A Visual Framework of Meta Genomic Analysis on Variations of Whole SARS-CoV-2 Sequences

Jeffrey Zheng (conjugatelogic@yahoo.com)

Key Laboratory of Quantum Information of Yunnan, Key Laboratory of Software Engineering of Yunnan, Yunnan University https://orcid.org/0000-0003-4225-7077

Jianzhong Liu

Key Laboratory of Quantum Information of Yunnan, Key Laboratory of Software Engineering of Yunnan

Research Article

Keywords: hierarchical organization, k-mers, probability distriution, eigenvalue, information entropy, clustering, combinatorial topology, geometric visualization, genomic index

Posted Date: August 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-65152/v1

A Visual Framework of Meta Genomic Analysis on Variations of Whole SARS-CoV-2 Sequences

Jeffrey Zheng, Jianzhong Liu

Abstract The worldwide outbreak of the CODIV-19 has become a global pandemic resulting in millions of confirmed cases and hundreds of thousands of deadths. To face such a global crisis, bioinformatics has played a key role in the diagnosis, follow-up, prognosis and treatment of COVID-19-infected patients.

A novel bioinformatic tool for metagenomic analysis of whole genomes is proposed in this paper that is composed of three projections: global, clustering and genomic index. For each projection, key modules are described. Global projection provides various combinatorial distributions for a whole genome of N length, and the m-mer scheme partitions this sequence as M segments on 1D, 2D and 3D density matrices for multiple projections. Clustering projections based on distributions from global projections make special filters extract specific parts as probability eigenvalues.

Genomic index projection provides comprehensive technologies under the theory of information entropy, and a list of measuring entropies are included, such as combinatorial entropy CE, integrated entropy IE, mean entropy ME and topological entropy TE. Three projections provide unified information to describe complicated functions, internal structures and refined variations for multiple groups of SARS-CoV-2 on variations and other genomes in comparisons.

The outputs of three projections are illustrated on variant maps to support category, clustering, classification and establishing root activities for refined quantitative operations from bottom to top strategy.

Keywords: hierarchical organization, k-mers, probability distriution, eigenvalue, information entropy, clustering, combinatorial topology, geometric visualization, genomic index

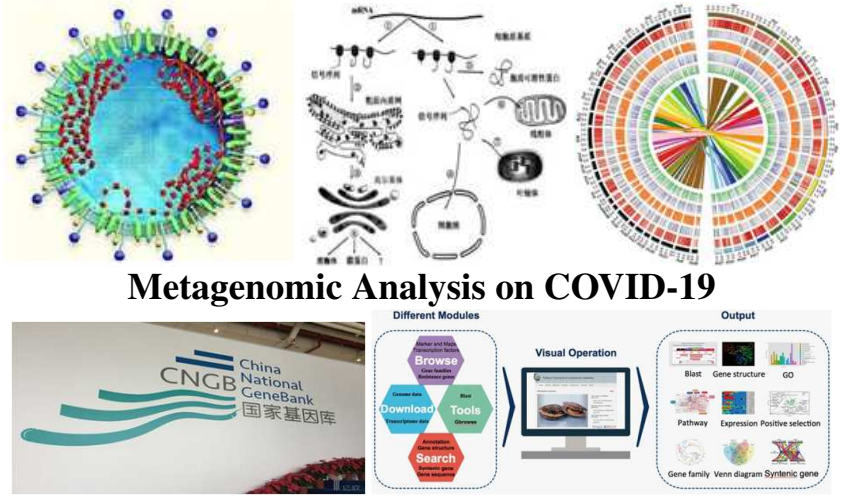
---

Jeffrey Zheng[1,2,3] e-mail: conjugatelogic@yahoo.com
· Jianzhong Liu[1,2] e-mail: liujianz6655@126.com
[1] Key Laboratory of Quantum Information of Yunnan
[2] Key Laboratory of Software Engineering of Yunnan
[3]Yunnan University, Kunming

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

Introduction

Initially from November 2019 or earlier, the worldwide spread outbreak of the CODIV-19 has turned into a global pandemic of over 200 countries resulting in more than 3 million confirmed cases and hundreds of thousands of deadths. Fighting this unexpected global crisis, bioinformatics has played a key role in the diagnosis, follow-up, prognosis and treatment of COVID-19-infected patients. It is important to attract more researchers and practitioners to this important field to stimulate worldwide collaboration and coordination in the direction.



Metagenomics for COVID-19

Metagenomics [1]-[14] represents a new approach in genomic analysis. This method accesses the potential reservoirs of novel genes in wider applications. To explore this reservoir, RNA genomes from SARS-CoV-2 samples are collected from thousands of COVID-19 patients worldwide. In addition to Koch's Postulate [15, 16] for pathogenicity of SARS-CoV-2, it is necessary to have an initial investigation of existing metagenomic technologies.

Using advanced metagenomic analysis methodologies, many statistical and computational tools and databases for metagenomics have been developed [20], such as functional and sequence-based analysis of the collective microbial genomes "full shotgun metagenomics" [21] and polymerase chain reaction (PCR) amplification of certain genes of interest "marker gene amplification metagenomics" (i.e., "full shotgun metagenomics" [21], 16S ribosomal RNA gene) or "meta-genomics" [22]. Two typical operations linked between next generation sequence data and data management storage and sharing are summarized in Fig 1.

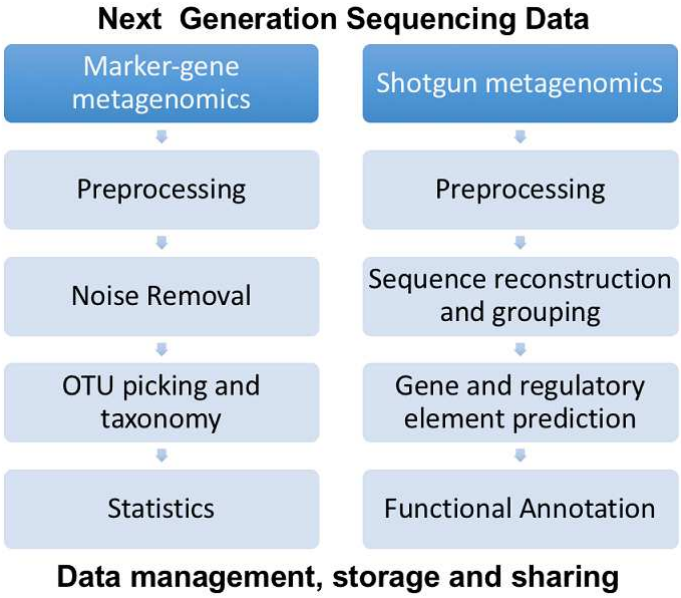


Fig. 1 Two workflows of analysis processes in metagenomics on next generation sequencing

Considering this advanced research direction on specific driving applications, it is worth further examining refined facts to understand specific advantages and weaknesses in existing metagenomic analysis tools to handle RNA virus genomes of SARS-CoV-2 in general.

Analysis Tools for Metagenomics

From a computational viewpoint, many analysis tools have been developed. Tools, such as EULER[23], Velvet[24], SOAP[26] and Abyss [27] were among the first to perform de novo assembly and are still widely used today.

The next generation of assembly tools, such as MetaVelvet/MetaVelvet-SL [25] and Meta-IDBA [28] create more accurate assemblies from datasets containing a mixture of multiple genomes. They use k-mer frequencies to detect kinks in the de-Bruijin graph and then use these k-mer thresholds to decompose the graph into subgraphs.

The IDBA-UD algorithm [29] addresses metagenomic sequencing technologies with uneven sequencing depths involving complicated processes for various k-mers in both low-depth and high-depth regions in multiple levels of hierarchy.

Binning is the process of grouping (binning) reads or contigs into individual genomes and assigning the groups to specific species, subspecies or genera. Binning methods are characterized in two different ways: 1) individual genomes have

a unique distribution of k-mer sequences (*genomic signature*); 2) similarity- or homology-based binning refers to BLAST or profile hidden Markov model pHMMs to obtain similarity information in available databases (NCBI or PFAM).

A list of tools have been developed, such as TETRA [30], S-GSOM, PhylophythiaS, ⋯, and ClaMS. Other similarity-based softwares are CARMA, MetaPhyler and SOrt-ITEMS. Some tools employ similarity-based binning algorithms in their mategenomics analysis piplines, such as IMG/MER 4, MG-RAST and MEGGAN.

Hybrid Approaches

Certain tools use a hybrid approach to employ both composition and similarity-based information to group sequences, such as PhymmBL [31] and MetaCluster [32] that cross a series of metagenomic samples, facilitating the assembly of micro genomes without the need for reference sequences.

Further Binning tools are characterized with category operations on 1) *ab initio* unsupervised classifiers and 2) supervised/training-based classifiers. For example, ClaMS is a classifier for metagenomic sequences [38].

Unsupervised binning refers to the process of pre-existing bins to classify a given dataset without user supervision. In contrast, supervised binning allows user interface and supervision in the training process per se. In general, Support Vector Machines SVM (PhylopythiaS), hidden Markov Models hMD (PhymmBL, TETRA) and Self Organizing Maps SOM (ESOMs) provide unsupervised classifiers, and PhylopythiaS and TETRA allow little user intervention, while ClaMS and ESOM provide a more supervised training approach to allow optimal classification for the specific dataset under consideration.

Optimal binning results are expected to combine both composition- and similarity-based approaches by hybrid tools, such as PhymmBL and MetaCluster.

Parallel Schemes

Advanced systems are Parallel-META: efficient metagenomic data analysis based on high-performance computation [33], MEGAN analysis of metagenomic data [34, 35], and Galaxy: a web-based genome analysis tool for experimentalists [36].

In relation to the identification of genes within the reads/assembled contig, genes are labeled as coding DNA sequence CDSs and noncoding RNA genes, and certain annotation piplines (e.g., IMG/MER) predict regulatory elements, such as clustered regularly interspaced short palindromic repeats (CRISPRs).

CDSs are identified by MetaGeneMark, Metagene, ⋯, and FragGeneScan. CRISPR elements are identified by CRT, OILER-CR and IMG/MER to retain the longest element prediction in case of overlap. Noncoding RNAs are predicted by tRNAscan.

Ribosomal RNA rRNA genes (5s, 16s and 23s) are predicted by rRNA models for IMG/MER and MG-RAST using similarity to compare three databases (SILVA, Greengene and the Ribosomal Database Project-RDP) to predict rRNA genes.

To predict protein coding genes, due to the large size of metagenomic datasets, very expensive computationally and highly automated operations are performed. BLAST or other sequence-similarity-based algorithms run on high-performance computer clusters under multithreading or other parallel computational approaches to divide jobs in multiple central/graphic processing units (CPUs/GPUs) to reduce the running time complexity and speed up execution time. Data repositories are metagenomic databases, such as KEGG, SEED, eggNOG, and COG/KOG. PFAM and TIGRFAM are protein domain databases.

BLAT (BLAST-like alignment tool) identifies the best homologs of those genes in the isolated genomes. It misses similarity below 70% identity, and many strong hits to other genes are missed.

The EBI metagenomics service is a web-based portal to use metadata structures and formats with the genomic standards consortium GSC guidelines. EBI uses FragGeneScan to obtain protein coding sequences. CAMERA is an online cloud computing service for the analysis of metagenomic data. MEGAN 5 is another tool to perform analysis of metagenomic data and offers a wide range of visualization tools for metagenomic annotation results to support multiple visualization schemes: functional or taxonomic dendrograms, tag clouds, bar charts and Krona taxonomic plots that allow hierarchical data to be explored in a zoomable pie chart.

Taxonomic analysis for prokaryotes (i.e., bacteria and archaca) is regularly performed using 16S data derived from sequencing technologies. In this area, due to the vast availability of algorithms and software for the analysis of 16S metagenomics datasets, QIIME seems to be established as the “gold standard”.

Statistical Analysis and Visualization of Results

Different tools provide comprehensive representation of a taxonomic tree to be visualized in applications, such as FigTree and a file in Biological Observation Matrix BIOM format representing OTU tables. Numerous tools and software packages are performing statistical analysis. The Primer-E package allows multiple multivariate statistical analyses, such as Multi-Dimensional Scaling MDS, analysis of similarities ANOSIM, and hypothesis testing. Using R statistical programming language, packages, such as Vegen, Phyloseq and Bioconductor provide multiple inbuilt functions and libraries to support a wider range of statistical analysis required for metagenomic datasets to thoroughly analyze visualization tools for genomic datasets.

Storage, Sharing and Minimum Information

Tools,such as IMG/MER, CAMERA, MG-RAST and EBI metagenomics provide an integrated environment for the analysis, management, storage and sharing of metagenome projects. For refined applications, minimum information about a metagenome sequence MIMS and minimum information about a MARKer sequence MIMARKS is devised to provide a scheme of standard languages for metadata annotation.

Difficulties in Virus Analysis, Big Data Visualization and Others

Analysis usually requires a reference database to find the closest match to an operational taxonomic unit (OTU) from a taxonomic lineage. Existing databases are less suitable for certain groups of organisms, such as protists and viruses which are extremely diverse and for which considerably less sequence information is available compared to bectoria.

Considering the special importance of Koch's postulate in the period of genomics [15, 16], it is necessary to find proper techniques to resolve this type of difficulty.

