Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Science Bulletin 67 (2022) 665–669

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

News & Views

An interactive viral genome evolution network analysis system enabling rapid large-scale molecular tracing of SARS-CoV-2

Yunchao Ling [a,1], Ruifang Cao [a,1], Jiaqiang Qian [a,1], Jiefu Li [a], Haokui Zhou [b], Liyun Yuan [a], Zhen Wang [a], Liangxiao Ma [a], Guangyong Zheng [a], Guoping Zhao [a,c], Zefeng Wang [a,d,e,*], Guoqing Zhang [a,*], Yixue Li [a,c,d,f,g,*]

[a] CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[b] Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[c] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
[d] Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai 200433, China
[e] CAS Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[f] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[g] Guangzhou Laboratory, Guangzhou 510005, China

The comprehensive analyses of the SARS-CoV-2 genomes could provide a global picture of how the virus was transmitted among different populations, which may help predict the oncoming trends of the pandemic. The main approach for the molecular tracing of viral transmission is to thoroughly compare the genomes of different viral strains, leading to a series of phylogenetic trees or evolution networks that can also help to interpret the genomic mutations along with transmission [1–3].

Previously, Lu and colleagues [4] used linkage disequilibrium and haplotype map to analyze genomes of 103 SARS-CoV-2 samples, and classified the viral genomes into type L and S. Similarly, a phylogenetic network was also constructed by Forster et al. using 160 viral genomes, which classified the viruses into three types (A/B/C) based on the nucleotide variants at five genomic loci [5]. However, construction and interpretation of the viral genome evolution network had become increasingly complicated with the rapid accumulation of available viral genomes, and therefore to artificially genotype and cluster the viruses from the network has become almost impossible. To meet the challenge of constructing a single high-quality tree from a huge number of sequences, Rochman et al. [6] developed a compromised method using a "divide and conquer" strategy to analyze over 300,000 SARS-CoV-2 genomes and reconstruct a global phylogeny. In addition, since the construction of the evolution network involves extensive calculations of large matrices, single-threaded analysis tools are unable to complete the analysis within a reasonable time limit in order to continuously track the dynamic changes of viral mutational patterns.

Facing these challenges, we developed a new software pipeline, the viral genome evolutionary analysis system (VENAS), which enables integrated analyses of genomic variations within a newly constructed evolution network. We also adopted a highly paralleled multi-processing strategy to maximize the computation efficiency in calculating the distance matrices. The VENAS further applied a community detection method to transform the evolution network into a two-dimensional isomorphic topological space, and used a network disassortativity trimming algorithm to extract the backbone network of the topological space. The development of VENAS enables researchers to rapidly construct a large-scale evolution network at a reduced time and computational cost, enabling interactive tracing of the viral evolution and core mutations among distinct strains. VENAS is cross-platform open-source software available at https://github.com/qianjiaqiang/VENAS. It can be easily installed using source code or pre-build binary.

To trace viral mutations along transmission routes using daily updated SARS-CoV-2 genomes, we seek to develop reliable computational algorithms to build an integrative genomic analysis system. The resulting system, VENAS, was constructed through a series of steps (Fig. 1). Briefly, we first collected all available SARS-CoV-2 genome datasets with stringent quality-control (Fig. 1a). Each genome was assigned with a series of binary labels (PIS, ν) that identify the parsimonious information sites (PISs) and their corresponding allele frequencies, where a unique set of PISs present a distinct genome type. The PISs with frequencies above a user-defined threshold (i.e., effective PISs, or ePIS) was used to calculate the Hamming distance between distinct genome types [7] (Fig. 1b and Supplementary methods online).

We further sorted the distances between different genome types and sequentially connected all types using the neighbor-joining method, resulting in a fully connected network with the shortest sum of distance (Fig. 1c, the nodes present genome types and the edges present adjusted Hamming distances). The pairs of genome types with the same distance were adjusted by the minor allele frequencies (MAFs) of differential ePISs, where the pairs with lower MAFs (i.e., higher conservation) were given a higher rank

* Corresponding authors.
E-mail addresses: wangzefeng@picb.ac.cn (Z. Wang), gqzhang@picb.ac.cn (G. Zhang), yxli@sibs.ac.cn (Y. Li).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2022.01.001

