, develop effective detection reagents, vaccines and therapeutic drugs, and effectively control and curb this epidemic, providing valuable data for decision-making information. Promote a large-scale calculation and mining research on the complete gene sequence of SARS-CoV-2 and related viral sequences, obtain a wide range of biological explanations, and analyze and mine potentially valuable information from the comparison calculation of genome sequence data.

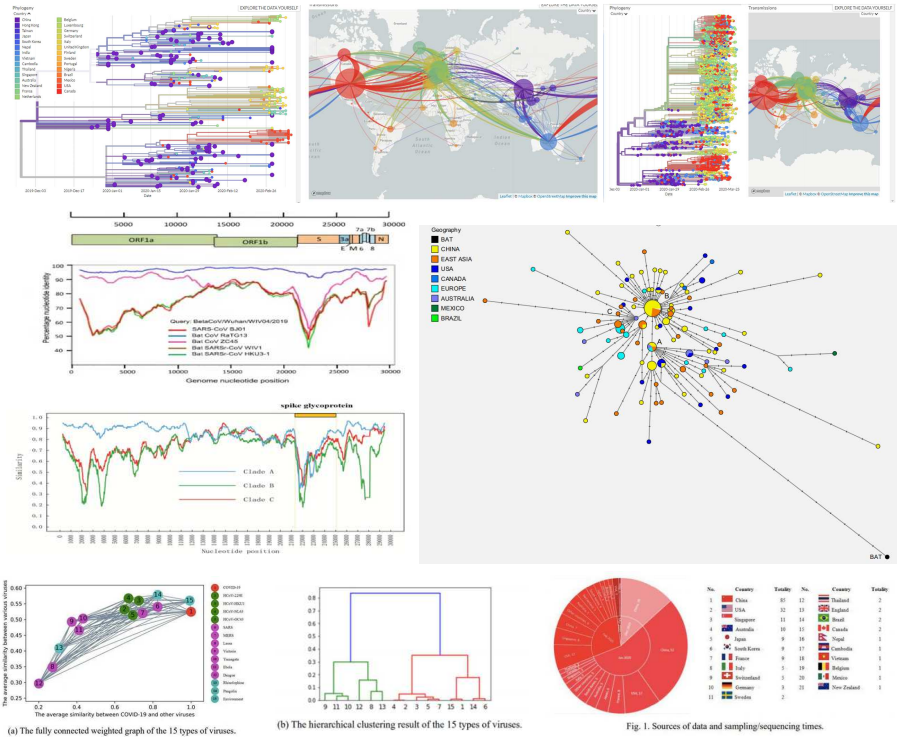
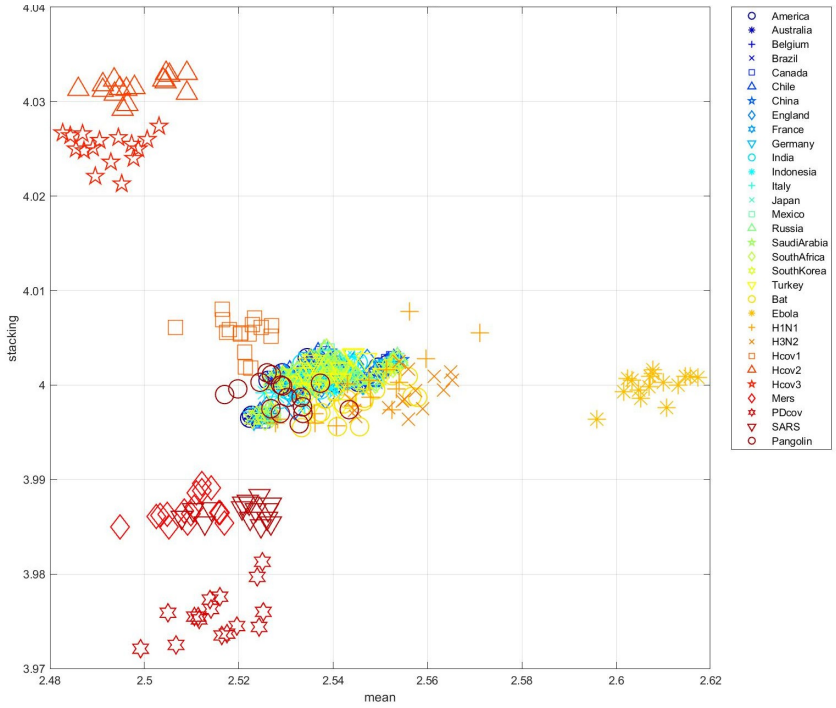


