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.

Molecular Phylogenetics and Evolution 175 (2022) 107583

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Exploring the dynamic variations of viral genomes via a novel genetic network

Yuyan Zhang [a,b,1], Jia Wen [b,c,1,2,*], Kun Xi [c], Qiuhui Pan [a,*]

[a] School of Mathematical Science, Dalian University of Technology, Dalian 116024, China
[b] School of Information Engineering, Suihua University, Suihua 152061, China
[c] Warshel Institute for Computational Biology, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China

ARTICLE INFO

Keywords:
Genetic network
Bayesian
K-mer model
Super-spreader
Vaccine

ABSTRACT

Exploring the dynamic variations of viral genomes utilizing with a phylogenetic analysis is vital to control the pandemic and stop its waves. Genetic network can be applied to depict the complicated evolution relationships of viral genomes. However, current phylogenetic methods cannot handle the cases with deletions effectively. Therefore, the k-mer natural vector is employed to characterize the compositions and distribution features of k-mers occurring in a viral genome, and construct a one-to-one relationship between a viral genome and its k-mer natural vector. Utilizing the k-mer natural vector, we proposed a novel genetic network to investigate the variations of viral genomes in transmission among humans. With the assistance of genetic network, we identified the super-spreaders that were responsible for the pandemic outbreaks all over the world and chose the parental strains to evaluate the effectiveness of diagnostics, therapeutics, and vaccines. The obtaining results fully demonstrated that our genetic network can truly describe the relationships of viral genomes, effectively simulate virus spread tendency, and trace the transmission routes precisely. In addition, this work indicated that the k-mer natural vector has the ability to capture established hotspots of diversities existing in the viral genomes and understand how genomic contents change over time.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread all over the world since the outbreak of coronavirus disease 2019 (COVID-19) in December 2019, which has wreaked havoc on public health in most countries (Wu et al., 2020a,b; Lam, 2020). Exploring the variations in viral genomes through a phylogenetic analysis is key to control the pandemic and stop its waves (Kissler et al., 2020; Oude Munnink et al., 2020).

Phylogenetic analysis always starts with the creation of a phylogenetic tree (Wu et al., 2020a,b; Zhou et al., 2020; Lu et al., 2020). However, the tree-building methods couldn't facilitate precise interpretation in relationships of SARS-CoV-2 genomes, as these genomes are shown with low genetic divergence (Li et al., 2020a,b; Wen et al., 2020). Moreover, different parts in a viral genome are thought as products of different genealogical trees (Rasmussen et al., 2014). Hence, several studies applied the genetic network, which is usually constructed with Maximum-likelihood (ML) or Bayesian methods, to depict the relationships among SARS-CoV-2 genomes (Nie et al., 2020; Gudbjartsson et al., 2020).

Coronaviruses are highly recombinogenic (Boni et al., 2020; Seemann et al., 2020). Closely related viral lineages might have different nucleotide substitution rates (Wertheim et al., 2012; Dearlove et al., 2020). To address these, more flexible models that make no assumption on nucleotide substitution rates and allow for variations in the substitution rate over time have to be developed (Li et al., 2020a,b). In addition, several SARS-CoV-2 isolates have been found to have deletions in the genome, which cannot well handled with current phylogenetic

Abbreviations: ACE2, angiotensin-converting enzyme 2; COVID-19, coronavirus disease 2019; dN, nonsynonymous substitution rate; dS, synonymous substitution rate; ML, Maximum-likelihood; MSA, Multiple sequence alignment; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; TMRCA, the most recent common ancestor; tSNE, t-Distributed Stochastic Neighbor Embedding.

* Corresponding authors.
E-mail addresses: wenjia@cuhk.edu.cn (J. Wen), qhpan@dlut.edu.cn (Q. Pan).

[1] These authors contributed equally to this work.

[2] Present address: National Clinical Research Center for Infectious Diseases, Shenzhen 518112, China

https://doi.org/10.1016/j.ympev.2022.107583
Received 25 January 2022; Received in revised form 31 May 2022; Accepted 15 June 2022
Available online 8 July 2022

methods (Young et al., 2020; Hu et al., 2021). It was verified that the k-mer model method can capture the recombination events and deal with the cases with deletions efficiently (Bauer et al., 2020). However, the k-mer approach is not suggested to tract transmission routes for its non-uniqueness. Therefore, the k-mer natural vector is employed to characterize the compositions and distribution features of k-mers occurring in a viral genome, and construct a one-to-one relationship between a viral genome and its k-mer natural vector (Wen et al., 2014). The k-mer natural vector makes no assumption on nucleotide substitution rates or base frequencies (Zhang et al., 2019, 2021).

Based on the k-mer natural vector, we developed a novel genetic network approach to investigate the variations of SARS-CoV-2 genomes in transmission among humans. It was demonstrated that our novel genetic network can effectively depict the genetic dynamics of viral genomes. In the following sections, we utilize three datasets of SARS-CoV-2 genomes to elucidate the validity of this new genetic network.

2. Results

2.1. Genetic cluster analysis for viral genomes

During transmission, viruses undergo adaptive evolution and generate genetic variants. To understand evolution selection of viral genomes, we performed a genetic cluster analysis for 158 SARS-CoV-2 genomes with a novel network representation (Fig. 1A–C). Four groups of G1 (Mann-Whitney U test, p = 4.6450e−69), G2 (Mann-Whitney U test, p = 0), G3 (Mann-Whitney U test, p = 3.3356e−05) and G4 (Mann-Whitney U test, p = 8.0578e−36) exist among 158 viral genomes (Fig. 1B). As shown in Fig. 2A and B and Table 1, the genomes in G1 are associated with the signature mutations of C8782T and T28144C; G2 includes the reference strain Wuhan/Hu-1; G3 has the changes of C241T, C3037T, and A23403G; G4 carries the G26144T mutation. Overall, these results are consistent with those obtained from ML or Bayesian methods (Li et al., 2020a,b; Yin, 2020).

In G1, two sub-clusters are distinguished with C29095T (Fig. 2B and Table S1). In the T-allele sub-cluster, Chinese individuals, who are all from Guangdong, have an average of 3.33 mutations, while the mutation rate for the others coming from Japan and the US is higher. In the C-allele sub-group, nearly half (15/33) of genomes are found outside Asia, mainly from the US (n = 7), Australia (n = 5), and Europe (n = 3). Genomes in G2, which are sampled from China and its adjacent regions (Table S2), have close connections and thus cluster together in the genetic network (Fig. 1B). Significantly, the genomes in G3 (Table S3), which were responsible for the explosive increase of COVID-19 in Europe and led to the global pandemic, couldn't be discerned with previous phylogenetic network (see Table 1, Forster et al., 2020). In addition, G4 is the major European type with representatives from England, France, Italy, and Sweden. Some have also been reported from Brazil, the US, Hong Kong, Singapore, South Korea, and Taiwan, but they are absent in the samples from China (Table S4).