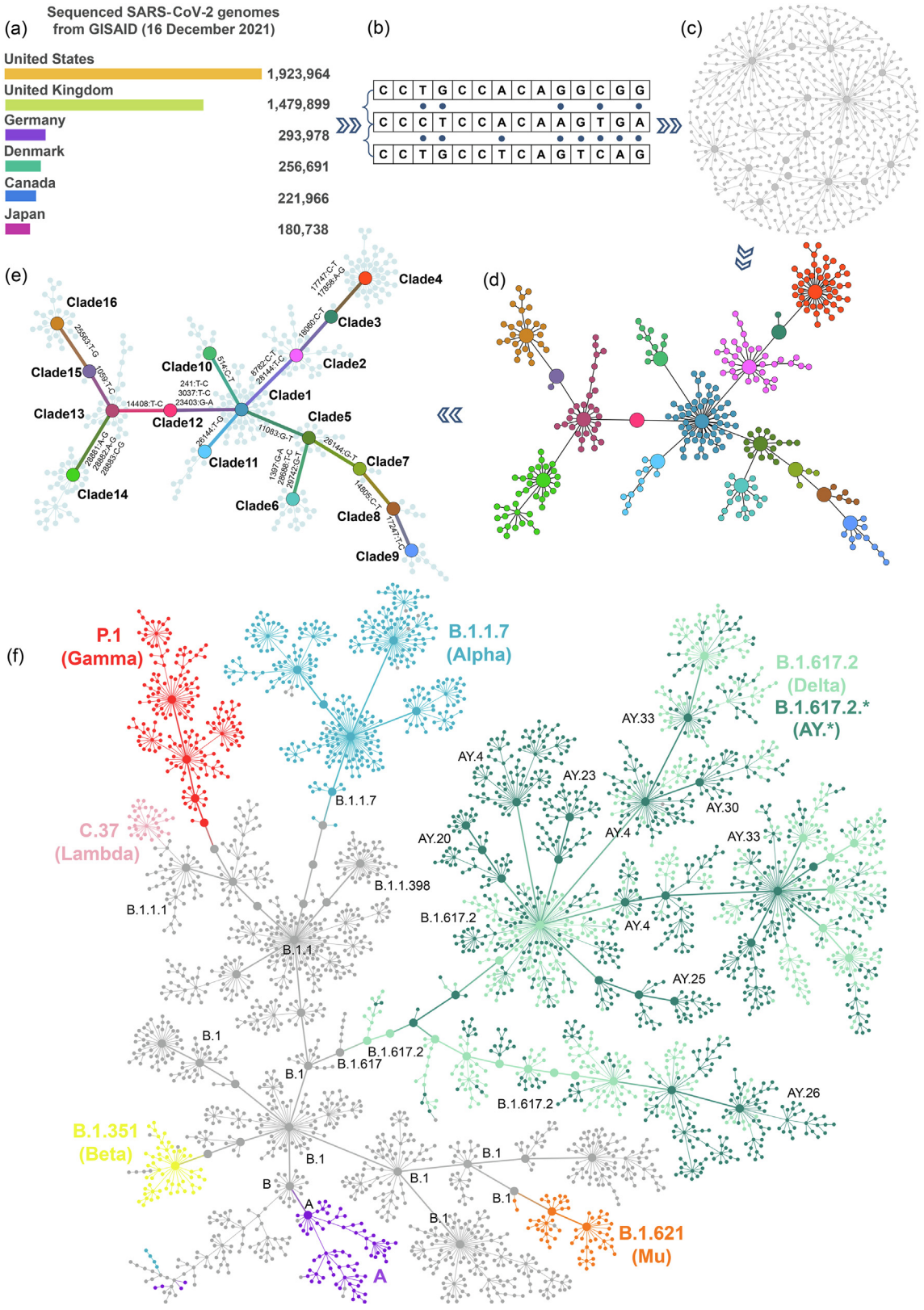

Fig. 1. The construction process and the latest version of the VENAS evolution network. (a) Collecting SARS-CoV-2 genomes from the GISAID database. (b) Generating a distance matrix between different genome types based on the Hamming distance, sorting, and adjusting the distances with MAFs. (c) Constructing the viral evolution network for genomic variation analysis. (d) Topologically clustering the network into subspaces using community detection. (e) Analyzing the evolution network to extract the major transmission paths that consist of core genome types with associated core mutations; the network is clustered into 16 topological clades, with self-increasing ordinal numbered clade names, and labeled links with differential variations between clades. (f) Current VENAS evolution network using fully sequenced SARS-CoV-2 genomes by June 2021. The backbone genome types were labeled by Pangolin lineage. The clade A was first detected during the early pandemic, and four variants of concern (VOCs) were highlighted in blue (Alpha), yellow (Beta), red (Gamma), and orange (Delta).