Facing a sea of biological data everywhere, Science magazine in 2005 asked the top 125 scientific problems [40] in

Problem 17: How will big pictures emerge from a sea of biological data?

In the current situation, it is truly a top challenge to generate meaningful pictures that emerge from a large number of biological datasets especially from genomes.

Variation results can occur due to inconsistencies in a number of factors, such as DNA extraction, primer pair and amplification region, sequencing platform and software used. Various variations make several difficulties compare and obtain trustworthy results. Through benchmarks, simulations and testing, an initiative would eliminate at least minimization, and biases can be generated by analyzing data using multiple methodologies.

In comparative metagenomic analyses, one can use tools to compare samples from ecological niches and extract information in common and/or unique to a specific environment.

New Genome Datasets on GISAID and Nextstrain Projects

GISAID - International Sharing Influenza Virus Sequences

The GISAID [18] initially promotes the international sharing of all influenza virus sequences, related clinical and epidemiological data associated with human viruses and geographical as well as species-specific data associated with avian and other

animal viruses to help research understand how the viruses evolve, spread and potentially become pandemics.

Up to April 25, 2020, over 12K viral genome sequences of hCoV-19 (SARS-CoV-2) shared unprecedented speed via GISAID. One application is shown in Fig 2.

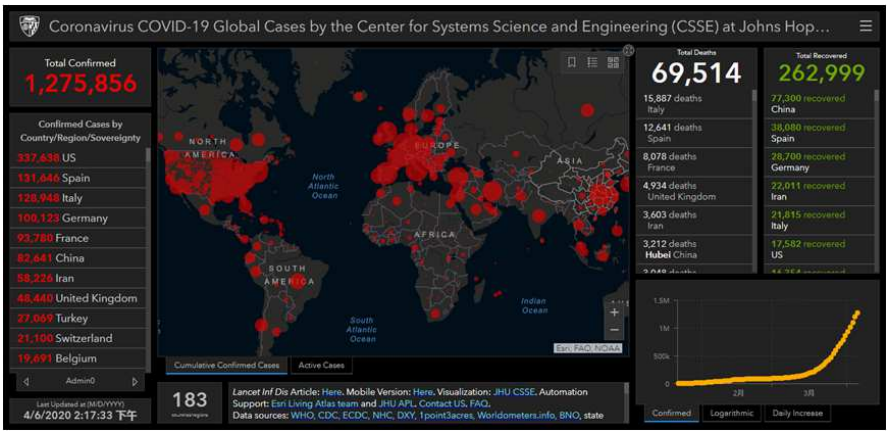


Fig. 2 COVID-19 Global Cases on Johns Hopkins University Website. Simulations and technical support by Nextstrain + GISAID

Nextstrain

Nextstrain is an open-source project [17] 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 shown in Fig 3.

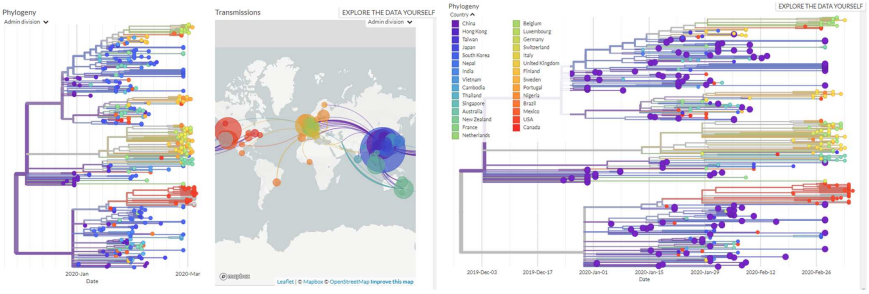


Fig. 3 Phylogeny of real cases over global on Nextstrain

Phylogeny Organization on Selected Root for COVID-19

The GISAID TreeTool uses the nextflu pipline to construct and visualize the tree. The pipline consists of an initial approximate maximum likelihood phylogeny reconstruction using FastTree, followed by refinement with RaXML (GISAID TreeTool in v2.0).

If the initial sequence is the origin sequence, there is no problem to make it the root and consequent new samples are gradually added in the tree in branches and intermediate nodes in a proper cluster based on the maximum likelihood relationship. However, if the first sequence is not a truly origin sequence and there would be potentially stigma [19] for COVID-19, then it is necessary for a fair whole phylogeny to change the root for the proper phylogeny.

From a comparison of machine-learning mechanisms, advanced TreeTool needs to be enhanced with both a semi- or full-supervised learning scheme and unsupervised schemes to organize datasets into multiple levels of hierarchy first and then based on well balanced clustering distributions to establish suitable phylogeny construction for further explorations. It is essential for the system to allow relocations of the root node consistently [99, 102].

The Newest Fighting Fields on CODID-19

Facing the complicated spread of COVID-19 worldwide, scientists and researchers in many countries are corporate to fight this type of invisible virus, and a set of research papers were published in wider areas, such as from SARS to MERS [41], possible origins of SARS-CoV-2 from {Bats, Pangolin ...}[42, 43, 44, 46, 49, 55, 66], human ACE2 receptors [45, 77], and the most important questions [47].

From an evolutionary viewpoint, studies have focused on genome composition and divergence [48] and evolution [50], from SARS to SARS-CoV-2 [51].

Successful practices were published as a summary report on 72314 patients in China on COVID-19 [52], fear versus data [53], patient pneumonia in China [54], full-genome evolutionary analysis [56], and global threat [57].

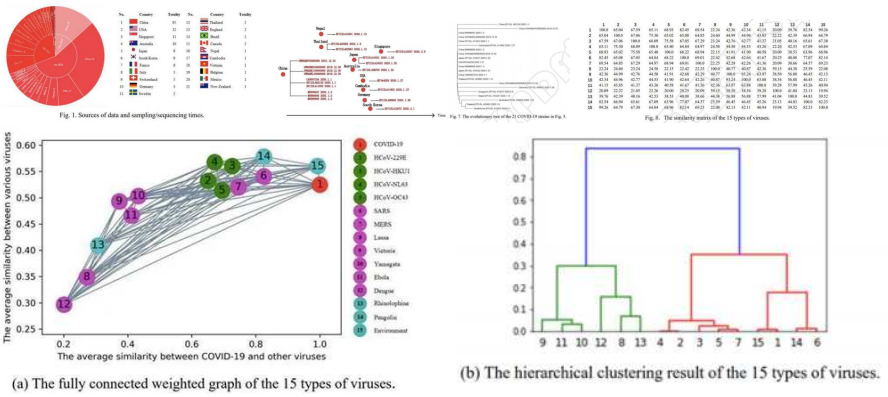
The first sets of genomes of SARS-CoV-2 were collected by new genomic samples {Wuhan, USA, Italy, Korea, India, Australia, North Italy} [58, 59, 60, 61, 62, 63, 64].

Simulation models and results were investigated in computational inference and evolution [67, 68], computers and viral diseases [69], identification [70], evolutionary perspective [71], biased codon usage [72], and variation analysis [73].

In relation to future development, interesting works have focused on recent trends [74], insight [75], animal to human transmission [76], classification [78], and HIV-1 [79].

Focusing on variation properties, a list of studies investigated genomic variance [80], emerging viruses [81], structure replication and pathogenesis [82], cross-species transmission [83], and virus evolution [84].

Applying similarity comparison, complex tree structures were investigated in similarity and evolutionary relationships [85]. The main steps in the processes are shown in Fig 4.



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

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

Fig. 4 Main steps of multiple coronaviruses on similarity networks [85]

The new 19 variations of SARS-CoV-2 genomes were identified in patient-driving mutations [86] with significant variation in cytopathic effects and viral load, up to 270-fold differences observed. The conclusion of such variations will strongly influence further medical practices in the treatment of COVID-19 patients.

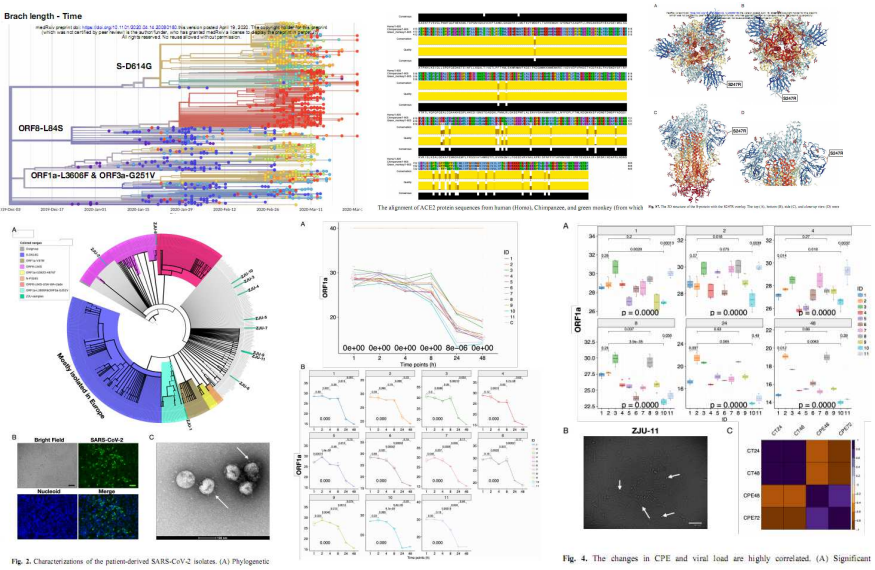


Fig. 5 Key steps of the phylogeny used in the treatment of COVIP-19 patients from [86]

The identification processes are briefly shown in Fig 5 to obtain initial information on genomic variations from Nextstrain's phylogeny, selected genomic samples through metagenomic analysis, 3D structural analysis, extractions, complicated biological quantitative measurements and comparisons for the results.

Variant Construction

Based on vector logic, modern matrix theory, geometric measure theory, combinatorial algebra and discrete mathematics [87]-[94], 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 [95]-[102], 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.

This novel theoretical construction provides a solid foundation of multiple hierarchies on multiple probability distributions and invariants to support the metagenomic analysis system from concept levels, design and engineering implementation. Since all relevant transformations and variations could be represented as maps, we emphasize our attention in this paper to show the base structure of this visual framework for SARS-CoV-2 genomes in a series of visual maps.

Aim of The Study

Different from specific targets in existing metagenomics, the metagenomic analysis system MAS focuses on exploring general information from collections of whole genomic sequences intrinsically included in virus RNA genomes on SARS-CoV-2 samples. Multiple distributions, curves and surfaces are illustrated and relevant quantitative measurements - invariants are represented as genomic indices to be mapped into a restricted geometric measurement region to support category, clustering, classification activities to view whole collections of genomes on variant maps. This powerful mapping mechanism can be further explored to resolve any types of big data collections for categories and content-based indexing to provide supersymmetric properties to manage giant data collection over the world to be a unified thermodynamic scheme under relevant entropy schemes. Further explorations are required.

In this paper, the hierarchical architecture of the MAS framework is briefly described without workflows and core equations. For readers interested in more techni-

cal information, please check the third paper of this special issue for detailed workfolows and main equations described.

Materials and Methods

Architecture of Metagenomic Analysis System MAS

The architecture of the metagenomic analysis system (MAS) is composed of three projections: A) global projection; B) clustering projection; C) genomic index projection shown in Fig. 6, and three projections are shown in Fig. 7 (a)-(c).

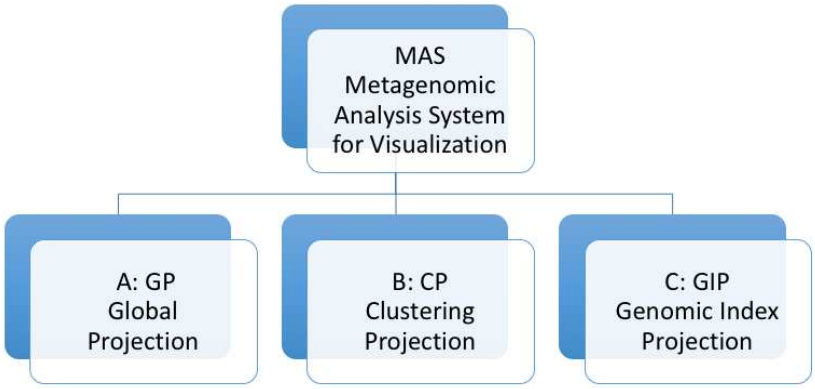
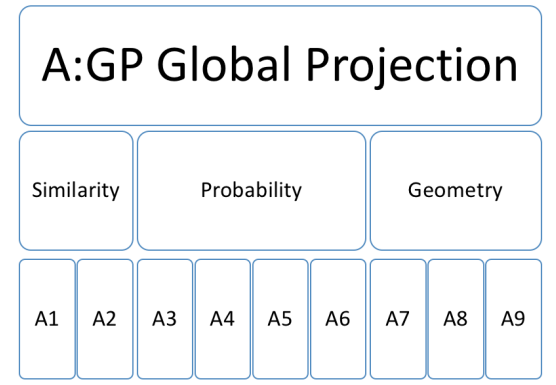


Fig. 6 Architecture of metagenomic analysis system MAS in three projections

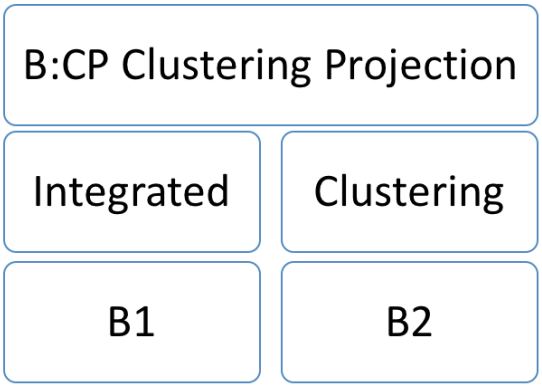
Three Projections

A: Global Projection

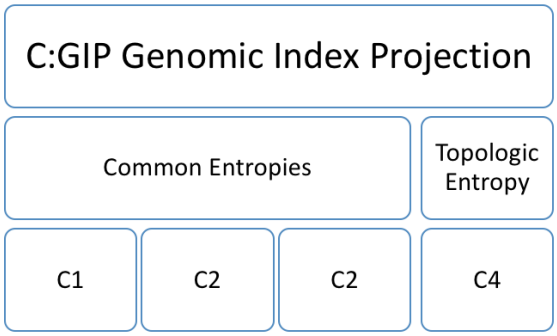
Global projection is composed of nine modules: $A := \{A_1, \cdots, A_9\}$ as follows.