Our genetic network can simulate the clinical representation and spread tendency effectively, which not only depicts transmission routes for documented cases, such as the infection paths of Mexico/CDMX-LnDRE_01-Italy/CDG1-Germany/BavPat1, Brazil/SPBR-02-Italy/SPL1, and Canada/On-PHL2445-USA/CA3 as stated in Forster et al. (2020), but also deduces undocumented transmission routes.

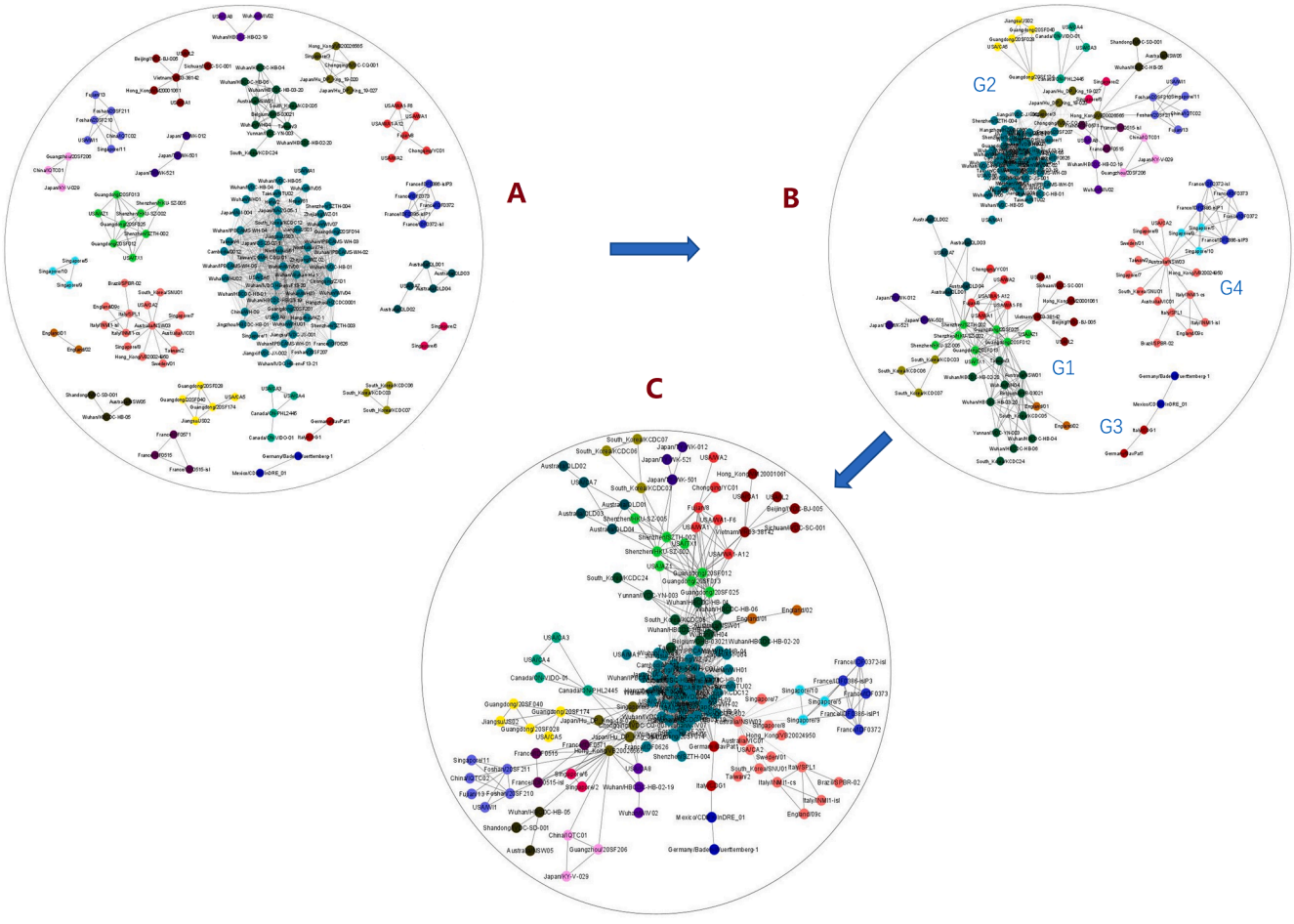

Fig. 1. Genetic cluster analysis for 158 SARS-CoV-2 genomes is performed with a novel network method that is shown with clustering process from A to B to C, in which four groups G1-G4 are clustered with the signature mutations.

