Visual Representation of SARS-CoV-2 Genomes in Multiple Regions on Integrated Maps

Minghan Zhu， Jeffrey Zheng

Abstract This paper represents visual results for the B2 module of the MAS. Using four meta genetic variables, multiple probability measures are extracted. Variations between pairs of virus genomes could be measured. From a macroscopic viewpoint, measures can be organized in a set of 16×16 with $(m+1)^2$ density matrix in combinations with supersymmetric properties. In view of the different types of coronary virus samples, various pairs of genomes make different projections under multiple levels of hierarchical forms of quantization matrix. Under proper selection, huge numbers of variations could be performed. Applying this transformation, a list of sample results are generated to illustrate intrinsic symmetric maps associated with selected parameters. Since this is an initial exploration, further explorations on theoretical foundation and specific applications are essential to support applicable theory and the systematic expansion on medical applications for COVID-19 patients.

Keyword SARS-COV-2, corona virus, COVID-19, probability measure, density matrix, intergrated distributions

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

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

Funding Supportted by the Key Project of Quantum Communication Technology (2018ZI002)

Introduction

At present, infectious diseases caused by various pathogens are prevalent all over the world. Most of the pathogens that can be transmitted from person to person, from animal to animal or from person to animal are microorganisms. For these infectious diseases, the epidemic prevention departments must timely master the incidence and adopt appropriate measures.

Among the many infectious diseases, there is a class of viruses that frighten humans - coronavirus (Coronaviruses). In 1937, coronavirus was first isolated from chickens. The first human coronavirus was isolated in 1965. Because of the obvious rod-like particle protrusions observed on the outer membrane under the electron microscope, which made it look like the crown of medieval European emperors, it was named "coronavirus ". In 1975, the virus naming committee officially named the coronavirus department. There are six known coronavirus infections in humans. Four common coronavirus HCOV-229E、HCOVHKU1、HCOV-OC43、HCOV-NL63 and 2 deadly coronaviruses MERS(Middle Eastern Respiratory Syndrome Coronavirus), SARS(Severe Acute Respiratory Syndrome).

Until early 2020, the international classification committee of viruses declared that the new coronavirus was named "SARS-CoV-2""(severe institute response syndrome coronavirus 2). The new coronavirus is a kind of RNA virus with capsule and linear single strand of genome. The particles are round or oval, and the diameter is about 60-140 nm.. Positive chain RNA means that the virus can direct protein synthesis when it enters the cell, and self-produce negative chains by RNA polymerase copy. While SARS-CoV-2 mortality rates are lower than those of SARS and MERS, transmission rates are much higher than those of these two deadly coronaviruses, causing concern worldwide. People infected with new coronavirus will have fever, coagulation symptoms, white lungs and other symptoms, serious may threaten life. Recent asymptomatic new coronavirus carriers may have mutated genes.

In order to study the possible variation of genome sequence, this paper randomly selected samples from 8 countries, using the visualization method under variant logic system, to quantify the four bases of virus genome sequence. 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 were developed for 40 years using variant construction [1]-[8] such as content-based image retrievel, medical image processing, Bat echo identifications, DNA maps, hierarchical organzition, phase space classification, feature extraction, filtering, combinations, projections and conjugate transformations [9]. It can from the perspective of overall invariance, compare the statistical distribution characteristics of invariance, and explore the possible variations of genome sequence between countries macroscopically, which lays the foundation for the study of the new coronavirus and typical coronavirus genomes.

Data Sources

The genome sequences data are downloaded from the open source databases called NCBI (National Center for Biotechnology Information) and GISAID (Global Shared Influenza Data Initiative) in this article [10]-[11].The data description is shown in Fig.1.

samples	NO.	*Locality*
SARS-COV-2	$(2019-nCoV)$	
	$NC-045512$	China
	$LC-528233$	Japan
	$EPI-ISL-412974$	Italy
	$EPI-ISL-410720$	France
	$EPI-ISL-4089771$	Australia
	$EPI-ISL-413014$	Canada
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
Animals Coronavirus		
	$KX-022602$	PDCOV
	$SL-CovZC45$	Bat
	$MT-084071$	Pangolin

Fig. 1 Datasets of SARS-CoV-2 and other cases worldwide

Distribution Characteristics and Method Description

Download the representative new coronavirus SARS-CoV-2 and various influenza virus-related data from NCBI and GISAID. First, the genome sequences are chose and cleaned carefully. It processed to calculate the number of A,T,C,G four bases in the corresponding section. Secondly, performed substitution and combination operations on the calculation results of the same genome sequence, and count the numbers respectively according to the same counting information contained in different segments. Then, recorded the results to form the 256 of quantization matrices, constituted to a huge space of $257 \times (m+1)^2$. Finally, the visual analysis method in variant measurement is provided.

The possible variation and differential characteristics of the base pairs in the genome sequence are displayed to form a distinguishable classification diagram with the characteristic of super symmetric reflection from macroscopic perspective. Specific formula derivation process please refer to the paper [13]. The flow of the method used is shown in Fig. 2.

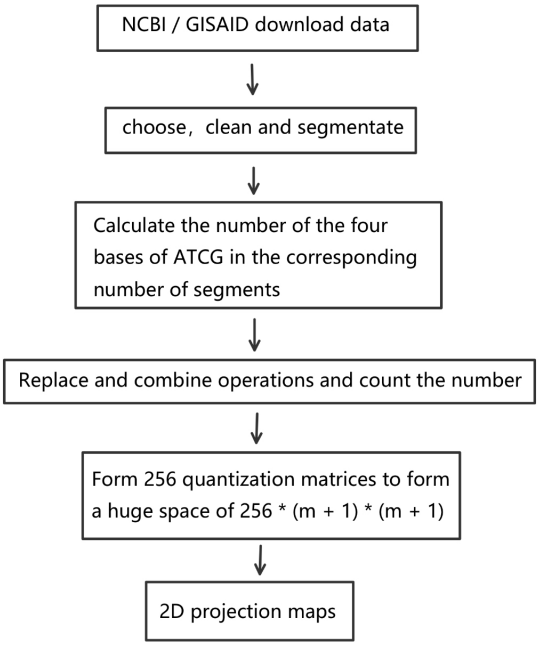


Fig. 2 The flow of the method maps

Results and Discussion

Comparison Results of SARS-COV-2 in Different Countries

The differences among the histograms are very small judging from the results of the two-dimensional projection, and it is difficult to distinguish the unique characteristic for the naked eye. The new coronavirus genome sequences data selected randomly belonging to six countries maps. Each genome sequence is about 30K,

