2019nCoVAS: Developing the Web Service for Epidemic Transmission Prediction, Genome Analysis, and Psychological Stress Assessment for 2019-nCoV

Ming Xiao, Guangdi Liu, Jianghang Xie, Zichun Dai, Zihao Wei, Ziyao Ren, Jun Yu, and Le Zhang

Abstract—Since the COVID-19 epidemic is still expanding around the world and poses a serious threat to human life and health, it is necessary for us to carry out epidemic transmission prediction, whole genome sequence analysis, and public psychological stress assessment for 2019-nCoV. However, transmission prediction models are insufficiently accurate and genome sequence characteristics are not clear, and it is difficult to dynamically assess the public psychological stress state under the 2019-nCoV epidemic. Therefore, this study develops a 2019nCoVAS web service (http://www.combio-lezhang.online/2019ncov/home.html) that not only offers online epidemic transmission prediction and lineage-associated underrepresented permutation (LAUP) analysis services to investigate the spreading trends and genome sequence characteristics, but also provides psychological stress assessments based on such an emotional dictionary that we built for 2019-nCoV. Finally, we discuss the shortcomings and further study of the 2019nCoVAS web service.

Index Terms—2019-nCoV, COVID-19, epidemic prediction models, LAUPs (lineage-associated underrepresented permutations), psychological stress assessment, genome analysis

1 Introduction

The COVID-19 epidemic, caused by the pathogenic virus 2019-nCoV, is still expanding around the world and poses a serious threat to human life and health [1]. Thus, it is critical for us to carry out epidemic transmission prediction [2], [3], genome sequence analysis [4], [5], and public psychological stress assessments [6], [7] for 2019-nCoV.

From the perspective of epidemic transmission, the basic reproduction number (R_0) is one of the key parameters for 2019-nCoV epidemic transmission prediction [8], [9], [10], [11], [12]. Although previously well-developed SIR [13] or SEIR [14] models can estimate the basic reproduction number (R_0), neither SIR nor SEIR consider the factors of suspected patient quarantine. Furthermore, the most commonly used web services of 2019-nCoV [5], [15], [16] only focus on the statistical analysis of real epidemic data, and there are only a few online predictive services with different epidemic transmission models.

From the perspective of genome sequence analysis, recent studies [17], [18], [19] performed sequence analysis and phylogenetic tree construction for 2019-nCoV. Most of these studies employed k-mer counting as a basic method to explore the frequent subsequence of genomes [20]. However, the k-mer counting method did not consider the characteristics of subsequences such as permutation specificity and CG content change from the perspective of lineage for 2019-nCoV, and sequencing errors cannot be avoided for the frequently mutated 2019-nCoV. Therefore, we cannot accurately and comprehensively describe the characteristics of the 2019-nCoV genome only by k-mer counting.

From the perspective of public psychological stress assessment, many previous studies [21], [22], [23] investigated the impact of 2019-nCoV on the public psychological stress state. For example, Chang *et al*. [21] evaluated the psychological health of 3881 students from Guangdong University using self-compiled 2019-nCoV scales. Fan *et al*. [22] assessed the psychological health status of people in Gansu Province through questionnaires. However, since we usually employ questionnaires to collect data, the population coverage is so narrow that it is difficult to dynamically assess the public psychological state and develop a professional emotional dictionary for 2019-nCoV. Furthermore, previous studies did not consider the connections among the public psychological stress state, real epidemic trends, and genome variation rate.

- *Ming Xiao, Jianghang Xie, Zichun Dai, Zihao Wei, Ziyao Ren, and Le Zhang are with the College of Computer Science, Sichuan University, Chengdu 610065, PR China. E-mail: {xiaoming, zhangle06}@scu.edu.cn, xjh0013@163.com, {daizichun, 2018141461086}@stu.scu.edu.cn, ziyaoren99@gmail.com.*
- *Guangdi Liu is with the College of Computer and Information Science, Southwest University, Chong-Qing 400715, PR China. E-mail: liuguangdi1103@126.com.*
- *Jun Yu is with the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, PR China. E-mail: junyu@big.ac.cn.*

Manuscript received 30 June 2020; revised 2 Oct. 2020; accepted 3 Jan. 2021. Date of publication 6 Jan. 2021; date of current version 6 Aug. 2021.
(Corresponding author: Le Zhang)
Digital Object Identifier no. 10.1109/TCBB.2021.3049617

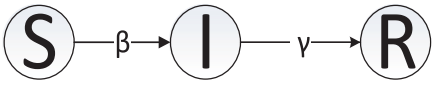


Fig. 1. SIR model.

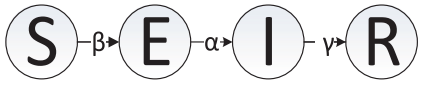


Fig. 2. SEIR model.

For these reasons, we developed an easy-to-use 2019 Novel Coronavirus Analysis Service (2019nCoVAS, http://www.combio-lezhang.online/2019ncov/home.html) with the following three major innovations.

First, 2019nCoVAS not only implements such a predictive model that considers the factors of suspected patient quarantine but also offers online epidemic transmission prediction and R_0 trend analysis.

Second, 2019nCoVAS downloads all open 2019-nCoV genomes from mainstream databases for sequence analysis, as well as uses JBLA [24] to count and analyze the common lineage-associated underrepresented permutations (LAUPs) for all 2019-nCoV genomes. Additionally, we introduce MOTIF discovery [25] to find the frequent permutation pattern of common LAUPs since it can successfully describe the sequence characteristics of the genome [24], [26], [27] from the perspective of never-existing permutations. Thus, 2019nCoVAS can help us to improve the accuracy of sequence and phylogenetic analysis.

Thrid, 2019nCoVAS not only builds up an emotional dictionary of 2019-nCoV by crawling big Weibo data [38] that can significantly expand the data size from the questionnaire but also provides related services such as high-frequency vocabulary visualization and public psychological stress assessments.

In general, 2019nCoVAS can provide epidemic transmission prediction, genome sequence analysis, and public psychological stress assessment for 2019-nCoV.

2 Implementations

2.1 Epidemic Transmission Prediction

In the beginning, we obtained the 2019-nCoV epidemic data for China by Akshare [28] from January 20 to May 1, 2020. The data consists of the number of confirmed cases, suspected cases, and deaths. We developed a Python script to preprocess the data. It should be noted that since Hubei Province added 14840 newly confirmed cases on February 12 due to a change in detection standards [29], we must proportion the newly confirmed cases on February 12 from February 7 to February 12.

To predict the epidemic transmission and basic reproduction number (R_0) for 2019-nCoV, we build up three epidemic transmission predictive models: SIR, SEIR, and SEIRQ. In particular, the SEIRQ model considers quarantined case factors, which can be used to predict epidemic situations under the condition of suspected quarantine cases. Next, we discuss these models.

2.1.1 SIR Model

Fig. 1 shows that the SIR model [30] classifies the total population into susceptible (S), infected (I), and recovered (R) populations. The susceptible (S) population transforms to infected (I) according to infection rate β (eq. (1) of the supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2021.3049671), and infected (I) gradually recovers (R) at recovery rate γ (Eq. 2 of the supplementary file, available online). Related methods are listed in supplementary file S1.1, available online.

2.1.2 SEIR Model

Since 2019-nCoV has a latent period, Fig. 2 introduces a SEIR model [14] that considers the exposed factor (E), which carries the virus without symptoms. Here, the susceptible (S) population gradually transforms into exposed (E) rather than infected (I). The exposed (E) gradually converts to infected (I) with the conversion ratio α, and the infected (I) eventually converts to recovered (R).

The conversion relationship between exposed (E) and infected (I) is described by 1 [14], and the rest of the information is listed in supplementary file S1.2, available online.

$$\frac{dE(t)}{dt} = \frac{\beta S(t) I(t)}{N} - \alpha E(t). \tag{1}$$

Here, α represents the probability that the exposed (E) will transform into infected (I). N is the sum of $S(t)$, $I(t)$, and $R(t)$.

2.1.3 SEIRQ Model

After we quarantined the suspected patients for 2019-nCoV outbreaks in China [1], we proposed a novel infectious SEIRQ model by considering the quarantined population, which is based on a modified SEIR model [31]. Fig. 3 shows a schematic diagram for SEIRQ.