Fig. 4 SARS-COV-2 related researches

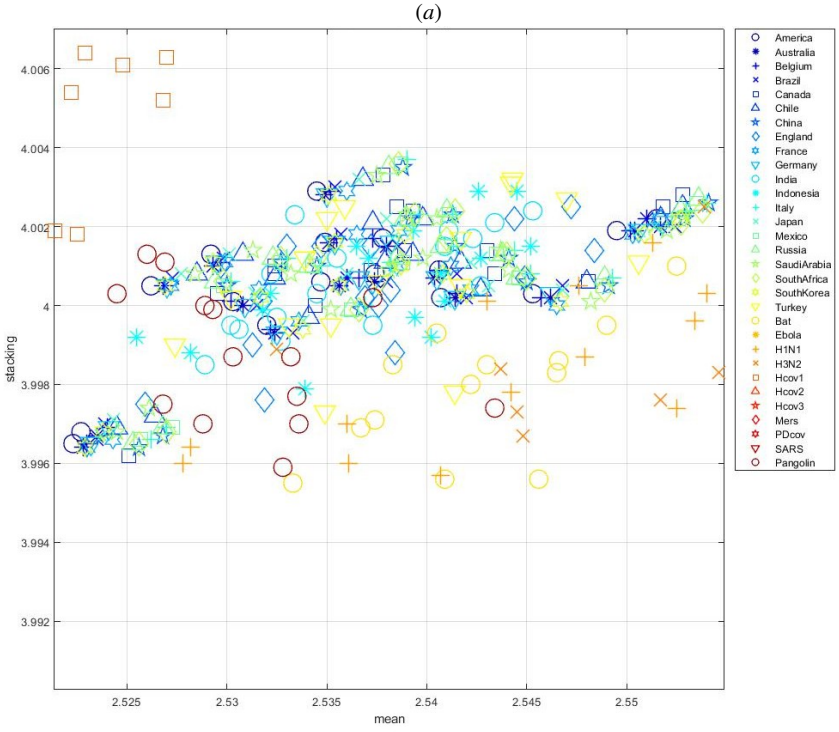
Results and Discussion

In this paper, novel coronavirus gene sequences and 11 other virus sequences belonging to 20 countries were selected, and the number of fragments was 16. According to the above method, the average entropy and the superposition entropy of 31 sequences are calculated. The average entropy is used as the X-axis, and the superposition entropy is used as the Y-axis. The results of Integrated Entropy Maps and enlarged 100 times for maps are shown in Fig. 5.

The novel coronavirus and 11 other viruses were clearly classified in the integrated entropy maps. Human coronaviruses HCOV-OC43 and hcov-nl63 are distributed in the upper left corner of the image, MERS and SARS are distributed in the down left corner, and partially intersected PDCOV (pig delta coronavirus) is concentrated in the lower left corner of the image, while Eloba virus is distributed in the upper center to the right. While 20 national samples of SARS-COV-2, human coronavirus named HCOV-HKU1, bat, pangolin carried coronavirus as well as H1N1 and H3N2 influenza viruses were distributed in the center region.



(*a*)



(*b*)

Fig. 5 Genomic Indices on Integrated Entropy Maps (a)-(b), $m = 16$; (a) Integrated Entropy Maps (b) Enlarged 100 times for Fig. 5(a)

A more detailed distribution can be observed from the magnification; the new coronavirus and bats, pangolin-carrying coronavirus and H1N1 and H3N2 influenza virus partially overlap, and there are cross regions. Eight other viruses have no cross effects on the new coronavirus.

The calculation results obtained by the shift are averaged, and the number of segment's length is 16. The average entropy is still used as the X-axis, and the stack entropy is used as the Y-axis. The results of genomic indices on mean entropy maps are shown in Fig. 6.

Genomic indices on mean entropy maps can cluster 11 other viruses with novel coronavirus more clearly. The distribution characteristics correspond to the integrated entropy maps.

Enlarged 100 times of images show the distribution of 20 new coronaviruses, using the X axis as a reference, followed by the United States, Belgium, Australia, Germany, India, South Korea, Brazil and France, the United Kingdom, Japan, Singapore, Chile, Canada, Saudi Arabia, Russia, Mexico, South Africa, China, and Italy.

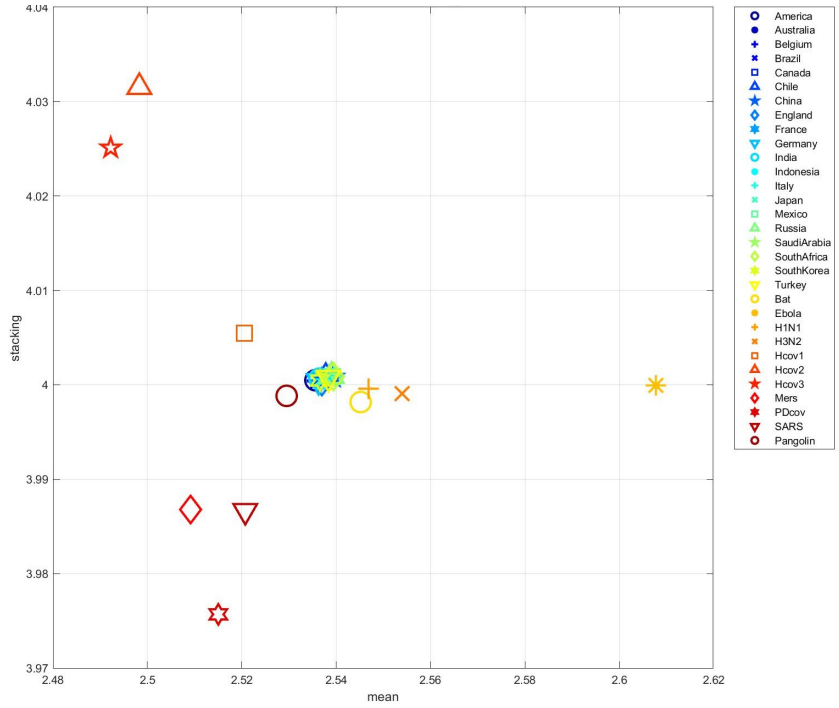
The samples were divided into 3 parts on average using the data of SARS-CoV-2 samples from 20 countries. Average entropy was used as A, B and C in red, green and blue, respectively. The details are shown in Fig. 7.