and the data segment number is selected to 512. The overall distribution image is white, blue, green, and yellow increasing accordingly. The white area represents no data and blue shows less data distribution. The yellow scattered dots represent the most dense data distribution and it has the largest number. Six images are very similar, it formed distinguishable classification diagrams with supersymmetric reflection characteristics, and all present square areas. The detail patterns are different slightly. It can be observed that China, Italy and France have high similarity figures. It presents the structure of square package with diamond, But the distribution diagram is more dense in Japan. And the middle diamond shape is more rounded. The patterns of Canada and Australia show large squares cover to small squares.The specific overall distribution is shown in the Fig. 3.

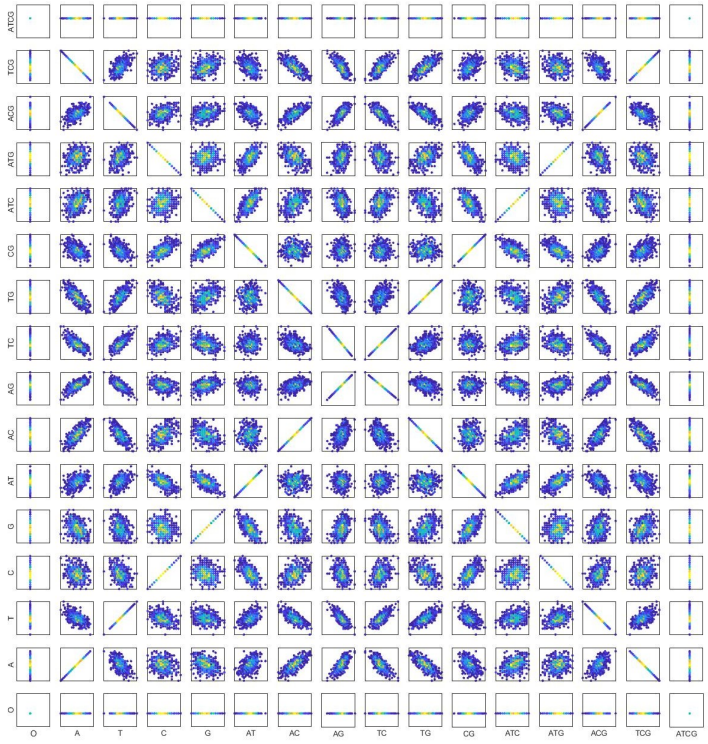
Compared Typical Coronavirus with SARS-COV-2

Normal Human Coronavirus

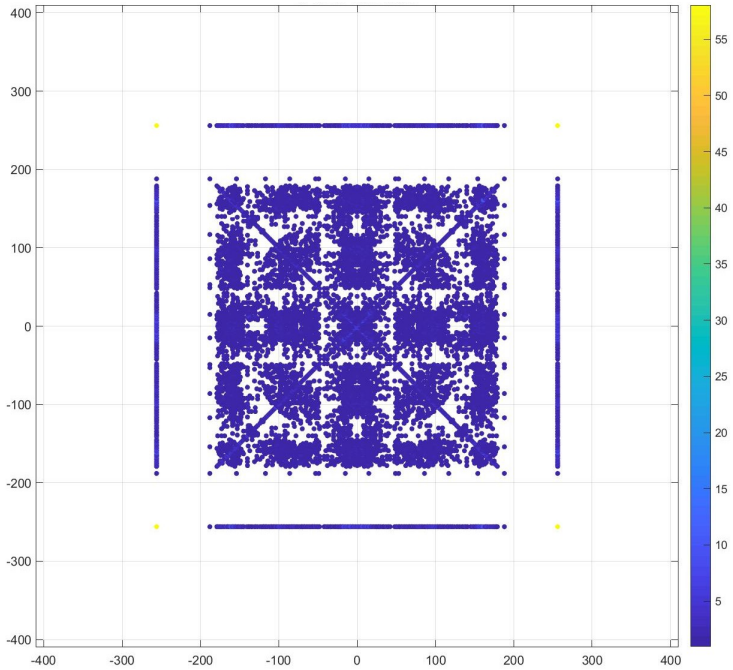
At present, there are four common human coronaviruses called HCOV-229E, HCOV-HKU1, HCOV-OC43 and HCOV-NL63. The sequence size of each data group is about 30K, and the number of segments is 59. According to the above conclusions, the difference between the histogram figures are not obvious, the following two-dimensional projection histogram figures are not shown. Judging from the two-dimensional projected superimposed image, although the four common human coronaviruses still present a supersymmetric graphic structure, the difference is obvious and the distribution characteristics are different.Among them, HCOV-NL63 is the closest to the distribution of the SARS-COV-2, and they all have same structure like a square.But it is obvious that it does not belong to the same category as the SARS-COV-2.The specific overall distribution is shown in the Fig. 4.