Based on SEIR model, SEIRQ model incorporates two types of quarantined populations. One is the quarantined patients in observe(O), and the other is the quarantined patients in treatment(T).

The susceptible (S) population will gradually change to the quarantined in observed (O) at a quarantine rate δ_1. Susceptible (S) is transformed into exposed (E) at infection rate β. The exposed (E) is quarantined into the quarantined in observed (O), and the quarantine rate is ε. At the same time, the exposed (E) is transformed into the infected (I) at

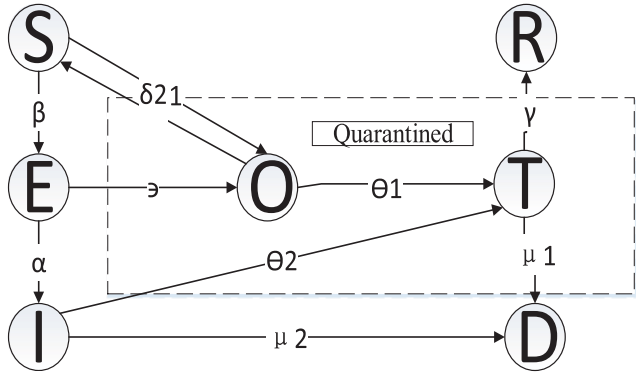


Fig. 3. SEIRQ model.

TABLE 1
Key Parameters of SEIRQ Model

Symbol	Significance	Symbol	Significance
S	Susceptible population	**I**	Infected population
R	Recovered population	**E**	Exposed population in latent period
O	Quarantined patients in observed	***T***	Quarantined patients in treatment
D	Deaths	α	Incidence rate of exposed (E)
β	Infection rate	γ	Cure rate
δ_1	Quarantine rate of susceptible (S)	δ_2	Quarantine rate of quarantined (Q)
ϵ	Quarantine rate of exposed (E)	θ_1	Rate at which quarantined in observed (O) become quarantined in treatment (T)
θ_2	Incidence rate of infected patients (I)	μ_1	Death rate of quarantined patients in treatment (T)
μ_2	Death rate of infected patients (I)		

rate α. After the quarantine period, the quarantined in observed (O) is quarantined into susceptible (S) at quarantine rate δ_2, which is considered to be uninfected. The quarantined in observed (O) will become the quarantined patients in treatment at rate θ_1. Infected (I) is confirmed in quarantined patients in treatment (T) at rate θ_2. Meanwhile, the infected die at rate μ_2. The quarantined patients in treatment (T) gradually recover (R) at a recovery rate γ. Finally, the quarantined patients in treatment (T) die at rate μ_1. Table 1 and supplementary file S1.3, available online list the key parameters and equations, respectively.

To validate the predictive capacity of the model, we selected real data from January 20 to January 30 as the training data to predict the number of infected populations for the SIR, SEIR, and SEIRQ models (Fig. 4A). We also use the cross-validation method [32] to compute the average root mean square error (AVG_RMSE) for the three models by Eq. (2). Fig. 4B shows the AVG_RMSE values that describe the deviation between the predictive and actual curves.

Specifically, we select the first two days of real data as training data and the next day as testing data to calculate the root mean square error (RMSE [33], eq. (2.1)). Then, we add the third-day data to the training dataset and compute the RMSE value by using the fourth-day data as the training data. Next, we use the same rule for three day-forward iterations.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(x_{real,i} - x_{model,i}\right)^2} \tag{2.1}$$

$$AVG\_RMSE = \frac{\sum_{j}^{D} RMSE_j}{D}. \tag{2.2}$$

Here, $x_{real,i}$ and $x_{model,i}$ represent the real confirmed case number and the number of infections predicted by the model, respectively. N is the number of days of testing data, and D represents the number of days of day-forward iterations.

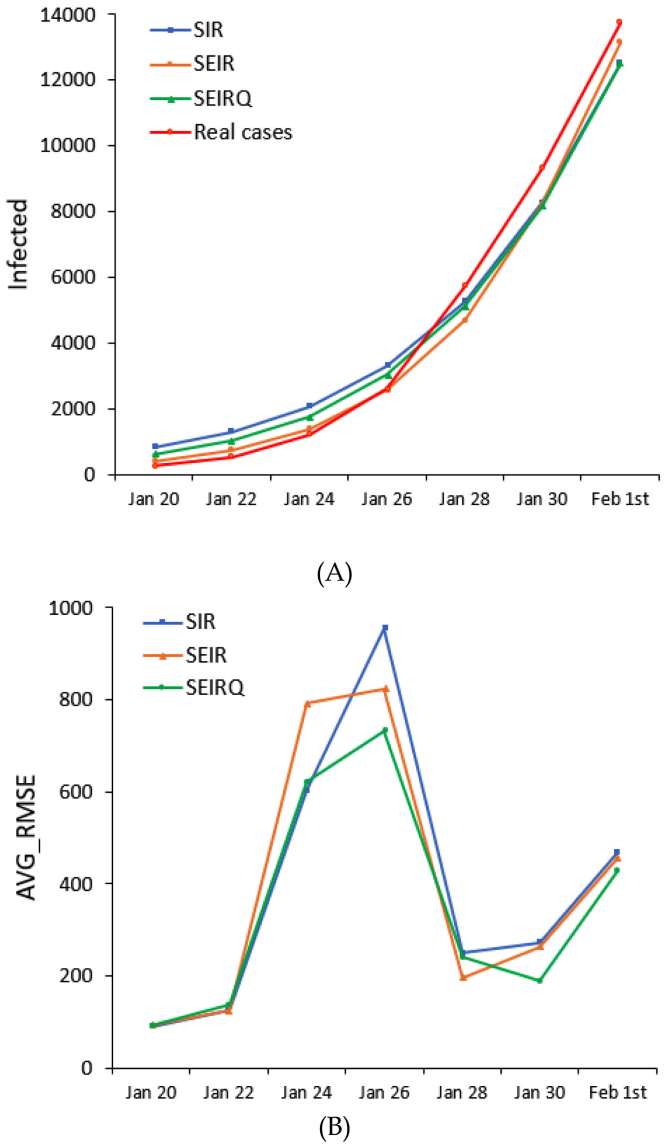


Fig. 4. Comparison between predicted and actual curves. (A) Predictive and actual curves for infected population. Horizontal and vertical axes represent date and average infected population value, respectively. (B) Average RMSE for SIR, SEIR, and SEIRQ models. Here, horizontal and vertical axes represent date and average RMSE value, respectively.

2.2 Basic Reproduction Number (R_0) Estimation

We usually employ R_0 to describe the dynamic change for the infective trend, which is computed by eq. (3) for each day [34]. R_0 represents the average number of people infected with one person who can be transmitted to others without external intervention and group immunity [35].

$$R_0 = \left(1 + \lambda T_g + \rho(1-\rho)\left(\lambda T_g\right)^2\right) \tag{3}$$

Here, ρ is the ratio of the incubation period to the generation time. λ is the growth rate during early exponential growth. T_g is the sum of the incubation period and the infection period [34].

2.3 Genome Sequence Analysis

For 2019-nCoV genome sequence analysis, we downloaded all 2019-nCoV genome sequences and variation

data from the 2019nCoVR [18] and GISAID [36] databases. We also compared 2019-nCoV with other coronavirus genomes by using all coronavirus sequence data from the NGDC [37].

To analyze the LAUP sequence characteristics for the 2019-nCoV genome, we used the Jellyfish-based LAUP analysis (JBLA) application [24] to compute the underrepresented sequence Eq. (4) and common LAUPs. We also introduced the MOTIF discovery method [25] to find the most frequent arrangement pattern in the common LAUPs and provide data download services to help users conduct further sequence and phylogenetic analysis. To analyze the connections among the variation rate, epidemic transmission trend, and public psychological stress state, we counted the average variation rate for 2019-nCoV genomes every day.