(see Supplementary methods online). Importantly, this evolution network was constructed with multiple linkages, whereas the conventional phylogenetic trees have bifurcate limitations. Therefore, unlike the phylogenetic tree, we provide additional spatial freedom to the network that is more consistent with the simultaneous mutation accumulation during the viral transmission. Since the number of genomic mutations that occurred during a single human to human transmission is relatively small, this method effectively combined viral genomic alterations with transmission events to reliably construct a network of viral evolution.

We next used a disjoint community detection method to cluster the evolution network into topologically linked subdomains, which represent different evolution clades containing many closely-connected genome types (Fig. 1d). Such segmentation enabled us to subjectively identify the topological clades with "tight" intraclade connectivity and the "sparse" interclade connectivity, which reflect the relationship of different genome types among viral communities formed during natural transmission. Finally, we used the network disassortativity trimming algorithm to extract the core nodes from the evolution network, and further calculated the shortest paths between the core nodes using the Dijkstra algorithm [8], generating a "backbone network" that recapitulates the main mutational paths (Fig. 1e and Supplementary methods online). Many SARS-CoV-2 genome samples also contain associated epidemiological information, including the sampling time, location and protocol, sequencing platform and organization, patient's travel and exposure history, etc., which was further integrated into the VENAS network (by labeling the relevant nodes) to directly infer the viral evolution patterns at critical points of transmission.

The computation of MAF-adjusted Hamming distance matrices is the rate-limiting step during the construction of the viral evolution network of VENAS. To improve the computing efficiency, we took advantage of the multi-core and multi-threaded features in high-performance computers and developed a highly parallel network construction pipeline (Supplementary results online). Compared with previous tools (POPART [9] and Pegas [10]), VENAS achieved a dramatic improvement of computational efficiency and completed the analysis of >10,000 genomes in ~10 min, making it possible to handle the future accumulation of big data that is beyond the capacity of existing analysis tools (Table S1 online). The resulting VENAS network can be visualized with a general relationship graph or force-directed graph tools, such as the web-based Apache Echarts (https://echarts.apache.org/), d3.js (https://d3js.org/), or the application-based Gephi.

Evolution networks have an advantage over phylogenetic trees due to additional spatial freedom of connectivity. We first examined if the VENAS can recapitulate the SARS-CoV-2 haplotype network in an early study using a small number of genomes before March 2020 [4], and confirmed that the community detection method in VENAS can reliably separate the L/S type into two clades with the same core mutations reported earlier (Fig. S1 online). We next applied VENAS to analyze the early stages of global transmission using a dataset containing all 1050 high-quality entries released by 25 March 2020. VENAS produced a "backbone network" with 16 major viral clades in the topological space that reflect the possible transmission paths (Fig. 1e). Each pair of viral clades was separated by core variants that also were identified by comparing PIS sites in these clades (Table S2 online). For example, the core variants C8782T/T28144C recognized as the PIS between clades 1 and 2 (i.e., separating the Clades 2, 3, and 4 from the other clades), which was also reported to distinguish the "type L/S" [4,5]; The variants G11083T between Clades 1 and 5 were first identified from the patients in Diamond Princess Cruise and were also reported by Sekizuka et al. [11]. Interestingly, Clades 1 and 12 were connected by a set of tightly-linked variants (C241T/C3037T/A23403G), which can cause single amino acid substitution in S protein. This mutational event coincided with a single patient who traveled from Shanghai to Bavaria for a business meeting of Webasto [12].

Comparing the VENAS network from different stages of the COVID-19 pandemic, we can trace the occurrence and development process of the variants (Fig. S3 online). The first clade detected in China (Clade 1 or Pango B) showed a much smaller growth in genome type diversity compared to Clade 13 (also known as Pango B.1) or Clade 14 (also known as Pango B.1.1), probably due to the strict confinement of the virus. In contrast, the clades mainly detected in Europe (Clades 12 and 13) mutated into additional genome variants, resulting in many novel types that were clustered into new clades and sub-branches (Clades 14–16), and were widespread in Europe and North/South America. Meanwhile, some of the clades (Clades 5–9, mainly detected on the Princess Diamond cruise) showed a linear sequential branching pattern with a small number of variations detected, likely reflecting some transmission patterns of isolated and localized space. In addition, several variations (e.g., G11083T, C21757T) were detected in multiple viral genome types from different clades, reflecting recurrent mutations independently evolved from multiple places without direct transmission. These genomic regions may be the mutational "hot-spots" that should be closely monitored in the future.