(a)



(b)



(c)

Fig. 7 Three projections in the MAS (a) nine GP modules (b) two CP modules, and (c) four GIP modules

A_1: 1D similarity	1DS	1D
A_2: 2D similarity	2DS	1D-2D
A_3: multiple and conditional probability	MCP	1D
A_4: protein coding	PC	1D
A_5: 2D to 1D linearizing	1DL	1D
A_6: momentum distribution	MD	1D
A_7: whole/parts Map	WPM	2D
A_8: K-mer process	KP	2D
A_9: 3D visual	3DV	2D-3D

B: Clustering Projection

Clustering projection is composed of two modules: $B := \{B_1, B_2\}$ as follows.

B_1: integrated density matrix	IDM	(256+1)	2D
B_2: clustering density matrix	CDM	(256+1)	2D

C: Genomic Index Projection

Genomic index projection is composed of four modules: $C := \{C_1, \cdots, C_4\}$ as follows.

C_1: combinatorial entropy	CE	scale
C_2: integrated entropy	IE	scale
C_3: mean entropy	ME	scale
C_4: topological entropy[39]	TE	scale

Fifteen Modules in MAS

As a list of fifteen functional modules in the MAS is described, their function, description, visual, and mode are shown in Table 1.

Because a series of visual maps are involved, it is convenient for readers to use relevant names for visual illustrations for specific purposes in science, technology, medicine, social activities and daily life.

Datasets

All datasets used in this special issue from various open source genomic banks, such as CNGBdb: Virus Database & SARS-CoV-2 Database; Database Commons; NCBI GenBank SARS-CoV-2-seqs and GISAID + GitHub + Nextstrain et al. More than two thousand whole sequences of SARS-CoV-2 over 200 countries have been

Table 1 Fifteen Modules in the MAS

Module	Function	Description	Visual	Mode
A_1:	1DS	similarity between two seq in k-mers	1D curves	16
A_2:	2DS	similarity among seg. parts in k-mers	$M \times M$ maps	32
A_3:	MCP	multiple and conditional probability	1D histograms	10
A_4:	PC	protein coding on seg. parts	1D curves	16
A_5:	1DL	from 2D to 1D linearizing transformation	1D curves	256
A_6:	MD	various momentum distributions	1D curves	16
A_7:	WPM	partition n parts in whole/part map	2D maps	256×2^n
A_8:	KP	one seg. on K-mer process	2D maps	$256 \times K$
A_9:	3DV	transforming 2D map to 3D visual	2-3D maps	256 3D trans.
B_1:	IDM	for m, $256 \times (m+1)^2$ density matrices & integration	2D maps	1
B_2:	CDM	features in $256 \times (m+1)^2$ density matrices & integration	2D maps	$2^{(m+1)^2}$
C_1:	CE	one of 16 selections on relevant entropy	scale	16
C_2:	IE	all 16 selections in an integrated entropy	scale	1
C_3:	ME	all 16 entropies for a mean entropy	scale	1
C_4:	TE	connectivity linkages among an entropy distribution	scale+1D	1D

collected mainly from the NCBI & GISAID site under the Nextstrain project from January to March 2020. In addition, a set of coronaviruses, H1N1 virus, Bats and Pangolins, MERS, Ebola, SARS and other sequences were collected as samples for comparisons.

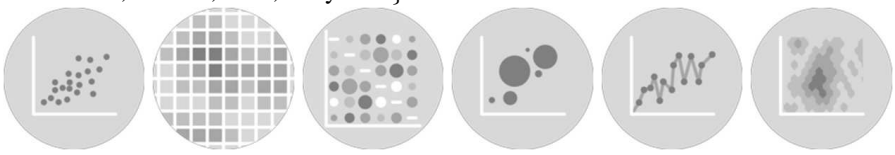
Typical samples are listed in Table 2.

Table 2 Earlier Datasets of SARS-CoV-2 & Other Cases Worldwide

Sample	No.	Collected Date	Locality
SARS-CoV-2	(2019-nCoV)	Earlier date	
	MN908947	2019-12	China
	LR757995	2019-12-26	China:Wuhan
	NC_045512	2020-01	China:Wuhan
	MN938384	2020-01-10	China:Shenzhen
	LC529905	2020-01	Japan
	MN985325	2020-01-19	USA
	MT007544	2020-01-25	Australia:Victoria
	MT012098	2020-01-27	India:Kerala State
	LC534418	2020-02-14	Japan
Normal coronavirus			
HCoV-HKU1	NC_006577		HK
HCoV-NL63	NC_005831		NL
HCoV-OC43	NC_006213		OC
HCoV-229E	NC_002645		
Beta coronavirus			
MERS-CoV	JX869059		
PDCoV	KX022602, KX022605		
SARS	AY2741193, NC_004718.3		

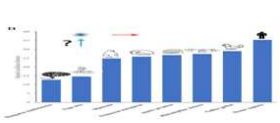
Names of Visual Maps

In metagenomic applications, there are many visual maps involved. It is convenient to make a list of six maps with common names of relevant maps: {Scatter, Heat, Correlation, Bubble, Line, Greyscale}.

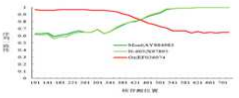


Scatter Heat Correlation Bubble Line Greyscale

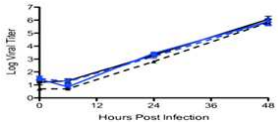
In addition, different maps were used in genomic analysis applications. The following eight maps are shown in {1D Histogram, 1D Curves, 1D Lines, 2D Heat + Tree categories; 2D Clustering, 2D Color, Taxonomic Tree, 3D Structure}.



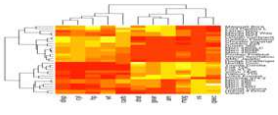
1D Histogram



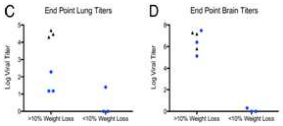
1D Curves



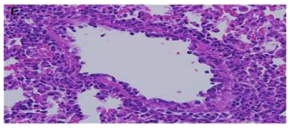
1D Lines



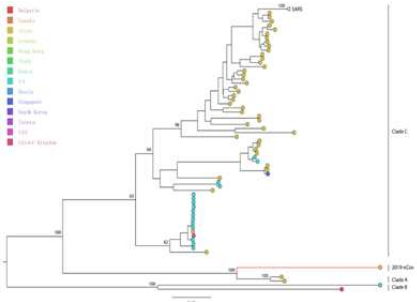
2D Heat + Tree



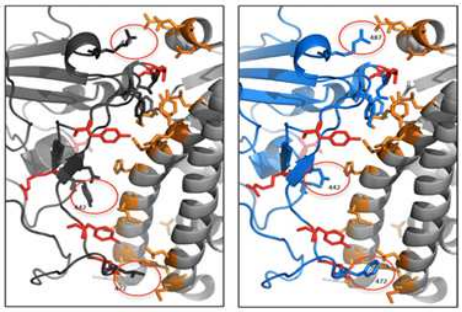
2D Clustering



2D Color



Taxonomic Tree



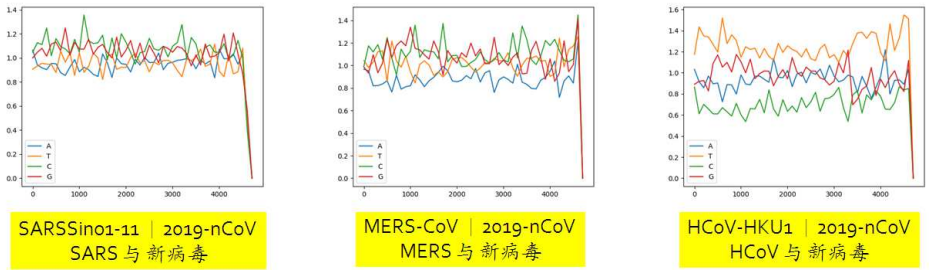
3D Structure

Results and Discussion

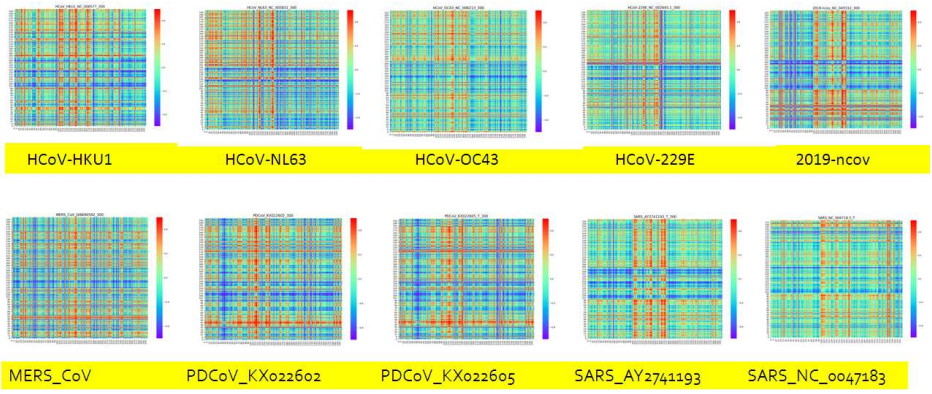
It is convenient to select one sample map for each module in the MAS framework.

Results of Global Projection Groups

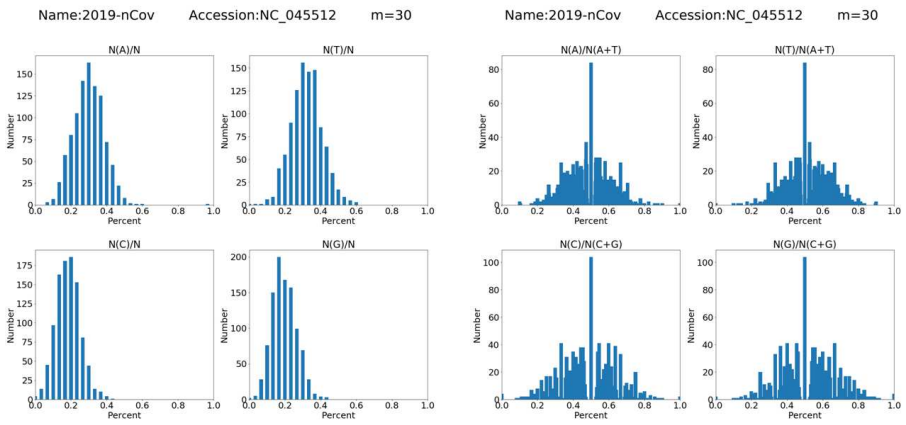
Nine modules {A1,..., A9} are selected to be represented as a set of visual maps as follows.