2.3.1 Multi-Genome LAUPs Analysis

To investigate the characteristics of the 2019-nCoV sequence, we used Jellyfish [38] to calculate all k-mers for 2019-nCoV and then calculated the LAUP [24] sequence for each 2019-nCoV genome by Eq. (4).

$$\mathrm{LAUPs}_k = \mathrm{C}_{\mathrm{Sim\_set}_k}(\mathrm{Kwg\_set}_k) \tag{4}$$

Here, $\mathrm{Sim\_set}_k$ includes all possible 4^k k-mers. $\mathrm{Kwg\_set}_k$ contains permutations in which all k-mers have appeared in the genome of 2019-nCoV. LAUPs are defined as complements of $\mathrm{Kwg\_set}_k$ on $\mathrm{Sim\_set}_k$.

In addition, we calculated the common LAUPs for the whole 2019-nCoV by Eq. (5) [24], determined the shortest common LAUPs, and analyzed the GC content of the common LAUPs by Eq. (6).

$$CommonLAUPs = \cap\left(\sum_{i=1}^{N} LAUPs_k^i\right). \tag{5}$$

Here, k is the length of LAUP and N is the total number of all new coronavirus genomes.

$$content_{GC(LAUP_k)} = \frac{num(G_k) + num(C_k)}{k}. \tag{6}$$

Here, k is the length of LAUP. The function num () indicates the number of bases. Content_ GC () is used to calculate the content of CG and LAUP.

As a single-strand positive-sense RNA virus, 2019-nCoV follows all the molecular rules of the RNA world. Two of the primary rules are U as a nuclear base, instead of T in DNA, and a secondary structure formed by single-strand RNA molecules that are mostly intramolecular [39]. To apply k-mer and LAUP concepts in 2019-nCoV data analysis, we first convert the full sequence of the virus into k-mers of various lengths and subsequently look for LAUPs by comparing them with two k-mer pools: one contains random k-mers generated in limited G+C and A+G content windows, and the other is generated from all unique and high-quality RNA vuses. In this protocol, LAUPs contain sequence-derived permutations that are excluded by RNA viruses. Therefore, in doing so, we are not only able to avoid the impact of genome sequencing errors, but also discover sequences that are negatively selected by viral populations [24].

Virus-infected humans, individual animals, and populations often serve as hosts for viruses to select their best fitness by forming quasi-species where deleterious mutations are excluded so that the viral fitness is evolved. In this process, LAUPs as a set of sequences are subjected to selection in terms of secondary structures and targets of cellular RNA surveillance and interactive systems (such as RNA degradation and miRNA targeting), which is complement to the selection of protein sequences. These analyses of LAUPs can help us to improve the accuracy of genome sequence and phylogenetic analyses as well as viral biology and host pathophysiology.

2.3.2 Statistic Analysis of Genome Variation Rate

To analyze the connections among the 2019-nCoV epidemic, genome variation, and public psychological stress state, Eq. (7) computes the median variation rate. Then, we can visualize the trend of the variation rate under the real epidemic situation and public psychological stress state (Section 2.3), and investigate their connections.

$$\mathrm{Variation\_median} = \begin{cases} \mathrm{Variation}_{\frac{N+1}{2}} & N\ mod\ 2 = 1 \\ \frac{1}{2}\left(\mathrm{Variation}_{\frac{N}{2}} + \mathrm{Variation}_{\left(\frac{N}{2}+1\right)}\right) & N\ mod\ 2 = 0 \end{cases}. \tag{7}$$

Here, N represents the number of all 2019-nCoV genomes on this day. Variation represents the sorted variation rate array of the day.

2.4 Psychological Stress Assesment

To obtain the Chinese public psychological stress data for 2019-nCoV, we built a crawler program based on the Weibo [40] API that crawled the data for all tweets and comments about 2019-nCoV from January 1 to March 31, 2020. Table 2 shows a set of searching Chinese strings of 2019-nCoV for our crawler program.

To dynamically assess the public psychological stress state during the epidemic, we proposed an automatic expansion method for the emotional dictionary in two steps. First, we employ a left and right information entropy algorithm [41] to locate the candidate words and construct the emotional dictionary. Second, we use the SO-PMI [42] and word2vec algorithms [43], [44] to determine the emotional polarity for the candidate words and screen new words for emotional polarity discrimination.

Thus, we developed two corresponding features for our web server. One builds an emotional dictionary for 2019-nCoV, which can highlight its high frequency vocabulary. The other assesses and visualizes the dynamic changes for the public's positive and negative emotions in response to 2019-nCoV at different time points. These features not only are able to provide a retrospective assessment function for psychologists but can also be used as references for national policy development.

TABLE 2
Searching Chinese Strings of 2019-nCoV

No.	Searching Chinese strings	English Translation
1	冠状病毒/新冠病毒	Coronavirus/COVID/ SARS-Cov-2/2019-nCoV
2	肺炎	Pneumonia
3	疫情	Epidemic
4	防控	Prevention and control
5	感染	Infect/Infection
6	医院	Hospital
7	确诊病例/患者	Confirmed cases/patients
8	新确诊病例/患者	Newly confirmed cases/ patients
9	武汉/湖北	Wuhan/Hubei
10	出院	Discharged
11	死亡	Death
12	密切接触	Close contact
13	口罩	Face mask

2.4.1 Exploring Candidate Word

We introduce left and right information entropy [45], [46] as a quantitative measure for the boundary degrees of freedom of candidate words by Eq. (8). The left and right information entropies for the candidate string (w) are labeled H_l and H_r, respectively.

$$H_l = -\sum_{w_l \in s_l} p(w_l|w)\log_2 p(w_l|w) \tag{8.1}$$

$$H_r = -\sum_{w_r \in s_r} p(w_r|w)\log_2 p(w_r|w). \tag{8.2}$$

where s_l is the left adjacency set of candidate word w, w_l is an element in s_l, s_r is the right adjacency set of candidate word w, and w_r is the element of s_r.

2.4.2 Discrimination of Emotional Polarity Based on PMI and Word2Vec

The SO-PMI algorithm [47], [48], [49] is mainly used to determine the degree of correlation between words. We employ it to compute the mutual information between words by Eq. (9).

$$PMI = \log_2\left(\frac{p(w_i, w_j)}{p(w_i) \times p(w_j)}\right). \tag{9}$$

Here, word probabilities $p(w_i)$, $p(w_j)$ and joint probabilities $p(w_i, w_j)$ can be estimated by counting the number of observations of w_i and w_j as well as the co-occurrence of w_i and w_j [50].

Here, the high mutual information indicates a high probability for the co-occurrence of two emotional words in many texts. We also introduced the Semantic Orientation (SO) [48] to determine whether a certain emotional word W is positive or negative by Eq. (10).

$$SO(W) = PMI(W, \beta^+) - PMI(W, \beta^-). \tag{10}$$

We need to select two sets of seed words to compute SO. One is the obvious positive tendency (β^+), and the other is the obvious negative tendency (β^-).

3 Performance

Fig. 5 shows the home page for 2019nCoVAS, the top of which is the functional navigation bar. The "home page" link shows the main features of 2019nCoVAS, followed by four drop-down menus: "infectious disease model," "genome sequence analysis," "psychological stress assessment," and "related links."

3.1 Infectious Disease Model

The "infectious disease model" offers two functional modes. One is "epidemic transmission prediction," which can predict the epidemic transmission for 2019-nCoV by SIR, SEIR, and SEIRQ. The other is "R_0 trend analysis," which can carry out trend analysis for the basic reproduction number (R_0).

3.1.1 Epidemic Transmission Prediction

After clicking the "epidemic transmission prediction" link, the user can employ the selective interface to input the start and end dates and choose the appropriate epidemic transmission predictive models (Fig. 6A), which are comprised of SIR, SEIR, and SEIRQ. Finally, users can view epidemic transmission predictions by clicking the "submit" button (Fig. 6A) or clicking the "reset" button to restore the parameters.