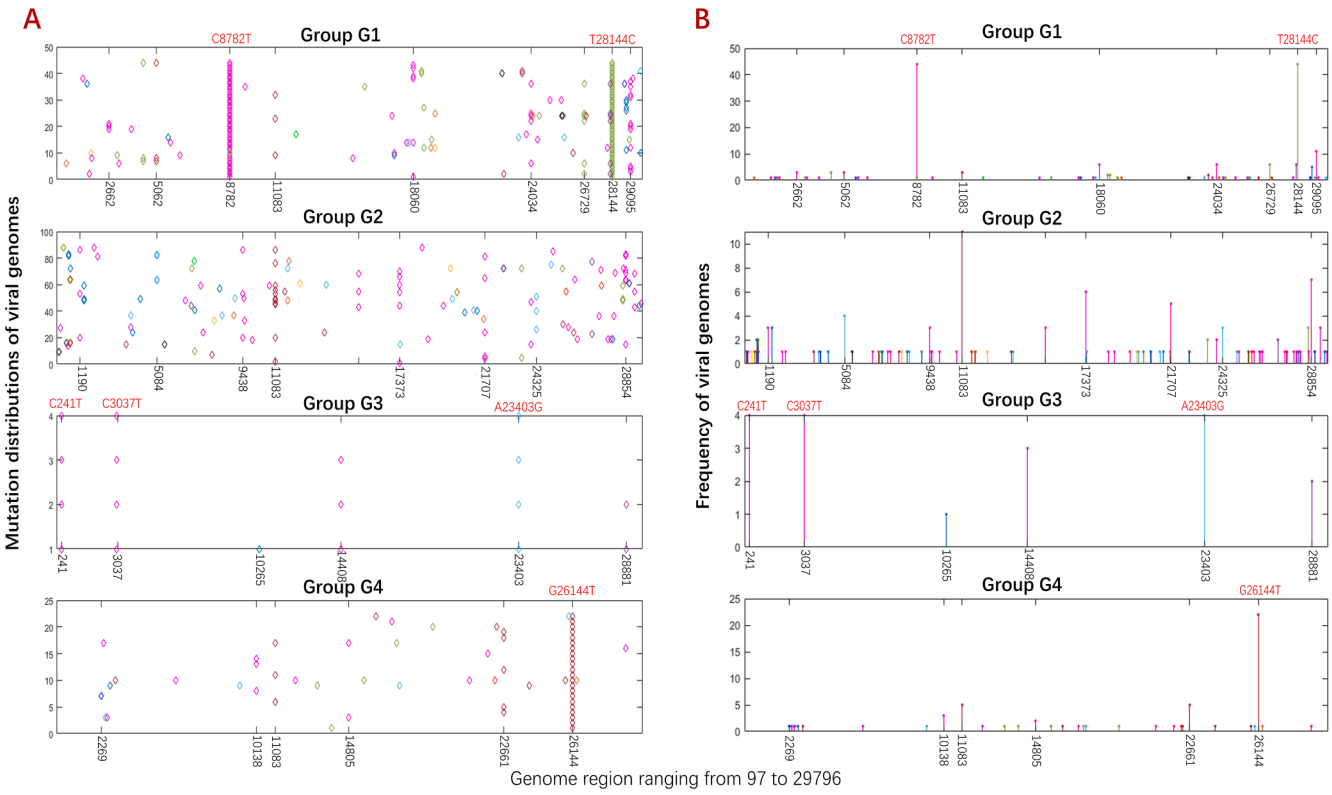

Fig. 2. Mutation profiles of 158 SARS-CoV-2 genomes, in which divergent mutations are depicted with different colors when compared with the reference strain Wuhan/Hu-1. (A) Mutation distributions in viral genomes from groups G1-G4 are shown, in which viruses are arranged in the column. (B) Frequencies of viral genomes for mutations occurring in groups G1-G4, respectively (see Tables S1–S4 in detail).

Table 1
Mutation analysis for genetic clusters of SARS-CoV-2 genomes in Datasets 1 and 2, respectively.

Cluster (Dataset 1)	Signature Mutation*	Cluster (Dataset 2)	Marker Mutation *	Cluster (Forster et al., 2020)
G1	C8782T, T28144C	C1	C8782T T28144C	Type A
		C2	C8782T C17747T A17858G, C18060T, T28144C	
G2	(Wuhan/Hu-1)	C3	(Wuhan/Hu-1)	Type B
G3	C241T, C3037T, A23403G	C4	C241T, C3037T, C14408T, A23403G	
G4	G26144T	C5	G26144T	Type C

* Signature Mutations and Marker Mutations are the common mutations for viral genomes in G1-G4 and C1-C5, respectively, when compared with the reference strain Wuhan/Hu-1.

2.2. Super-spreader identification and transmission route inference for viral genomes

It is important to trace transmission routes in control of the pandemic. Yang et al. (2020) probed the genetic dynamics of 247 SARS-CoV-2 genomes, in which four virus clusters were inferred to lead to the pandemic outbreaks worldwide. However, this result is not precise, since viruses with differential virulence have played different roles in transmission. We classified this dataset into five clades labelled as C1-C5 with our novel genetic network (Fig. 3A and Table 1). C1 has the marker mutations of C8782T and T28144C; the genomes in C2 have the changes of C8782T, C17747T, A17858G, C18060T, and T28144C, so C2 can be thought as a sub-clade of C1. C3 is equivalent to G2 as mentioned above; C4 has the mutations of C241T, C3037T, C14408T, and A23403G; C5 has the G26144T variation. It has again been verified that four clusters exist in SARS-CoV-2 genomes, which coincides with our former conclusion and results obtained by other researchers (Li et al., 2020a,b; Yin, 2020; Yang et al., 2020; Zhang et al., 2020).

Assisted with the genetic network for SARS-CoV-2 genomes, we identified the super-spreaders of viral genomes that are not only linked with components in its clade but also connected to elements from other clades. Obviously, Germany/BavPat1 and Australia/NSW03 are the super-spreaders from clades C4 and C5, respectively (Fig. 3A). The super-spreaders having contacts across different clades are thought to be responsible for the pandemic outbreaks worldwide (Fig. 3B). In addition, we performed a root-to-tip regression analysis of temporal signals for the collection dates of the super-spreaders (see Dataset 2) and inferred that the time of the most recent common ancestor (TMRCA) for SARS-CoV-2 was around November-December 2019 (Fig. 3C), indicating that these viral genomes shared a recent common ancestor but entered into different evolution clades.

Most mutations existing in viral genomes randomly occur in transmission without the function changes (Fig. 4A (see Tables S5-S9 in detail) and 4B and Table 2). Inspecting the genetic variations enables us to trace transmission routes. For example, USA/WA1 was found to be closely related to Fujian/8 through a genetic mutation analysis (Yang et al., 2020). This relationship can be directly observed from our genetic network without mutation analysis (Fig. 3A). Taking advantage of this new genetic network, we can trace transmission routes locally. Several strains obtained from the patients in South Korea (on 27 February 2020) are associated with mutations of T4402C and G5062T, meanwhile the isolates released in Beijing (on 28 January 2020) carry these mutations. Hence, it is inferred that the viruses in Koreans are in the same transmission routes with those from Beijing.