By October 2021, more than 4 million SARS-CoV-2 genomes have been sequenced worldwide, with several variants of concern (VOC) and variants of interest (VOI) showing higher transmission ability than the original type. We applied VENAS to analyze the up-to-date data, and found that VENAS can efficiently generate a comprehensive evolution network with precise annotation of the evolutional relationship between different genome types (Fig. 1f). We also mapped the 4 VOCs and 2 VOIs on the viral evolution network with PANGO nomenclature [13] to trace the mutational routes. The early type A virus named by pangolin first evolved into the type B virus through the S84L variation on ORF8 (first observed around 5 January 2020); then, the type B.1 virus evolved through a series of variations represented by D614G on the S protein (first observed around 28 January 2020). Subsequently, the virus began to mutate in different directions and branched stepwise into many variants, including 4 VOCs and 2 VOIs. The network suggests that the VOCs and VOIs arise independently during transmission from weaker strains, which is in line with the general understanding.

The new network showed that the Delta variant is replacing Alpha as the most dominant viral type globally, with a very large viral diversity and new variants being generated (Fig. 1f dark green). Topologically, the VENAS network directly showed the complex evolutionary relationship between various branches/sub-branches, such as the multilayer relationships between Beta (B.1.351), Alpha (B.1.1.7), and Lambda (C.37 or B.1.1.1.37). The VENAS network can also accurately classify the variants that cannot be identified by existing virus nomenclature software. For example, the variants on the edge of Delta clade were classified as B.1.617.2 but not AY.* by Pangolin software [14]. Moreover, the evolutionary relationships from VENAS have important clinical implications, as the treatment option of new variants can be developed based on the studies from its neighbor or nearby branches.

Previously, the common method to study viral genome variations was through the construction of phylogenetic trees, often using a neighbor-joining algorithm [15]. We made an important improvement in VENAS by changing the binary structure into a multi-dimensional force-directed graph and adjusting the order of ambiguous connections with MAFs. The VENAS effectively clustered the viral genome types into clades by the topological layout of the graph, which reflected the major evolutionary patterns between virus genomes within each clade and provided a basic data model for the study of virus transmission trends.

We filtered PISs with a threshold of variant frequency to obtain ePISs, which increased the confidence level and enabled a direct comparison of MAFs even with sequencing errors or incomplete sequencing. VENAS combined a biological measurement of variation frequency (MAF) with the Hamming distance by assigning the MAF as a weight of Hamming distance, and thus the differential genome types with the maximum conservation were firstly connected to enhance the robustness of the network.

It should be noted that the VENAS network is a non-directional acyclic graph, and thus people should combine the epidemiological information when inferring transmission path from the network. For example, the sampling times in the VENAS network may reveal the direction of virus transmission, the locations of the earliest sample may help to trace the possible origin of the local outbreak, and the sampling organization or sequencing method can help us to exclude potential artifacts and sequence errors. Another caveat during network interpretation is that the "variant reversion" could be mistakenly called in some areas because the neighbor-joining method tends to force a fully connected network despite missing samples. In such a case, a base at a specific locus may appear to be mutated back and forth on two adjacent or nearby edges. The insufficient sampling of some specific genome types can lead to a missing node that is required to construct a coherent path of nodes reflecting the real transmission events. In such a situation, the algorithm will "enforce" the network construction through a neighbor node that is the closest to the real node. This problem may potentially be mitigated by multiple rounds of network reconstruction.

The increasing amount of genome sequences has surpassed the capacity of existing analysis tools such as Pegas and POPART. By optimizing the data analysis process, intermediate data storage structure, and parallel computing efficiency, VENAS can compute millions of sequences simultaneously in a reasonable time (Supplementary material and Table S1 online). The accumulation of new data also enables the comparison of NEVAS networks at different transmission stages, which may allow us to generate a predictive model on how the network is changing along the time and thus to forecast the oncoming trends of transmission. We speculate that the most valuable information will be the topological change of the evolution network, which can be simulated with new algorithms. Such simulation could also be used to identify branches/sub-branches as the potential initiation site for new VOC.

In conclusion, our platform can handle the massive amount of viral genomic data and rapidly generate the evolution network using a highly paralleled computation protocol. The topology-based community detection and the network disassortativity trimming algorithm also enabled the identification of critical viral groups that formed a backbone network of virus transmission. The interactive user interface of VENAS allows us to identify known branches in the early stage of pandemic and detect additional new branches and sub-branches that reflect the transmission trends. A close examination of the latest evolution network shows that many variants, including VOCs, have arisen independently during transmission, implying that new virulent variants may emerge from weaker strains again. As a general platform, VENAS can serve as a valuable tool for data analysis and visualization that supports virus tracing and drug development for further pandemics, as well as for retrospective analysis of other viral pandemics like SARS-CoV or Influenza (see Supplementary results online).