Conclusion

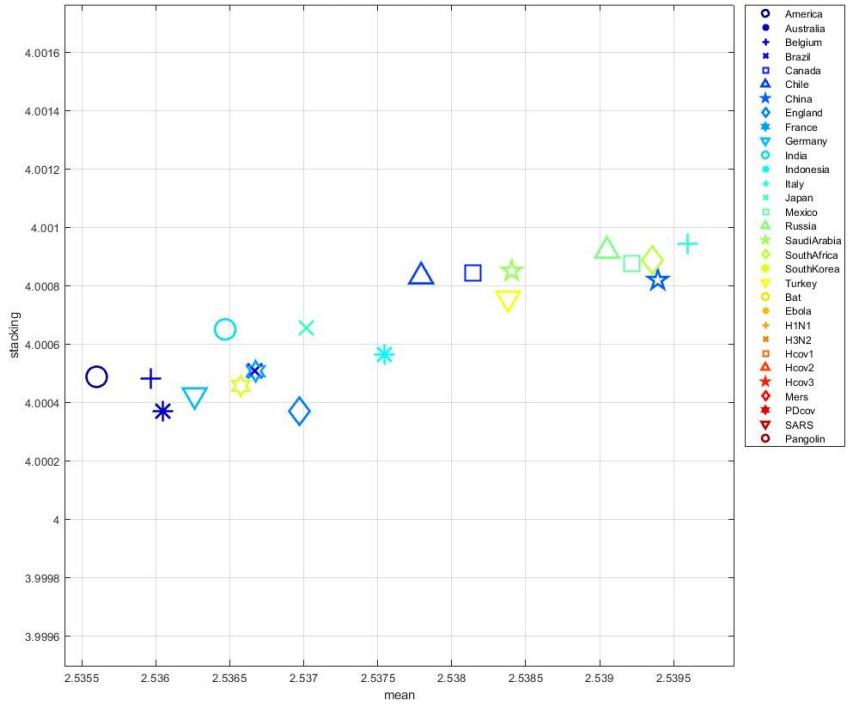
In this paper, SARS-CoV-2 of sequences from different samples are used as well as 11 other viral genome sequences, each of which acts as a coordinate scatter only at a specific location in the restricted geometric region. Based on the principle of entropy invariance, according to the visual analysis method, we can clearly identify the changes of SARS-CoV-2 samples in the global scope. Compared with the traditional analysis method, it shows the variation characteristics of different virus genome sequences from the macro point of view and has the super ability of differential identification to provide new ideas for the development of computer information, and to provide effective data support for medical workers.

Conflict Interest

No conflict of interest has been claimed.

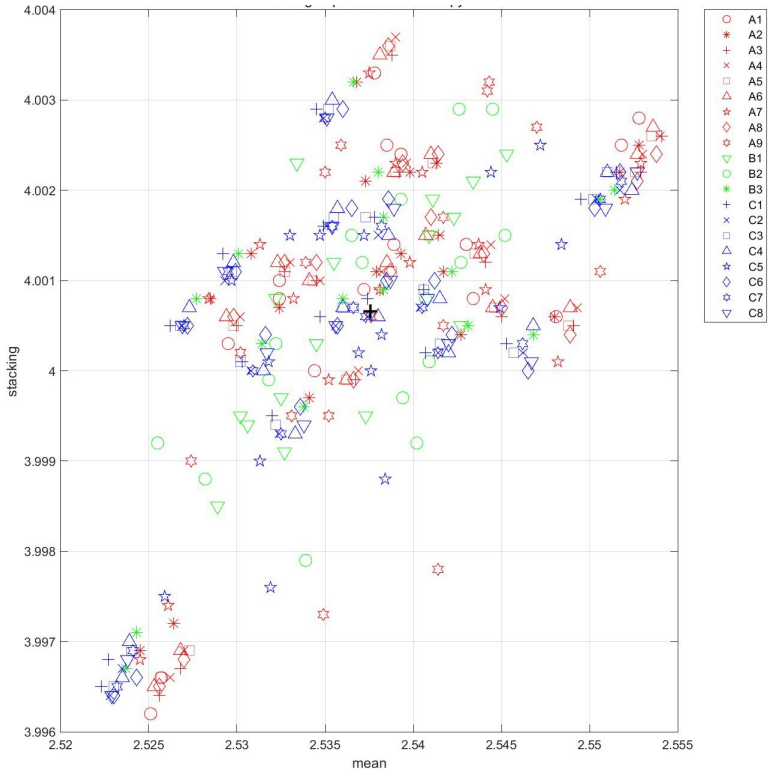


(*a*)

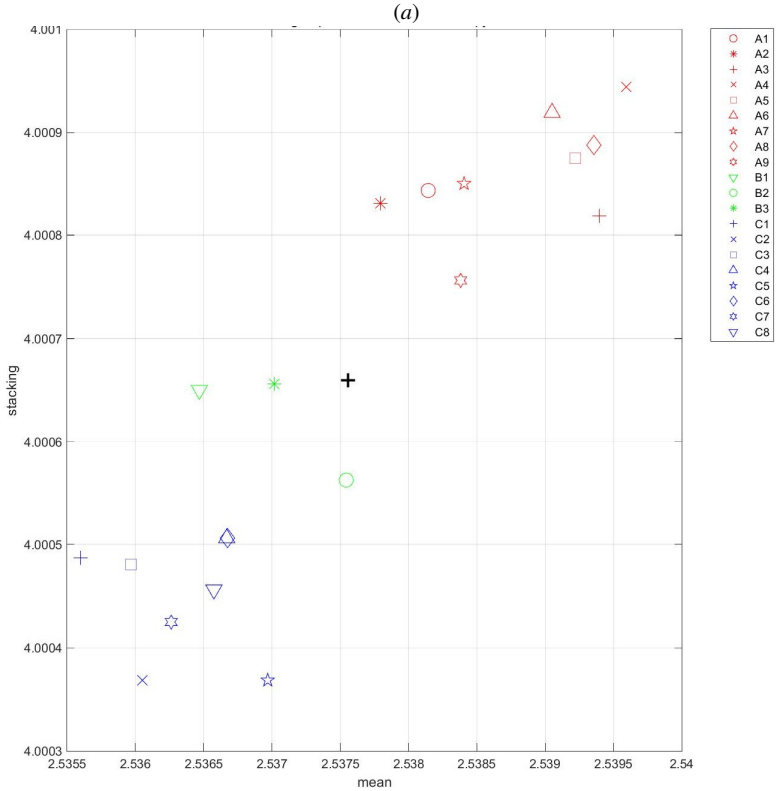


(*b*)

