Infection, Genetics and Evolution 96 (2021) 105106

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Alignment-free sequence comparison for virus genomes based on location correlation coefficient

Lily He[a,*], Siyang Sun[b], Qianyue Zhang[b], Xiaona Bao[a], Peter K. Li[c,*]

[a] *School of Science, Beijing University of Civil Engineering and Architecture, Beijing 102616, PR China*
[b] *The High School Affiliated to Renmin University of China, Beijing 100080, PR China*
[c] *School of Life Sciences, Tsinghua University, Beijing 100084, PR China*

ARTICLE INFO

Keywords:
SARS-CoV-2
Alignment-free
Correlation measure
DNA sequence

ABSTRACT

Coronaviruses (especially SARS-CoV-2) are characterized by rapid mutation and wide spread. As these characteristics easily lead to global pandemics, studying the evolutionary relationship between viruses is essential for clinical diagnosis. DNA sequencing has played an important role in evolutionary analysis. Recent alignment-free methods can overcome the problems of traditional alignment-based methods, which consume both time and space. This paper proposes a novel alignment-free method called the correlation coefficient feature vector (CCFV), which defines a correlation measure of the L-step delay of a nucleotide location from its location in the original DNA sequence. The numerical feature is a $16 \times L$-dimensional numerical vector describing the distribution characteristics of the nucleotide positions in a DNA sequence. The proposed L-step delay correlation measure is interestingly related to some types of $L+1$ spaced mers. Unlike traditional gene comparison, our method avoids the computational complexity of multiple sequence alignment, and hence improves the speed of sequence comparison. Our method is applied to evolutionary analysis of the common human viruses including SARS-CoV-2, Dengue virus, Hepatitis B virus, and human rhinovirus and achieves the same or even better results than alignment-based methods. Especially for SARS-CoV-2, our method also confirms that bats are potential intermediate hosts of SARS-CoV-2.

1. Introduction

The worldwide outbreak of the SARS-CoV-2 virus has necessitated a deeper understanding of the transmission path, evolutionary process, and other dynamics of viruses. Since the first appearance of novel pneumonia (COVID-19) in Wuhan, Hubei province, China, there has been a lot of discussion on the origin of the causative virus, SARS-CoV-2 (Sironi et al., 2020). SARS-CoV-2 is the seventh coronavirus known to infect humans: SARS-CoV, MERS-CoV and SARS-CoV-2 can cause severe disease. Hepatitis B virus (HBV) is another infectious virus that can establish a persistent and chronic infection in humans through immune energy. In 2016, worldwide estimates suggest that 257 million people are chronically infected with the Hepatitis B virus (HBV). About 15% to 25% of them may die from cirrhosis or liver cancer (Nelson et al., 2016). HBV is a partially double-stranded DNA virus and a member of the hepadnaviridae family. Dengue viruses can also spread to people via the bite of an infected Aedes species mosquito. About 40% of the world's population living in areas with a risk of dengue (Tsang et al., 2019). Human rhinoviruses is one of the most common viruses in humans and is the predominant cause of the common cold. Sequence analysis is a popular and effective technique for studying these viruses, and genetic sequence comparison has become a crucial procedure in many modern biological techniques.

Sequence comparison is traditionally performed by alignment-based methods. These methods often attempt to maximize an alignment score calculated as the sum of substitution scores minus gap penalties. The popular algorithms include Smith-Waterman, Needleman-Wunch, Muscle, etc (Edgar, 2004; Chookajorn, 2020). However, alignment-based methods are burdened by large memory and time consumption, and may be affected by high mutation and recombination rates (Vinga, 2014). These problems can be overcome by alignment-free methods, which are gaining attraction in biological fields. Alignment-free methods quantify sequence similarity/dissimilarity that does not use or produce alignment at any step of algorithm application. Since alignment-free methods do not rely on dynamic programming, they are computationally less expensive and therefore suitable for whole

* Corresponding authors.
E-mail addresses: lilyhe6@163.com (L. He), peter_wondrrt@outlook.com (P.K. Li).

https://doi.org/10.1016/j.meegid.2021.105106
Received 20 June 2021; Received in revised form 8 September 2021; Accepted 3 October 2021
Available online 6 October 2021

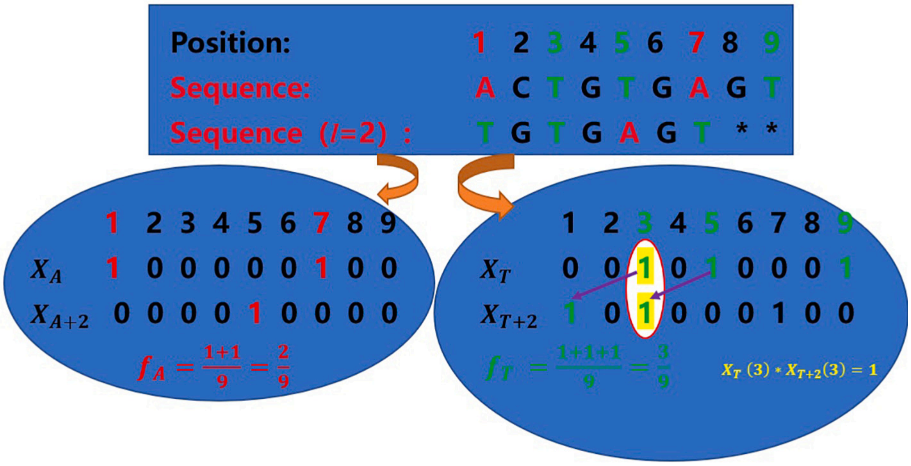
Position: 1 2 3 4 5 6 7 8 9
Sequence: A C T G T G A G T
Sequence (l=2) : T G T G A G T * *
1 2 3 4 5 6 7 8 9
X_A 1 0 0 0 0 0 1 0 0
X_{A+2} 0 0 0 0 1 0 0 0 0
$f_A = \frac{1+1}{9} = \frac{2}{9}$
1 2 3 4 5 6 7 8 9
X_T 0 0 1 0 1 0 0 0 1
X_{T+2} 1 0 1 0 0 0 1 0 0
$f_T = \frac{1+1+1}{9} = \frac{3}{9}$
$X_T(3) * X_{T+2}(3) = 1$

Fig. 1. The process of calculating the X_t, X_{t+2} and f_t when t is A and T, respectively.

genome comparisons.

An alignment-free method called Subsequence Natural Vector (SNV) was proposed to subtyping HIV-1 genomes and achieved very high sensitivity and specificity (He et al., 2020). The alignment-free method Natural Vector has achieved very high accuracy in classification of viruses including Ebola, Zika and HBV, etc. (Li et al., 2017; Deng et al., 2011). The Chaos Game Representation has succeeded in species classification (Joel, 1990; Randhawa et al., 2020b). A machine learning method combined with digital signal processing was used for rapid classification of novel pathogens especially SARS-CoV-2 (Randhawa et al., 2020a). Among the alignment-free approaches, $k-$mer methods based on the frequencies of subsequences of a defined length k were widely used in sequence analysis (Kurtz et al., 2008). A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes (Zielezinski et al., 2019; Bernard et al., 2019). Such as, feature frequency profiles (FFP) is one popular k-mer based method that considers the frequencies, positions, and other properties of a k bp-length substring in a genome (Wu et al., 2009), and recently a new method which is based on k-mer is proposed, this method can be effectively used for sequence comparison (Sarkar et al., 2021).

This paper proposes a numerical feature vector that describes the nucleotide distribution of viral gene sequences. A DNA sequence is first mapped to four binary vectors, whose elements are assigned 0 or 1 depending on the locations of the four nucleotides along the sequence. Second, these four vectors are processed by our proposed correlation coefficient feature vector (CCFV) method, which calculates the correlation coefficient and auto-correlation coefficient of an L-step delay from the original sequence. Third, from 1 to L-steps, these correlation coefficients are compiled into a $4^2 \times L$-dimensional vector. This method represents a sequence by a low-dimensional numerical vector, which greatly improves the construction speed of phylogenetic tree. The similarity between two viruses is then determined by the Euclidean distance between their $4^2 \times L$-dimensional vectors. Finally, a detailed evolutionary tree is drawn by the unweighted pair group method with arithmetic mean (UPGMA). Our method is shown to obtain the correct evolutionary relationship. According to biological features of bat and the very close evolutionary relationship between bat-nCoV and SARS-CoV-2 as shown in the phylogenetic tree constructed by our method, we demonstrate that bats seem to be the natural reservoir of SARS-CoV-2 causing the recent COVID-19 outbreak.

2. Methods

2.1. Location correlation coefficient

Let $S = s_1 s_2 \cdots _N$ be a DNA sequence with $s_i \in \{A, C, G, T\}$. To obtain the important information of a nucleotide t, $t \in \{A, C, G, T\}$ in the sequence, we first convert the sequence into a indicator vector as follows: Let $X_t = (x_t(1), x_t(2), \cdot, x_t(N))$, $t \in \{A, C, G, T\}$, where

$$x_t = \begin{cases} 1, & \text{when } s_i = t \\ 0, & \text{else} \end{cases} \tag{1}$$

We define a L-step delay numerical sequence X_{t+L} of X_t as:

$$X_{t+L} = (x_t(L+1), x_t(L+2), \cdots, x_t(L+N)) \tag{2}$$

where $x_t(N+1) = \cdots = x_t(l+N) = 0$, especially, $X_t = X_{t+0}$. Then we define the average occurrence number of $t \in \{A, C, G, T\}$ as:

$$f_t = \frac{1}{N}(x_t(1) + x_t(2) + \cdots + x_t(N)). \tag{3}$$

The L-step delay correlation of locations in nucleotide, $\rho_{tt}(L)$, is defined as:

$$\rho_{tt}(L) = \frac{1}{N}\sum_{i=1}^{N}(x_t(i) - f_t)(x_{t+L}(i) - f_t), \tag{4}$$

Example: Given a sequence="ACTGTGAGT", we have N=9, and suppose that $L = 2$ and t is A and T, respectively. Then:

$X_A = 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0; X_T = 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 1;$

$X_{A+2} = 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0; X_{T+2} = 1\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0;$

$$f_A = \frac{1}{9}(x_A(1) + x_A(2) + \cdots + x_A(9)) = \frac{1}{9}(1+1) = \frac{2}{9};$$

$$f_T = \frac{1}{9}(x_T(1) + x_T(2) + \cdots + x_T(9)) = \frac{1}{9}(1+1+1) = \frac{3}{9};$$

The detailed calculation process is shown in as shown in Fig. 1.