Deadly Coronavirus

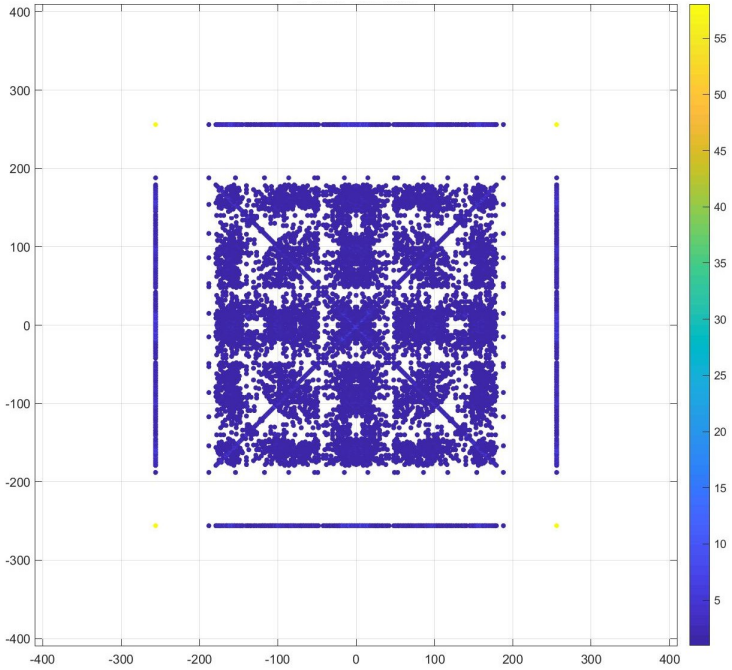
A joint study found that the average genetic sequences of SARSCoV-2, SARS and MERS viruses are more than percent of 70 and 40 sequence similarity by the Chinese Academy of Sciences, the Academy of Military Medical Sciences, and the Chinese Academy of Sciences Biological Laboratory [1]. According to the above conclusions, we selected the genome sequences of the deadly coronavirus MERS and SARS In this paper. Each data size is about 30K, and the number of segments is 512. The images show that MERS and SARS are very different from the SARS-COV-2.The pattern of the projected overlays made with SARS genome sequence data is clear and beautiful. But the MERS has more regular distribution, which divides the square area into 9 parts. The specific overall distribution is shown in the Fig. 5.


ATCG
TCG
ACG
ATG
ATC
CG
TG
TC
AG
AC
AT
G
C
T
A
O
O
A
T
C
G
AT
AC
AG
TC
TG
CG
ATC
ATG
ACG
TCG
ATCG

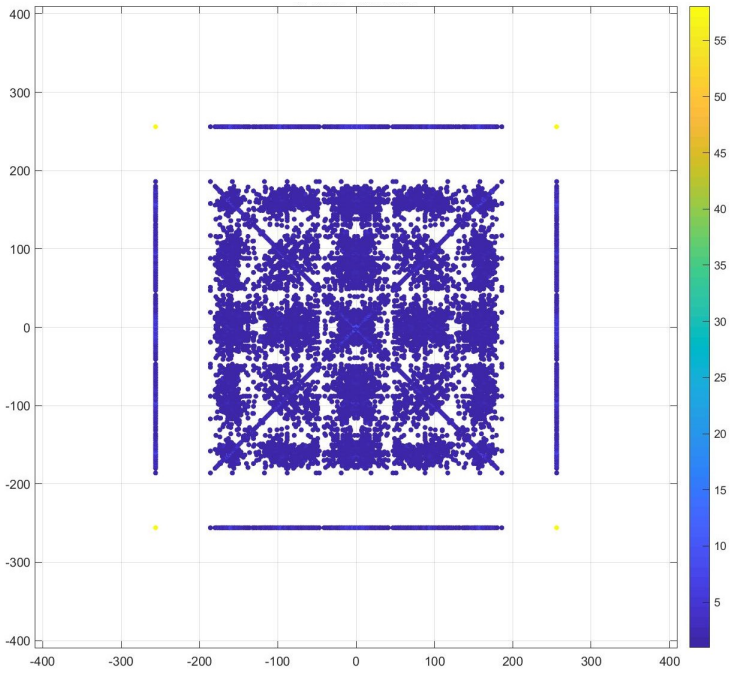
(*a*)


400
300
200
100
0
-100
-200
-300
-400
-400
-300
-200
-100
0
100
200
300
400
55
50
45
40
35
30
25
20
15
10
5

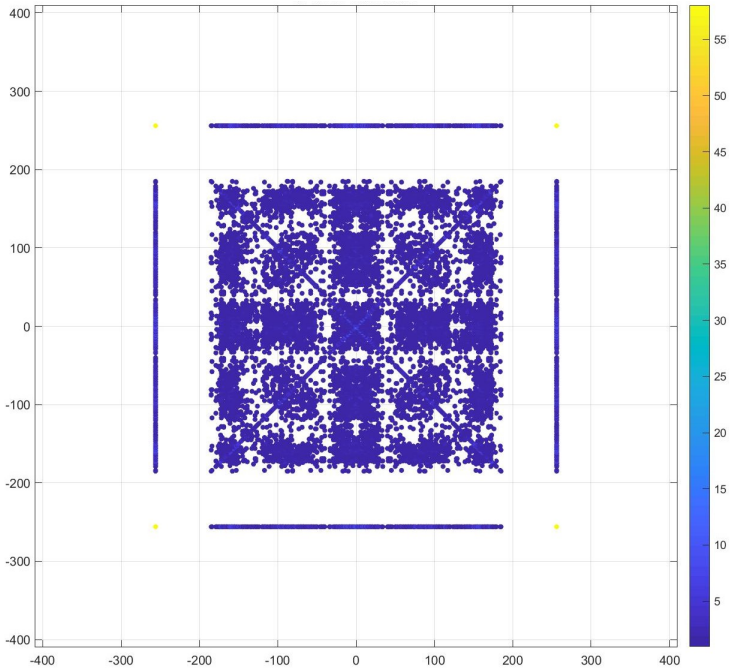
(*b*)


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
55
50
45
40
35
30
25
20
15
10
5