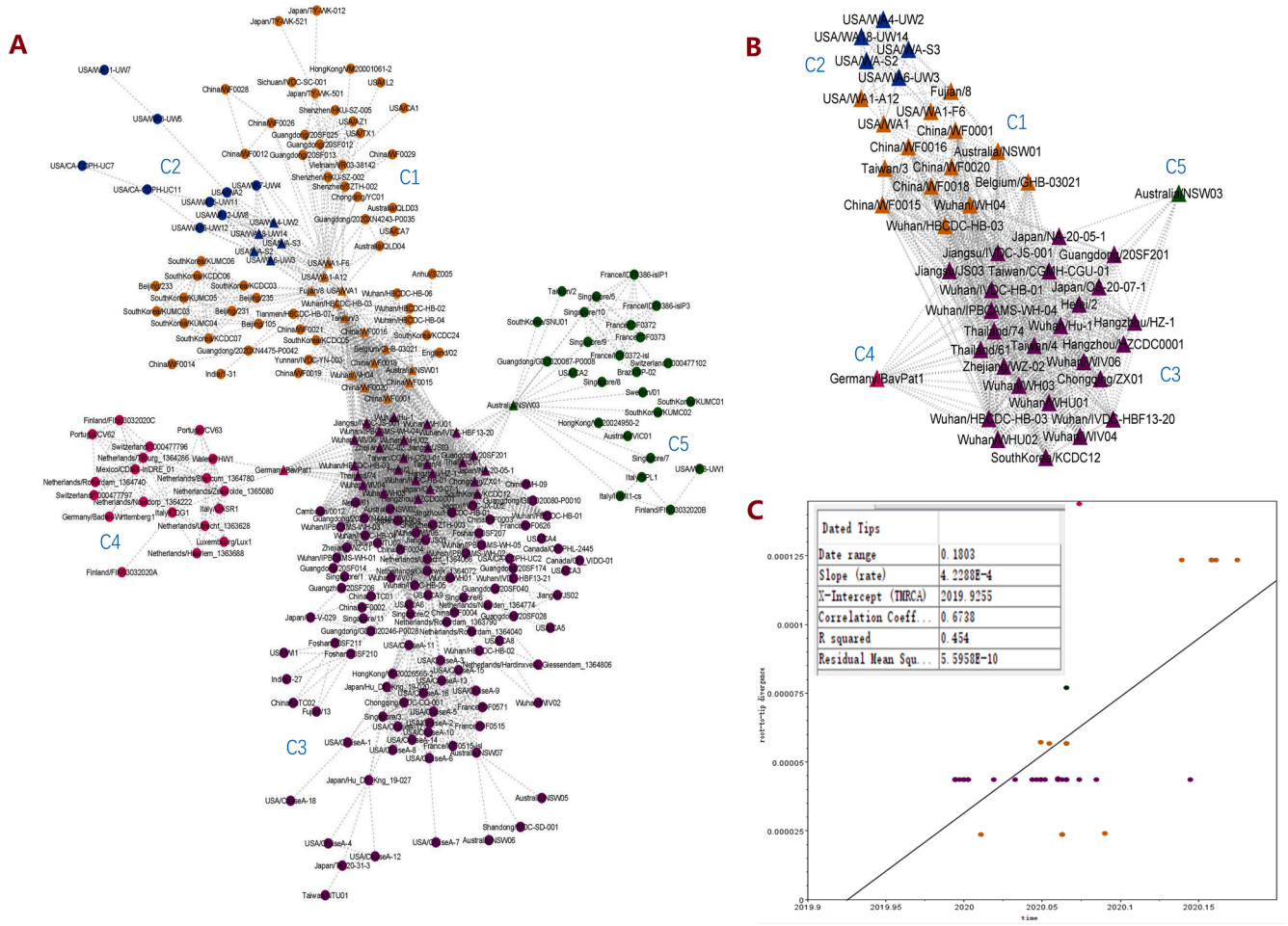

Fig. 3. (A) Genetic network of 247 SARS-CoV-2 genomes, in which five clades labelled as C1-C5 are classified with the marker mutations. (B) Assisted with the genetic network, we identified the super-spreaders of viral genomes that are not only linked with components in its clade but also connected to elements from other clades. (C) TMRCA was predicted with a root-to-tip regression analysis of temporal signals for the collection dates of the super-spreaders (Dataset 2).

2.3. Clinical consideration for viral genomes

Clinical response to a fast-moving infectious disease heavily depends on choosing the best virus strains in animal models (Bauer et al., 2020; Wohl et al., 2016). It is highly desirable that vaccine could be effective against all circulating virus strains. Although most mutations are random accumulation of genetic errors, several genetic variations have already been reserved. At present, nearly all vaccine candidates select Wuhan/Hu-1 as the target strain for the COVID-19, but a lot of evolution activities in viral genomes have not been considered in animal models (Gudbjartsson et al., 2020). It is preferred to choose the representative strains that can evaluate the effectiveness of diagnostics, therapeutics, and vaccines efficiently and comprehensively.

Genetic network of 184 SARS-CoV-2 isolates reveals four clusters (Fig. S1). Associating the genetic network with each cluster, we determined the parental strains to be the ones with the most linkages in the network. Moreover, these parental strains were shown with the least divergences for viral genomes from the same cluster. Hence, these parental strains were suggested as representatives of viral genomes to evaluate the effectiveness of diagnostics, therapeutics, and vaccines. Spike protein is a vital driver in virus evolution through binding with human receptor protein, so a molecular simulation for Spike protein from the parental strain contacting with human receptor protein ACE2 (angiotensin-converting enzyme 2) was modeled with PyMOL (Fig. 5A). Until now, several crystal structures for Spike protein complexed with the antibodies have been released in the PDB database. Using these structures, we could construct different Spike-antibody structure models, in which the mutation in Spike protein can be introduced by PyMOL. Since vaccine takes effect through its antibody in our body, the binding free energy between Spike protein and antibody, which can be calculated with conventional molecular dynamic simulation tools, such as Amber (Lee et al., 2020), Gromacs (Makarewicz and Kaźmierkiewicz, 2013), Molaris (Sham and Warshel, 1998), etc., is utilized to verify the feasibility of vaccine candidate. As shown in Fig. 5B, except one mutation (D614G in S1/S2), there is no variation occurring in Spike protein, meaning that our current vaccine candidates are effective to all circulating virus strains, which is identical to the result obtained from (Dearlove et al., 2020). In addition, the k-mer natural vector space for SARS-CoV-2 genomes was projected into a 2D image by tSNE (t-Distributed Stochastic Neighbor Embedding). Similarly, 184 viral genomes are grouped into four clusters, of which the parental strains are located at the central positions in each cluster (Fig. 6), indicating that the parental strains chosen as representatives for all circulating virus strains are feasible.