Similarly, we define the correlation $\rho_{st}(L)$ of nucleotides s and t in a sequence as follows:

$$\rho_{st}(L) = \frac{1}{N}\sum_{i=1}^{N}(x_s(i) - f_s)(x_{t+L}(i) - f_t) \tag{5}$$

To normalize the above correlation, we finally define the correlation coefficient $\tau_{st}(L)$ between nucleotides s and t as:

$$\tau_{st}(L) = \frac{\rho_{st}(L)}{\sqrt{\rho_{ss}(0) * \rho_{tt}(0)}} \tag{6}$$

The auto-correlation coefficient $\tau_{tt}(l)$ of nucleotide t itself is defined as

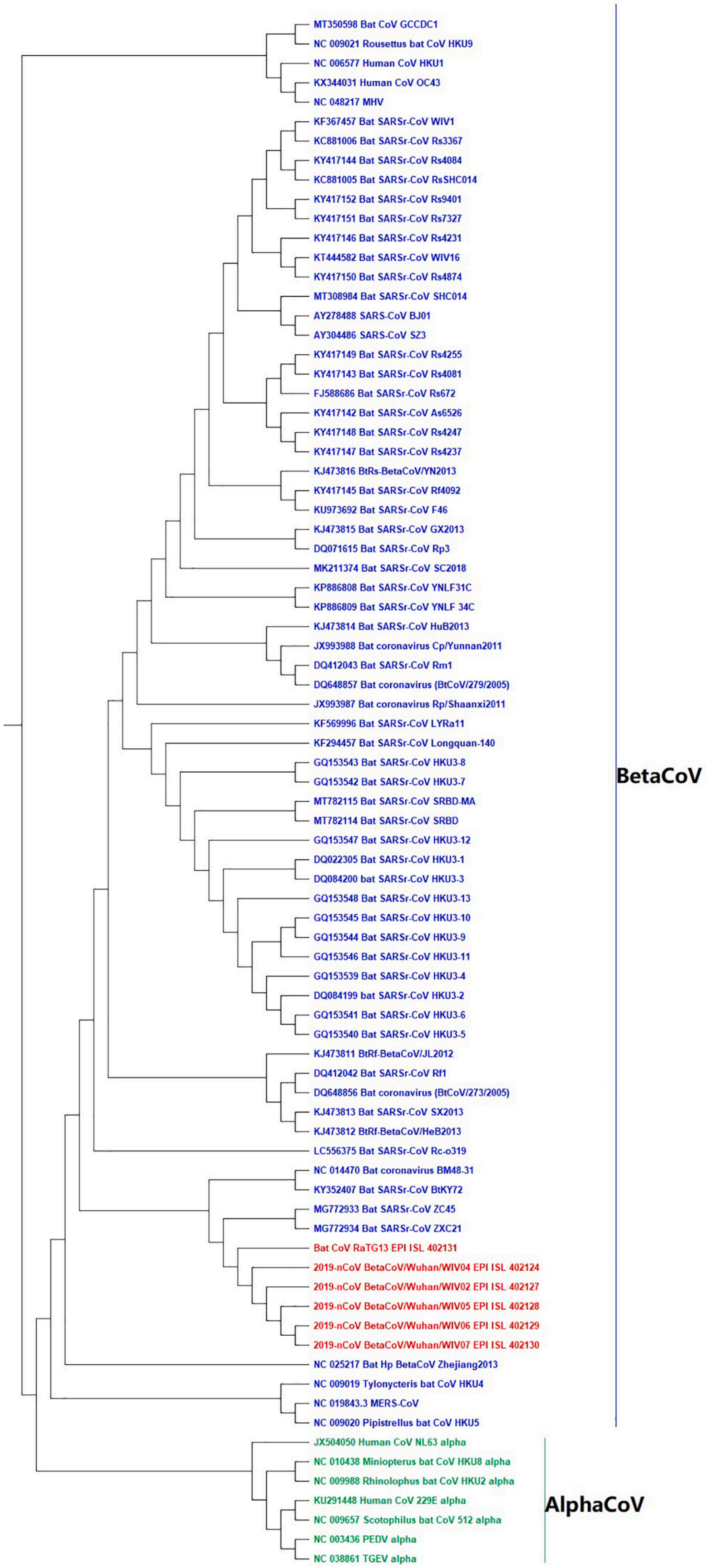
MT350598 Bat CoV GCCDC1
NC 009021 Rousettus bat CoV HKU9
NC 006577 Human CoV HKU1
KX344031 Human CoV OC43
NC 048217 MHV
KF367457 Bat SARSr-CoV WIV1
KC881006 Bat SARSr-CoV Rs3367
KY417144 Bat SARSr-CoV Rs4084
KC881005 Bat SARSr-CoV RsSHC014
KY417152 Bat SARSr-CoV Rs9401
KY417151 Bat SARSr-CoV Rs7327
KY417146 Bat SARSr-CoV Rs4231
KT444582 Bat SARSr-CoV WIV16
KY417150 Bat SARSr-CoV Rs4874
MT308984 Bat SARSr-CoV SHC014
AY278488 SARS-CoV BJ01
AY304486 SARS-CoV SZ3
KY417149 Bat SARSr-CoV Rs4255
KY417143 Bat SARSr-CoV Rs4081
FJ588686 Bat SARSr-CoV Rs672
KY417142 Bat SARSr-CoV As6526
KY417148 Bat SARSr-CoV Rs4247
KY417147 Bat SARSr-CoV Rs4237
KJ473816 BtRs-BetaCoV/YN2013
KY417145 Bat SARSr-CoV Rf4092
KU973692 Bat SARSr-CoV F46
KJ473815 Bat SARSr-CoV GX2013
DQ071615 Bat SARSr-CoV Rp3
MK211374 Bat SARSr-CoV SC2018
KP886808 Bat SARSr-CoV YNLF31C
KP886809 Bat SARSr-CoV YNLF 34C
KJ473814 Bat SARSr-CoV HuB2013
JX993988 Bat coronavirus Cp/Yunnan2011
DQ412043 Bat SARSr-CoV Rm1
DQ648857 Bat coronavirus (BtCoV/279/2005)
JX993987 Bat coronavirus Rp/Shaanxi2011
KF569996 Bat SARSr-CoV LYRa11
KF294457 Bat SARSr-CoV Longquan-140
GQ153543 Bat SARSr-CoV HKU3-8
GQ153542 Bat SARSr-CoV HKU3-7
MT782115 Bat SARSr-CoV SRBD-MA
MT782114 Bat SARSr-CoV SRBD
GQ153547 Bat SARSr-CoV HKU3-12
DQ022305 Bat SARSr-CoV HKU3-1
DQ084200 bat SARSr-CoV HKU3-3
GQ153548 Bat SARSr-CoV HKU3-13
GQ153545 Bat SARSr-CoV HKU3-10
GQ153544 Bat SARSr-CoV HKU3-9
GQ153546 Bat SARSr-CoV HKU3-11
GQ153539 Bat SARSr-CoV HKU3-4
DQ084199 bat SARSr-CoV HKU3-2
GQ153541 Bat SARSr-CoV HKU3-6
GQ153540 Bat SARSr-CoV HKU3-5
KJ473811 BtRf-BetaCoV/JL2012
DQ412042 Bat SARSr-CoV Rf1
DQ648856 Bat coronavirus (BtCoV/273/2005)
KJ473813 Bat SARSr-CoV SX2013
KJ473812 BtRf-BetaCoV/HeB2013
LC556375 Bat SARSr-CoV Rc-o319
NC 014470 Bat coronavirus BM48-31
KY352407 Bat SARSr-CoV BtKY72
MG772933 Bat SARSr-CoV ZC45
MG772934 Bat SARSr-CoV ZXC21
Bat CoV RaTG13 EPI ISL 402131
2019-nCoV BetaCoV/Wuhan/WIV04 EPI ISL 402124
2019-nCoV BetaCoV/Wuhan/WIV02 EPI ISL 402127
2019-nCoV BetaCoV/Wuhan/WIV05 EPI ISL 402128
2019-nCoV BetaCoV/Wuhan/WIV06 EPI ISL 402129
2019-nCoV BetaCoV/Wuhan/WIV07 EPI ISL 402130
NC 025217 Bat Hp BetaCoV Zhejiang2013
NC 009019 Tylonycteris bat CoV HKU4
NC 019843.3 MERS-CoV
NC 009020 Pipistrellus bat CoV HKU5
JX504050 Human CoV NL63 alpha
NC 010438 Miniopterus bat CoV HKU8 alpha
NC 009988 Rhinolophus bat CoV HKU2 alpha
KU291448 Human CoV 229E alpha
NC 009657 Scotophilus bat CoV 512 alpha
NC 003436 PEDV alpha
NC 038861 TGEV alpha
BetaCoV
AlphaCoV

Fig. 2. UPGMA phylogenetic tree constructed from the nucleotide sequences of 80 complete coronavirus genomes ($L = 5$). MHV: murine hepatitis virus; PEDV: porcine epidemic diarrhea virus; TGEV: porcine transmissible gastroenteritis virus, SARS-CoV-2 and bat CoV RaTG13 are shown in bold and in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