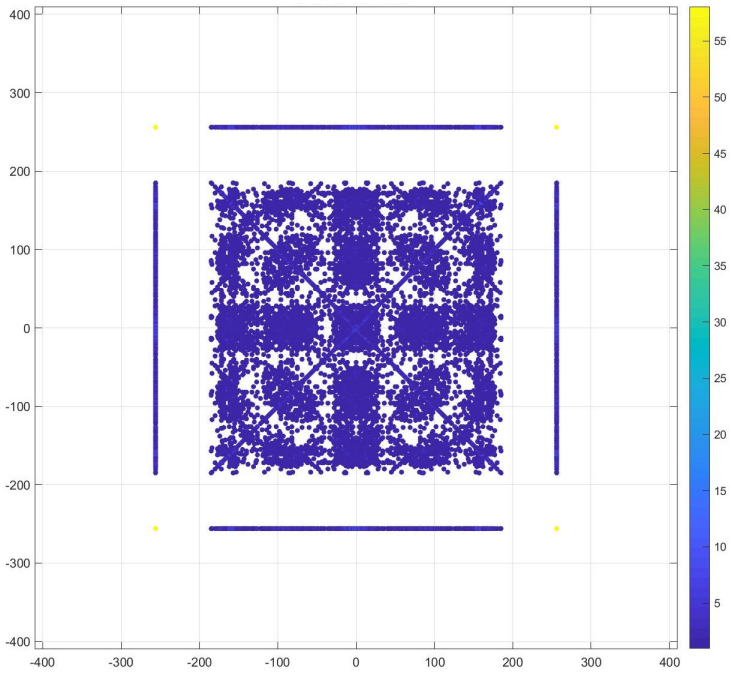
(*c*)


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
55
50
45
40
35
30
25
20
15
10
5

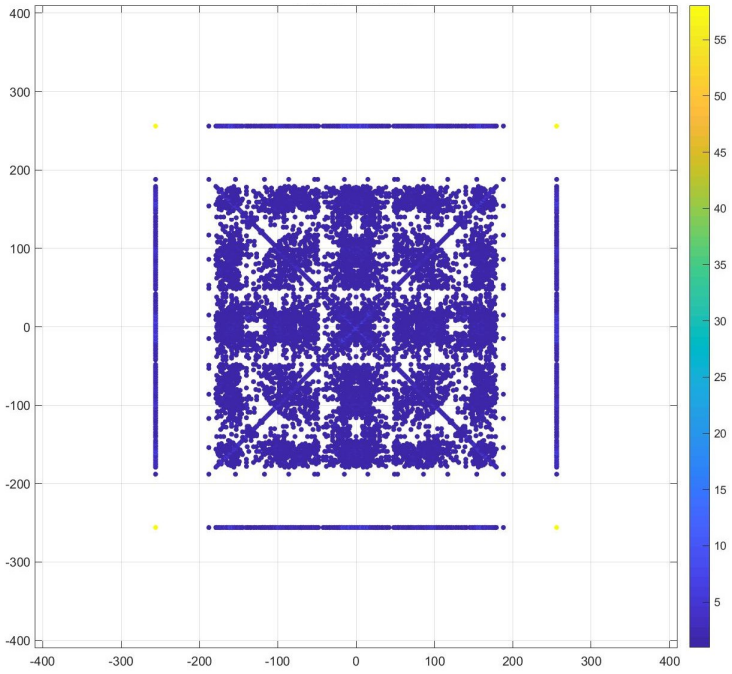
(*d*)


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
55
50
45
40
35
30
25
20
15
10
5

(*e*)


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
55
50
45
40
35
30
25
20
15
10
5

(*f*)

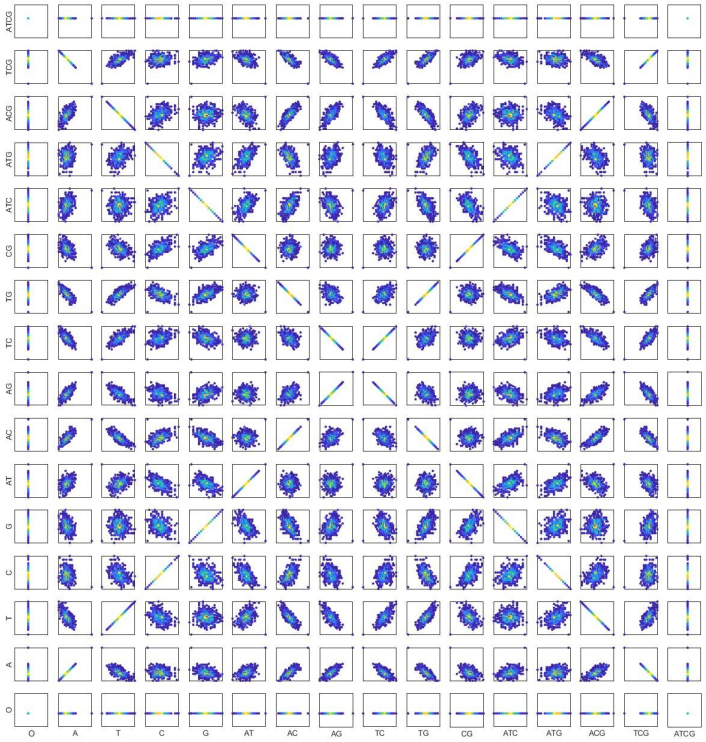


(*g*)