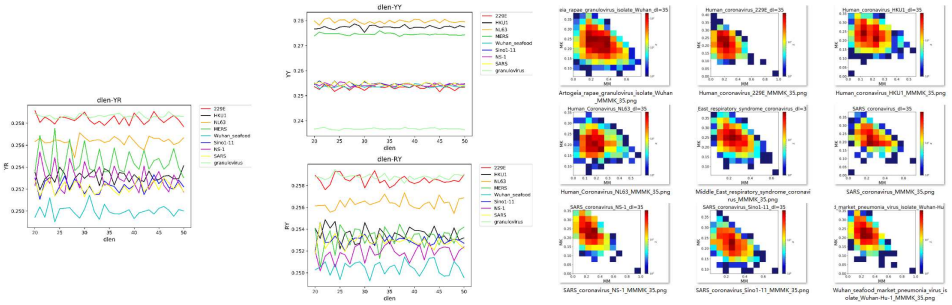
A1. SARS-CoV-2: SARS and others



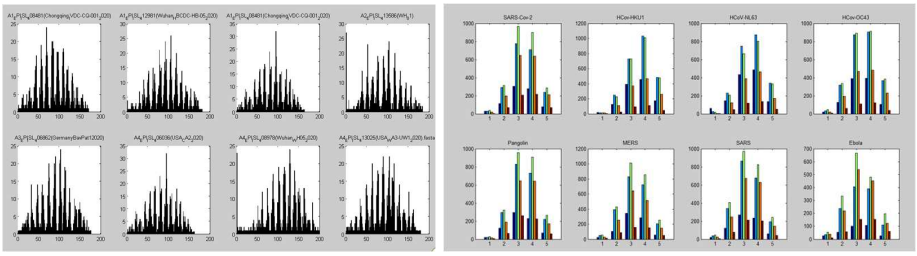
A2: 2D similarity maps for SARS-CoV-2



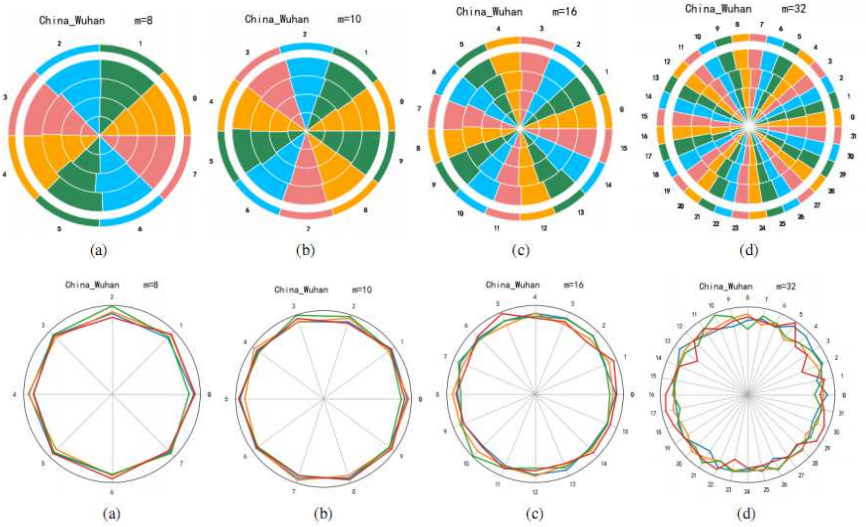
A3. multiple and conditional maps of SARS-CoV-2



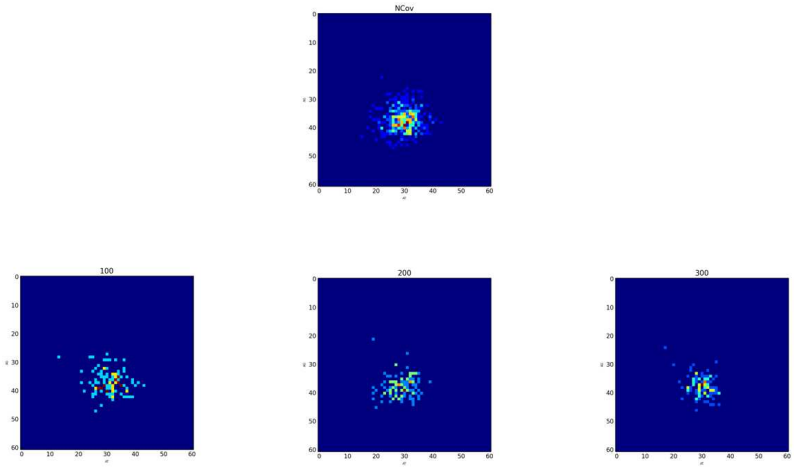
A4. 1D protein coding of SARS-CoV-2 2D SARS-CoV-2: others protein coding



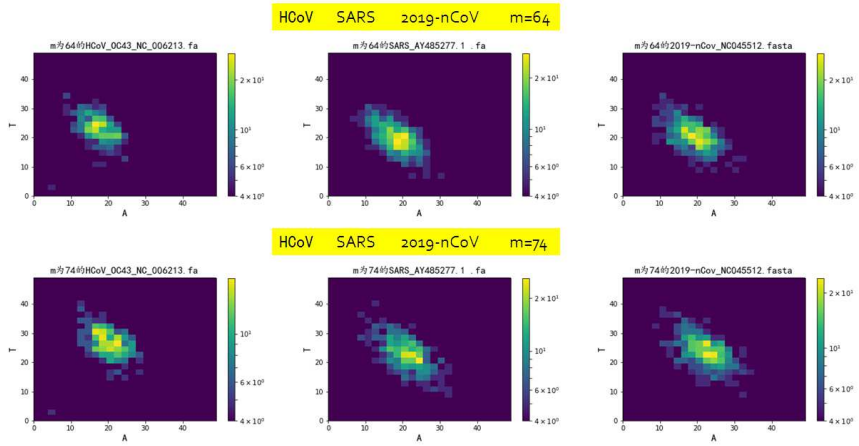
A5. CoV-2:2D in linearized K=16 CoV-2:2D in linearized K=4



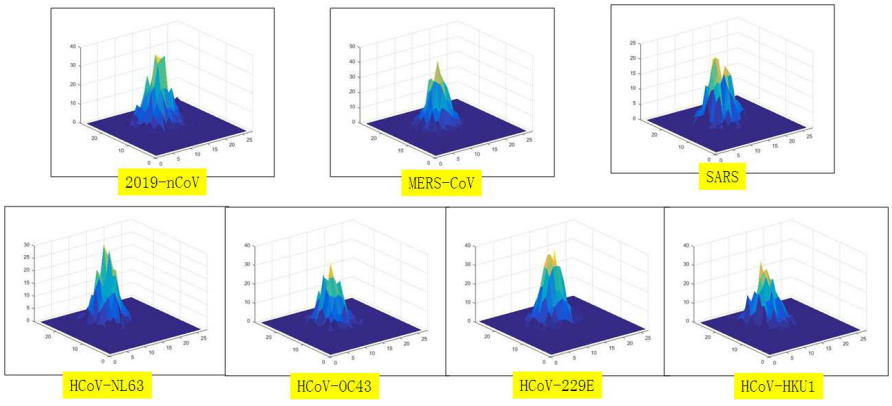
A6: momentum probability of SARS-CoV-2



A7. SARS-CoV-2:whole/parts



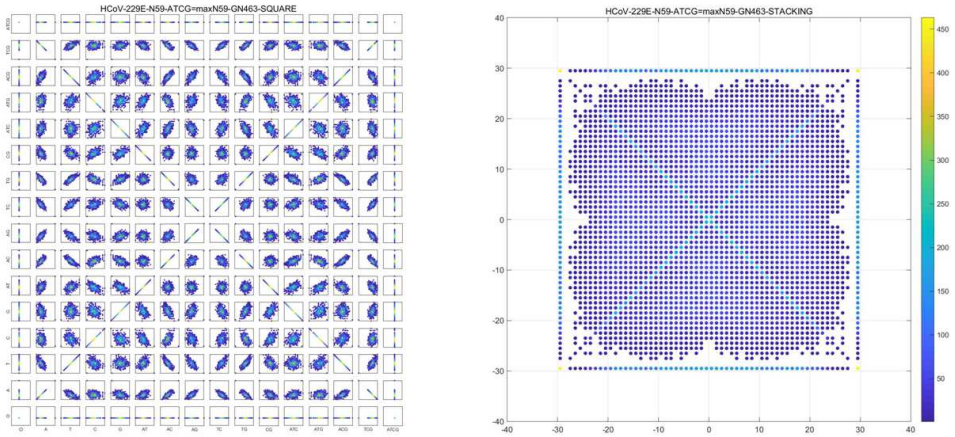
A8: K-mers of SARS-CoV-2



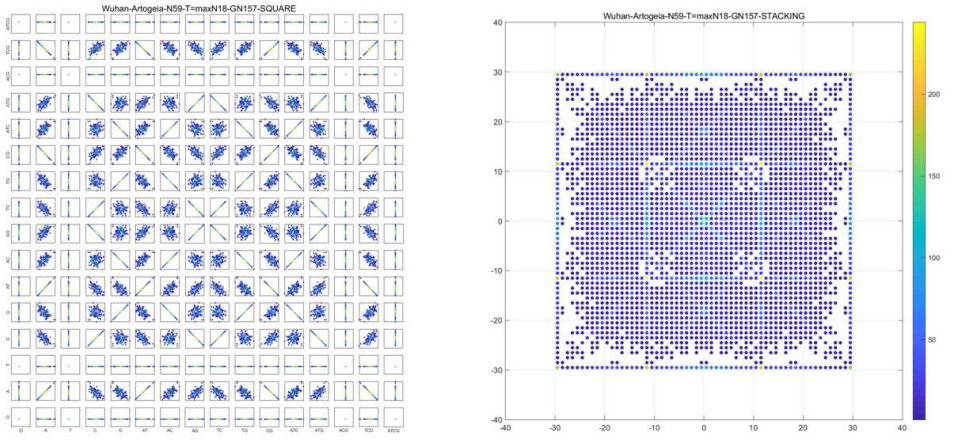
A9. SARS-CoV-2 and others:3D visualizations

Results of Clustering Projection Groups

Two modules {B1,B2} are selected to be represented as a set of visual maps as follows.



B1. SARS-CoV-2 on integrated maps

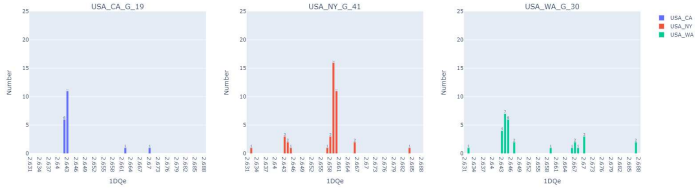


B2. SARS-CoV-2 on clustering maps

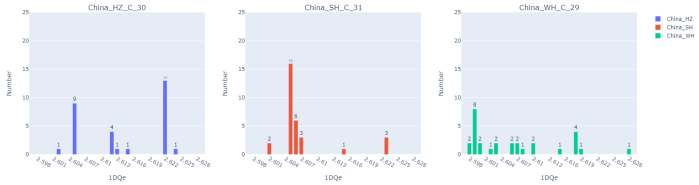
Results of Genomic Index Projection Groups

Four modules {C1,..., C4} are selected to be represented as a set of visual maps as follows.

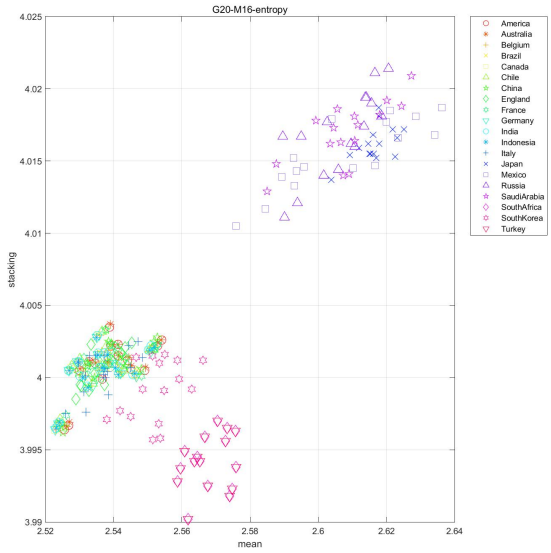
C1. SARS-CoV-2 of 2D genomic indices on combinatorial entropy maps in four countries



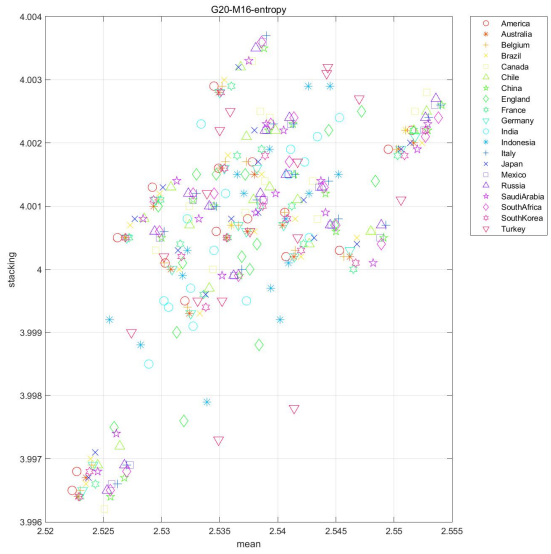
C1. SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in the USA



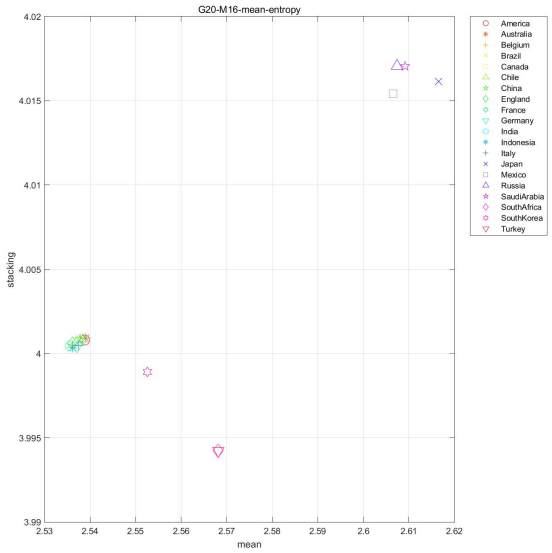
C1. SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in the China