Viruses are associated unknown pathogen properties with genomic modifications, which complicates the evaluation of animal models and clinical research. The k-mer natural vector can reflect the fluidity of mutations and captures the recombination events efficiently. For instance, Singapore/4 and Taiwan/NTU01 that are located far from Wuhan/Hu-1 have found deletions in the core genome (Fig. 6). Scotland/CVR01 and Canada/BC_37_0-2 contain ambiguous bases that are not identified in sequencing process, which shorten the length of viral

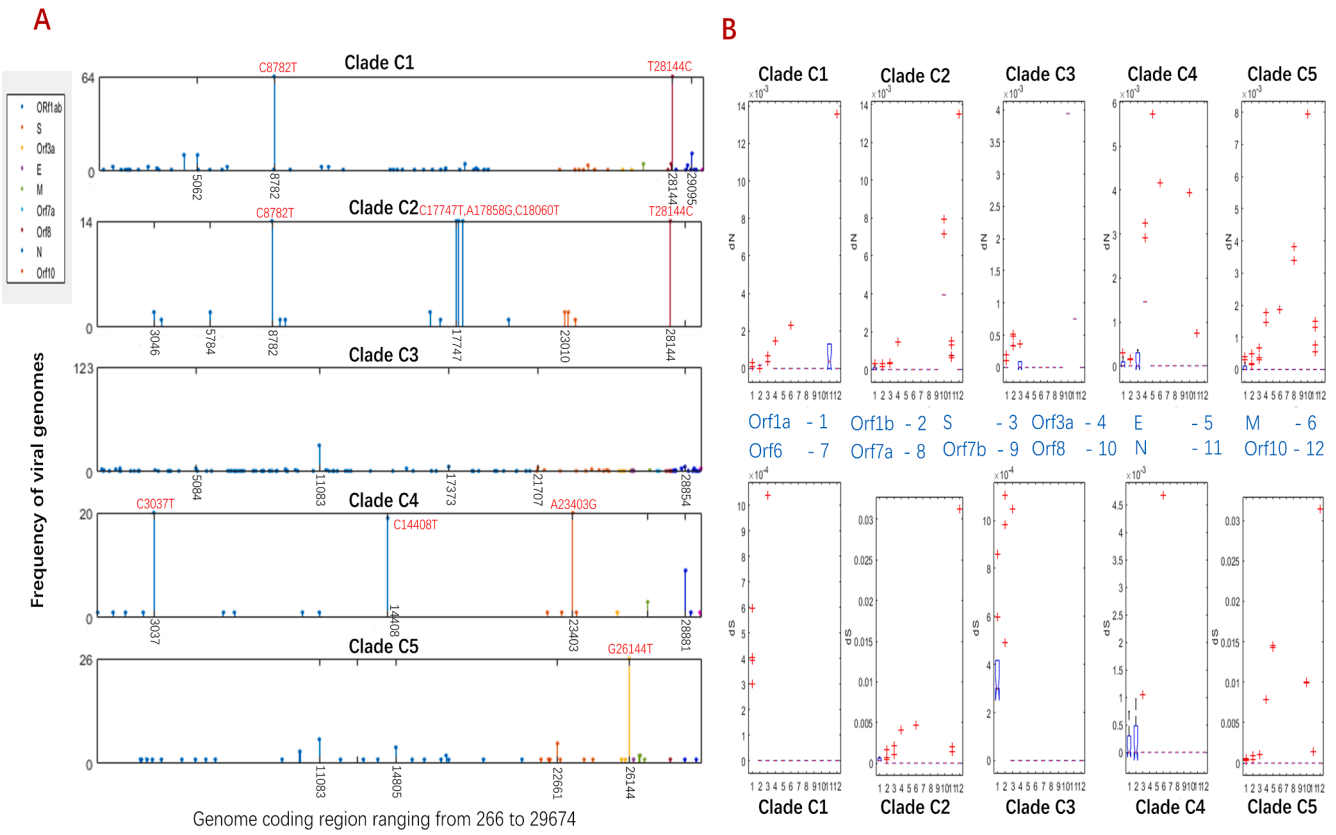

Fig. 4. Variations existing in the coding regions of SARS-CoV-2 genomes from clades C1-C5. (A) Mutations identified in the regions of Orf1ab[266:21555], S [21563:25384], Orf3a[25393:26220], E[26245:26472], M[26523:27191], Orf7a[27394:27759], Orf8[27894:28259], N[28274:29533], and Orf10[29558:29674] (Tables S5–S9); (B) Comparisons of dN (nonsynonymous substitution rate) and dS (synonymous substitution rate) in the coding regions of SARS-CoV-2 genomes.

Table 2
Mutation statistics for SARS-CoV-2 genomes in Dataset 2.

Clade	Number (n)	Nucleotide level		Amino acid level	
		Repetition	Non-repetition	Repetition	Non-repetition
C1	64	267/139	74/72	129/65	42/41
C2	14	86/16	16/11	50/9	9/6
C3	123	211	128	132	75
C4	20	125/86	24/20	71/52	16/15
C5	26	84/58	42/41	61/35	25/24

*(With/without the marker mutations) In a clade (C1-C5), one mutation may appear several times. The repeated mutations are only counted as 1 in the Non-repetition mode, while the Repetition model calculates its sum.

genome, so they are far from other viral gnomes from the same cluster, as well as Beijing/IVDC-BJ-005 and Shenzhen/SZTH-001 related to sequencing errors. It indicates that our k-mer natural vector has the ability to capture the established hotspots of diversities existing in viral genome and understand how genomic contents change over time.

3. Discussion

Based on the k-mer natural vector, a novel genetic network was conducted to explore the variations of viral genomes in transmission process. It has been verified that four genetic clusters exist among SARS-CoV-2 genomes. This result is consistent with those obtained from ML or Bayesian methods. Although several studies support the existence of four genetic clusters, their results are not totally correct. For example, the genomes having G11083T mutation exist in different clusters, so one cluster reported to carry G11083T as the divergent mutation is not credible (Yang et al., 2020). Moreover, our genetic cluster method can effectively simulate virus spread tendency, and identify the strains that tend to break out in later development. Genomes carrying the mutations of C241T, C3037T, and A23403G were rare at the beginning but led to an explosive increase of COVID-19 in Europe and developed into a global pandemic that cannot be discerned with the former phylogenetic network analysis (Forster et al., 2020).

It is elucidated that our genetic network can truly describe the relationships of viral genomes. Genomes with close linkages group together in the genetic network. Assisted with the genetic network, we identified the super-spreaders of viral genomes that are thought to be responsible for the pandemic outbreaks all over the world. In addition, this new genetic network enables us to understand mutation traits in virus adaptive evolution and trace transmission routes precisely. Currently, developing the effective vaccines is a higher priority in preventing and mitigating the waves of the pandemic. Viruses are associated with genetic variants in evolution process, and some mutations maybe prevail at local transmission regions, so it is necessary to monitor genetic mutations in viral genomes. Utilized the genetic network, the parental strains were chosen as representatives of viral genomes to evaluate the effectiveness of diagnostics, therapeutics, and vaccines. According to the genetic characteristics of parental strains, we could improve the diagnostics, guide therapeutics programs and vaccine development, and take timely adjustments when facing the changes of the pandemic.