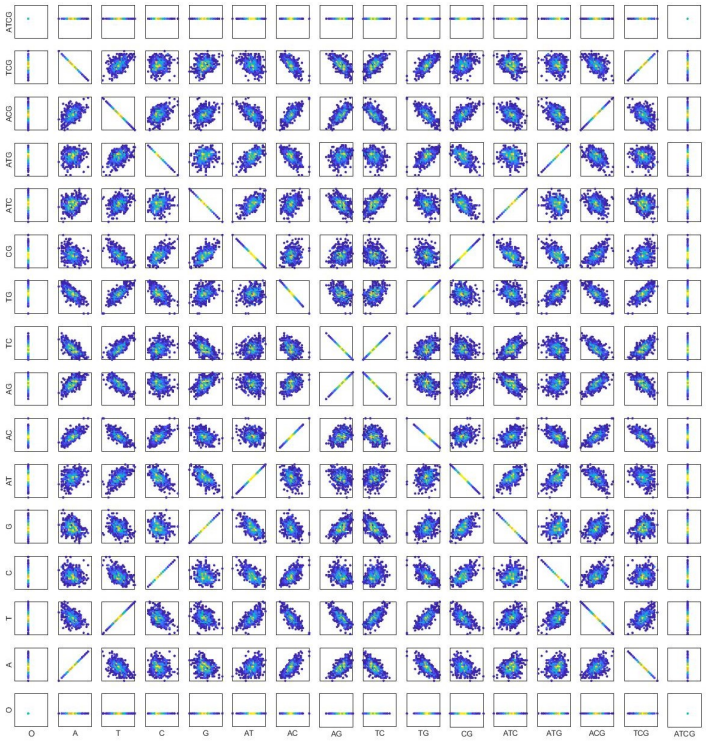
Fig. 3 Integrated density matrices in eight maps (a)-(h), m=512; (a) square of SARS-COV-2, m=59 (b) China (c) Italy (d) France (e) Japan (f) Canada (g) Austrilia

Corona Viruseses of Animals

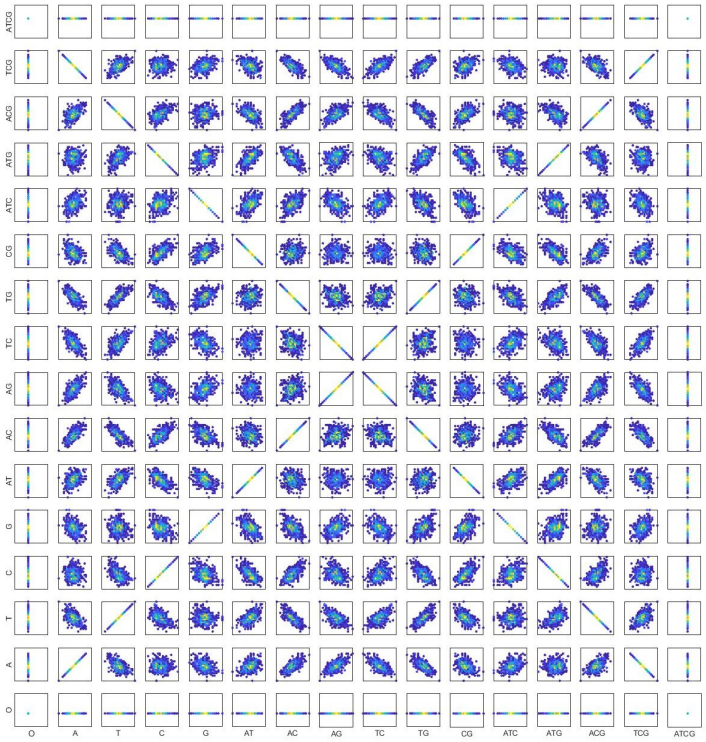
Shi Zhengli's team believes that bats are the most likely wild animals which carrying new coronaviruses at the Wuhan Institute of Virology, Chinese Academy of Sciences [14]. And the total gene level is over percent of 90. The University of Hong Kong and South China Agricultural University had analyzed more than 1,000 metagenome samples. They believed that pangolin may be potential intermediate hosts for new coronaviruses [15]. Therefore, we selected three animal coronavirus genome sequences, which were PDCOV (porcine delta coronavirus), coronavirus carried by bat and pangolin in this paper. A 256 matrix with a data size of 30K and a segment number of 512 is established for two-dimensional projection on this basis. The overall distribution of pangolin is more similar to the SARS-COV-2. The specific overall distribution is shown in the Fig. 6.


ATCG
TCG
ACG
ATG
ATC
CG
TG
TC
AG
AC
AT
G
C
T
A
O
O A T C G AT AC AG TC TG CG ATC ATG ACG TCG ATCG

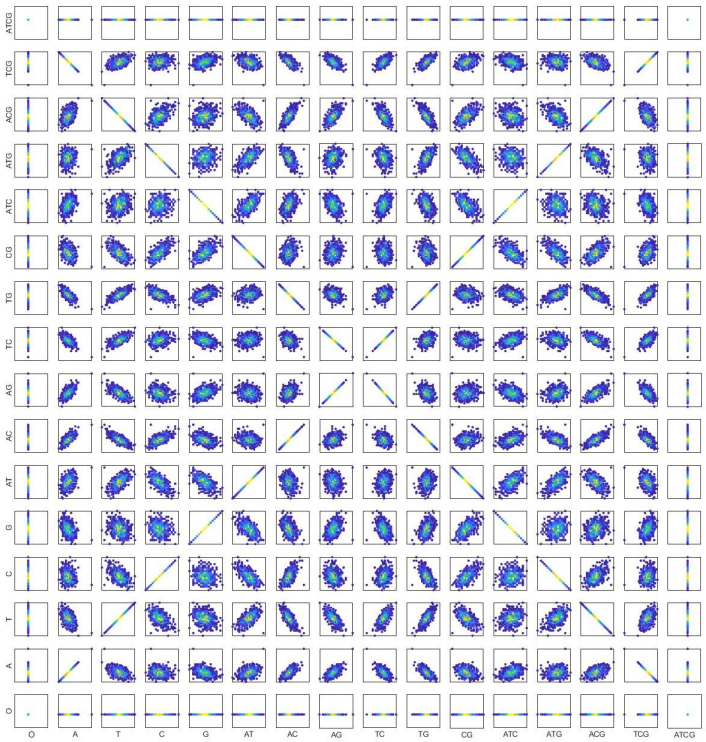
(*a*)


ATCG
TCG
ACG
ATG
ATC
CG
TG
TC
AG
AC
AT
G
C
T
A
O
O A T C G AT AC AG TC TG CG ATC ATG ACG TCG ATCG

(*b*)