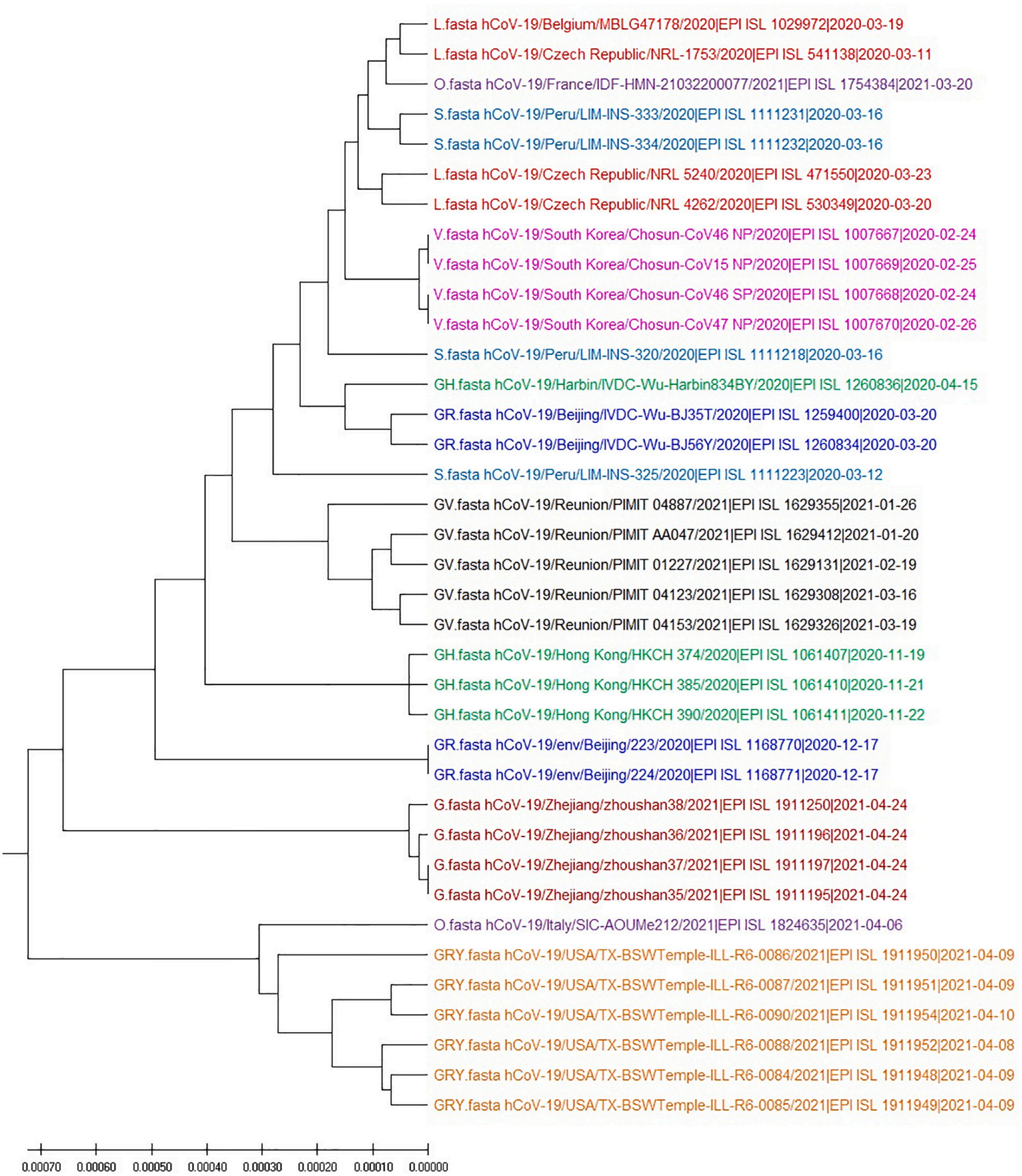
L.fasta hCoV-19/Belgium/MBLG47178/2020|EPI ISL 1029972|2020-03-19
L.fasta hCoV-19/Czech Republic/NRL-1753/2020|EPI ISL 541138|2020-03-11
O.fasta hCoV-19/France/IDF-HMN-21032200077/2021|EPI ISL 1754384|2021-03-20
S.fasta hCoV-19/Peru/LIM-INS-333/2020|EPI ISL 1111231|2020-03-16
S.fasta hCoV-19/Peru/LIM-INS-334/2020|EPI ISL 1111232|2020-03-16
L.fasta hCoV-19/Czech Republic/NRL 5240/2020|EPI ISL 471550|2020-03-23
L.fasta hCoV-19/Czech Republic/NRL 4262/2020|EPI ISL 530349|2020-03-20
V.fasta hCoV-19/South Korea/Chosun-CoV46 NP/2020|EPI ISL 1007667|2020-02-24
V.fasta hCoV-19/South Korea/Chosun-CoV15 NP/2020|EPI ISL 1007669|2020-02-25
V.fasta hCoV-19/South Korea/Chosun-CoV46 SP/2020|EPI ISL 1007668|2020-02-24
V.fasta hCoV-19/South Korea/Chosun-CoV47 NP/2020|EPI ISL 1007670|2020-02-26
S.fasta hCoV-19/Peru/LIM-INS-320/2020|EPI ISL 1111218|2020-03-16
GH.fasta hCoV-19/Harbin/IVDC-Wu-Harbin834BY/2020|EPI ISL 1260836|2020-04-15
GR.fasta hCoV-19/Beijing/IVDC-Wu-BJ35T/2020|EPI ISL 1259400|2020-03-20
GR.fasta hCoV-19/Beijing/IVDC-Wu-BJ56Y/2020|EPI ISL 1260834|2020-03-20
S.fasta hCoV-19/Peru/LIM-INS-325/2020|EPI ISL 1111223|2020-03-12
GV.fasta hCoV-19/Reunion/PIMIT 04887/2021|EPI ISL 1629355|2021-01-26
GV.fasta hCoV-19/Reunion/PIMIT AA047/2021|EPI ISL 1629412|2021-01-20
GV.fasta hCoV-19/Reunion/PIMIT 01227/2021|EPI ISL 1629131|2021-02-19
GV.fasta hCoV-19/Reunion/PIMIT 04123/2021|EPI ISL 1629308|2021-03-16
GV.fasta hCoV-19/Reunion/PIMIT 04153/2021|EPI ISL 1629326|2021-03-19
GH.fasta hCoV-19/Hong Kong/HKCH 374/2020|EPI ISL 1061407|2020-11-19
GH.fasta hCoV-19/Hong Kong/HKCH 385/2020|EPI ISL 1061410|2020-11-21
GH.fasta hCoV-19/Hong Kong/HKCH 390/2020|EPI ISL 1061411|2020-11-22
GR.fasta hCoV-19/env/Beijing/223/2020|EPI ISL 1168770|2020-12-17
GR.fasta hCoV-19/env/Beijing/224/2020|EPI ISL 1168771|2020-12-17
G.fasta hCoV-19/Zhejiang/zhoushan38/2021|EPI ISL 1911250|2021-04-24
G.fasta hCoV-19/Zhejiang/zhoushan36/2021|EPI ISL 1911196|2021-04-24
G.fasta hCoV-19/Zhejiang/zhoushan37/2021|EPI ISL 1911197|2021-04-24
G.fasta hCoV-19/Zhejiang/zhoushan35/2021|EPI ISL 1911195|2021-04-24
O.fasta hCoV-19/Italy/SIC-AOUMe212/2021|EPI ISL 1824635|2021-04-06
GRY.fasta hCoV-19/USA/TX-BSWTemple-ILL-R6-0086/2021|EPI ISL 1911950|2021-04-09
GRY.fasta hCoV-19/USA/TX-BSWTemple-ILL-R6-0087/2021|EPI ISL 1911951|2021-04-09
GRY.fasta hCoV-19/USA/TX-BSWTemple-ILL-R6-0090/2021|EPI ISL 1911954|2021-04-10
GRY.fasta hCoV-19/USA/TX-BSWTemple-ILL-R6-0088/2021|EPI ISL 1911952|2021-04-08
GRY.fasta hCoV-19/USA/TX-BSWTemple-ILL-R6-0084/2021|EPI ISL 1911948|2021-04-09
GRY.fasta hCoV-19/USA/TX-BSWTemple-ILL-R6-0085/2021|EPI ISL 1911949|2021-04-09
0.00070
0.00060
0.00050
0.00040
0.00030
0.00020
0.00010
0.00000

Fig. 3. UPGMA phylogenetic tree of 37 SARS-CoV-2 based on the Muscle method.

the normalized auto-correlation:

$$\tau_{\mathrm{tt}}(L) = \frac{\rho_{\mathrm{tt}}(L)}{\rho_{\mathrm{tt}}(0)} \tag{7}$$

Finally, we get a 16 × L feature vector of nucleotides called the correlation coefficient feature vector (CCFV):

$$\begin{aligned}&(\tau_{\mathrm{AA}}(1), \tau_{\mathrm{AA}}(2), \cdots, \tau_{\mathrm{AA}}(L), \tau_{\mathrm{AC}}(1), \tau_{\mathrm{AC}}(2), \cdots, \tau_{\mathrm{AC}}(L),\\ &\tau_{\mathrm{AG}}(1), \tau_{\mathrm{AG}}(2), \cdots, \tau_{\mathrm{AG}}(L), \tau_{\mathrm{AT}}(1), \tau_{\mathrm{AT}}(2), \cdots, \tau_{\mathrm{AT}}(L),\\ &\cdots\cdots\cdots\cdots\cdots, \tau_{\mathrm{TT}}(1), \tau_{\mathrm{TT}}(2), \cdots, \tau_{\mathrm{TT}}(L)).\end{aligned}$$