The predictive results are composed of two parts. One is Fig. 6B, which shows the estimated parameters such as the infective rate (β), incidence rate of the exposed (α), and cure rate (γ). The other is Fig. 6C, which shows the predicted epidemic transmission curve. For example, after choosing

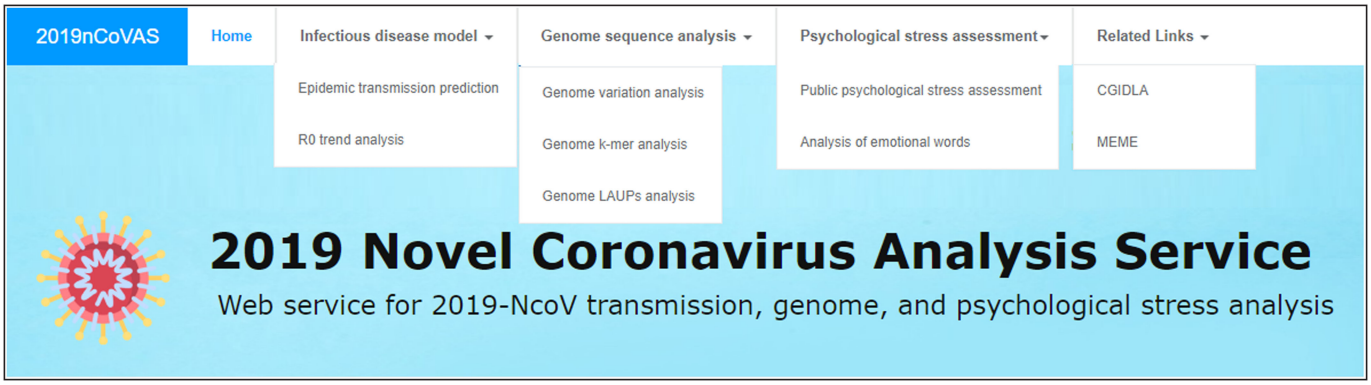


Fig. 5. Home page of 2019nCoVAS.

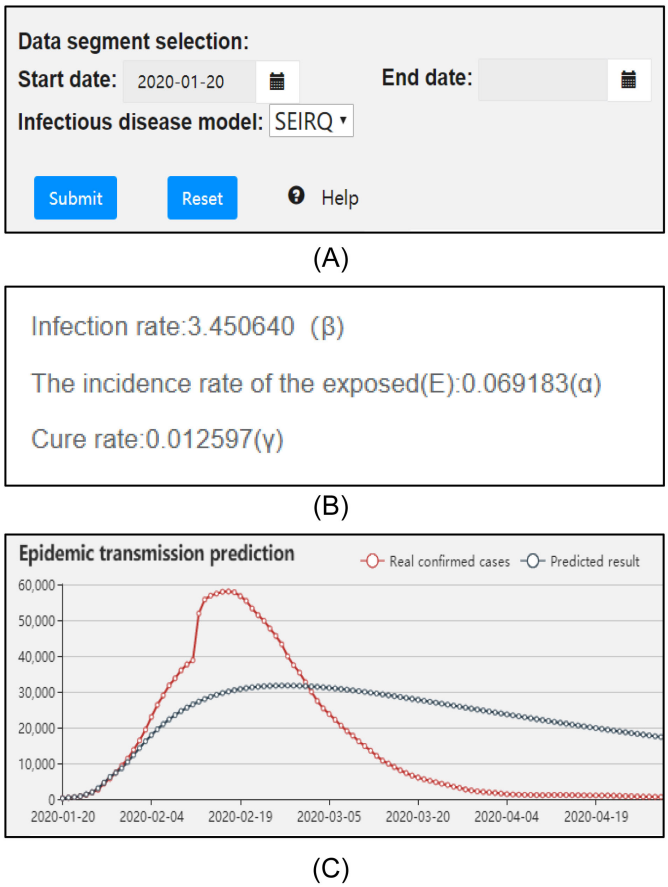


Fig. 6. Epidemic transmission prediction. (A) Infectious disease model selective interface. (B) Estimated parameters. (C) Predicted result. Horizontal and vertical axes represent date and number of infected people, respectively. Red and blue represent real confirmed cases and predicted number of infected people, respectively.

"2020-1-20" and "SEIRQ," Fig. 6B lists the estimated parameters. Fig. 6C shows the predicted infected case curve and the real confirmed case curve.

3.1.2 R_0 Trend Analysis

After clicking the "R_0 trend analysis" link, the user can employ a selective interface to choose the start date (Fig. 7A) and the epidemic transmission predictive model (Fig. 7A), which includes SIR, SEIR, and SEIRQ. The user can obtain the dynamic trend for R_0 by clicking the "submit" button (Fig. 7A) or clicking the "reset" button to restore the parameters to their default values. Here, R_0 indicates the infectivity of the disease. We usually consider that the epidemic situation is well controlled when R_0 is less than 1 [51].

Fig. 7B shows the daily R_0 trend Eq. (3). For example, after choosing "2020-1-20" and "SEIRQ," Fig. 7B shows that R_0 continues to decrease, but it is still greater than 1 until the end of April.

3.2 Genome Sequence Analysis

The "genome sequence analysis" has three features. The first is "genome k-mer analysis," which can count k-mers and analyze the genome of 2019-nCoV. The second is "genome LAUP analysis," which can explore LAUPs for the 2019-nCoV genome. Third is "genome variation analysis,"

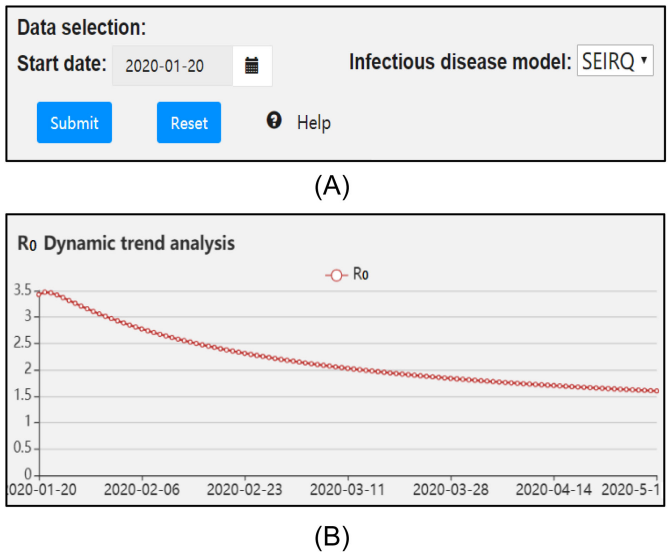


Fig. 7. R_0 trend analysis: (A) Selective interface. (B) R_0 trend; horizontal and vertical axes represent date and value of R_0, respectively.

which can investigate the connections among the genomic variation rate, real number of infection cases, and public positive emotional rate.

3.2.1 Genome K-Mer Analysis

After clicking the "genome k-mer analysis" link, the user can employ the selective interface to input the start and end dates and choose the length of the k-mer (Fig. 8A). The user can click the "submit" button to obtain the k-mer counting results for all 2019-nCoV genomes in the selected time period (Figs. 8B and 8C), or click the "reset" button to restore the parameters to their default values.

The k-mer analysis consists of two figures. One is Fig. 8B, which shows the abundance histogram [52] of the k-mer counting results. The other is Fig. 8C, which shows the top 10 k-mer permutations with the most frequent occurrences for all 2019n-CoV genomes in the selected time period. Additionally, the user can click the "data download" button (Fig. 8C) to download the detailed k-mer counting file.

For example, after choosing start data "2020-01-23," end date "2020-04-30," and length of k-mer "12" (Fig. 8A), Fig. 8B shows that the peak of the 12-mer frequencies of 2019-nCoV appeared at abundances of 6, whereas that of all coronaviruses appeared at abundances of 12. Fig. 8C shows that the most frequent 12-mer permutation for 2019n-CoV is "TTTTTTTTTTTT."

3.2.2 Genome LAUPs Analysis

After clicking the "genome LAUP analysis" link, the user can employ a selective interface to select the start and end dates (Fig. 9A) and then click the "submit" button to obtain the LAUPs analysis for all 2019-nCoV genomes in the selected time period (Figs. 9B and 9C), or click the "reset" button to restore the parameters to their default values.