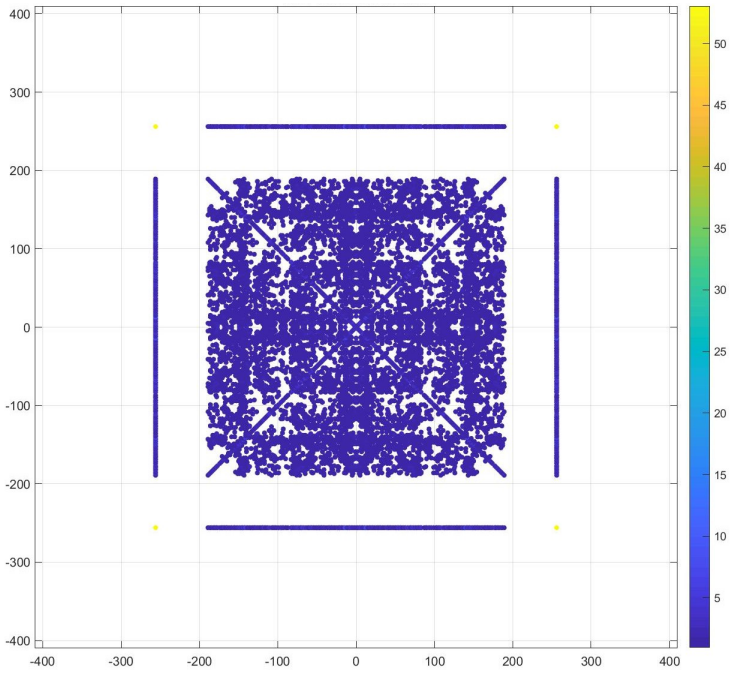


(*c*)

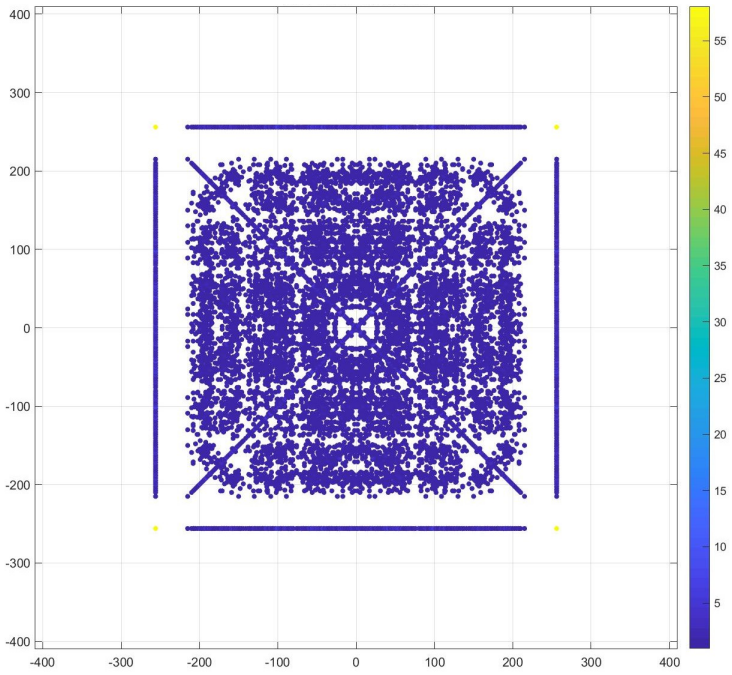


(*d*)

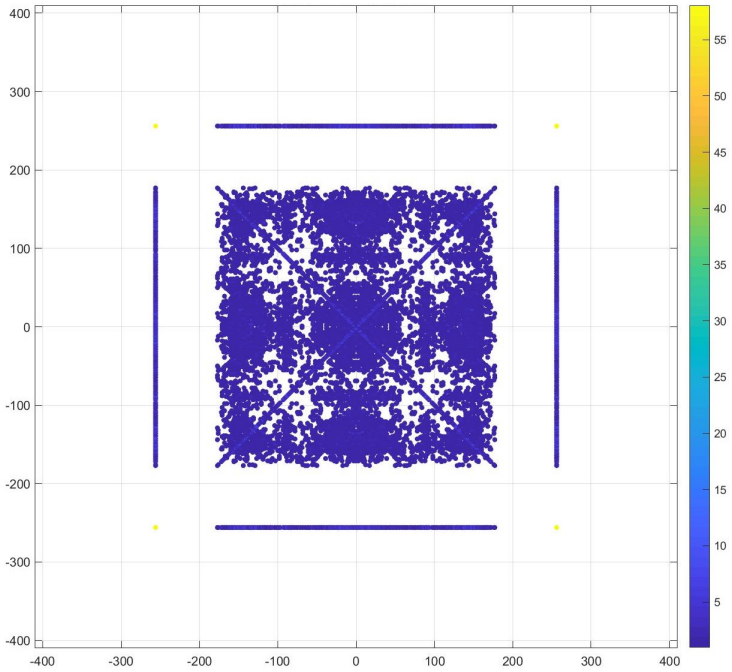
Fig. 4 Integrated density matrices in four histograms (a)-(d), m=59; (a) HCoV-229E (b) HCoV-HUK1 (c) HCoV-NL63 (d) HCoV-OC43


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
50
45
40
35
30
25
20
15
10
5