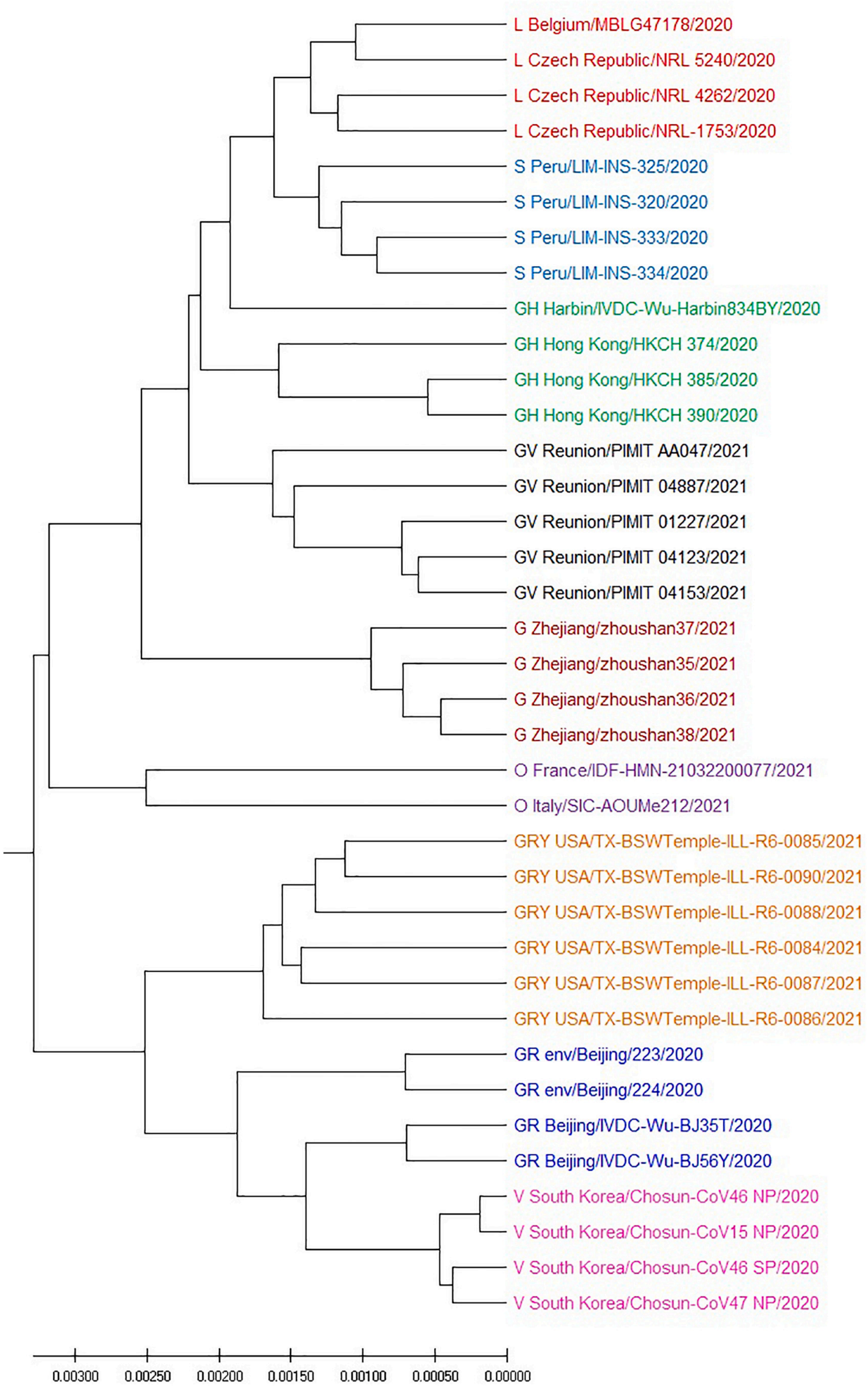
L Belgium/MBLG47178/2020
L Czech Republic/NRL 5240/2020
L Czech Republic/NRL 4262/2020
L Czech Republic/NRL-1753/2020
S Peru/LIM-INS-325/2020
S Peru/LIM-INS-320/2020
S Peru/LIM-INS-333/2020
S Peru/LIM-INS-334/2020
GH Harbin/IVDC-Wu-Harbin834BY/2020
GH Hong Kong/HKCH 374/2020
GH Hong Kong/HKCH 385/2020
GH Hong Kong/HKCH 390/2020
GV Reunion/PIMIT AA047/2021
GV Reunion/PIMIT 04887/2021
GV Reunion/PIMIT 01227/2021
GV Reunion/PIMIT 04123/2021
GV Reunion/PIMIT 04153/2021
G Zhejiang/zhoushan37/2021
G Zhejiang/zhoushan35/2021
G Zhejiang/zhoushan36/2021
G Zhejiang/zhoushan38/2021
O France/IDF-HMN-21032200077/2021
O Italy/SIC-AOUMe212/2021
GRY USA/TX-BSWTemple-ILL-R6-0085/2021
GRY USA/TX-BSWTemple-ILL-R6-0090/2021
GRY USA/TX-BSWTemple-ILL-R6-0088/2021
GRY USA/TX-BSWTemple-ILL-R6-0084/2021
GRY USA/TX-BSWTemple-ILL-R6-0087/2021
GRY USA/TX-BSWTemple-ILL-R6-0086/2021
GR env/Beijing/223/2020
GR env/Beijing/224/2020
GR Beijing/IVDC-Wu-BJ35T/2020
GR Beijing/IVDC-Wu-BJ56Y/2020
V South Korea/Chosun-CoV46 NP/2020
V South Korea/Chosun-CoV15 NP/2020
V South Korea/Chosun-CoV46 SP/2020
V South Korea/Chosun-CoV47 NP/2020
0.00300
0.00250
0.00200
0.00150
0.00100
0.00050
0.00000

Fig. 4. UPGMA phylogenetic tree of 37 SARS-CoV-2 based on the correlation coefficient feature vector (CCFV) method with $L = 5$.

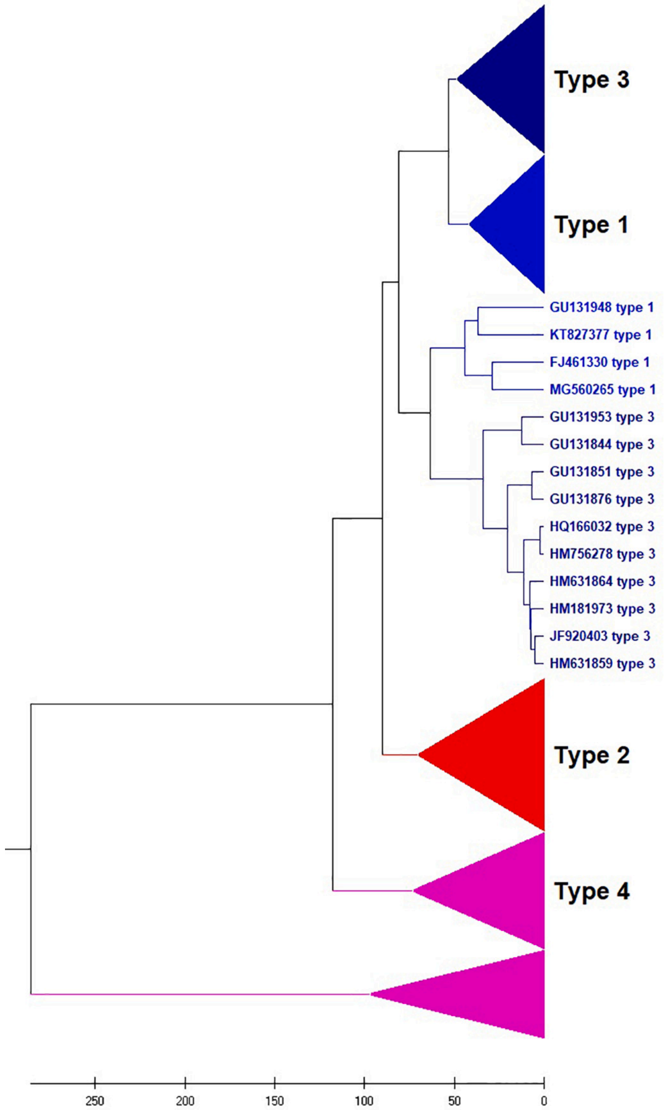

Fig. 5. UPGMA phylogenetic tree of 330 Dengue viruses based on the NV method.

Therefore, if there're n genomes, a $n \times (16 \times L)$ matrix can be construct. Meanwhile, Euclidean distance is used to get the similarity matrix. **Example**: Continue with the example above, for sequence="ACTGTGAGT", we have N=9, and suppose that $L = 2$ and t is A and T, respectively. Then:

$$\rho_{\mathrm{AA}}(0) = \frac{1}{9}\sum_{i=1}^{9}(x_A(i) - f_A)(x_A(i) - f_A) = \frac{1}{9}[(1 - \frac{2}{9}) \times 2] = \frac{14}{81};$$

$$\rho_{\mathrm{AA}}(2) = \frac{1}{9}\sum_{i=1}^{9}(x_A(i) - f_A)(x_{A+2}(i) - f_A)$$

$$= \frac{1}{9}[(1 - \frac{2}{9}) \times (0 - \frac{2}{9}) \times 2 + (0 - \frac{2}{9}) \times (0 - \frac{2}{9}) \times 6 + (0 - \frac{2}{9}) \times (1 - \frac{2}{9})]$$
$$= -\frac{2}{81};$$

$$\rho_{\mathrm{TT}}(0) = \frac{1}{9}\sum_{i=1}^{9}(x_T(i) - f_T)(x_T(i) - f_T) = \frac{1}{9}[(1 - \frac{3}{9}) \times 3] = \frac{2}{9};$$