The LAUP analysis consists of two figures. One is Fig. 9B, which shows the statistics of common LAUPs. The other is Fig. 9C, which shows the statistics of CG content Eq. (6) for the common LAUPs of all 2019n-CoV genomes in the

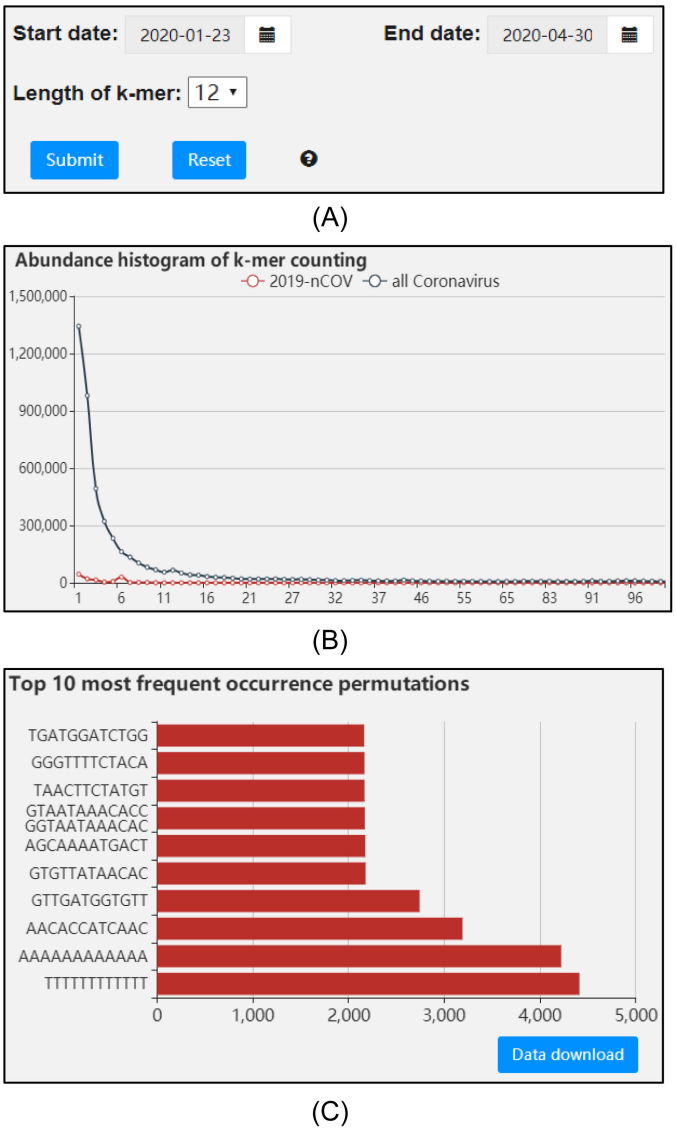


Fig. 8. Genome k-mer analysis: (A) Selective interface. (B) Abundance histogram of k-mer counting results. Horizontal axis represents abundance of k-mers, and vertical axis represents frequency of k-mers with that abundance in the relevant genome [52]; red and blue lines represent 2019-nCoV and all coronavirus sequences, respectively. (C) Top 10 k-mer permutations with most frequent occurrence. Vertical and horizontal axes represent k-mer permutation and counting value of k-mer, respectively.

selected time period. Additionally, the user can click the "data download" button (Fig. 9C) to download the detailed LAUP analytical file.

For example, after choosing start date "2020-01-23" and end date "2020-06-01" (Fig. 9A), Fig. 9B shows that the length of the shortest common LAUPs of 2019-nCoV is 6, whereas the length of the shortest common LAUPs for all coronavirus sequences is 8. If the 6-mers of a genome sequence have any common LAUPs of 2019-nCoV when constructing the phylogenetic tree, this indicates that a distant genetic connection between this genome and the genome of 2019-nCoV. Since the number of common LAUPs is small, its computing cost is much lower than when using the normal sequence alignment method to investigate the connections between the candidate genome and the genomet of 2019-nCoV. Fig. 9C shows that the LAUPs' CG

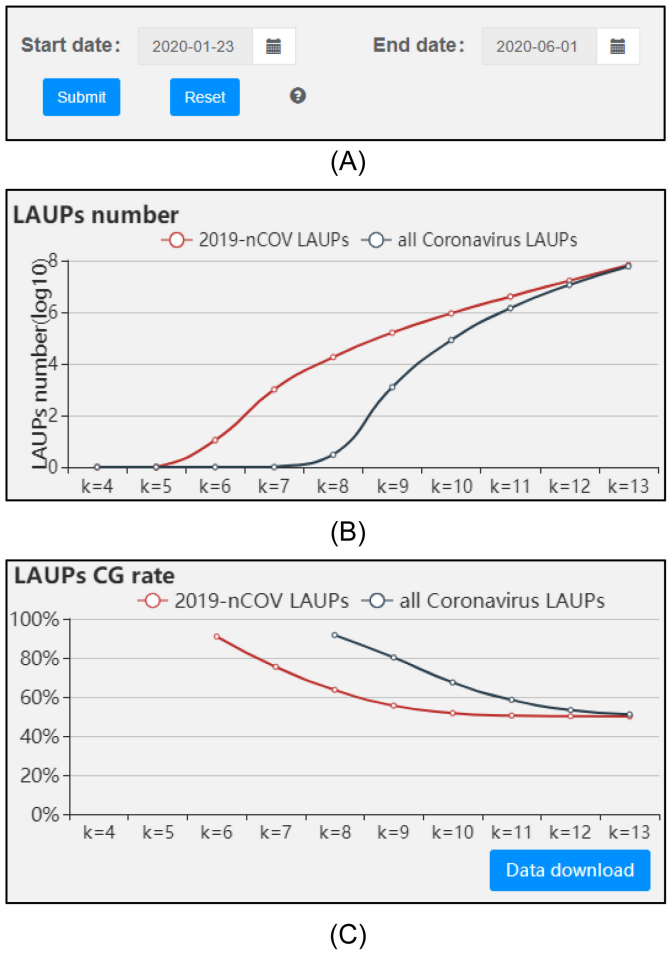


Fig. 9. LAUP analysis: (A) Selective interface (b) LAUP number statistics. (c) CG content statistics of LAUPs; red and blue lines represent 2019n-CoV and all Coronavirus sequences, respectively.

content of 2019-nCoV is greater than that of all coronaviruses under the same length of K.

3.2.3 *Genome Variation Analysis*

After clicking the "Genome variation analysis" link, the user can employ a selective interface to select the start and end dates (Fig. 10A) and then click the "Submit" button to obtain the dynamic visualization of the genomic variation rate for all 2019-nCoV genomes in the selected time period (Fig. 10A) or click the "Reset" button to restore the parameters to default.

Fig. 10B visualizes the daily median genome variation rate eq. (7), number of new infections in the real epidemic, and public positive emotions eq. (10). Additionally, the user can click the "data download" button (Fig. 10B) to download the detailed variation rate result file.

For example, after choosing start date "2020-01-23" and End date "2020-03-15" (Fig. 10A), Fig. 10B shows that the virus variation rate gradually increases, the new confirmed cases continue to decrease, and the public positive emotion rate gradually returned to zero after February 19.