Fig. 6 Genomic indices on mean entropy maps (a)-(b), $m = 16$; (a) mean entropy maps (b) Enlarged 100 times for Fig. 6(a)



(*a*)



(*b*)

Fig. 7 Genomic indices on mean entropy maps by three parts (a)-(b), $m = 16$; (a) mean nntropy maps (b) Average mean entropy for Fig. 7(a)

Acknowledgements

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

References

1. Z. J. Zheng, A. Maeder, The conjugate classification of the kernel form of the hexagonal grid, Modern Geometric Computing for Visualization, Springer-Verlag, 73-89, 1992.
2. Z. J. Zheng,Conjugate transformation of regular plan lattices for binary images, PhD Thesis, Monash University, 1994.
3. Jeffrey Z. J. Zheng, Christian H. H. Zheng, A framework to express variant and invariant functional spaces for binary logic, Frontiers of Electrical and Electronic Engineering in China, 5(2):163-172, Higher Educational Press and Springer-Verlag, 2010.
4. Jeffrey Z.J. Zheng, Christian H.H. Zheng and Tosiyasu L. Kunii. A Framework of Variant Logic Construction for Cellular Automata, Cellular Automata - Innovative Modeling for Science and Engineering, Dr. Alejandro Salcido (Ed.), InTech Press, 2011.
5. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019 https://www.springer.com/in/book/9789811322815
6. Jeffrey Zheng, Variant Construction Theory and Applications, Vol.1: Theoretical Foundation and Applications, Science Press 2020 (Chinese, Formal Publishing Soon). 郑智捷, 变值体系理论及其应用 第 1 册：理论基础及其应用，科学出版社 2020 (即将正式发行)
7. Jeffrey Zheng, Research Gate: http://researchgate.net/pprofile/JeffreyZheng
8. Jeffrey Zheng, Chris Zheng, Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019, 193-202 https://link.springer.com/chapter/10.1007/978-981-13-2282-2 11 被斯普林格－自然杂志出版社，选入抗击新型冠状病毒肺炎研究(Research of COVID-19)资料汇集。推荐给 PMC 和 WHO（PubMed Central PMC and the World Health Organization WHO）以方便全球科学研究人员免费使用。
9. Jeffrey Zheng, Jianzhong Liu, A Visual Framework of Meta Genomic Analysis on Variations of Whole SARS-CoV-2 Sequences, https://www.researchsquare.com/article/rs-65152/v1
10. Jeffrey Zheng, Minghan Zhu, Input-Outpout Types of Fifteen Modules on Discrete and Real Measurements for COVID-19, https://www.researchsquare.com/article/rs-65158/v1
11. GISAID: Open access to influenza virus data https://gisaid.org
12. NCBI: Open access to dataes https://www.ncbi.nlm.nih.gov
13. Xu et al., Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission[J], SCIENCE CHINA Life Sciences, 2020.
14. Ji W, Wang W, Zhao X, Zai J, Li X.Cross-species transmission of the newly identified coronavirus 2019-nCoV[J]. J Med Virol. 2020;92:433440.
15. Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in 3 humans and its potential bat origin[J]. BioRxiv. 2020.
16. Lam,T. T. et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins[J]. Nature.2020.
17. Yanni Li et al. Similarities and Evolutionary Relationships of COVID-19 and Related Viruses[J].arxiv.2020.
18. Yu WB, Tang GD, Zhang L, et al. Decoding evolution and transmissions of novel pneumonia coronavirus using the whole genomic data[J]. ChinaXiv. 2020:202002.00033.
19. Peter Forster, Lucy Forster, Colin Renfrew, et al. Phylogenetic network analysis of SARS-CoV-2 genomes[J]. PANS. 2020.