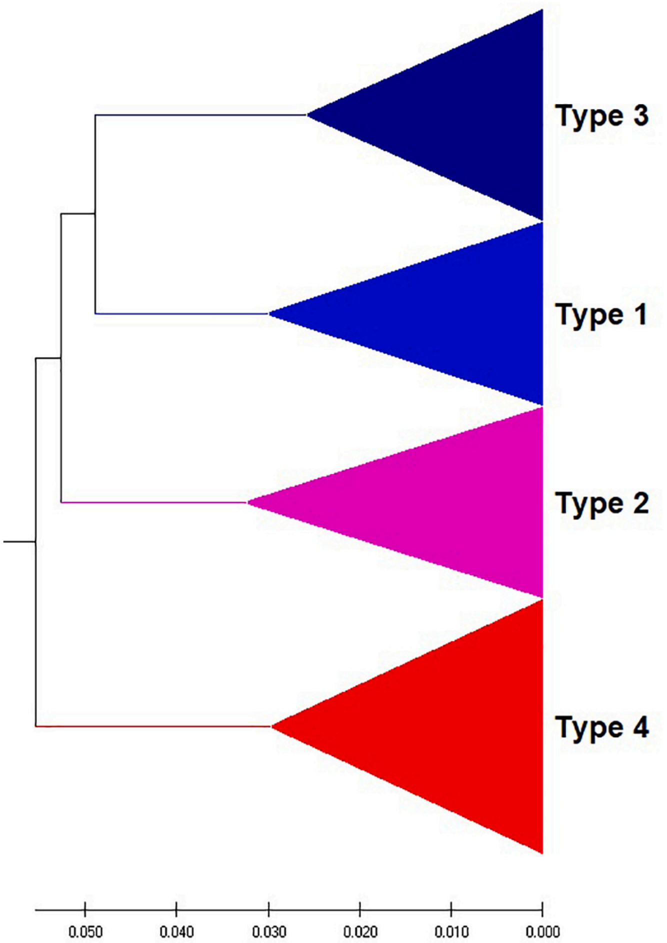

Fig. 6. UPGMA phylogenetic tree of 330 Dengue viruses based on the CCFV method with $L = 5$.

$$\rho_{\mathrm{TT}}(2) = \frac{1}{9}\sum_{i=1}^{9}(x_T(i) - f_T)(x_{T+2}(i) - f_T) = \frac{1}{9}[(0 - \frac{3}{9}) \times (1 - \frac{3}{9}) \times 2$$

$$+(0 - \frac{3}{9}) \times (0 - \frac{3}{9}) \times 4 + (1 - \frac{3}{9}) \times (1 - \frac{3}{9}) + (1 - \frac{3}{9}) \times (0 - \frac{3}{9}) \times 2] = 0;$$

$$\rho_{\mathrm{AT}}(2) = \frac{1}{9}\sum_{i=1}^{9}(x_A(i) - f_A)(x_{T+2}(i) - f_T)$$

$$= \frac{1}{9}[(1 - \frac{2}{9}) \times (1 - \frac{3}{9}) \times 2 + (0 - \frac{2}{9}) \times (0 - \frac{3}{9}) \times 6 + (0 - \frac{2}{9}) \times (1 - \frac{3}{9})]$$
$$= \frac{4}{27};$$

$$\tau_{\mathrm{AA}}(2) = \frac{\rho_{\mathrm{AA}}(2)}{\rho_{\mathrm{AA}}(0)} = \frac{1}{7}; \tau_{\mathrm{TT}}(2) = \frac{\rho_{\mathrm{TT}}(2)}{\rho_{\mathrm{TT}}(0)} = 0; \tau_{\mathrm{AT}}(2) = \frac{\rho_{\mathrm{AT}}(2)}{\sqrt{\rho_{\mathrm{AA}}(0)}\sqrt{\rho_{\mathrm{TT}}(0)}}$$
$$= \frac{2}{\sqrt{7}}.$$

3. Results

The SARS-CoV-2 datasets were derived from GISAID (https://www.gisaid.org/) and NCBI (https://www.ncbi.nlm.nih.gov/). The Dengue virus, Hepatitis B virus (HBV) and Human rhinovirus virus (HRV) datasets were from NCBI. Our CCFV method was applied to the evolutionary analysis for these viruses. L was determined as 5 in all datasets.

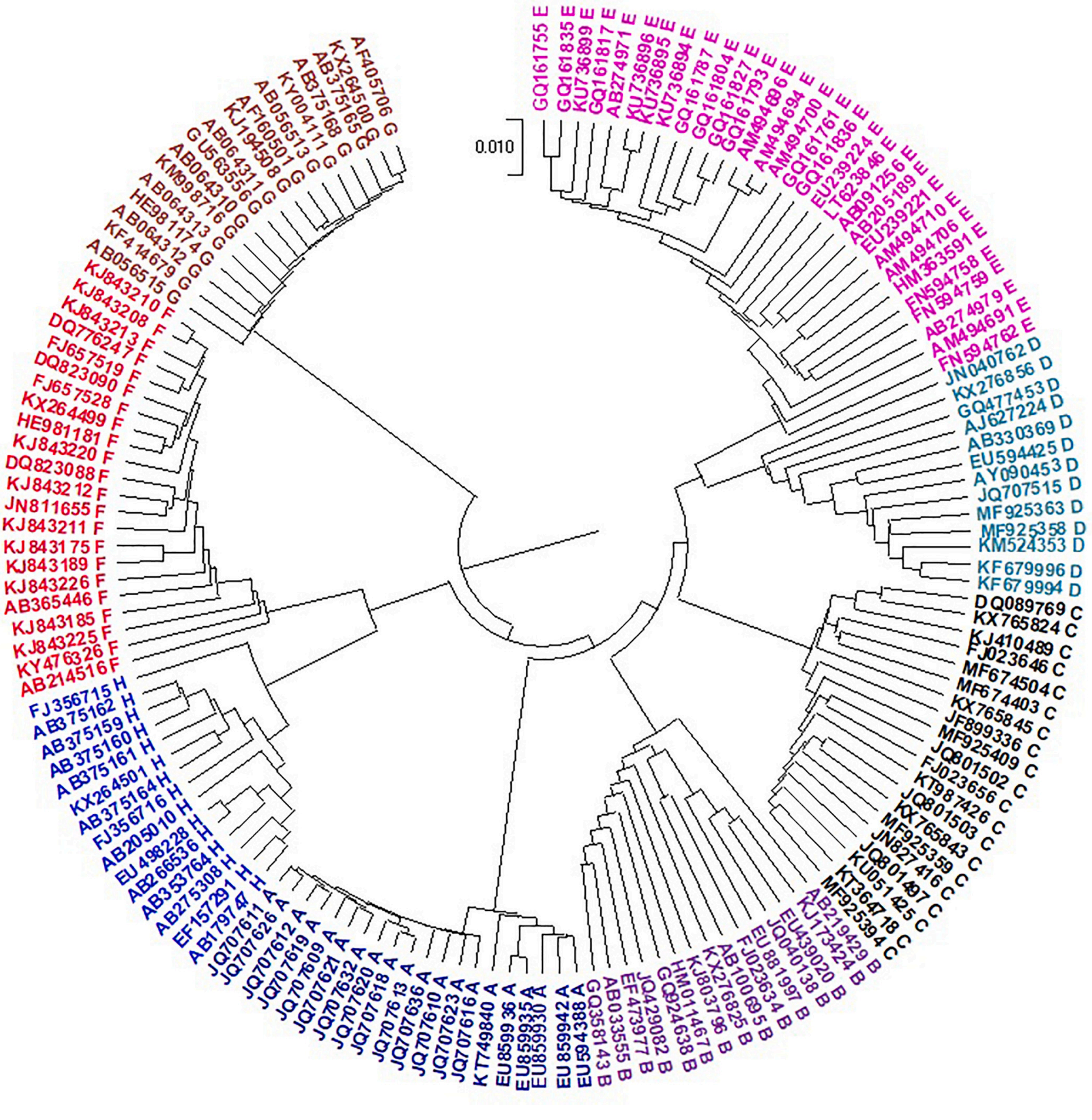

Fig. 7. UPGMA phylogenetic tree of 152 HBVs based on the CCFV method with $L = 5$.

Computations were performed on a personal computer with an Intel Core i5-10210U CPU @ 1.60 GHz and 16 GB RAM.

3.1. SARS-CoV-2

China's first COVID-19 outbreak in Wuhan, Hubei Province, was reported in late 2019. The virus can cause severe pneumonia and has become a major global health threat (Chang et al., 2020; Benvenuto et al., 2020). The COVID-19 virus has been officially termed as SARS-CoV-2. The clinical symptoms of COVID-19 infection include (but are not limited to) dyspnea, fever, pneumonia, and renal failure. The SARS-CoV-2 virus rapidly mutates, has strong transmission ability, and produces similar symptoms to influenza. For these reasons, COVID-19 is difficult to control and diagnose. For example, COVID-19 recurred in Beijing's Xinfadi Market in July of 2020. As of August 29 2021, more than 216 million cases have been diagnosed and 4.4 million deaths have been reported worldwide, and the data continue to rise according to Johns Hopkins Coronavirus Resource Center (CRC) (https://coronavirus.jhu.edu/map.html).

3.1.1. Coronaviruses

According to epidemiological investigation and gene sequence comparisons, SARS-CoV-2 is 87.5% similar to Bat-SL-CoVZC45 and 87.6% similar to Bat-SL-CoVZXC21. Both variants are coronaviruses found in bats. Previous studies have shown that SARS-CoV-2 infection is transmitted by an intermediate host (bats or the civet Paguma larvata) before mutating and spreading among humans (Zhou et al., 2020; Zhu et al., 2020). In the present study, 80 whole genomes of coronaviruses