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFC0863300, 2021YFF0703703, and 2020YFC0845900), CAS Strategic Priority Research Program (XDB38060100, XDB38030100, XDB38050000, XDB38040100, and XDC01040100), and Science and Technology Service Network Initiative of Chinese Academy of Sciences (KFJ-STS-QYZD-126). We want to thank GISAID EpiFlu™ Database, the NCBI Genbank Database, and all contributors for selflessly sharing the SARS-CoV-2 sequences and metadata.

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2022.01.001.

References

[1] Kafetzopoulou LE, Pullan ST, Lemey P, et al. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. Science 2019;363:74–7.
[2] Grubaugh ND, Saraf S, Gangavarapu K, et al. Travel Surveillance and Genomics Uncover a Hidden Zika Outbreak during the Waning Epidemic. Cell 2019;178:1057–71.
[3] Chinese SMEC. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 2004;303:1666–9.
[4] Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 2020;7:1012–1023.
[5] Forster P, Forster L, Renfrew C, et al. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA 2020;117:9241–9243.
[6] Rochman ND, Wolf YI, Faure G, et al. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proc Natl Acad Sci USA 2021;118:e2104241118.
[7] Hamming RW. Error detecting and error correcting codes. Bell Syst Tech J 1950;29:147–60.
[8] Dijkstra EW. A note on two problems in connexion with graphs. Numer Math 1959;1:269–71.
[9] Leigh JW, Bryant D. Popart: Full-feature software for haplotype network construction. Methods Ecol Evol 2015;6:1110–6.
[10] Paradis E. pegas: An R package for population genetics with an integrated-modular approach. Bioinformatics 2010;26:419–20.
[11] Sekizuka T, Itokawa K, Kageyama T, et al. Haplotype networks of SARS-CoV-2 infections in the *Diamond Princess cruise* ship outbreak. Proc Natl Acad Sci USA 2020;117:20198–201.
[12] Böhmer MM, Buchholz U, Corman VM, et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. Lancet Infect Dis 2020;20:920–8.
[13] Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5:1403–7.
[14] O'Toole Á, Scher E, Underwood A, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol 2021;7:veab064.
[15] Doolittle WF. Phylogenetic classification and the universal tree. Science 1999;284:2124–8.

Yunchao Ling is a researcher at Bio-Med Big Data Center of Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences (CAS). He graduated from Beijing Institute of Genomics, CAS with a Ph.D. degree in 2014. He is mainly engaged in research and development activities like multiple omics database construction, artificial intelligence-based biomedical knowledge mining and integration, biological data analysis and visualization in the fields of human phenomics, precision medicine, and clinical healthcare.

Ruifang Cao is working as an engineer at Bio-Med Big Data Center of Shanghai Institute of Nutrition and Health, CAS. She received her M.S. degree from East China Normal University in 2015. Her research mainly focuses on the collection, management, and visulation of multi-omics data.

Jiaqiang Qian is an engineer at Bio-Med Big Data Center of Shanghai Institute of Nutrition and Health, CAS. He graduated from Fudan University with M.S. degree in 2018. His current research interest involves the development of multi-omics data processing methods and bioinformatics algorithms efficiency improvement.

Zefeng Wang is a researcher at Shanghai Institute for Nutrition and Health, CAS. He obtained his Ph.D. degree in biological chemistry from Johns Hopkins Medical School. His main research interest is to study the regulation of RNA processing in eukaryotic cells using a combination of computational and experimental approaches.

Guoqing Zhang is the vice director and principal investigator of Bio-Med Big Data Center of Shanghai Institute of Nutrition and Health, CAS. He received his Ph.D. degree from Huazhong University of Science and Technology. His research interest includes bioinformatics database and knowledge base, focusing on the integration and mining of omics data, literature data, and clinical data in the fields, such as precision medicine, large population cohort, the development of personalized drugs, microbiome, and synthetic biology.

Yixue Li is currently the Director of Bio-Med Big Data Center, and visiting professor of Shanghai Institute of Nutrition and Health, CAS. He is also working in Guangzhou Laboratory as the Strategic Consulting Scientist. He received his Ph.D. degree from Heidelberg University (Germany). His research interest focuses on bioinformatics and systems biology, cancer genomics, biomedical database construction and data analysis algorithms, disease animal model genomics, and precision medicine.