3.3 Psychological Stress Assessment

The "psychological stress assessment" offers two functional modes. One is the "public psychological stress assessment,"

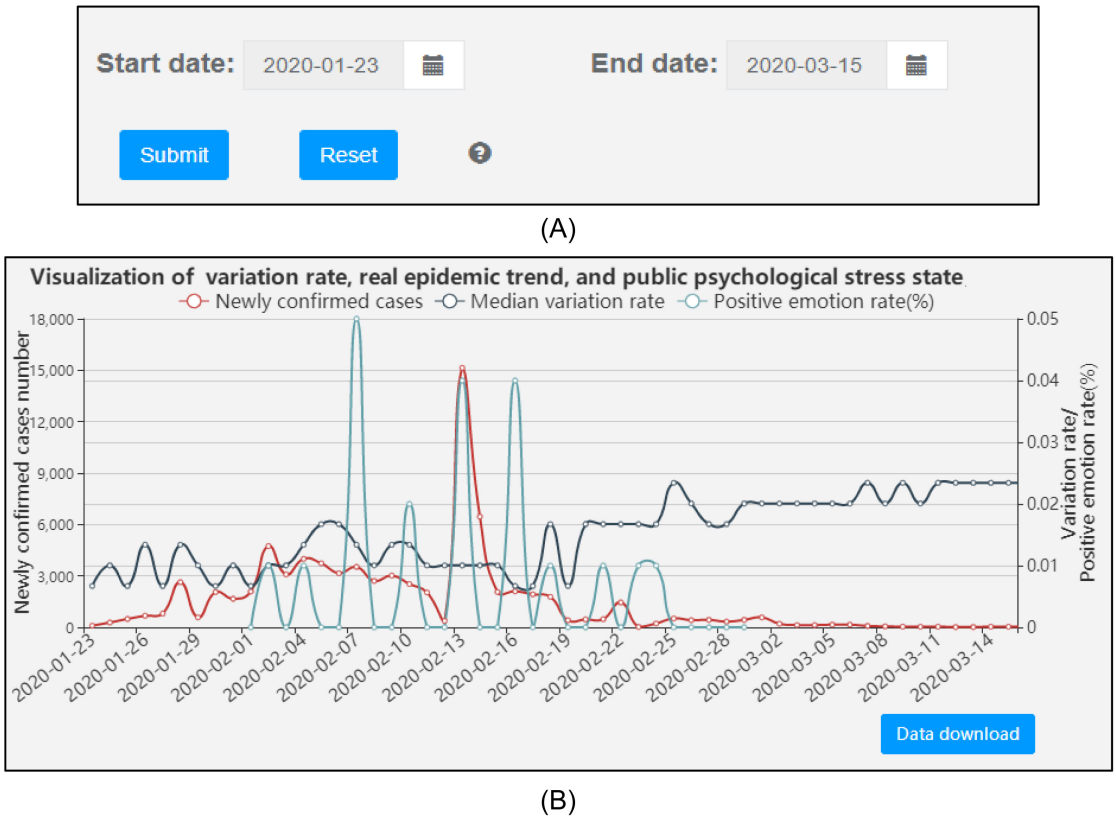


Fig. 10. Genome variation analysis0 (A) Selective interface. (B) Visualization of genome variation rate, real confirmed cases, and public positive emotion. Vertical axis represents date, and horizontal axis represents number of confirmed cases, rate of genomic variation, and proportion of public positive emotions (%). Red, green, and blue represent newly confirmed cases, genome variation rate, and public positive emotion rate, respectively.

which can analyze the dynamic public psychological stress state under the epidemic situation; the other is "analysis of emotional words," which can visualize the high frequency of emotional words for 2019-nCoV.

3.3.1 *Public Psychological Stress Assessment*

After clicking the "public psychological stress assessment" link, the interface displays the rate of positive and negative emotion per day (Fig. 11A). The user can select the period of observation by dragging the scroll bar from left to right. Additionally, the user can select the start and end dates of statistics from two drop-down boxes (Fig. 11B). Finally, the user can click the "download" button to obtain the analytic results based on the selected date.

The analytic results are composed of two components (Fig. 11A). The top of Fig. 11A shows the proportional distribution of the positive emotion rate over time (in days), and the bottom of Fig. 11A shows that of the negative emotions. For example, the proportion of positive emotion is low before February 5, whereas the negative emotion is high before February 3. However, positive emotion gradually increases and negative emotion gradually decreases after February 5.

3.3.2 *Analysis of Emotional Words*

After clicking the "analysis of emotional words" link, the interface displays the word clouds (Fig. 12A) for all emotional words generated during the epidemic period (February 2020 to April 2020). Additionally, the user can query words through the text box (Fig. 12B) and click the "submit" button to have emotional polarity for the word from the text box.

The word clouds (Fig. 12A) include positive words, negative words, and neutral words. Here, the font size of a word is positively related to its occurrence frequency. After typing "加油" and clicking the submit button, Fig. 12B shows that the emotional polarity of the word is positive. In addition, considering that the words in the source data are in Chinese, we translated the top five emotional words in Fig. 12C.

3.4 Related Links

Finally, "related links" include well-developed tools such as "motif discovery" [25] and "CGIDLA" [26], in which "motif discovery" can provide an online service for motif discovery [25], and "CGIDLA" [26] can help us further analyze the CG permutation specificity of LAUPs.

4 Conclusions

2019nCoVAS provides an informative and interactive platform for the analysis and visualization of epidemic transmission prediction, genome sequence analysis, and public psychological stress assessments.

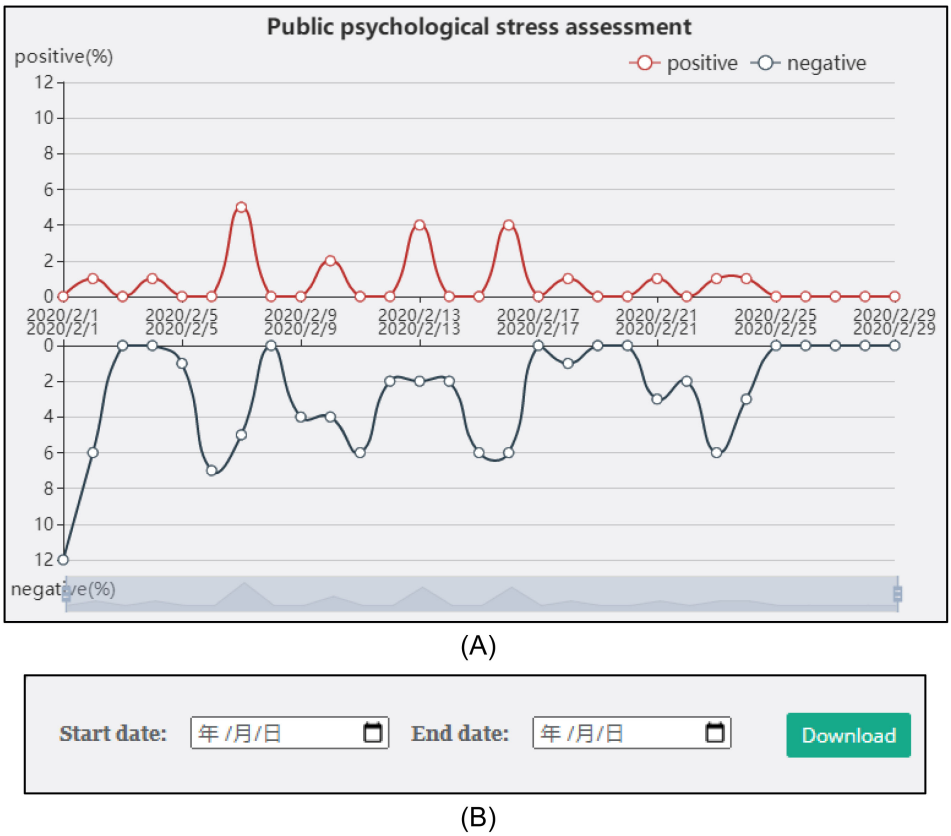


Fig. 11. Public psychological stress assessment: (A) Visualization of positive and negative emotional trends; horizontal and vertical axes represent date and emotion rate, respectively. Red and blue lines represent positive and negative emotion rates, respectively. (B) Data download interface.

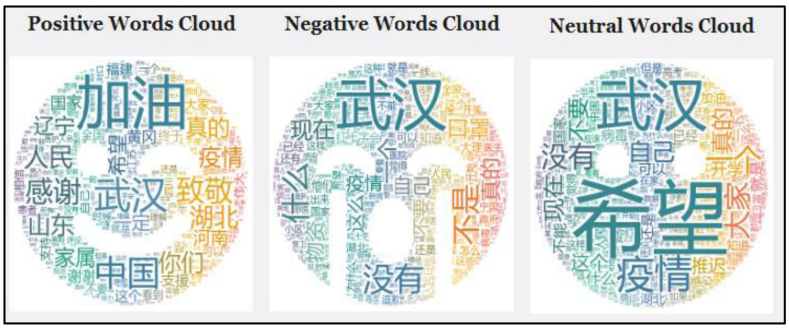
Fig. 4 demonstrates that the average RMSE (Eq. (2)) of SEIRQ is significantly less than that of the SIR and SEIR models between January 24 and January 30. This occurred because Wuhan started using strict isolation measures on January 24 [53] and most Chinese provinces started the level-one response to public health emergencies on January 30 [54]. Thus, the SEIRQ model, which has isolation features, is more suitable for epidemic cases with isolation measures. Fig. 6C demonstrates that the number of infective cases predicted by SEIRQ is less than the number of real confirmed cases in February, implying that the epidemic may be underestimated in February. Additionally, Fig. 7B shows that R_0 estimated by SEIRQ continues to decrease but is still greater than 1 until the end of April, indicating that the epidemic will continue.