20. Nextstrain Real time tracking of pathogen evolution https://nextstrain.org

Figures

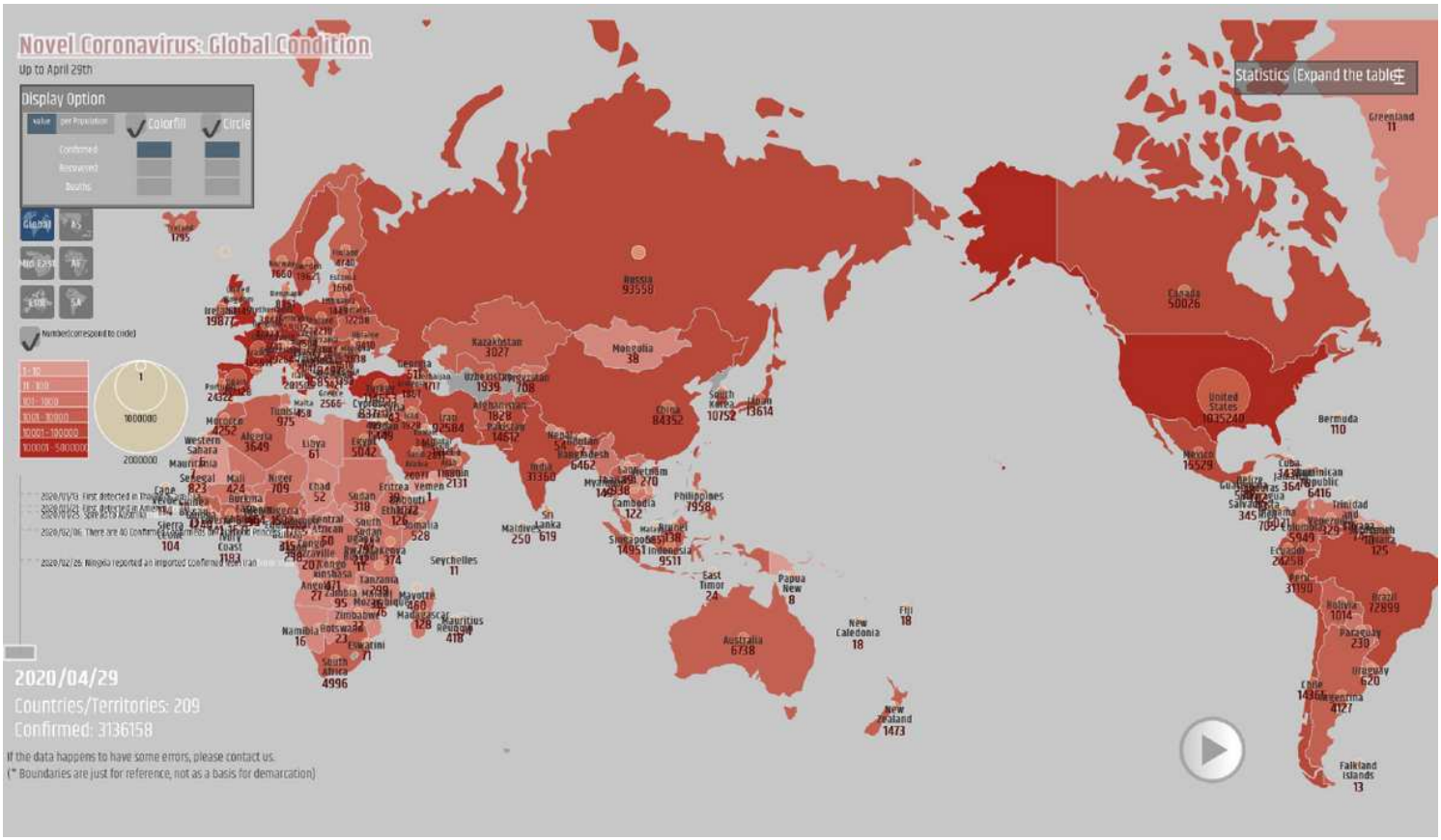


Figure 1

Global outbreak analysis maps. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Samples	NO.	*Locality*
SARS-COV-2	$(2019-nCoV)$	
	$EPI-ISL-412978$	China
	$EPI-ISL-417310$	England
		Turkey
		South Africa
		Singapore
	$EPI-ISL-416521$	SaudiArabia
		Russia
		America
		Mexico
	$LC-528233$	Japan
	$EPI-ISL-412974$	Italy
	$MT-012098$	India
	$EPI-ISL-410720$	France
		Chile
	$EPI-ISL-413014$	Canada
	$EPI-ISL-413016$	Brazil
	$EPI-ISL-417426$	Belgium
	$EPI-ISL-407193$	South Korea
	$EPI-ISL-408977$	Australia
	$A3-EPI-ISL-406862$	Germany
Human Coronavirus		
	$NC-002645$	HCOV-229E
	$NC-006577$	HCOV-HKU1
	$NC-006213$	HCOV-OC43
	$NC-005831$	HCOV-NL63
Deadly Coronavirus		
	$AY-508724$	SARS
	$JX-869059$	MERS
	$NC-002549$	EBOLA
Animals Coronavirus		
	$KX-022602$	PDCOV
	$SL-CovZC45$	Bat
	$MT-084071$	Pangolin
Other Virus		
		H1N1
		H3N2