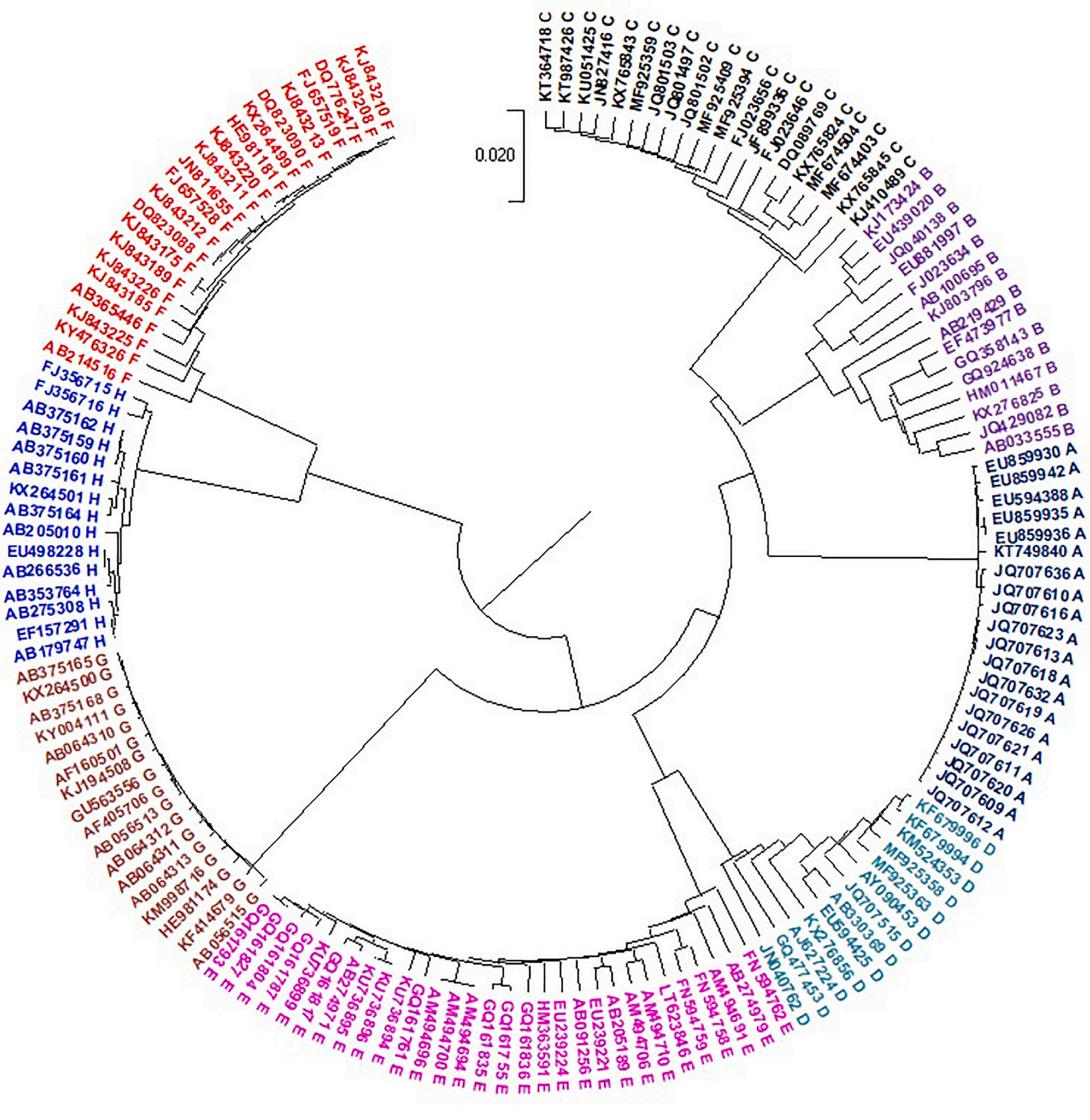

Fig. 8. UPGMA phylogenetic tree of 152 HBVs based on Muscle method.

were sourced from the GISAID and NCBI databases. The phylogenetic tree built from our CCFV using the UPGMA method is given in Fig. 2. Among the 73 BetaCoV and 7 AlphaCoV, 5 SARS-CoV-2 from Wuhan and the bat coronavirus BatCoV RaTG13 belong to sister branches (red branches in Fig. 2), indicating maximum similarity among these 6 genomes. (Zhou et al., 2020) found that SARS-CoV-2 in early patient samples shared 96.2% sequence identity with bat-coronavirus (bat-nCoV) RaTG13 at the whole-genome level. In their phylogenetic tree obtained by sequence homology comparison, seven SARS-CoV-2 viruses from seven patients clustered together with bat-coronavirus (bat-nCoV) RaTG1 (Zhou et al., 2020). According to biological features of bat and the high identity sequence between bat-nCoV and SARS-CoV-2 demonstrated by a lot of additional research, bats are considered as the natural reservoir of SARS-CoV-2 (Wu et al., 2020). In our phylogenetic tree, the 5 SARS-CoV-2 viruses from Wuhan cluster with RaTG13 as well. Thus we conclude that bats seems to be the natural reservoir of SARS-CoV-2 by our method.

3.1.2. *SARS-CoV-2 variant*

With the emergence of SARS-CoV-2 mutants, people began to pay attention to its variants. On the one hand, for the development of vaccine, on the other hand, it can provide beneficial help for clinic. GISAID divided SARS-CoV-2 into nine clades: S, O, L, V, G, GH, GR, GV, and GRY. The S and L clades were around at the beginning of the pandemic. But then later there were other variants.

We obtain 37 SARS-CoV-2 genomes from GISAID which include nine different clades. Using Muscle method, O,S,GR,SH are mixed with other clades, see Fig. 3. The variant of SARS-CoV-2 usually contain only a few

Table 1
Running time for Muscle, and CCFV method. “s”: seconds; “min”: minute.

Method	Muscle	CCFV
HBV (152)	11 min	0.93 s
Dengue (330)	Larger than 30 min	4.92 s
HRV (116)	Larger than 30 min	1.2 s

mutations, the model-based algorithm can not identify difference of variants in many cases. Our technique consider location correlation coefficient of the different nucleotide and achieve the better result shown in Fig. 4.

3.2. Dengue virus

DENV is transmitted to humans through the bite of an infected mosquito of the Aedes genus. Dengue fever infects humans in more than 100 countries worldwide (https://www.cdc.gov/dengue/about/index.html) (Sirisena and Noordeen, 2014). Approximately three billion people are at risk of contracting DENV, and 400 million people are infected with the virus each year. Huma infection is caused by four serotypes called DENV 1, 2, 3, and 4 (Tsang et al., 2019). Therefore, research on this virus is urgently demanded. To predict the serotypes of an unknown virus, the evolutionary analysis of this virus should be performed. The phylogenetic trees constructed from the genomes of 330 Dengue viruses by NV and our CCFV method are presented in Figs. 5 and 6, respectively. The viruses are correctly placed into four types, but the type 3 doesn’t come together by the NV method. Our method obviously outperformed

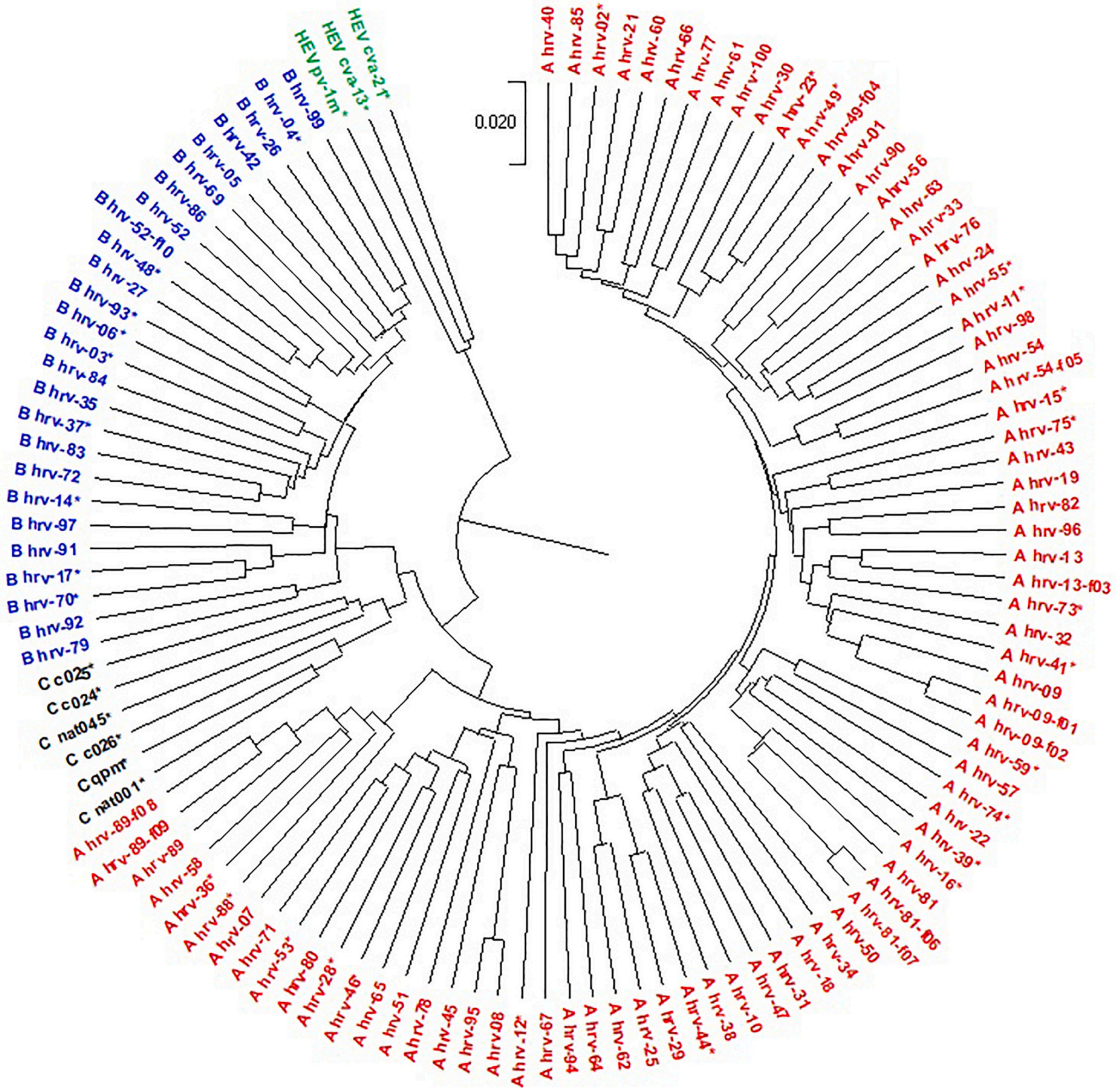

Fig. 9. UPGMA phylogenetic tree of 116 HRVs based on the CCFV method with $L = 5$.