Second, Fig. 8 shows that 2019-nCoV has a strong specificity of permutation and base content compared with other coronaviruses by genome sequence analysis. For example, Fig. 8B shows that the peak of the 12-mer frequency of 2019-nCoV appeared at different abundances from those of other coronaviruses. Additionally, Fig. 9 demonstrates that the length of the shortest common LAUPs of 2019-nCoV is different than that of other coronaviruses (Fig. 9B), and the CG content of LAUPs of 2019-nCoV is greater than that of other coronaviruses under the same length of K (Fig. 9C).

Additionally, through dynamic visualization of variation, we found that the variation rate is gradually increasing, the number of newly confirmed cases is decreasing, and the public emotion gradually calmed (Fig. 10B) after February 19, which indicates that the government's prevention and control measures quickly controlled the epidemic and effectively stabilized the public mood.

Third, Fig. 11 shows that the positive mood gradually increased and the negative mood gradually decreased (Fig. 11A) after February 5. We consider that this phenomenon may be related to the government's prevention, control measures, and major news events such as the completion of Leishenshan Hospital on February 8 [55]. According to the emotional dictionary analysis (Fig. 12A), the highly frequent positive emotional part includes words such as "come on" and "thanks," and the negative part includes words such as "face masks," "supplies," and other words related to medical supplies. This indicates that people are still worried about the provision of medical supplies during the epidemic of 2019n-CoV.

In general, 2019nCoVAS is an effective web service for 2019-nCoV. However, due to the limitations of computing power, it does not provide real-time simulations. In the distant future, we will not only employ high-performance computing technology [56], [57] to realize real-time simulations but will also develop more in-depth LAUP analysis methods [26], [27] to further analyze the sequence characteristics of the 2019-nCoV genome. Finally, since the 2019- nCoV epidemic is still spreading around the world, we will carry out genomic sequence analysis, epidemic transmission prediction, and public psychological stress assessments for different countries.



Category	Chinese	English
Positive	加油	Come on
	武汉	Wuhan
	中国	China
	致敬	Salute
	感谢	Thanks
Negative	没有	No
	口罩	Face mask
	不是	Not
	物资	Goods and materials
	不要	Don't
Neutral	希望	Hope
	疫情	Epidemic
	开学	Term begins
	大家	We
	自己	Myself

(C)

Fig. 12. Analysis of emotional words: (A) High-frequency word cloud. (B) Emotional word query interface. (C) Top five emotional words in Chinese and English.

ACKNOWLEDGMENTS

This work was supported by National Science and Technology Major Project [2018ZX10201002], China Postdoctoral Science Foundation [2020M673221], and Sichuan University Postdoctoral Research and Development Foundation [2020SCU12056].

REFERENCES

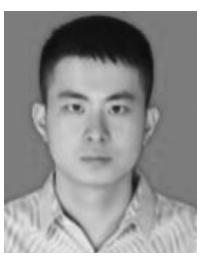
[1] A. D.. Toit, "Outbreak of a novel coronavirus," *Nat. Rev. Microbiol.*, vol. 18, no. 3, pp. 123–123, 2020.
[2] T. Li, Z. Cheng, and L. Zhang, "Developing a novel parameter estimation method for agent-based model in immune system simulation under the framework of history matching: A case study on influenza a virus infection," *Int. J. Mol. Sci.*, vol. 18, no. 12, Dec./ Jan. 2017.
[3] W. Wu, L. Song, Y. Yang, J. Wang, H. Liu, and L. Zhang, "Exploring the dynamics and interplay of human papillomavirus and cervical tumorigenesis by integrating biological data into a mathematical model," *BMC Bioinf.*, vol. 21, no. Suppl 7, May 2020, Art. no. 152.
[4] L. Zhang, W. Bai, N. Yuan, and Z. Du, "Comprehensively benchmarking applications for detecting copy number variation," *PLoS Comput. Biol.*, vol. 15, no. 5, May 28, 2019, Art. no. e1007069.
[5] L. Zhang *et al.*, "Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model," *Bioinformatics*, Jul. 26, 2019, doi: 10.1093/bioinformatics/btz542.
[6] G.-D. Liu, Y.-C. Li, W. Zhang, and L. Zhang, "A brief review of artificial intelligence applications and algorithms for psychiatric disorders," *Engineering*, vol. 6, no. 4, pp. 462–467, 2020.
[7] G. Liu *et al.*, "Research on psychological scales based on multi-theory fusion," *Curr. Bioinf.*, vol. 15, pp. 1–9, 2019.
[8] B. Ivorra, M. R. Ferrandez, M. Vela-Perez, and A. M. Ramos, "Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of china," *Commun. Nonlinear Sci. Numer. Simul.*, pp. 105303–105303, 2020.
[9] W. C. Roda, M. B. Varughese, D. Han, and M. Y. Li, "Why is it difficult to accurately predict the COVID-19 epidemic?," *Infect. Dis. Modelling*, vol. 5, pp. 271–281, 2020.
[10] Z. Yang *et al.*, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in china under public health interventions," *J. Thoracic Dis.*, vol. 12, no. 3, pp. 165–174, Mar. 2020.
[11] C. Hou *et al.*, "The effectiveness of quarantine of wuhan city against the corona virus disease 2019 (COVID-19): A well-mixed SEIR model analysis," *J. Med. Virol.*, vol. 92, pp. 841–848, 2020.
[12] K. Prem *et al.*, "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in wuhan, china: A modelling study," *Lancet Public Health*, vol. 5, no. 5, pp. E261–E270, May 2020.
[13] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
[14] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Rev.*, vol. 42, no. 4, pp. 599–653, Dec. 2000.
[15] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 533–534, 2020.
[16] J. Valls, A. Tobias, P. Satorra, and C. Tebe, "COVID19-Tracker: A shiny app to analise data on SARS-CoV-2 epidemic in spain," *Gaceta Sanitaria*, vol. 35, pp. 99–101, 2020.
[17] R. Lu *et al.*, "Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding," *Lancet*, vol. 395, no. 10224, pp. 565–574, 2020.
[18] W.-M. Zhao *et al.*, "The 2019 novel coronavirus resource," *Yi Chuan = Hereditas*, vol. 42, no. 2, pp. 212–221, 2020.
[19] N. Zhu *et al.*, "A novel coronavirus from patients with pneumonia in china, 2019," *New Engl. J. Med.*, vol. 382, no. 8, pp. 727–733, 2020.
[20] M. Xiao *et al.*, "K-mer counting: Memory-efficient strategy, parallel computing and field of application for bioinformatics," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2018, pp. 2561–2567.
[21] C. Jinghui, Y. Yuxin, and W. Dong, "Mental health status and its influencing factors among college students during the epidemic of COVID-19," *J. South Med. Univ.*, no. 02, pp. 171–176, 2020.
[22] F. peng *et al.*, "Analysis of public psychological behavior and countermeasures during the epidemic of COVID-19, " *Soc. Sci. Rev.*, no. 2, pp. 1–5, 2020.
[23] J. Zhang, W. Wu, X. Zhao, and W. Zhang, "Recommended psychological crisis intervention response to the 2019 novel Coronavirus pneumonia outbreak in China: A model of West China hospital," *Precis. Clin. Med.*, vol. 3, no. 1, pp. 3–8, 2020.
[24] L. Zhang, M. Xiao, J. Zhou, and J. Yu, "Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA)," *Bioinformatics*, vol. 34, no. 21, pp. 3624–3630, Nov. 1, 2018.
[25] T. L. Bailey *et al.*, "MEME SUITE: Tools for motif discovery and searching," *Nucl. Acids Res.*, vol. 37, no. Web Server issue, pp. W202–W208, 2009.