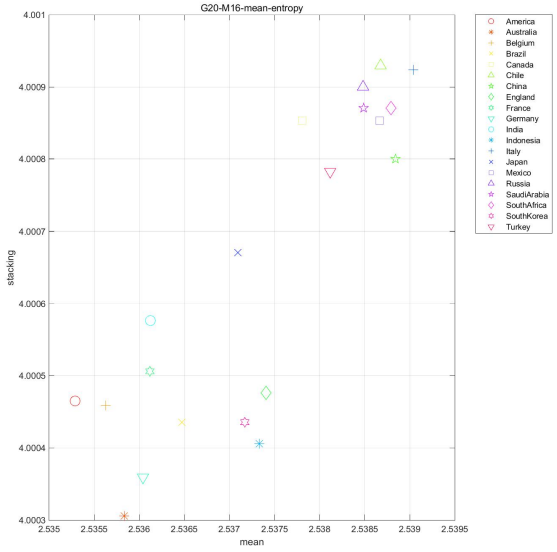
C2:C3. SARS-CoV-2 of genomic indices on integrated entropy maps



Enlarged 100 times for C2:C3. SARS-CoV-2 of genomic indices on integrated entropy maps

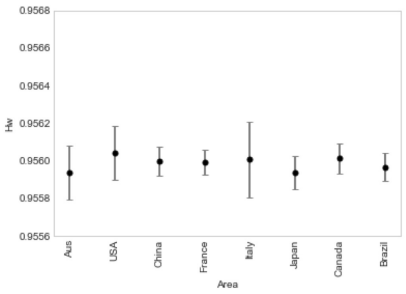


C2:C3. SARS-CoV-2 of genomic indices on entropy mean maps



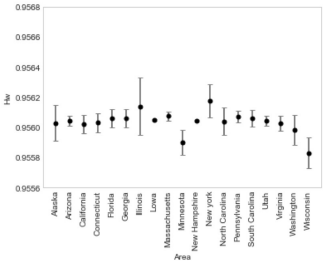
Enlarged 100 times for C2:C3. SARS-CoV-2 of genomic indices on mean entropy maps

不同国家拓扑熵误差棒图

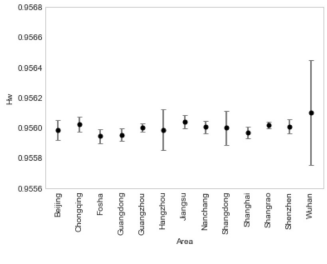


C4. SARS-CoV-2 of genomic indices on 1D topological entropy maps in eight Countries

美国不同地区拓扑熵误差棒图

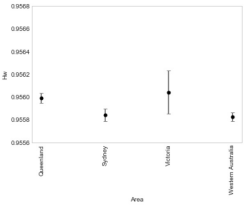


中国不同地区拓扑熵误差棒图

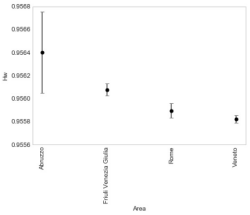


C4. SARS-CoV-2 of genomic indices on 1D topological entropy maps in the USA and China

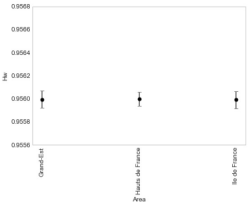
澳大利亚误差棒图



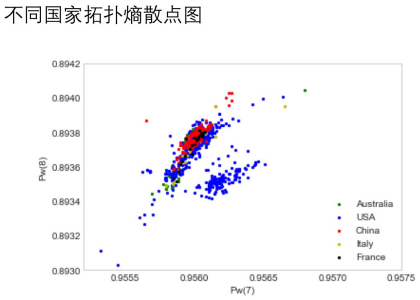
意大利误差棒图



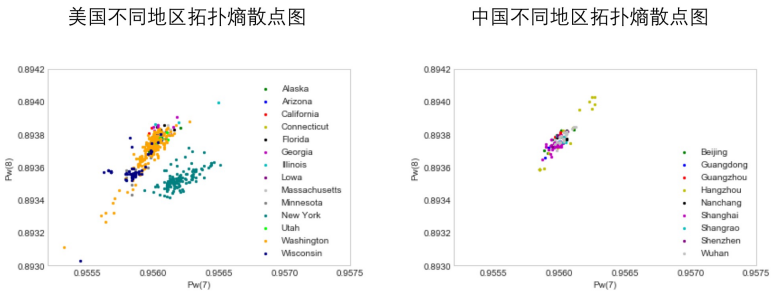
法国误差棒图



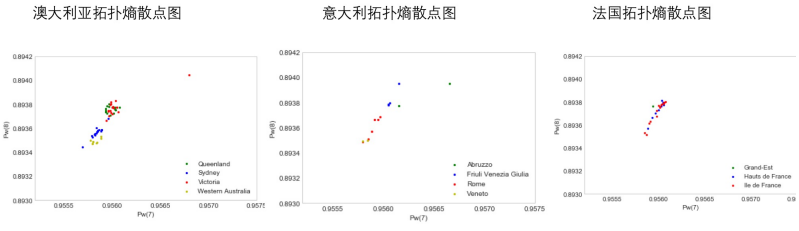
C4. SARS-CoV-2 of genomic indices on 1D topological entropy maps in the Australia, Italy and France



C4. SARS-CoV-2 of genomic indices on 2D topological entropy maps in eight countries



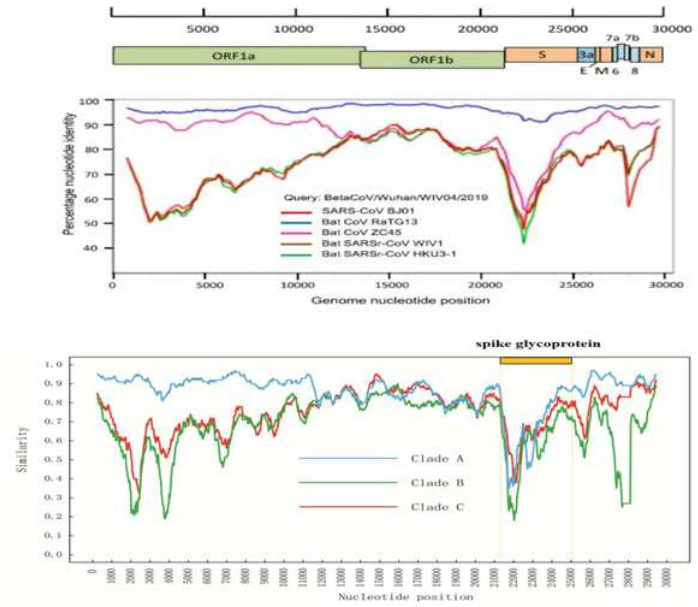
C4. SARS-CoV-2 of genomic indices on 2D topological entropy maps in the USA and China



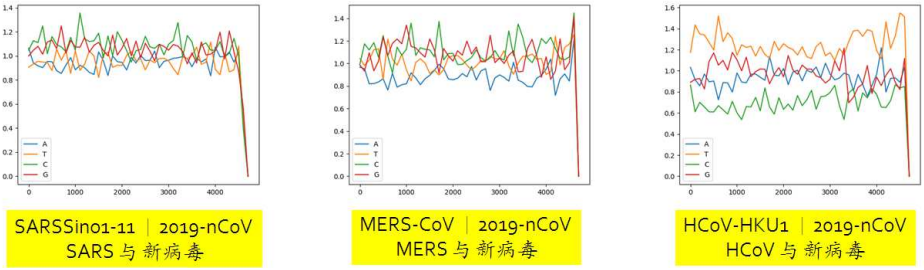
C4. SARS-CoV-2 of genomic indices on 2D topological entropy maps in the Australia, Italy and France

Comparisons

Similarity Comparisons

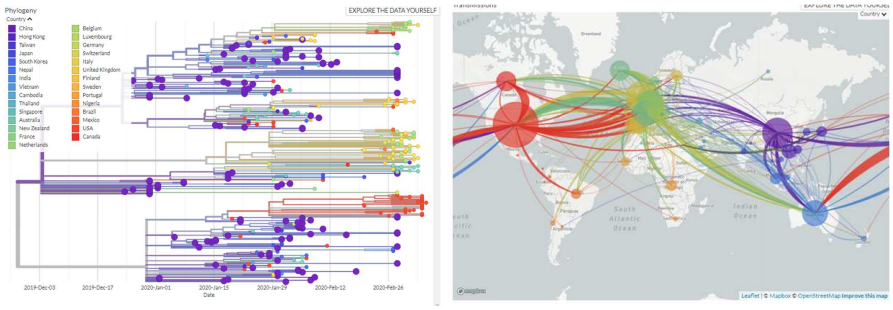


Similarities between CoV-2 and bats & snake virus in published papers [42]



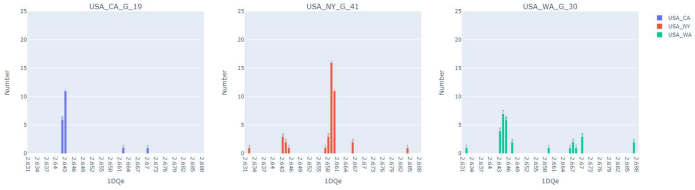
MAS: 1D similarity maps for SARS-CoV-2 and others

Phylogeny & Genomic Index Maps

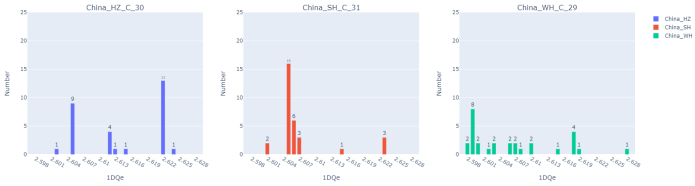


A Nextstrain phylogeny map [17]

C1. SARS-CoV-2 of 2D genomic indices on combinatorial entropy maps in four countries



C1. SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in the USA



C1. SARS-CoV-2 of 1D genomic indices on combinatorial entropy maps in the China

Both Nextstrain phylogeny maps and genomic index maps provide invaluable categorical information. Only relative differences among clusters are contained in the Nextstrain phylogeny with the nearest likelihood relationships in limited discrete evaluating generations.

However, genomic index maps provide invariant position information for all metagenomic sequences on a flexible-scaling region to support infinite evolutions of different variations from foundation levels in general.

Future Explorations

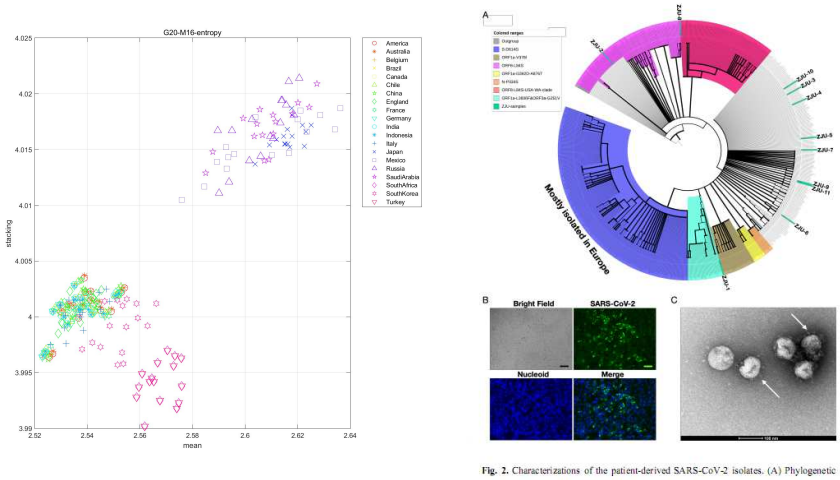


Fig. 2. Characterizations of the patient-derived SARS-CoV-2 isolates. (A) Phylogenetic

From genomic index maps to general medical practices.

Extracting precise categorical information from hierarchical genomic index maps provides invaluable information on the classification of genomic sequences for advanced practices of genomics, transcriptomics, proteomics, metabolomics and phonemics using metamodel organisms representing viruses, plants and animals in addition to medical, pharmacological, pharmaceutical, pathogenic, neurologic and other applications for COVID-19 patients.

Discussion

A series of visual results for three groups of functions A-C are presented. Among groups A & B, it is feasible to transform one or two genomic sequences as distinct probability distributions via relevant schemes. Multiple distributions from 1D to 3D visual maps are illustrated. From one input sequence, corresponding distributions can be generated.

However, the C group provides further integrations to make each sequence as only one point located in a certain position of a restricted geometric region, and further hierarchical scaling into infinite small is possible. Under this superinvariant framework of the information entropy family, global variations of SARS-CoV-2

samples over the world can be clearly identified in one unified map. This provides superpowerful capacities to consistently compare with all SARS-CoV-2 genomes.

There are potentially infinite variations for this type of dynamic system. Further detailed investigations are required.

From relevant results in comparisons, similarity properties can be performed between two genomes. A list of modules provides different transformations to make various distributions and specific genomic indices. The most important invariants of the MAS are four entropy parameters to provide global invariant parameters to describe any types of variations among multiple species of genomes globally.

This is the most important contribution for the MAS to support genomics in general. Further explorations are required.

Conclusion

It is important to have an integrated framework to analyze RNA viruses to overcome intrinsically stronger variations extremely associated with hierarchical time, locations, from micro to macro environments and other complicated conditions to spread, carry, transmit, prevent and detect activities involved.

Based on variant construction of hierarchical organizations from meta levels of analysis to apply quantum thermodynamics and information entropy facilities, it is feasible to transfer various virus genomic sequences as unique sets of genomic indices to organize all relevant information mapped into a restricted geometric region. The foundation of thermodynamic variations and global invariant properties for quantitative characteristics support a universal usefulness of this supersymmetric framework in future exploration.