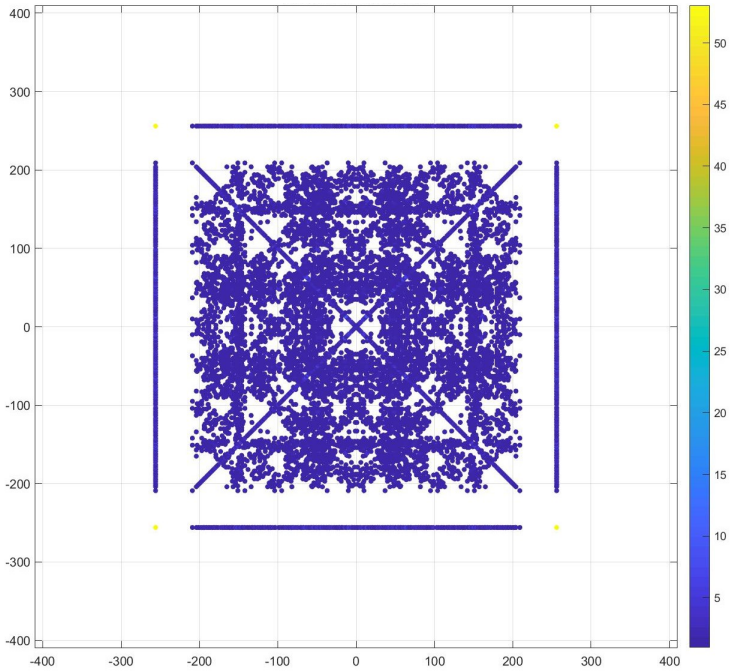
(*a*)


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
55
50
45
40
35
30
25
20
15
10
5

(*b*)

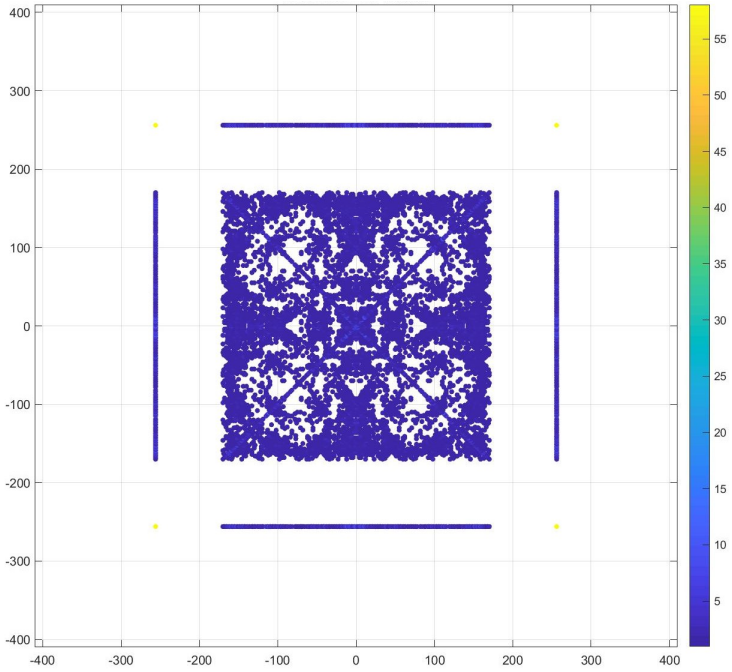


(*c*)

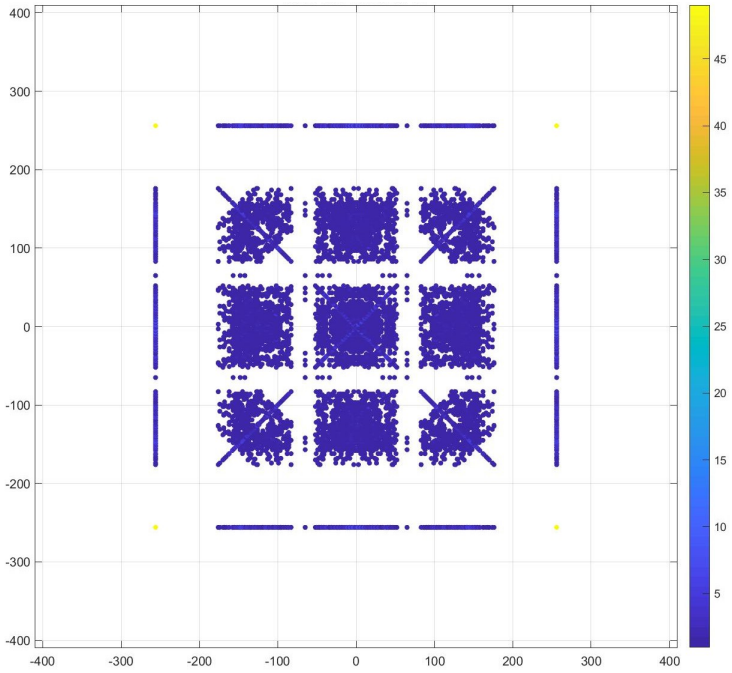


(*d*)

Fig. 5 Integrated density matrices in four maps (a)-(d), m=512; (a) HCoV-229E (b) HCoV-HUK1 (c) HCoV-NL63 (d) HCoV-OC43

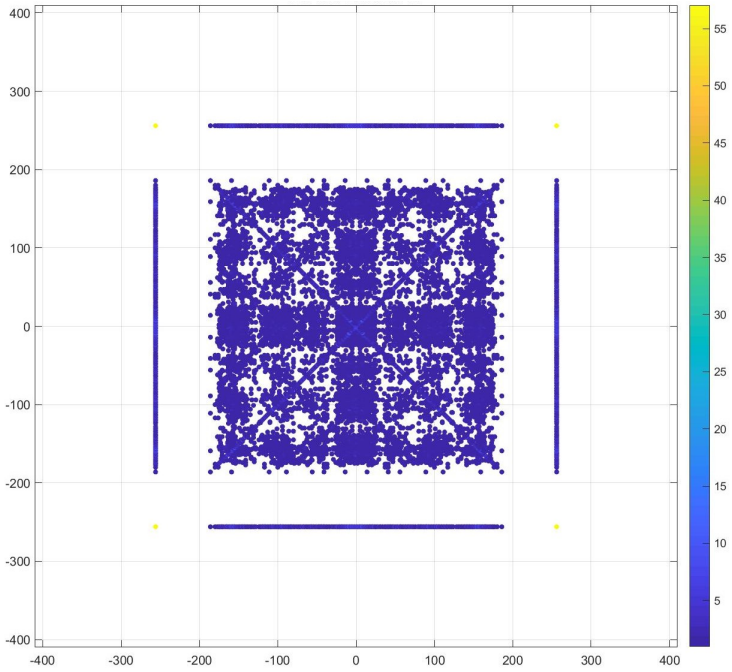


(*a*)