Figure 2

Datasets of SARS-CoV-2 and other viruses

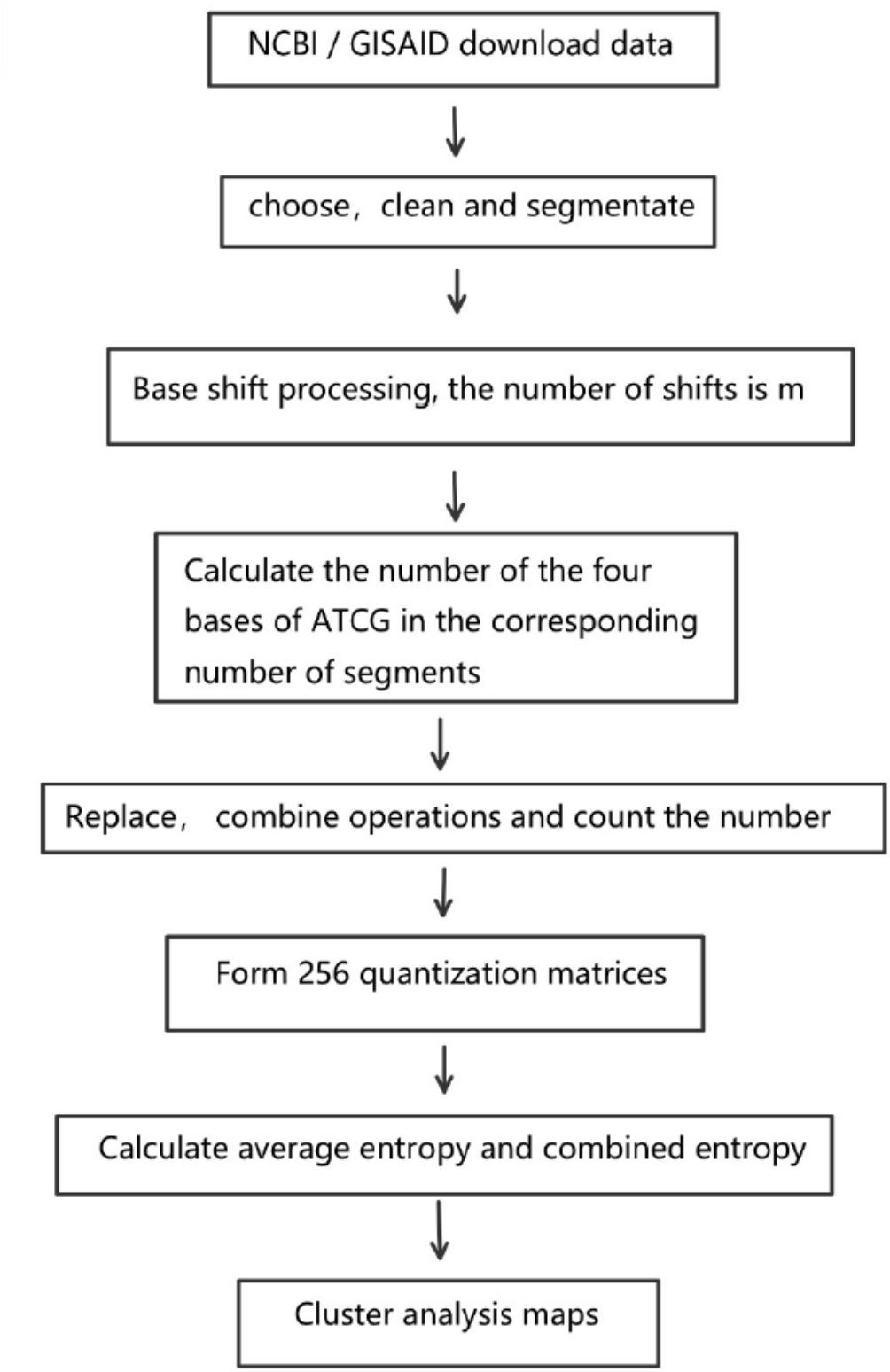
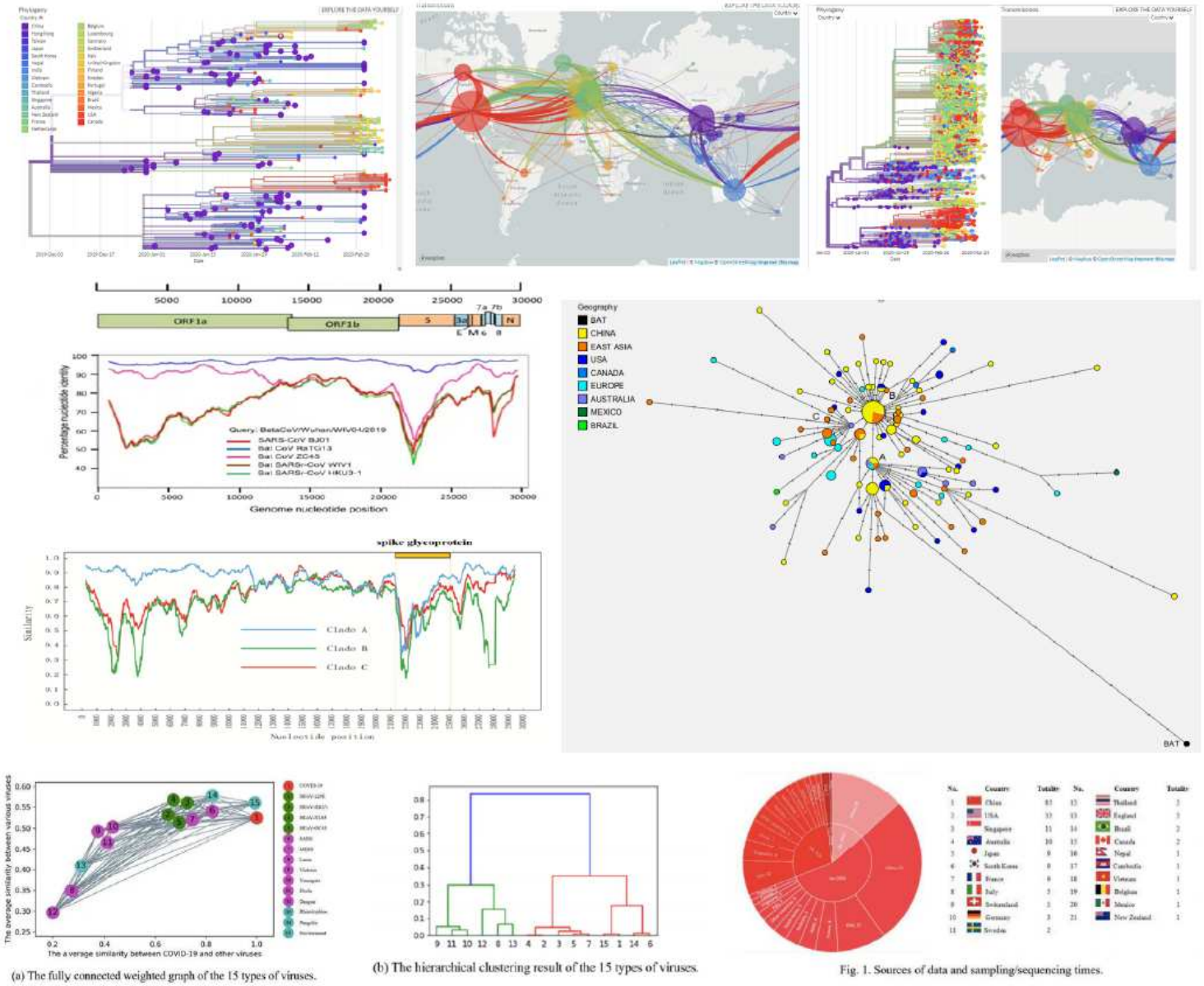


Figure 3

The flow of the method maps



(a) The fully connected weighted graph of the 15 types of viruses.