In the second paper "Input-Output Types of Fifteen Modules on Discrete and Real Measurements for COVID-19" in this special issue, further discussions of relevant input-output types are discussed, and the main equations are described. It is a complementary documentation of MAS for COVID-19.

Since only brief contents are described in this paper, please find refined illustrations with further detailed information on each module in other supporting papers of this special issue. We look forward to obtaining real metagenomic analysis applications for MAS in the near future.

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

Conflict of Interest

No conflict of interest has been claimed.

References

1. FZ Song, Genomics, Military Medical Science Press 2011 (Chinese) 宋方洲，基因组学，军事医学科学出版社 2011
2. C, Saccone and G. Pesole. Handbook of Comparative Genomics, John Wiley & Sons Inc. 2003 C.萨科内，G.佩索莱， 比较基因组学手册，化学工业出版社 2008
3. QY He et al. Research on Functional Protein, Science Press 2012 (Chinese) 何庆瑜 等，功能蛋白质研究，科学出版社 2012
4. S Klusmann, The Aptamer Handbook: Functional Oligonucleotides and Their Applications, John Wiley & Sons Inc. 2005 斯文.克卢斯曼，核酸适配体手册，化学工业出版社 2013
5. ZH Yang, Computational Molecular Evolution, Fudan University Press 2008 (Chinese) 杨子恒，计算分子进化，复旦大学出版社 2008
6. BL Hao, Chaos and Fractals, Shanghai Science and Technology Press 2015 (Chinese) 郝柏林，混沌与分形， 上海科学技术出版社 2015
7. Y Huang, Generating Science of Molecular System, Science Press 2012 (Chinese) 黄原，分子系统发生学，科学出版社 2012
8. Alexander Isaev, Introduction to Mathematical Methods in Bioinformatics, Springer-Verlag 2006 生物信息学中的数学方法引论， 科学出版社 2011
9. Michael Yarus, Life From An RNA World, Harvard University Press 2010
10. P. Baldi & S. Brunak, Bioinformatics The Machine Learning Approach, The MIT Press 2002
11. J. Collado-Vides & R. Hofestadt, Gene Regulation and Metabolism, The MIT Press 2001
12. J. Barciszewski, V. Erdmann, Noncoding RNAs, Kluwer Academic Publishers 2003
13. R. Durbin, S. Eddy, K. Krogh, G. Mitchison, Biological Sequence Analysis, Cambridge University Press 2010
14. R. Durrentt, Probability Models for DNA Sequence Evolution, Springer 2008
15. JX Yan, Koch's Postulate in the Period of Genomics, Universal Science 2013 (Chinese) 严家新 基因组时代的科赫法则， 环球科学 2013 https://huanqiukexue.com/plus/view.php?aid=23170
16. D.N. FREDRICKS & D.A. RELMAN,Sequence-Based Indentification of Microbial Pathogenes: a Reconsideration of Koch's Postulates, CLINICAL MICROBIOLOGY REVIEWS, 9(1) 1996 18-33 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC172879/pdf/090018.pdf
17. Nextstrain Real time tracking of pathogen evolution https://nextstrain.org
18. GISAID: Open access to influenza virus data https://gisaid.org
19. Editorial, Stop the Coronavirus Stigma Now, Nature 580, 165 2020 doi:10.1038/d41586-020-01009-0
20. Oulas et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from Biodiversity studies. Bioinformatics and Biology Insights 2015:9 7588 doi: 10.4137/BBi.s12462.
21. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet. 2004;38:52552.
22. Handelsman J. Metagenomics: spending our inheritance on the future. Microb Biotechnol. 2009; 2(2):138-9
23. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001;98(17):974853.
24. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):8219.
25. Afiahayati, Sato K, Sakakibara Y. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res. 2014;22(1):6977.
26. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24(5):7134.
27. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):111723.
28. Peng Y, Leung HC, Yiu SM, Chin FY. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics. 2011;27(13):i94101.

29. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):14208.
30. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetra-nucleotide usage patterns in DNA sequences. BMC Bioinformatics. 2004;5:163.
31. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic clas- sification with interpolated Markov models. Nat Methods. 2009;6(9):6736.
32. Wang Y, Leung H, Yiu S, Chin F. MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. BMC Genomics. 2014;15(suppl 1):S12.
33. Su X, Xu J, Ning K. Parallel-META: efficient metagenomic data analysis based on high-performance computation. BMC Syst Biol. 2012;6(suppl 1):S16.
34. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):37786.
35. Huson DH, Mitra S. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol Biol. 2012;856:41529.
36. Blankenberg D, Von Kuster G, Coraor N, et al. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010;Chapter 19:.121.
37. Wang Y, Leung HC, Yiu SM, Chin FY. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. Bioinformatics. 2012;28(18):i35662.
38. Pati A, Heath LS, Kyrpides NC, Ivanova N. ClaMS: a classifier for metagenomic sequences. Stand Genomic Sci. 2011;5(2):24853.
39. Koslicki D, Toplological entropy of DNA sequences, Bioinformatics, Vol.27(8) 1061-1067, 2011.
40. E Pennisi, Problem 17: How Will Big Pictures Emerge from a Sea of Biological Data?, Science Vol 309:94 2005 in SCIENCE Magazine: Top 125 Scientific Problems http://science.sciencemag.org/content/sci/309/5731/78.2.full.pdf Science公布全世界最前沿125个科学问题：(问题17: 怎样从海量生物数据中产生大的可视化图片？)
41. Song Z, Xu Y, Bao L, et al. From SARS to MERS, thrusting coronaviruses into the spotlight. Viruses. 2019; 11: 59. https://doi.org/10.3390/v11010059
42. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 Mar;579(7798):270-273. doi: 10.1038/s41586-020-2012-7. Epub 2020 Feb 3. PMID: 32015507
43. Lam, T.T., Shum, M.H., Zhu, H. et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature (2020). https://doi.org/10.1038/s41586-020- 2169-0
44. K.G. Andersen, A. Rambaut, W.I. Lipkin, et al. The proximal origin of SARS- CoV-2. Nat Med (2020). https://doi.org/10.1038/s41591-020-0820-9
45. R. Yan, Y. Zhang, Y. Li, et al. Structural basis for the recognition of SARS-CoV- 2 by full-length human ACE2, Science 27 Mar 2020: Vol. 367, Issue 6485, pp. 1444-1448 DOI: 10.1126/science.abb2762
46. Zhang YZ, Holmes EC. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell. 2020 Mar 26. pii: S0092-8674(20)30328-7. doi: 10.1016/j.cell.2020.03.035. [Epub ahead of print] PMID: 32220310
47. Yuen KS, Ye ZW, Fung SY, Chan CP, Jin DY. SARS-CoV-2 and COVID-19: The most important research questions. Cell Biosci. 2020;10:40. Published 2020 Mar 16. doi:10.1186/s13578-020-00404-4
48. Wu A, Peng Y, Huang B, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe. 2020 Mar 11;27(3):325-328. doi: 10.1016/j.chom.2020.02.001. Epub 2020 Feb 7. PMID: 32035028
49. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020 Feb 22;395(10224):565-574. doi: 10.1016/S0140-6736(20)30251-8. Epub 2020 Jan 30. PMID: 32007145
50. X Tang, C Wu, X Li, et al. On the origin and continuing evolution of SARS-CoV- 2, National Science Review DOI: 10.1093/nsr/nwaa036 2020-03-03

51. K.K. To, I.F. Hung, J.F. Chan, K.Y. Yuen. From SARS coronavirus to novel animal and human coronaviruses, J Thorac Dis, 5 (Suppl. 2)(2013), pp. S103-S108
52. Z. Wu, J.M. McGoogan et al. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention [published online ahead of print, 24 February 2020] JAMA (2020), 10.1001/jama.2020.2648
53. Y Roussel, A Giraud-Gatineau, M Jimeno, et al. SARS-CoV-2: Fear Versus Data, International Journal of Antimicrobial Agents, Available online 19 March 2020, https://doi.org/10.1016/j.ijantimicag.2020.105947
54. N. Zhu, D. Zhang, W. Wang, et al. China Novel Coronavirus Investigating and Research Team, A novel coronavirus from patients with pneumonia in China 2019. N. Engl. J. Med. 382, 727-733(2020). DOI:10.1056/NEJMoa2001017pmid:31978945
55. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr Biol. 2020 Mar 19. pii: S0960-9822(20)30360-2. doi: 10.1016/j.cub.2020.03.022. [Epub ahead of print] PMID: 32197085
56. Paraskevis D, Kostaki EG, Magiorkinis G, et al. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result 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
57. Zheng J. SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat. Int J Biol Sci 2020; 16(10):1678-1685. doi:10.7150/ijbs.45053. Available from http://www.ijbs.com/v16p1678.htm
58. Chan JF, Kok KH, Zhu Z, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect. 2020 Jan 28;9(1):221-236. doi: 10.1080/22221751.2020.1719902. eCollection 2020. Erratum in: Emerg Microbes Infect. 2020 Dec;9(1):540. PMID: 31987001
59. Holshue ML, DeBolt C, Lindquist S, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med. 2020 Mar 5;382(10):929-936. doi: 10.1056/NEJMoa2001191. Epub 2020 Jan 31. PMID: 32004427
60. Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019-nCoV in Italy: Where they come from? J Med Virol. 2020 May;92(5):518-521. doi: 10.1002/jmv.25699. Epub 2020 Feb 12. PMID: 32022275
61. Park WB, Kwon NJ, Choi SJ, et al. Virus Isolation from the First Patient with SARS-CoV-2 in Korea. J Korean Med Sci. 2020 Feb 24;35(7):e84. doi: 10.3346/jkms.2020.35.e84. PMID: 32080990
62. Yadav PD, Potdar VA, Choudhary ML, et al. Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res. 2020 Mar 27. doi: 10.4103/ijmr.IJMR_663_20. PMID: 32242873
63. Caly L, Druce J, Roberts J, Bond K, et al. Isolation and rapid sharing of the 2019 novel coronavirus (SARS-CoV-2) from the first patient diagnosed with COVID-19 in Australia. Med J Aust. 2020 Apr 1. doi: 10.5694/mja2.50569. [Epub ahead of print] PMID: 32237278
64. Licastro D, Rajasekharan S, Dal Monego S, et al. Isolation and full- length genome characterization of SARS-CoV-2 from COVID-19 cases in Northern Italy. J Virol. 2020 Apr 1. pii: JVI.00543-20. doi: 10.1128/JVI.00543-20. PMID: 32238585
65. Shen Z, Xiao Y, Kang L, et al. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. Clin Infect Dis. 2020 Mar 4. pii: ciaa203. doi: 10.1093/cid/ciaa203. [Epub ahead of print] PMID: 32129843
66. Li C, Yang Y, Ren L. Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. Infect Genet Evol. 2020 Mar 10;82:104285. doi: 10.1016/j.meegid.2020.104285. [Epub ahead of print] PMID: 32169673
67. Cagliani R, Forni D, Clerici M, Sironi M. Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2. J Virol. 2020 Apr 1. pii: JVI.00411-20. doi: 10.1128/JVI.00411-20. PMID: 32238584

68. Rehman SU, Shafique L, Ihsan A, Liu Q. Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. Pathogens. 2020 Mar 23;9(3). pii: E240. doi: 10.3390/pathogens9030240. PMID: 32210130
69. Robson B. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Comput Biol Med. 2020 Apr;119:103670. doi: 10.1016/j.compbiomed.2020.103670. Epub 2020 Feb 26. PMID: 32209231
70. Cleemput S, Dumon W, Fonseca V, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics. 2020 Feb 28. pii: btaa145. doi: 10.1093/bioinformatics/btaa145. [Epub ahead of print] PMID: 32108862
71. Malik YS, Sircar S, Bhat S, et al. Emerging novel coronavirus (2019-nCoV)- current scenario, evolutionary perspective based on genome analysis and recent developments. Vet Q. 2020 Dec;40(1):68-76. doi: 10.1080/01652176.2020.1727993. PMID:32036774
72. Kandeel M, Ibrahim A, Fayez M, Al-Nazawi M. From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. J Med Virol. 2020 Mar 11. doi: 10.1002/jmv.25754. [Epub ahead of print] PMID: 32159237
73. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, Zhang Z. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020 Mar 13. doi: 10.1002/jmv.25762. [Epub ahead of print] PMID: 32167180
74. Kannan S, Shaik Syed Ali P, Sheeza A, Hemalatha K. COVID-19 (Novel Coronavirus 2019) - recent trends. Eur Rev Med Pharmacol Sci. 2020 Feb;24(4):2006- 2011. doi: 10.26355/eurrev_202002_20378. Review. PMID: 32141569
75. Ashour HM, Elkhatib WF, Rahman MM, Elshabrawy HA. Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. Pathogens. 2020 Mar 4;9(3). pii: E186. doi: 10.3390/pathogens9030186. Review. PMID: 32143502
76. Xu J, Zhao S, Teng T, et al. Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV. Viruses. 2020 Feb 22;12(2). pii: E244. doi: 10.3390/v12020244. PMID: 32098422
77. Xu H, Zhong L, Deng J, et al. High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. Int J Oral Sci. 2020 Feb 24;12(1):8. doi: 10.1038/s41368-020-0074-x. PMID: 32094336
78. Wassenaar TM, Zou Y. 2019-nCoV/SARS-CoV-2: rapid classification of betacoronaviruses and identification of Traditional Chinese Medicine as potential origin of zoonotic coronaviruses. Lett Appl Microbiol. 2020 May;70(5):342-348. doi: 10.1111/lam.13285. Epub 2020 Feb 28. PMID: 32060933
79. Xiao C, Li X, Liu S, Sang Y, Gao SJ, Gao F. HIV-1 did not contribute to the 2019- nCoV genome. Emerg Microbes Infect. 2020 Feb 14;9(1):378-381. doi: 10.1080/22221751.2020.1727299. eCollection 2020. PMID: 32056509
80. Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. J Med Virol. 2020 May;92(5):522-528. doi: 10.1002/jmv.25700. Epub 2020 Feb 19. PMID: 32027036
81. Liu SL, Saif L. Emerging Viruses without Borders: The Wuhan Coronavirus. Viruses. 2020 Jan 22;12(2). pii: E130. doi: 10.3390/v12020130. PMID: 31979013Gralinski LE, Menachery VD. Return of the Coronavirus: 2019-nCoV. Viruses. 2020 Jan 24;12(2). pii: E135. doi: 10.3390/v12020135. PMID: 31991541
82. Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. J Med Virol. 2020 Apr;92(4):418-423. doi: 10.1002/jmv.25681. Epub 2020 Feb 7. Review. PMID: 31967327
83. Ji W, Wang W, Zhao X, et al. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 2020 Apr;92(4):433-440. doi: 10.1002/jmv.25682. PMID:31967321
84. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. J Med Virol. 2020; 92: 455- 459. https://doi.org/10.1002/jmv.25688
85. Yanni Li, Bing Liu, Jiangtao Cui, et al. Similarities and Evolutionary Relationships of COVID-19 and Related Viruses, https://arxiv.org/pdf/2003.05580.pdf