In this work, the k-mer natural vector was used to deal with the recombination and deletions existing in viral genome and overcome the deficiencies of previous k-mer models. The k-mer natural vector has the ability to capture the established hotspots of diversities and understand how genomic contents change over time. One significant novelty of our k-mer natural vector is that each viral genome can be rigorously

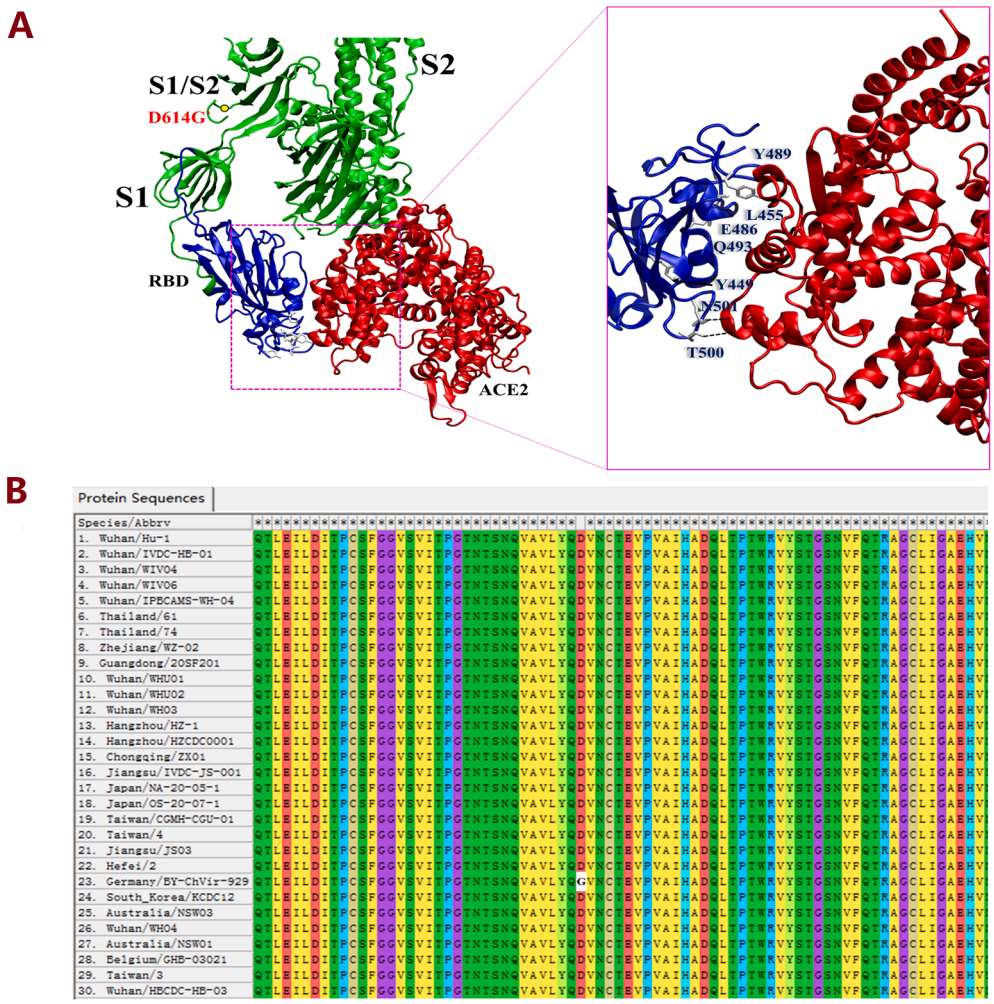

Fig. 5. (A) A molecular simulation for protein–protein interaction between Spike protein (PDB ID: 6M0J) and human receptor protein ACE2 (PDB ID: 7DZW) is modeled with PyMOL, in which the critical contact sites (Y449, L455, E486, Y489, Q493, T500, and N501) in ACE2 are shown; (B) The MSA for Spike proteins from the parental strains is presented, of which only one mutation (D614G in S1/S2 region) is observed from the strain of Germany/BY-ChVir-929.

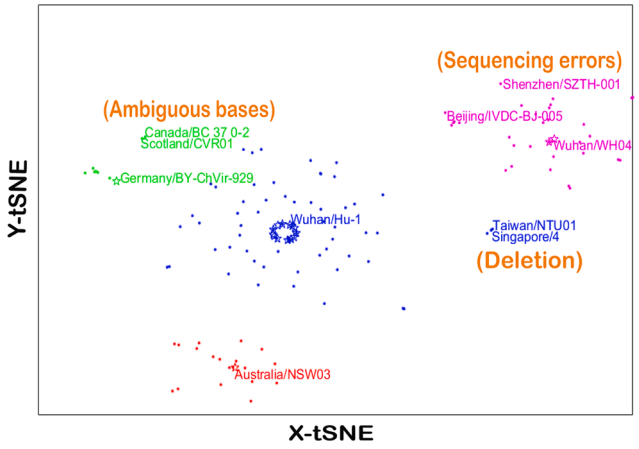

Fig. 6. The k-mer natural vector space for 184 SARS-CoV-2 genomes is projected into a 2D image by tSNE. All viruses are grouped into four clusters (denoted with different colors). The parental strains signed with pentagram are located at the central positions for each cluster, of which four parental strains of Wuhan/Hu-1, Wuhan/WH04, Australia/NSW03, and Germany/BY-ChVir-929 are labelled. Singapore/4 and Taiwan/NTU01 have deletions in the genome; Scotland/CVR01 and Canada/BC_37_0-2 contain ambiguous bases; Beijing/IVDC-BJ-005 and Shenzhen/SZTH-001 are related to sequencing errors.

recovered by its k-mer natural vector. Compared with alignment-based methods, the k-mer natural vector concerns the global similarities of viral genomes, such as the changes averaged across whole genome rather than at specific locations and require no evolution model or human intervention.

4. Material and methods

4.1. Data collection

Three datasets for SARS-CoV-2 genomes collected from GISAID database with basic sequence information in Dataset.xls were applied to elucidate the validity of our novel genetic network.

Dataset 1: 158 SARS-CoV-2 genomes were used to understand the evolution selection of viral genomes and simulate their spread tendencies among humans (Forster et al., 2020).

Dataset 2: Using the genetic network of 247 SARS-CoV-2 genomes, we identified the super-spreaders and traced the transmission routes for viral genomes (Yang et al., 2020).

Dataset 3: According to the genetic network for 184 SARS-CoV-2 genomes, in which the genomes with deletions, ambiguous bases, or sequencing errors were included, we determined the parental strains of viral genomes (Bauer et al., 2020).