[26] M. Xiao, X. Yang, J. Yu, and L. Zhang, "CGIDLA:Developing the web server for CpG island related density and LAUPs (Lineage-associated underrepresented permutations) study," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2148–2154, Nov./Dec. 2020.
[27] L. Zhang, Z. Dai, J. Yu, and M. Xiao, "CpG-Island-based annotation and analysis of human house-keeping genes," *Brief. Bioinf.*, vol. 22, no. 1, pp. 515–525, Jan. 18, 2021.
[28] A. King. "AkShare," 2019. [Online]. Available: https://github.com/jindaxiang/akshare.
[29] C. Ren, S. Ren, Y. Chai, and Y. Liu, "Modeling agile supply chain dynamics: A complex adaptive system perspective," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2002, Art. no. 6.
[30] F. J. Sedlazeck *et al.*, "Accurate detection of complex structural variations using single-molecule sequencing," *Nat. Methods*, vol. 15, no. 6, pp. 461–468, Jun. 2018.
[31] S. CAO, P. FENG, and P. SHI, "Study on the epidemic development of COVID-19 in Hubei province by a modified SEIR model.," *J. Zhejiang Univ. (Med. Sci.)*, vol. 49, no. 2, pp. 178–184, 2020.
[32] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Inf. Sci.*, vol. 191, no. none, pp. 192–213, 2012.
[33] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.
[34] J. F.-W. Chan *et al.*, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster," *Lancet*, vol. 395, no. 10223, pp. 514–523, Feb. 15, 2020.
[35] K. Dietz, "The estimation of the basic reproduction number for infectious diseases," *Statist. Methods Med. Res.*, vol. 2, no. 1, pp. 23–41, 1993.
[36] Y. Shu and J. Mccauley, "GISAID: Global initiative on sharing all influenza data – From vision to reality," *Euro Surveillance*, vol. 22, no. 13, 2017, Art. no. 30494.
[37] N. G. D. C. Members, and Partners, "Database resources of the national genomics data center in 2020," *Nucleic Acids Res.*, vol. 48, pp. D24–D33, 2019.
[38] C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, 2011.
[39] T. J. Green, R. Cox, J. Tsao, M. Rowse, S. Qiu, and M. Luo, "Common mechanism for RNA encapsidation by negative-strand RNA viruses," *J. Virol.*, vol. 88, no. 7, pp. 3766–3775, 2014.
[40] T. S. Xu, X. Yang, H. C. Zhang, and J. Zhang, "An advanced data capture method based on sina weibo" pp. 1489–1492, 2015.
[41] R. YAO, G. XU, and J. SONG, "Micro-blog new word discovery method based on improved mutual information and branch entropy," *J. Comput. Appl.*, vol. 36, no. 10, pp. 2772–2776, 2016.
[42] P. R. Kanna and P. Pandiaraja, "An efficient sentiment analysis approach for product review using turney algorithm," *Procedia Comput. Sci.*, vol. 165, pp. 356–362, 2019.
[43] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1857–1863, 2015.
[44] X. Rong, "Word2vec Parameter Learning Explained," pp. 1–19, 2014, *arXiv:1411.2738*.
[45] G. Liu, Y. Xia, C. Yang, and L. Zhang, "The review of the major entropy methods and applications in biomedical signal research," in *Proc. Int. Symp. Bioinf. Res. Appl.*, pp. 87–100, 2018.
[46] C.-Y. Lin, N. Xue, D. Zhao, X. Huang, and Y. Feng, "Natural Language Understanding and Intelligent Applications," in *Proc. Int. Conf. Comput. Process. Oriental Lang.*, 2016., pp. 175–177.
[47] A. Salle and A. Villavicencio, "Why so down? The role of negative (and positive) pointwise mutual information in distributional semantics," pp. 1–6, 2019, *arXiv:1908.06941*.
[48] A. Toprak and M. Turan, "The positive effect of PMI on the selection of meaningful words," in *Proc. 11th Int. Conf. Elect. Electronics Eng.*, 2019, pp. 911–915.
[49] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Proc. GSCL*, 2009, pp. 31–40.
[50] Y. Matsumoto, R. Sproat, K.-F. Wong, and M. Zhang, "Computer processing of oriental languages beyond the orient: The research challenges ahead," in *Proc. 21st Int. Conf. Comput. Orient. Lang.*, 2006, pp. 256–277.
[51] S. Zhao *et al.*, "Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak," *Int. J. Infect. Dis.*, vol. 92, pp. 214–217, Mar. 2020.
[52] B. Chor, D. Horn, N. Goldman, Y. Levy, and T. Massingham, "Genomic DNA k -mer spectra: Models and modalities," *Genome Biol.*, vol. 10, no. 10, pp. 1–10, 2009.
[53] I. c. c. i. Wuhan, "Notice on epidemic prevention and control of 2019nCov in Wuhan," 2020. [Online]. Available: http://www.gov.cn/xinwen/2020-01/23/content_5471751.htm
[54] Caixin, "All 31 provinces in mainland China launched the first level response to public health emergencies," 2020. [Online]. Available: http://china.caixin.com/2020-01-29/101509411.html
[55] C. Daily. "Leishenshan hospital ready to receive patients," Feb. 8, 2020; http://www.chinadaily.com.cn/a/202002/08/WS5e3e7810a310128217275fb3.html
[56] L. Shi, Z. Huang, N. Hu, Z. Yang, and L. Zhang, "Integrating semantic query function into D-NetWeaver," *J. Med. Imag. Health Inform.*, vol. 5, no. 5, pp. 982–986, 2015.
[57] B. Jiang, W. Dai, A. Khaliq, X. Zhou, and L. Zhang, "Accelerating 3D diffusion model in a cylindrical coordinate system by graphics processing unit (GPU) techniques," *Math. Comput. Simul.*, vol. 109, pp. 1–19, 2015.

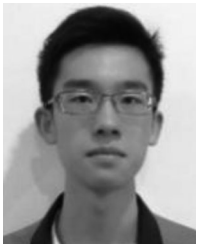
Ming Xiao (Member, IEEE) received the BS, MS, and PhD degrees from Southwest University, China, in 2007, 2010 and 2018, respectively. Currently he is a postdoctor with the College of Computer Science, Sichuan University. His research interests involve bioinformatics, data mining, and parallel computing.

Guangdi Liu received the MS degrees from Xihua University, China, in 2014. Currently he is working toward the doctoral degree in the college of computer and information science, Southwest University. His research interests include bioinformatics, artificial intelligence, and data mining.

Jianghang Xie received the BS degree from the Chongqing University of Posts and Telecommunications, in 2019. Currently, he is working toward the master's degree at the College of Computer Science, Sichuan University. His research interests include machine learning and parallel computing.



Zichun Dai received the BS degree from Sichuan University, China, in 2018. Currently he is working toward the master's degree at the College of Computer Science, Sichuan University. His research interests include bioinformatics, parallel computing, and data mining.



Zihao Wei is currently working toward the graduate degree at the College of Computer Science, Sichuan University. His research interests include web services development and data mining.

Ziyao Ren is currently working toward the graduate degree at the College of Computer Science, Sichuan University. His research interests include data mining and parallel computing.

Jun Yu received the BS degree from Jilin University, China, in 1983, and the MS and PhD degrees from the New York University School of Medicine, in 1986 and 1990. From 1998 to 2003, he worked as a research scientist at the Human Genome Center, Institute of Genetics, Chinese Academy of Sciences. From 2003 to 2012, he served as the deputy director of Beijing Institute of Genomics, Chinese Academy of Sciences. Currently, he is a research scientist with the Beijing Institute of Genomics, Chinese Academy of Sciences, His research interests include genomics and bioinformatics.

Le Zhang received BS degree from the Beijing Institute of Technology, China, in 1999, and the MS and PhD degrees from Louisiana Tech University in 2005, completed his postdoctoral training from 2005 to 2008 in Harvard Medical School. Currently he is a full professor at the College of Computer Science, Sichuan University. His research interests include bioinformatics, computational biology, artificial intelligence, and high-performance computing.