86. HP Yao, XY Lu, ···, LJ Li, Patient-drived mutations impact pathogenicity of SARS-CoV-2, DOI: https://doi.org/10.1101/2020.04.14.20060160 https://www.medrxiv.org/10.1101/2020.04.14.20060160v2
87. L. K. Hua, *Loo-Keng Hua Selected Papers*, Springer, 1982.
88. L. K. Hua, *Selected Work of Hua Loo-Keng on Popular Sciences*, Shanghai Education Press, 1984 (in Chinese).
89. D. E. Knuth, *The Art of Computer Programming*, Vol. 1, 3rd edition, Addison-Wesley, 1998.
90. D. E. Knuth, *The Art of Computer Programming*, Vol. 4A: Combinatorial Algorithms, Part 1, Addison-Wesley, 2011.
91. F. Morgan, *Geometric Measure Theory*, 4th edition, Elsevier 2009.
92. G.Z. Tu, *Combinatorial Enumeration Methods & Applications*, Science Press,1981 (in Chinese).
93. A. Tucker, *Applied Combinatorics*, John Wiley & Sons, 2007.
94. L. Z. Xu, M. S. Jiang and Z. Q. Zhu, *Combinatorial Mathematics of Computation*, Shanghai Science & Technology Press, 1983 (in Chinese).
95. Z. J. Zheng, A. Maeder, The The conjugate classification of the kernel form of the hexagonal grid, *Modern Geometric Computing for Visualization*, Springer-Verlag, 73-89, 1992.
96. Z. J. Zheng,*Conjugate transformation of regular plan lattices for binary images*, PhD Thesis, Monash University, 1994.
97. 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.
98. 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.
99. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019 https://www.springer.com/in/book/9789811322815
100. 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)
101. Jeffrey Zheng, ResearchGate: http://researchgate.net/pprofile/Jeffrey_Zheng
102. 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。

Figures

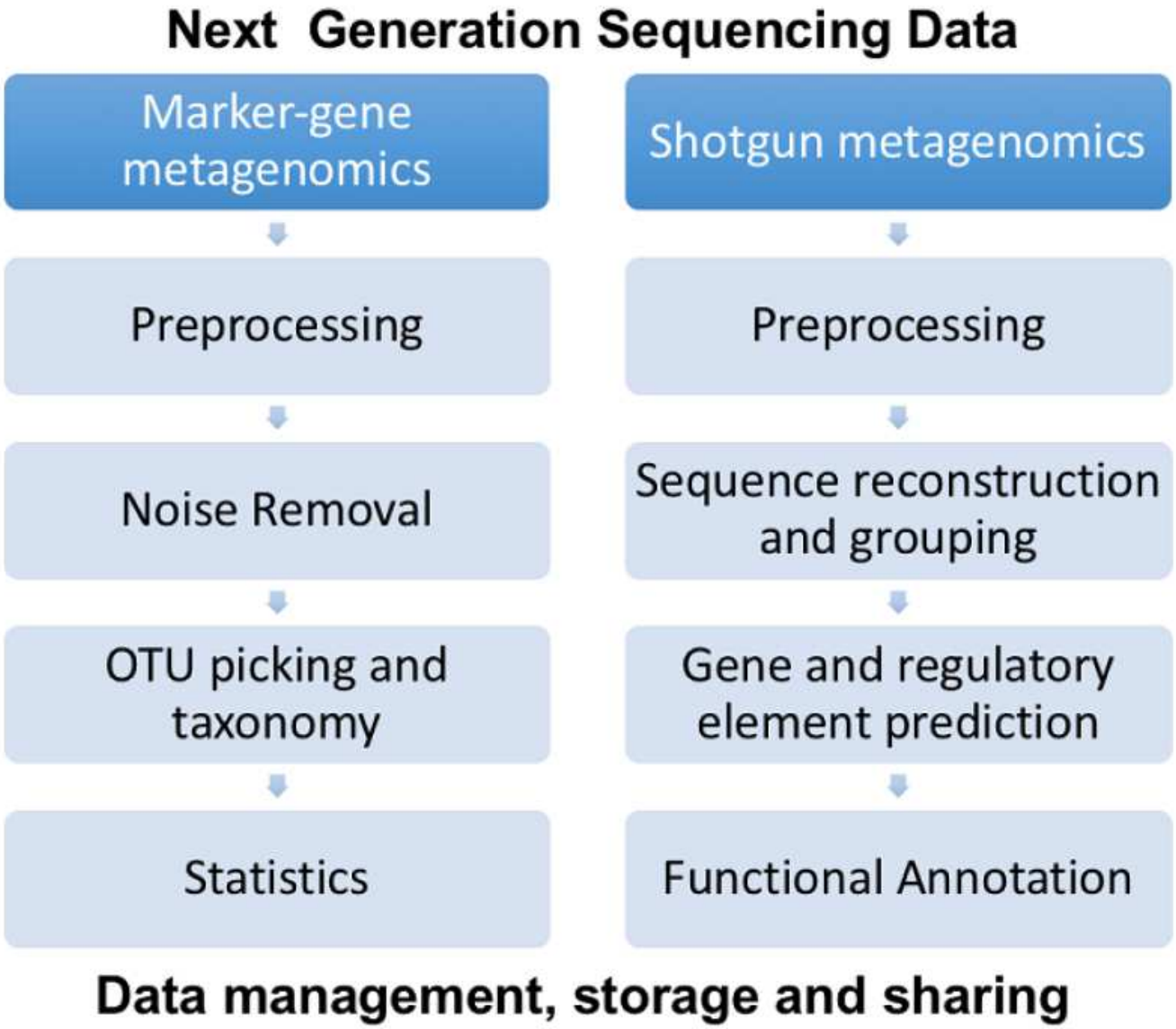


Figure 1

Two workflows of analysis processes in metagenomics on next generation sequencing

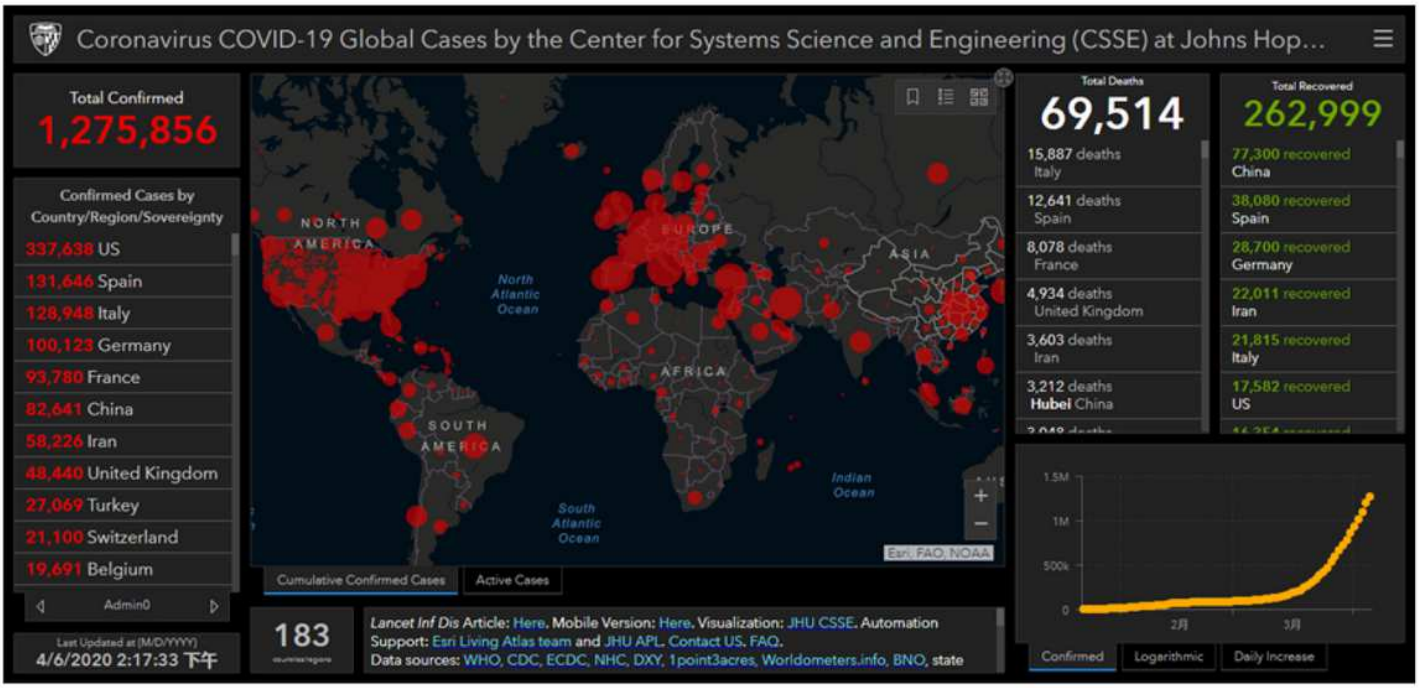


Figure 2

COVID-19 Global Cases on Johns Hopkins University Website. Simulations and technical support by Nextstrain + GISAID