(b) The hierarchical clustering result of the 15 types of viruses.

Fig. 1. Sources of data and sampling/sequencing times.

Figure 4

SARS-COV-2 related researches. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

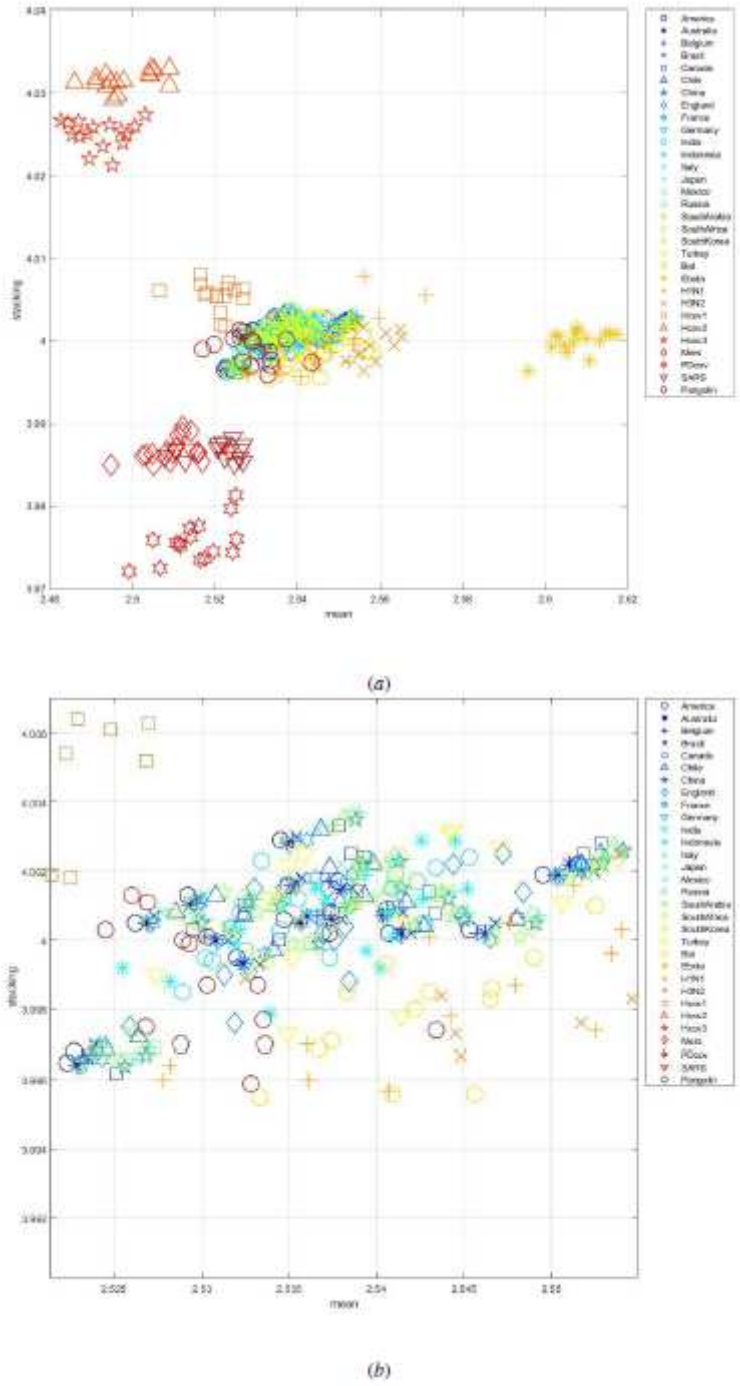


Figure 5

Genomic Indices on Integrated Entropy Maps (a)-(b), m = 16; (a) Integrated Entropy Maps (b) Enlarged 100 times for Fig. 5(a)