4.2. K-mer natural vector for viral genome

Let s = 'N$_1$N$_2$⋯N$_L$' be a virus genome with s length L, L where N$_l$ ∈ {A,C,G,T}, $l = 1,2,\cdots,L$, and s[j][i] be the location of the i-th occurrence of a k-mer s[j] in s, $j = 1,2,\cdots,4^k$. For each given k, the distributions of a k-mer s[j] can be described by three quantities:

$n_{s[j]}$: Number of s [j] occurrences in s;

$\mu_{s[j]}$: Mean distance of s[j] from the first position of s;

$D_m^{s[j]}$: Central moment of s[j], that is,

$$D_m^{s[j]} = \sum_{i=1}^{n_{s[j]}} \frac{(s[j][i] - \mu_{s[j]})^m}{n_{s[j]}^{m-1}(L-k+1)^{m-1}}, \quad m = 1,2,\cdots,n_{s[j]}.$$

Hence, the k-mer natural vector for viral genome s could be defined by $(n_{s[j]}, \mu_{s[j]}, D_m^{s[j]})$, $j = 1,2,\cdots,4^k$.

By the definition, the k-mer natural vector concatenates the numbers of occurrence and mean distance for k-mer with its central moments, it thus contains the information of k-mers and avoids the deficiencies of previous k-mer model methods. Moreover, the relationship between a virus genome and its k-mer natural vector is one-to-one for each given k, which has been mathematically proved in Text S1. In addition, it has been verified that a k-mer natural vector with order two central moment is 3×4^k $n_{[i]}, \mu_{[i]}, D_2^{[i]}$ enough to represent a viral genome, so $(n_{s[j]}, \mu_{s[j]}, D_2^{s[j]})$ is sufficient to depict a viral genome, which still satisfies the one-to-one mapping.

4.3. Genetic network of viral genomes

For k-mer model method, the parameter k has a great influence on obtaining results and computational complexities. Following our former work (Wen et al., 2014), the optimal k should be within the range of $[\text{ceil}(\log_4 \min(L)), \text{ceil}(\log_4 \max(L)) + 1]$, where L is the set of lengths for viral genomes. Based on this, the value k chosen for the full-length SARS-CoV-2 genomes is 8. Once e $[\text{ceil}(\log_4 \min(L)), \text{ceil}(\log_4 \max(L)) + 1]$, $[\text{ceil}(\log_4 \min(L)), \text{ceil}(\log_4 \max(L)) + 1]$, ach SARS-CoV-2 genome is uniquely represented by a k-mer natural vector, the pairwise distance (d_{st}) for SARS-CoV-2 genomes can be calculated with Spearman distance:

$$d_{st} = 1 - \frac{(r_s - \overline{r_s})(r_t - \overline{r_t})}{\sqrt{(r_s - \overline{r_s})(r_s - \overline{r_s})'}\sqrt{(r_t - \overline{r_t})(r_t - \overline{r_t})'}}$$

where r_{sj} is the rank of x_{sj} taken over $x_{1j}, x_{2j}, \cdots, x_{mj}$, r_s and r_t are the coordinate-wise rank vectors of x_s and x_t, i.e. $r_s = (x_{s1}, x_{s2}, \cdots, x_{sn})$, $\overline{r_s} = \frac{1}{n}\sum_j r_{sj}$, $\overline{r_t} = \frac{1}{n}\sum_j r_{tj}$.

Genetic network for viral genomes in Dataset 1 can be built as follows:

(1) Genomes are connected if they have the shortest pairwise distance. Among 158 SARS-CoV-2 genomes, Italy/CDG1 and Germany/BavPat1 are closest with the shortest pairwise distance, so they are connected. There are 23 genetic clusters built in Fig. 1A.

(2) The distance between genetic clusters is the mean of pairwise distances. For example, the distance between genetic clusters of Germany/BavPat1-Italy/CDG1 and Germany/Baden-Wuerttemberg-1-Mexico/CDMX-InDRE_01 is the mean of four pairwise distances.

(3) Two genetic clusters should be linked if the distance between genetic clusters is the shortest, of which the genomes with the shortest pairwise distance are connected. Genetic clusters of Germany/Baden-Wuerttemberg-1-Mexico/CDMX-InDRE_01 and Germany/BavPat1-Italy/CDG1 are shown with the shortest distance. Meanwhile, Italy/CDG1 and Mexico/CDMX-InDRE_01 are closest among four genomes. Therefore, two genetic clusters form Group G3 in Fig. 1B, in which Italy/CDG1 and Mexico/CDMX-InDRE_01 are linked. Four groups G1-G4 in Fig. 1B are constructed from 23 genetic clusters.

(4) The distances between groups G1-G4 are updated as (2). The genetic network of viral genomes is completed when all the genomes are linked together (Fig. 1C).

4.4. Phylogenetic analysis

Multiple sequence alignment (MSA) for SARS-CoV-2 genomes was conducted with MAFFT (Katoh and Standley, 2013). Mutation analysis for nucleotide or amino acid was inspected with the reference strain Wuhan/Hu-1 using MEGA7 (Kumar et al., 2016). A root-to-tip regression analysis of temporal signals for the collection dates of the super-spreaders was visualized with TempEst (Rambaut et al., 2016). Genetic network of viral genomes was illustrated with Cytoscape 3.8.0 (Shannon et al., 2003). In addition, we applied the Mann-Whitney U test to verify the validity of the cluster results.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yuyan Zhang: Data curation, Methodology, Formal analysis, Writing – original draft. **Jia Wen:** Conceptualization, Methodology, Formal analysis, Validation, Writing – original draft. **Kun Xi:** Software, Visualization. **Qiuhui Pan:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files.

Acknowledgments

We sincerely thank the authors of the coronavirus related data from GISAID database.

Funding

This work is partially supported by Natural Scientific Research Funding of Heilongjiang (Grant No. LH2019A031).

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.ympev.2022.107583.

References

Bauer, D.C., Tay, A.P., Wilson, L.O.W., Reti, D., Hosking, C., McAuley, A.J., Pharo, E., Todd, S., Stevens, V., Neave, M.J., Tachedjian, M., Drew, T.W., Vasan, S.S., 2020. Supporting pandemic response using genomics and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. Transbound Emerg. Dis. 67 (4), 1453–1462.

Boni, M.F., Lemey, P., Jiang, X., Lam, T.-Y., Perry, B.W., Castoe, T.A., Rambaut, A., Robertson, D.L., 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. 5 (11), 1408–1417.

Dearlove, B., Lewitus, E., Bai, H., Li, Y., Reeves, D.B., Joyce, M.G., Scott, P.T., Amare, M. F., Vasan, S., Michael, N.L., Modjarrad, K., Rolland, M., 2020. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc. Natl. Acad. Sci. U. S. A. 117 (38), 23652–23662.

Forster, P., Forster, L., Renfrew, C., Forster, M., 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc. Natl. Acad. Sci. U. S. A. 117 (17), 9241–9243.