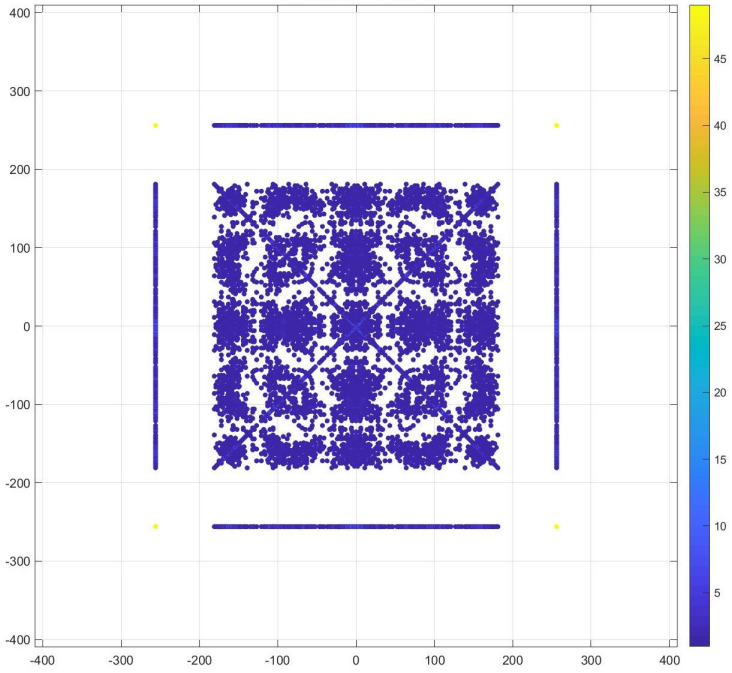


(*b*)

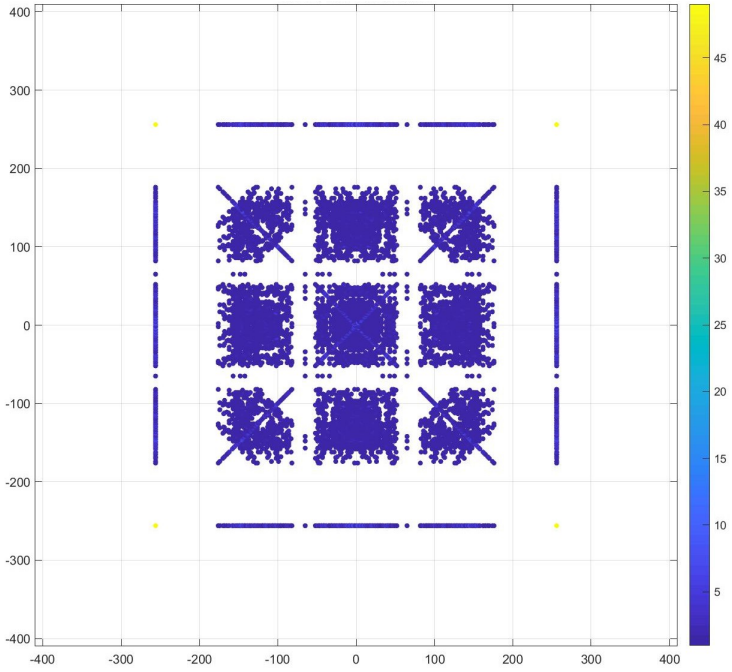
Fig. 6 Integrated density matrices in two maps (a)-(d), m=512; (a) SARS (b) MERS


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
55
50
45
40
35
30
25
20
15
10
5

(*a*)


400
300
200
100
0
-100
-200
-300
-400
-400 -300 -200 -100 0 100 200 300 400
45
40
35
30
25
20
15
10
5

(*b*)



(*c*)

Fig. 7 Integrated density matrices in three maps (a)-(c), m=512; (a) Bat (b) Pangolin (c) PDCOV

Purpose and Significance of Research

The in-depth exploration of the intrinsic information of the genome sequence will facilitate medical personnel to have a more comprehensive understanding of the information of viral inheritance, genetic characteristics and evolution.Tracing the possible virus source and host and exploring the possible relationship between host and virus are helpful for epidemiological investigation. Whole genome sequencing and data analysis of the virus provide data support for the development of relevant diagnostic reagents, specific drugs and vaccines, and lay an efficient foundation for the study of the distribution pattern of virus gene variation.

Conclusion

By analyzing coronavirus related genome sequences and comparing the two-dimensional projection characteristics of different genome sequences, a large spatial geometry was established. Under the overall distribution angle, the perfect supersymmetric distribution is presented.With the change of the sequence of four bases in the genome sequence, the unique graphic features are displayed, and the kaleidoscopic structure is formed. In the field of computer, artificial intelligence and big data analysis, presenting a large amount of space provides new methods and lays a theoretical foundation. In the medical field, the whole genome sequencing base sequence is used to form a contrast map with distinct distinguishing features, which provides data support for medical workers and provides information reference for the study of gene variation characteristics.

Conflict Interest

No conflict of interest has claimed.

Acknowledgements

The authors would like to thank NCBI,GISAID, Nextstrain to provide 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[J]
10. Jeffrey Zheng, Minghan Zhu, Input-Outpout Types of Fifteen Modules on Discrete and Real Measurements for COVID-19[J]
11. GISAID: Open access to influenza virus data https://gisaid.org
12. NCBI: Open access to dataes https://www.ncbi.nlm.nih.gov
13. Jeffrey Zheng, Minghan Zhu, Input-Outpout Types of Fifteen Modules on Discrete and Real Measurements for COVID-19[J]
14. 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.
15. Ji W, Wang W, Zhao X, Zai J, Li X.Cross-species transmission of the newly identifiedcoronavirus 2019-nCoV[J]. J Med Virol. 2020;92:433440.