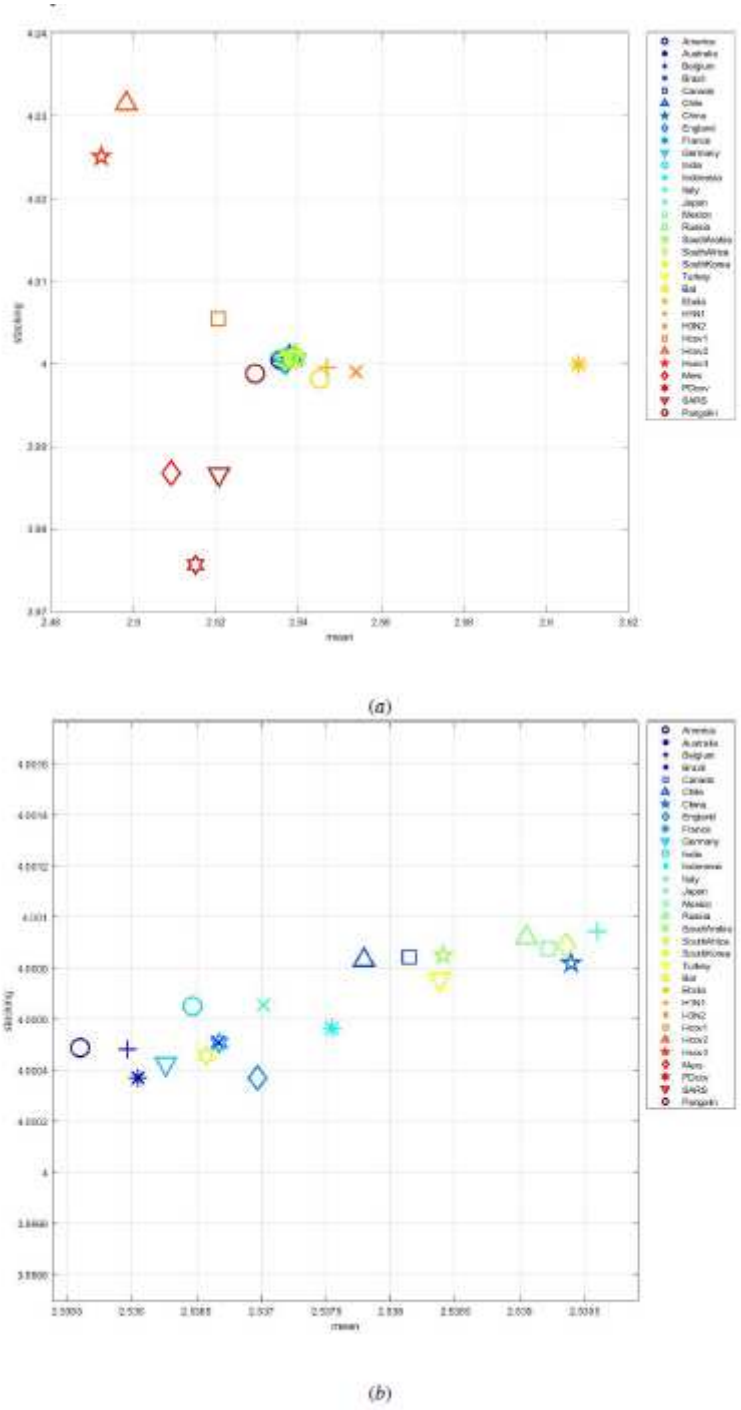


Figure 6

Genomic indices on mean entropy maps (a)-(b), m = 16; (a) mean entropy maps (b) Enlarged 100 times for Fig. 6(a)

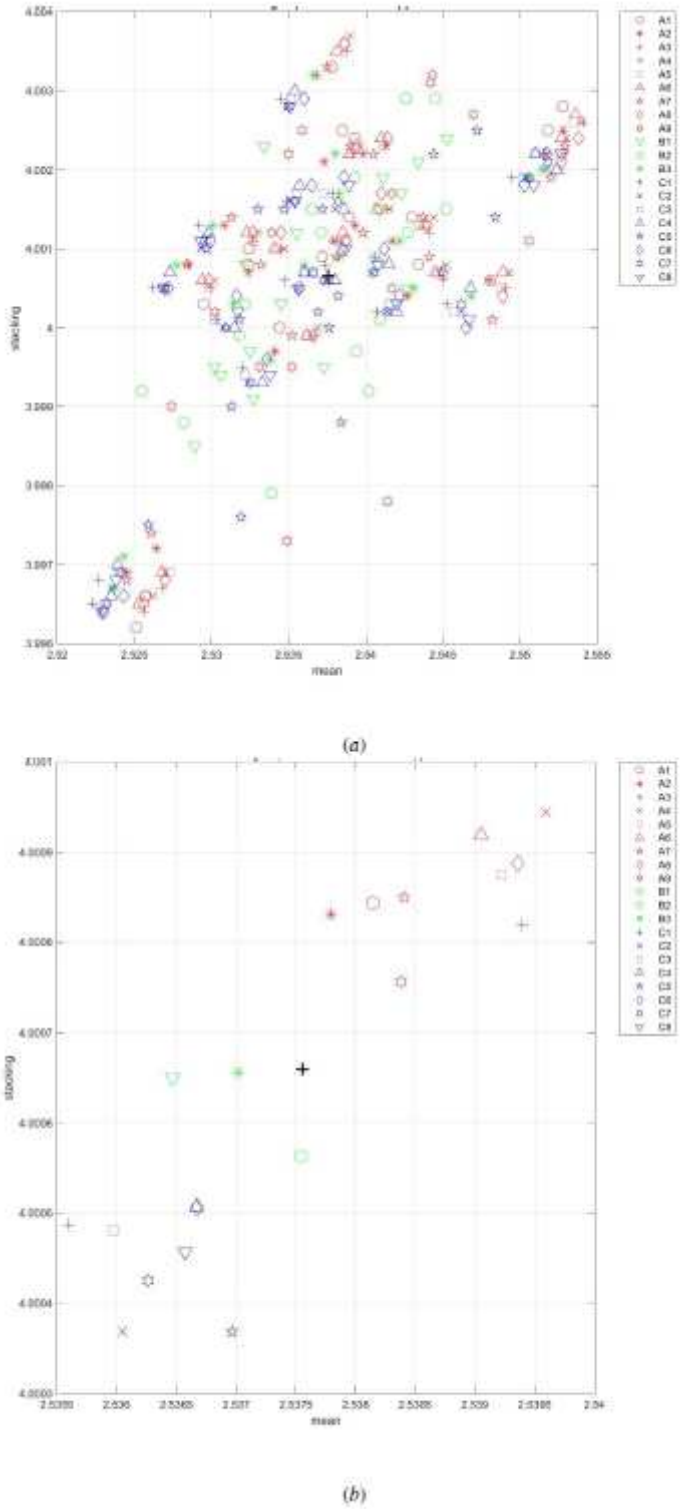


Figure 7

Genomic indices on mean entropy maps by three parts (a)-(b), m = 16; (a) mean nntropy maps (b) Average mean entropy for Fig. 7(a)