Gudbjartsson, D.F., Helgason, A., Jonsson, H., Magnusson, O.T., Melsted, P., Norddahl, G.L., Saemundsdottir, J., Sigurdsson, A., Sulem, P., Agustsdottir, A.B., Eiriksdottir, B., Fridriksdottir, R., Gardarsdottir, E.E., Georgsson, G., Gretarsdottir, O.S., Gudmundsson, K.R., Gunnarsdottir, T.R., Gylfason, A., Holm, H., Jensson, B.O., Jonasdottir, A., Jonsson, F., Josefsdottir, K.S., Kristjansson, T., Magnusdottir, D.N., le Roux, L., Sigmundsdottir, G., Sveinbjornsson, G., Sveinsdottir, K.E., Sveinsdottir, M., Thorarensen, E.A., Thorbjornsson, B., Löve, A., Masson, G., Jonsdottir, I., Möller, A.D., Gudnason, T., Kristinsson, K.G., Thorsteinsdottir, U., Stefansson, K., 2020. Spread of SARS-CoV-2 in the Icelandic Population. N. Engl. J. Med. 382 (24), 2302–2315.

Hu, B., Guo, H., Zhou, P., Shi, Z.-L., 2021. Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol. 19 (3), 141–154.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780.

Kissler, S.M., Tedijanto, C., Goldstein, E., Grad, Y.H., Lipsitch, M., 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science 368 (6493), 860–868.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33 (7), 1870–1874.

Lam, T.-Y., 2020. Tracking the Genomic Footprints of SARS-CoV-2 Transmission. Trends. Genet. 36 (8), 544–546.

Lee, T.S., Allen, B.K., Giese, T.J., Guo, Z., Li, P., Lin, C., McGee Jr., T.D., Pearlman, D.A., Radak, B.K., Tao, Y., Tsai, H.C., Xu, H., Sherman, W., York, D.M., 2020. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. J. Chem. Inf. Model 60, 5595–5623.

Li, J., Li, Z., Cui, X., Wu, C., 2020a. Bayesian phylodynamic inference on the temporal evolution and global transmission of SARS-CoV-2. J. Infect. 81, 318–356.

Li, X., Zai, J., Zhao, Q., Nie, Q., Li, Y., Foley, B.T., Chaillon, A., 2020b. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J. Med. Virol. 92, 602–611.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W.J., Wang, D., Xu, W., Holmes, E.C., Gao, G.F., Wu, G., Chen, W., Shi, W., Tan, W., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395, 565–574.

Makarewicz, T., Kaźmierkiewicz, R., 2013. Molecular dynamics simulation by GROMACS using GUI plugin for PyMOL. J. Chem. Inf. Model. 53, 1229–1234.

Nie, Q., Li, X., Chen, W., Liu, D., Chen, Y., Li, H., Li, D., Tian, M., Tan, W., Zai, J., 2020. Phylogenetic and phylodynamic analyses of SARS-CoV-2. Virus Res. 287, 198098.

Oude Munnink, B.B., Nieuwenhuijse, D.F., Stein, M., O'Toole, Á., Haverkate, M., Mollers, M., Kamga, S.K., Schapendonk, C., Pronk, M., Lexmond, P., van der Linden, A., Bestebroer, T., Chestakova, I., Overmars, R.J., van Nieuwkoop, S., Molenkamp, R., van der Eijk, A.A., GeurtsvanKessel, C., Vennema, H., Meijer, A., Rambaut, A., van Dissel, J., Sikkema, R.S., Timen, A., Koopmans, M., 2020. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat. Med. 26, 1405–1410.

Rambaut, A., Lam, T.T., Max Carvalho, L., Pybus, O.G., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2, vew007.

Seemann, T., Lane, C.R., Sherry, N.L., Duchene, S., Gonçalves da Silva, A., Caly, L., Sait, M., Ballard, S.A., Horan, K., Schultz, M.B., Hoang, T., Easton, M., Dougall, S., Stinear, T.P., Druce, J., Catton, M., Sutton, B., van Diemen, A., Alpren, C., Williamson, D.A., Howden, B.P., 2020. Tracking the COVID-19 pandemic in Australia using genomics. Nat. Commun. 11, 4376.

Rasmussen, M.D., Hubisz, M.J., Gronau, I., Siepel, A., 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10, e1004342.

Sham, Y.Y., Warshel, A., 1998. The surface constrained all atom model provides size independent results in calculations of hydration free energies. J. Chem. Phys. 109, 7940–7944.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504.

Wen, F., Yu, H., Guo, J., Li, Y., Luo, K., Huang, S., 2020. Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2. J. Infect. 80, 671–693.

Wen, J., Chan, R.H., Yau, S.C., He, R.L., Yau, S.S., 2014. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene 546, 25–34.

Wertheim, J.O., Fourment, M., Kosakovsky Pond, S.L., 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. Mol. Biol. Evol. 29, 451–456.

Wohl, S., Schaffner, S.F., Sabeti, P.C., 2016. Genomic Analysis of Viral Outbreaks. Annu. Rev. Virol. 3, 173–195.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G., Jiang, T., 2020a. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe 27, 325–328.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C., Zhang, Y.Z., 2020b. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269.

Yang, X., Dong, N., Chan, E.W., Chen, S., 2020. Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. Emerg. Microbes Infect. 9, 1287–1299.

Yin, C., 2020. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics 112, 3588–3596.

Young, B.E., Fong, S.W., Chan, Y.H., Mak, T.M., Ang, L.W., Anderson, D.E., Lee, C.Y., Amrun, S.N., Lee, B., Goh, Y.S., Su, Y.C.F., Wei, W.E., Kalimuddin, S., Chai, L.Y.A., Pada, S., Tan, S.Y., Sun, L., Parthasarathy, P., Chen, Y.Y.C., Barkham, T., Lin, R.T.P., Maurer-Stroh, S., Leo, Y.S., Wang, L.F., Renia, L., Lee, V.J., Smith, G.J.D., Lye, D.C., Ng, L.F.P., 2020. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet 396, 603–611.

Zhang, J., Ma, K., Li, H., Liao, M., Qi, W., 2020. The continuous evolution and dissemination of 2019 novel human coronavirus. J. Infect. 80, 671–693.

Zhang, Y., Wen, J., Li, X., Li, G., 2021. Exploration of hosts and transmission traits for SARS-CoV-2 based on the k-mer natural vector. Infect. Genet. Evol. 93, 104933.

Zhang, Y., Wen, J., Yau, S.S., 2019. Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. Genomics 111, 1298–1305.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.