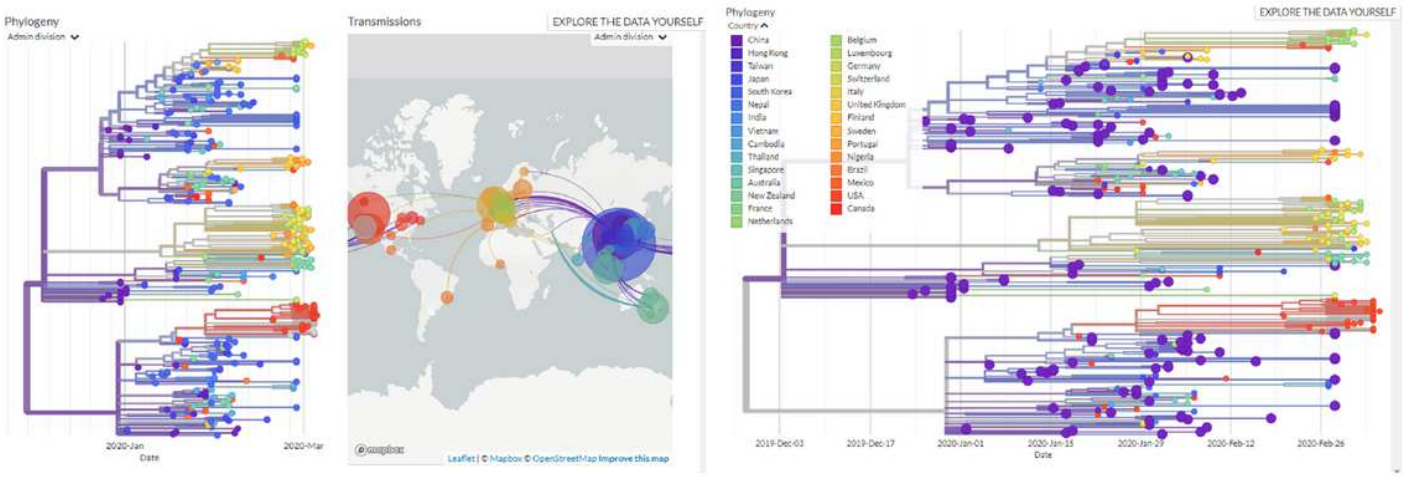
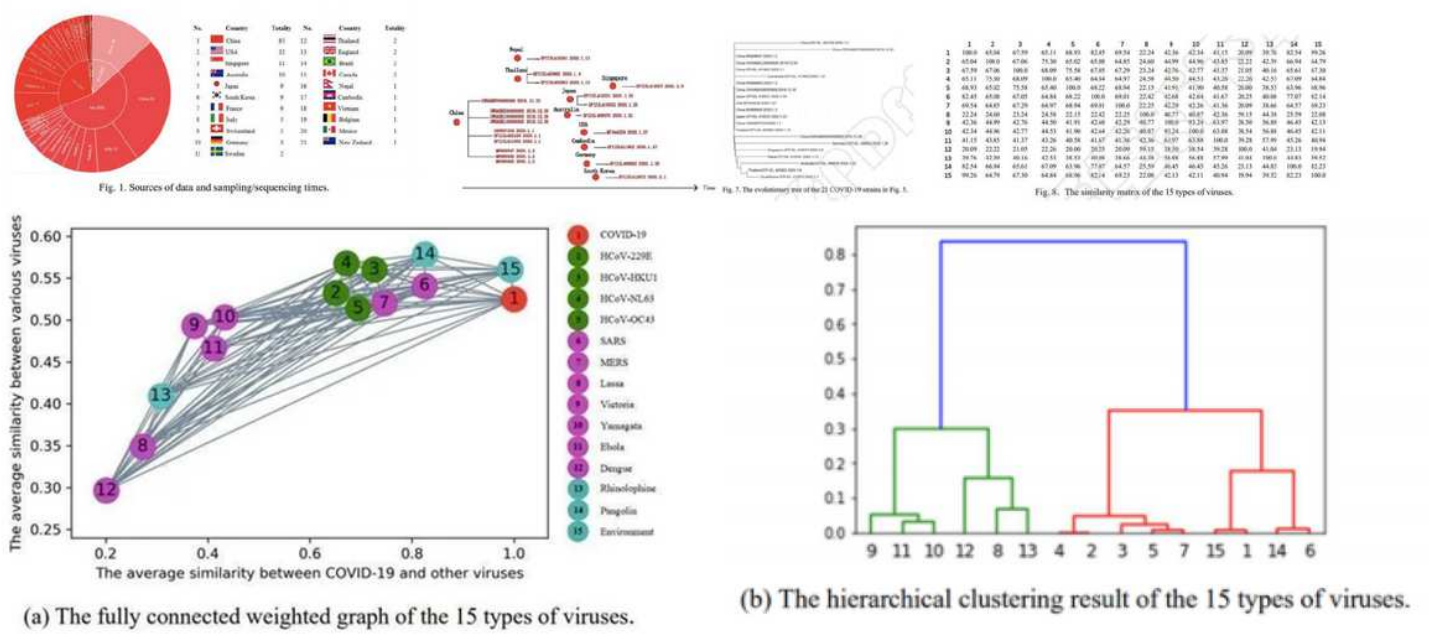


Figure 3

Phylogeny of real cases over global on Nextstrain



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

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

Figure 4

Main steps of multiple coronaviruses on similarity networks [85]

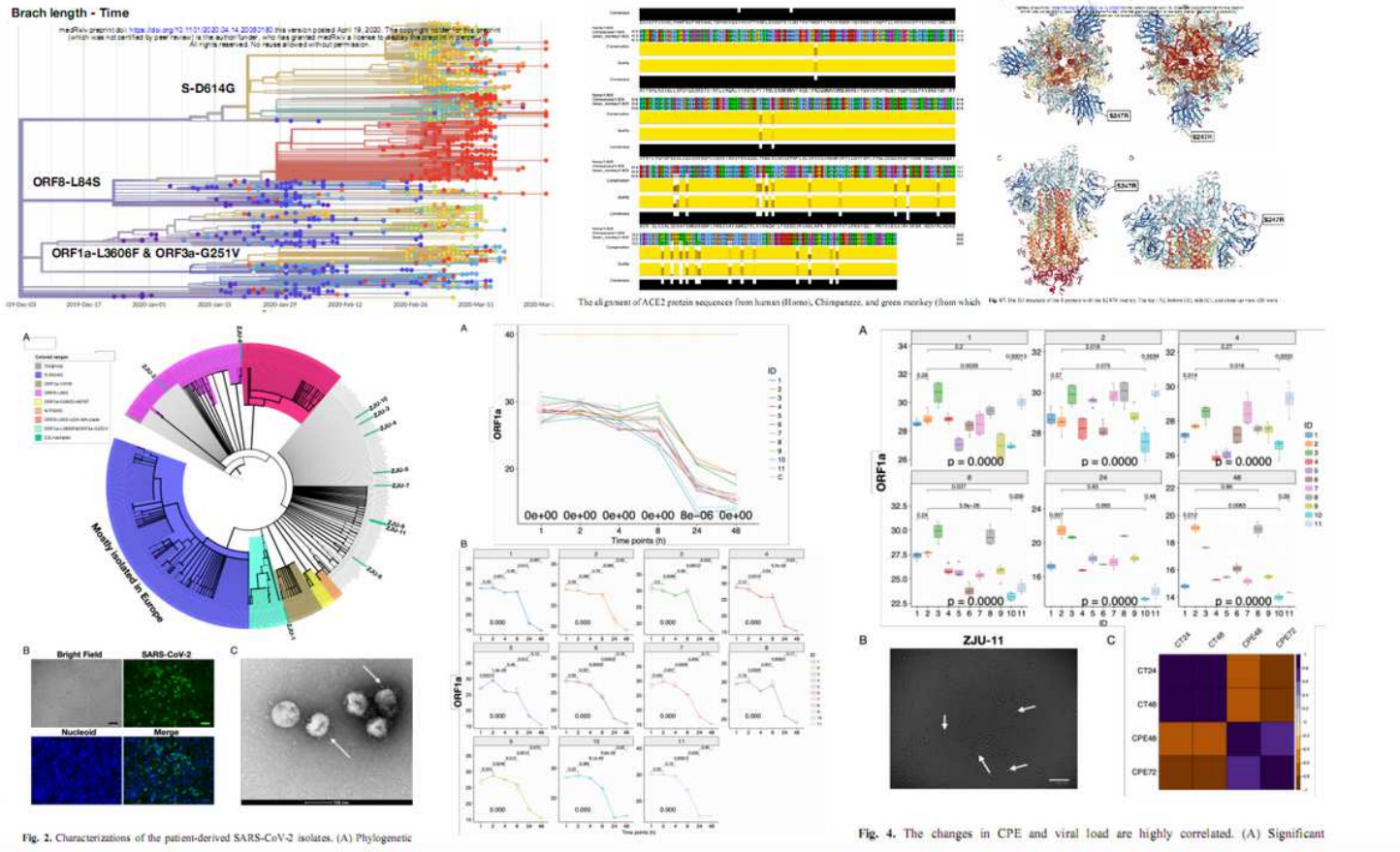


Fig. 2. Characterizations of the patient-derived SARS-CoV-2 isolates. (A) Phylogenetic

Fig. 4. The changes in CPE and viral load are highly correlated. (A) Significant

Figure 5

Key steps of the phylogeny used in the treatment of COVIP-19 patients from [86]

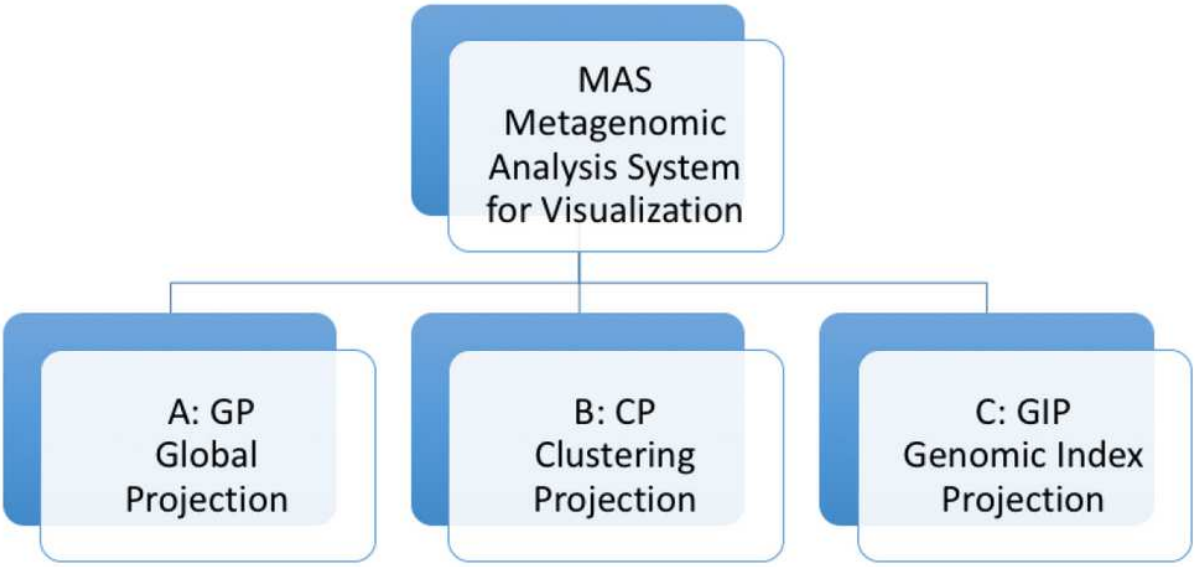


Figure 6

Architecture of metagenomic analysis system MAS in three projections

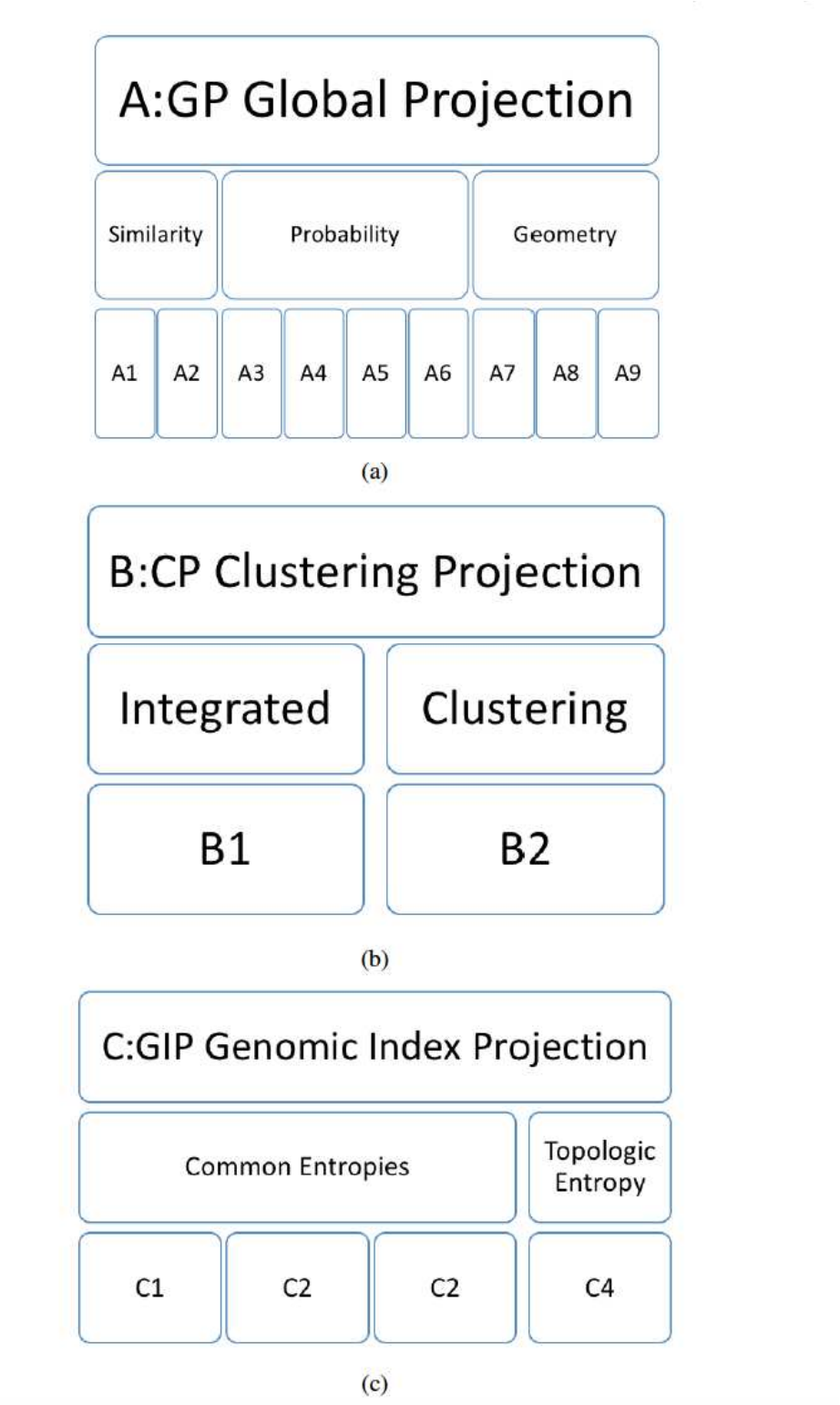


Figure 7

Three projections in the MAS (a) nine GP modules (b) two CP modules, and (c) four GIP modules