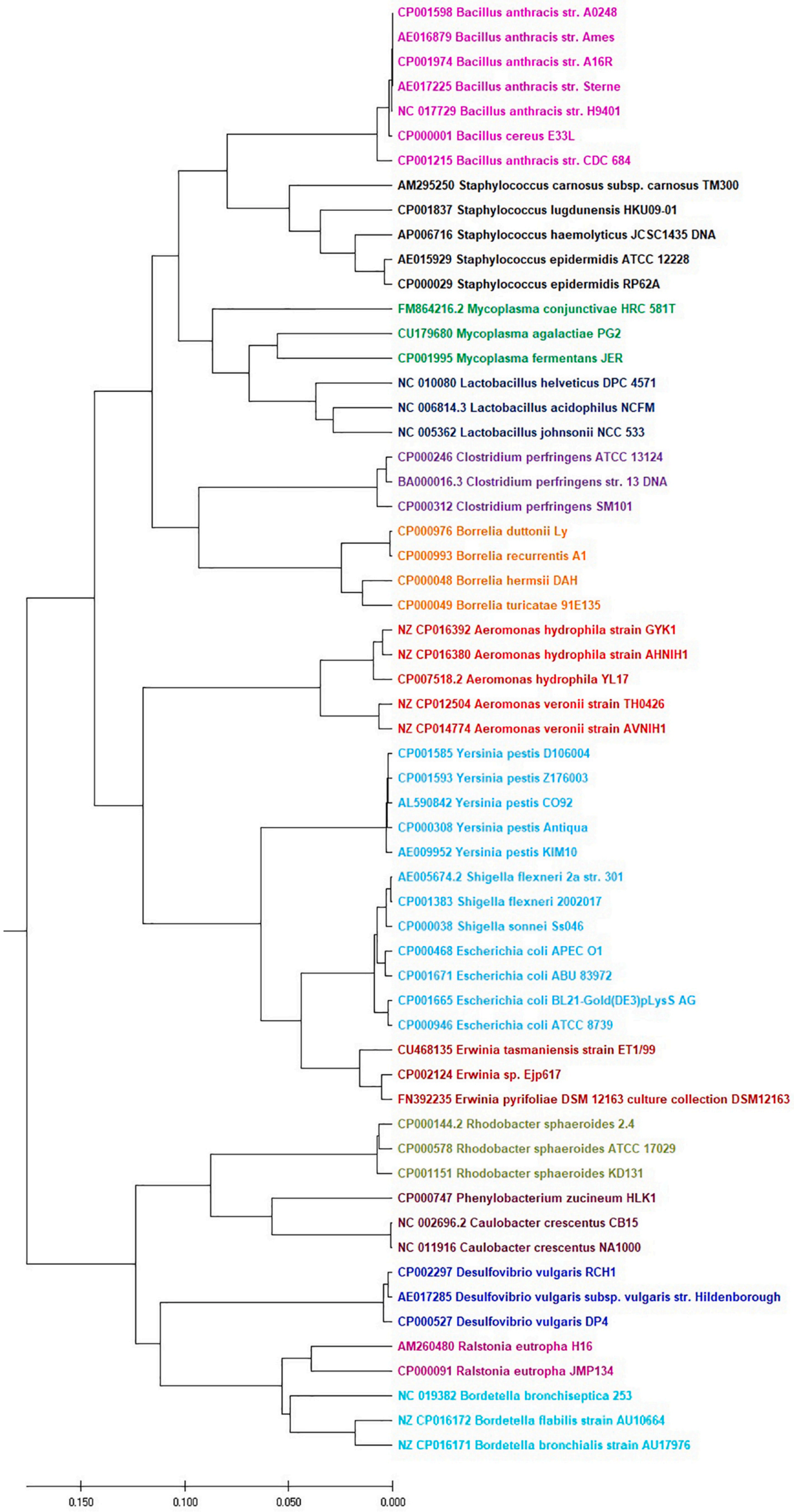
CP001598 Bacillus anthracis str. A0248
AE016879 Bacillus anthracis str. Ames
CP001974 Bacillus anthracis str. A16R
AE017225 Bacillus anthracis str. Sterne
NC 017729 Bacillus anthracis str. H9401
CP000001 Bacillus cereus E33L
CP001215 Bacillus anthracis str. CDC 684
AM295250 Staphylococcus carnosus subsp. carnosus TM300
CP001837 Staphylococcus lugdunensis HKU09-01
AP006716 Staphylococcus haemolyticus JCSC1435 DNA
AE015929 Staphylococcus epidermidis ATCC 12228
CP000029 Staphylococcus epidermidis RP62A
FM864216.2 Mycoplasma conjunctivae HRC 581T
CU179680 Mycoplasma agalactiae PG2
CP001995 Mycoplasma fermentans JER
NC 010080 Lactobacillus helveticus DPC 4571
NC 006814.3 Lactobacillus acidophilus NCFM
NC 005362 Lactobacillus johnsonii NCC 533
CP000246 Clostridium perfringens ATCC 13124
BA000016.3 Clostridium perfringens str. 13 DNA
CP000312 Clostridium perfringens SM101
CP000976 Borrelia duttonii Ly
CP000993 Borrelia recurrentis A1
CP000048 Borrelia hermsii DAH
CP000049 Borrelia turicatae 91E135
NZ CP016392 Aeromonas hydrophila strain GYK1
NZ CP016380 Aeromonas hydrophila strain AHNIH1
CP007518.2 Aeromonas hydrophila YL17
NZ CP012504 Aeromonas veronii strain TH0426
NZ CP014774 Aeromonas veronii strain AVNIH1
CP001585 Yersinia pestis D106004
CP001593 Yersinia pestis Z176003
AL590842 Yersinia pestis CO92
CP000308 Yersinia pestis Antiqua
AE009952 Yersinia pestis KIM10
AE005674.2 Shigella flexneri 2a str. 301
CP001383 Shigella flexneri 2002017
CP000038 Shigella sonnei Ss046
CP000468 Escherichia coli APEC O1
CP001671 Escherichia coli ABU 83972
CP001665 Escherichia coli BL21-Gold(DE3)pLysS AG
CP000946 Escherichia coli ATCC 8739
CU468135 Erwinia tasmaniensis strain ET1/99
CP002124 Erwinia sp. Ejp617
FN392235 Erwinia pyrifoliae DSM 12163 culture collection DSM12163
CP000144.2 Rhodobacter sphaeroides 2.4
CP000578 Rhodobacter sphaeroides ATCC 17029
CP001151 Rhodobacter sphaeroides KD131
CP000747 Phenylobacterium zucineum HLK1
NC 002696.2 Caulobacter crescentus CB15
NC 011916 Caulobacter crescentus NA1000
CP002297 Desulfovibrio vulgaris RCH1
AE017285 Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
CP000527 Desulfovibrio vulgaris DP4
AM260480 Ralstonia eutropha H16
CP000091 Ralstonia eutropha JMP134
NC 019382 Bordetella bronchiseptica 253
NZ CP016172 Bordetella flabilis strain AU10664
NZ CP016171 Bordetella bronchialis strain AU17976
0.150
0.100
0.050
0.000

Fig. 10. UPGMA phylogenetic tree of 59 bacteria genome sequences based on the CCFV method with $L = 5$.

the NV method in prediction accuracy.

3.3. Hepatitis B virus

The HBV causes a persistent chronic infection in immunocompromised humans (Lazarus et al., 2018). The HBV includes ten genotypes assigned A-J (Yuen et al., 2018; Juliette et al., 2013). Therefore, the rapid and accurate identification of HBV genotypes is a significant goal in clinical diagnosis. Here, we performed an evolutionary analysis of 152 complete HBV genomes extracted from the Hepatitis B virus Database (HBVdb). The phylogenetic tree by our CCFV method are shown in Fig. 7. CCFV clearly assigned each genotype to a correct branch. As comparison, the UPGMA trees built by the currently popular alignment method Muscle are shown in Fig. 8. Both methods give the same result, however, CCFV is much faster than Muscle, see Table 1. This result demonstrates that the CCFV is superior than Muscle in terms of runtime.

3.4. Human rhinovirus

Human rhinovirus (HRV) was first isolated in the 1950s from nasopharyngeal secretions of patients with the common cold. HRV belongs to the Picornaviridae family and genus Enterovirus. Past studies have shown that HRV consists of three genetically distinct groups: HRV-A, HRV-B, and HRV-C. In a previous study (Palmenberg et al., 2009), 116 complete genomes including 113 HRV and 3 outgroup HEV-C genomes were studied. The result for the 116 genomes based on CCFV is shown in Fig. 9. The result based our technique is the same as the result in the previous study. However, our running time is only 1.20 seconds and the running time in the previous paper is very high because of usage of multiple sequence alignment for building the phylogenetic tree.

4. Discussion and conclusion

The proposed CCFV method maps a DNA sequence to four binary vectors, then determines the correlation coefficient and auto-correlation coefficient of an k step delay from the original sequence. The correlation coefficients after delaying the sequence by 1 to L steps are concatenated into a $(16 \times L)$-dimensional vector. Describing each DNA sequence in this vector form, we measured the similarity between two sequences by calculating the Euclidean distance between the vectors representing those sequences. Using the UPGMA method, we finally constructed the phylogenetic trees of the genomes contained in two SARS-CoV-2 datasets, a DENV dataset, and a HBV dataset. Our method clustered together five SARS-CoV-2 variants from Wuhan. Moreover, this group formed two adjacent branches with the bat coronavirus suggesting that bats are an natural reservoir of SARS-CoV-2. The CCEV successfully established the evolutionary relationships for viruses SARS-CoV-2, DENV, HBV and HRV. In addition, we derive some interesting formulas to approximate the proposed correlation/auto-correlation coefficient. These formulas indicate the L step delay correlation coefficient is directly proportional to the frequency of some types of spaced $L+1$-mers. We compare the approximate correlation coefficient with the true correlation coefficient in all datasets. The approximate error is very small and may be ignorable.

To illustrate the effectiveness of our approach for larger genomes, we collect 59 bacterial genomes with length ranging from 0.8 to 5 million bp. Then we build a phylogenetic tree by the CCFV technique. As shown in Fig. 10, these bacterial genomes are separated into 14 families which are clearly separated from each other.

For genome sequence S, NV method consider the number, the mean position and the mean position of nucleotide $\alpha \in A, C, G, T$ in S. For each nucleotide, our method considers not only the correlation of appearance of nucleotides, but also the correlation due to L-steps delay. Thus the NV method only contains the distribution of single nucleotide, but ignores the correlation of nucleotides. Therefore, we can extract more information of nucleotides in a genome.

Our correlation coefficient measure has some limitations as well. First, the defined delay step number L is a variable parameter, which must be determined in multiple trial-and-error tests. Second, some viruses in the same region do not form a branch in the evolutionary tree by our method, possibly because they have mutated from viruses in other regions, or the dimension of the numerical vector is insufficiently high to capture the information loss. These problems should be investigated in further study.

CRediT author statement

Lily He and Peter K. Li conceived the initial planning, implemented the code and analysed the results. All authors wrote and reviewed the manuscript.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest related to this manuscript.

Acknowledgements

This study is supported by Promotion plan for young teachers' scientific research ability of Beijing University of Civil Engineering and Architecture (X21026).

Appendix

A.1. Location correlation coefficient approximate formula

That is:

$$\begin{aligned}\rho_{tt}(0) &= \frac{1}{N}\sum_{i=1}^{N}(x_t(i)-f_t)(x_t(i)-f_t)\\ &= \frac{1}{N}\sum_{i=1}^{N}x_t^2(i)-f_t^2\\ &= \frac{1}{N}\sum_{i=1}^{N}x_t(i)-f_t^2\\ &= f_t-f_t^2\end{aligned} \tag{8}$$

According to definition of the auto-correlation $\rho_{tt}(L)$ of nucleotide t, we obtain

$$\begin{aligned}\rho_{\text{tt}}(L) &= \frac{1}{N}\sum_{i=1}^{N}(x_t(i)-f_t)(x_{t+L}(i)-f_t)\\ &= \frac{1}{N}\sum_{i=1}^{N}x_t(i)x_{t+L}(i)-f_t^2+\frac{1}{N}\sum_{i=1}^{L}x_t(i)*f_t.\end{aligned}$$

When $L = 1$, $\sum_{i=1}^{N}x_t(i)*x_{t+L}(i) = \sum_{i=1}^{N}x_t(i)*x_{t+L}(i)$, and the following relationship holds:

$$x_t(i)*x_{t+L}(i) = \begin{cases}1, & \text{when } x_t(i)=1=x_{t+L}(i)\\ 0, & \text{else}\end{cases}$$

In this case, the value of $\sum_{i=1}^{N}x_t(i)*x_{t+L}(i)$ represents the number of "*tt*" (**2-mer**) present. For general L, the value of $\sum_{i=1}^{N}x_t(i)*x_{t+L}(i)$ represents the number of "$t*\cdots*t$" L**+1-mer** present, denoted as $n_{t*\ldots*t}$. Here each $t*\cdots*t$-mer has identical head and tail nucleotide t separated by L-1 undetermined nucleotides. In some studies, this kind of k-mer is called spaced k-mer.

Let $n_t(L)$ be the number of t occurrences in the first L locations of a DNA sequence. Then

$$\begin{aligned}\rho_{\text{tt}}(L) &= \frac{1}{N}\sum_{i=1}^{N}x_t(i)x_{t+L}(i)-f_t^2+\frac{1}{N}\sum_{i=1}^{L}x_t(i)*f_t\\ &= \frac{n_{t*\cdots*t}}{N}-f_t^2+\frac{n_t(L)}{N}*f_t.\end{aligned} \tag{9}$$

Note that $\frac{n_t(L)}{N}$ is less than $\frac{L}{N}$. Since L is very small and the genome length N is very large, for example in this paper $L = 5$ and N is more than 10000, the last term is ignorable. Therefore the approximate formula for $\rho_{tt}(L)$ is as follows:

$$\begin{aligned}\rho_{\text{tt}}(L) &\approx \frac{1}{N}\sum_{i=1}^{N}x_t(i)x_{t+L}(i)-f_t^2\\ &= \frac{n_{t*\cdots*t}}{N}-f_t^2.\end{aligned} \tag{10}$$

The $\rho_{st}(L)$ can be approximated as:

$$\begin{aligned}\rho_{\text{st}}(L) &\approx \frac{1}{N}\sum_{i=1}^{N}x_s(i)x_{t+L}(i)-f_s*f_t\\ &= \frac{n_{s*\cdots*t}}{N}-f_s*f_t.\end{aligned} \tag{11}$$

The l-step delay correlation coefficient $\tau_{st}(L)$ can be approximated as:

$$\begin{aligned}\tau_{\text{st}}(L) &= \frac{\rho_{\text{st}}(L)}{\sqrt{\rho_{\text{ss}}(0)\rho_{\text{tt}}(0)}}\\ &\approx \frac{\frac{n_{s*\cdots*t}}{N}-f_s*f_t}{\sqrt{f_s(1-f_s)*f_t(1-f_t)}}.\end{aligned} \tag{12}$$

The auto-correlation coefficient $\tau_{tt}(L)$ of nucleotide t can be approximated as:

$$\tau_{\text{tt}}(L) \approx \frac{\frac{n_{t*\cdots*t}}{N}-f_t^2}{f_t(1-f_t)}. \tag{13}$$

According to above approximate formulas, we can see that the proposed correlation coefficient is directly proportional to the ratio $\frac{n_{s*\cdots*t}}{N}$, which is the frequency of spaced $L+1$-mer $s*\cdots*t$ in a sequence.

References

Benvenuto, D., Giovanetti, M., Salemi, M., Prosperi, M., Flora, C.D., Alcantara, L.C.J., Angeletti, S., Ciccozzi, M., 2020. The global spread of 2019-ncov: a molecular evolutionary analysis. Pathog. Global Health 114, 64–67.

Bernard, G., Chan, C.X., Chan, Y.B., Chua, X.Y., Cong, Y., Hogan, J.M., Maetschke, S.R., Ragan, M.A., 2019. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Brief. Bioinform. 20, 426–435.

Chang, T.J., Yang, D.M., Wang, M.L., Liang, K.H., Wang, C.T., 2020. Genomic analysis and comparative multiple sequence of sars-cov2. J. Chin. Med. Assoc. 83, 1.

Chookajorn, T., 2020. Evolving COVID-19 conundrum and its impact. Proc. Natl. Acad. Sci. U.S.A. 117, 12520–12521.

Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.-T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6, e17293.

Edgar, R.C., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

He, L., Dong, R., He, R.L., Yau, S.T., 2020. A novel alignment-free method for HIV-1 subtype classification. Infect. Genet. Evol. 77.

Joel, J.H., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 2163–2170.

Juliette, H., Fanny, J., Gilbert, D., Alan, K., Fabien, Z., Christophe, C., 2013. HBVdb: a knowledge database for Hepatitis B virus. Nucleic Acids Res. 41, D566–D570.

Kurtz, S., Narechania, A., Stein, J.C., Ware, D., 2008. A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9, 517.

Lazarus, J.V., Timothy, B., Brchot, C., Anna, K., Veronica, M., Michael, N., Capucine, P., Ulrike, P., Homie, R., Thomas, L.A.a., 2018. The Hepatitis B epidemic and the urgent need for cure preparedness. Nat. Rev. Gastroenterol. Hepatol. 15, 517–518.

Li, Y., He, L., He, R.L., Yau, S.S.T., 2017. Zika and flaviviruses phylogeny based on the alignment-free natural vector method. DNA Cell Biol. 36, 1–8.

Nelson, N.P., Easterbrook, P.J., McMahon, B.J., 2016. Epidemiology of Hepatitis B virus infection and impact of vaccination on disease. Clinics Liver Dis. 20, 607–628.

Palmenberg, A.C., Spiro, D., Kuzmickas, R., Wang, S., Djikeng, A., 2009. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. Science 324, 55–59.

Randhawa, G., Soltysiak, M., Roz, H., Souza, C., Hill, K., Kari, L., 2020a. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS One 15, e0232391.

Randhawa, G.S., Hill, K.A., Kari, L., 2020b. MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis. Bioinformatics 36, 2258–2259.

Sarkar, B.K., Sharma, A.R., Bhattacharya, M., Sharma, G., Lee, S.S., Chakraborty, C., 2021. Determination of k-mer density in a DNA sequence and subsequent cluster formation algorithm based on the application of electronic filter. Sci. Rep. 11, 13701.

Sirisena, P.D.N.N., Noordeen, F., 2014. Evolution of dengue in Sri Lanka changes in the virus, vector, and climate. Int. J. Infect. Dis. 19, 6–12.

Sironi, M., Hasnain, S.E., Phan, T., Luciani, F., Gonzlez-Candelas, F., 2020. SARS-CoV-2 and COVID-19: a genetic, epidemiological, and evolutionary perspective. Infect. Genet. Evol. 84, 104384.

Tsang, T.K., Ghebremariam, S.L., Gresh, L., Gordon, A., Halloran, M.E., Katzelnick, L.C., Rojas, D.P., Kuan, G., Balmaseda, A., Sugimoto, J.a., 2019. Effects of infection history on dengue virus infection and pathogenicity. Nat. Commun. 10, 1246.

Vinga, S., 2014. Editorial: alignment-free methods in computational biology. Brief. Bioinform. 15, 341–342.

Wu, F., Su, Z., Bin, Y., Zhang, Y.Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269.

Wu, G.A., Jun, S.R., Sims, G.E., Kim, S.H., 2009. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. Proc. Natl. Acad. Sci. U.S.A. 106, 12826–12831.

Yuen, M.F., Chen, D.S., Dusheiko, G.M., Hla, J., Dty, L., Locarnini, S.A., Peters, M.G., Lai, C.L., 2018. Hepatitis B virus infection. Nat. Rev. Dis. Primers 4, 18035.

Zhou, P., Yang, X.L., Shi, Z.L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.

Zhu, N., Wang, W., Liu, Z., Liang, C., Tan, W., 2020. Morphogenesis and cytopathic effect of SARS-CoV-2 infection in human airway epithelial cells. Nat. Commun. 11, 3910.

Zielezinski, A., Girgis, H.Z., Bernard, G., Leimeister, C.A., Karlowski, W.M., 2019. Benchmarking of alignment-free sequence comparison methods. Genome Biol. 20, 144.