Cognitive Computation
https://doi.org/10.1007/s12559-020-09790-w

Principal Component Analysis Applications in COVID-19 Genome Sequence Studies

Bo Wang[1] · **Lin Jiang**[2]

Received: 7 July 2020 / Accepted: 6 November 2020

Abstract
RNA genomes from coronavirus have a length as long as 32 kilobases, and the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that caused the outbreak of coronavirus disease 2019 (COVID-19) pandemic has long sequences which made the analysis difficult. Over 20,000 sequences have been submitted to GISAID, and the number is growing fast each day which increased the difficulties in data analysis; however, genome sequence analysis is critical in understanding the COVID-19 and preventing the spread of the disease. In this study, a principal component analysis (PCA) was applied to the aligned large size genome sequences and the numerical numbers were converted from the letters using a published method designed for protein sequence cluster analysis. The study initialized with a shortlist sequence testing, and the PCA score plot showed high tolerance with low-quality data, and the major virus sequences from humans were separated from the pangolin and bat samples. Our study also successfully built a model for a large number of sequences with more than 20,000 sequences which indicate the potential mutation directions for the COVID-19 which can be served as a pretreatment method for detailed studies such as decision tree-based methods. In summary, our study provided a fast tool to analyze the high-volume genome sequences such as the COVID-19 and successfully applied to more than 20,000 sequences which may provide mutation direction information for COVID-19 studies.

Keywords Principle Component Analysis · COVID-19 · Genome Sequences · Mutation

Introduction

Coronavirus is an RNA virus that contains large known RNA genomes with 27 to 32 kilobases in length, and coronavirus disease 2019 (COVID-19) is caused by a novel coronavirus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1, 2]. Coronavirus evolution has been studied using a phylogenetic network [3] to reconstruct their evolutionary paths. Phylogenetic tree analysis [4–6] was used to investigate the identification of coronavirus. Dimension reduction algorithm Uniform Manifold Approximation and Projection (UMAP) [7, 8] has also been applied for COVID-19 protein studies [9], but the genome sequence application is rare. A potential problem for the phylogenetic relationship-based method is that only largely complete genomes are preferred with limited noise suppression [3]. Principal component analysis (PCA) is a powerful tool for reducing the dimensionality of complex data sets [10, 11]. The PCA study has been reported to have high performance in noise suppression and fast data process for various applications [12–14] and showed high performance in excluding outliers or artifacts [15, 16]. When over 20,000 viral genome sequences have been shared in GISAID (www.gisaid.org), PCA is one of the best options to study the sample clustering information and provides the potential virus mutation information for further studies. Though GISAID is considered as an effective and trusted database for sharing of both published and "unpublished" influenza data [17, 18], it is still possible to get low quality of data

✉ Bo Wang
bwang1@ncat.edu

✉ Lin Jiang
ljiang@ncf.edu

[1] Department of Chemistry, College of Science and Technology, North Carolina Agricultural and Technical State University, Greensboro, NC 27411, USA

[2] Division of Natural Sciences, New College of Florida, 5800 Bay Shore Road, Sarasota, FL 34243, USA

Published online: 13 January 2021

due to various types of errors. Therefore, PCA could also be served as good noise suppression and data-pretreatment tool for other popular genome sequences analysis methods.

PCA has been applied to study both protein sequences [19–21] and whole-genome [22] to simplify the highly complex data. A frequency-rank-based method was specially designed for PCA on protein sequence [23] which can efficiently reserve the variance information. In the PCA study, when projecting the large size data set into an eigenspace that represents the directions of greatest variation, complex samples such as large size protein sequences could be represented by principal component [23, 24]. The method has been successfully applied to distinguish protein clusters; however, it has not been applied to DNA or RNA sequences that are much longer than the protein sequences. Though DNA or RNA sequences have fewer variables (possible nucleotides) than that of protein sequences (20 amino acids), the idea of converting letter sequences to the numerical data matrix is similar. While the RNA genome has fewer options than proteins, the core idea of reserve sequence variance using frequency is the same. Since the ranks are used as the final data (same frequency data are ranked alphabetically), the difference between protein and RNA sequences is limited after mean-centered scaling. PCA is a suitable tool to work with higher volumes of features than observations [25], and the RNA genomes like the coronavirus which is as large as 30 k are appropriate for PCA analysis. The GISAID database has collected more than 20,000 sequences from clinicians and researchers all over the world from December 2019 till mid-May 2020, and the number is growing fast each day which provides a great opportunity to study the potential application of frequency-based PCA analysis. The method in this study would provide a fast tool to analyze the extremely large amount of sequences with regular lab computers and have a high tolerance to the potential noises. More important, the clustering analysis generated by PCA could help to track the COVID-19 genome mutations.

Materials and Methods

Genome Sequence Alignment

The genome sequences were obtained from the GISAID database (www.gisaid.org) in fasta format. The data were downloaded on May 25, 2020, and the laboratories contributed to the database are listed in Dataset S1. The genome sequences were then aligned using MAFFT [26] version 7 (https://mafft.cbrc.jp/), and the whole genome was aligned using a lightweight option.

Sequence Conversion

The aligned sequences were imported to Matlab 2019 (Mathworks) for the frequency conversion analysis. The aligned sequence was also visualized in UGENE V34.0 (Unipro) [27, 28]. The sequence conversion was carried out using the algorithm reported before [23] in a Matlab script.

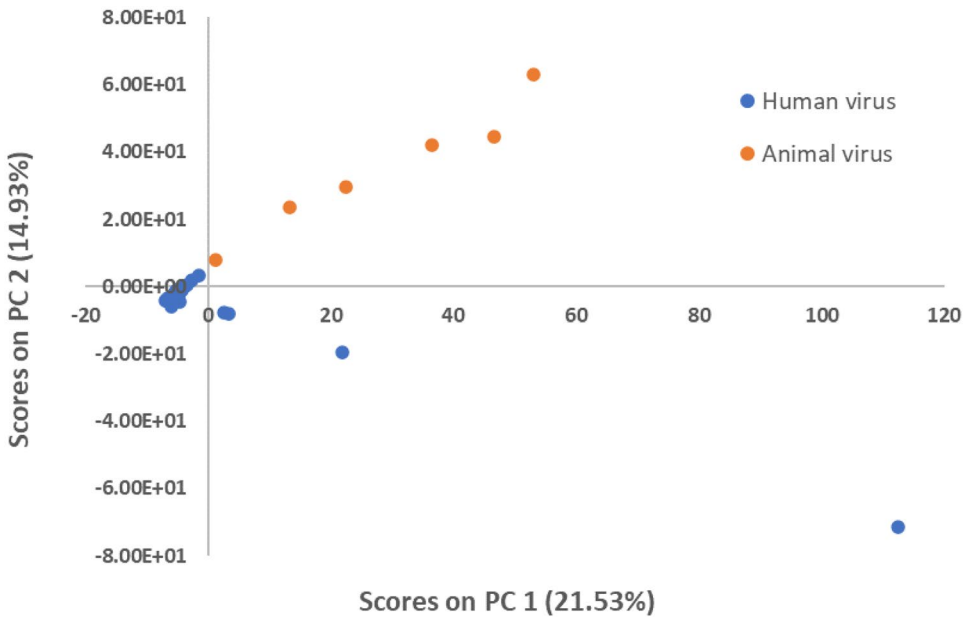


Fig. 1 The PCA score plot of the shortlist sequences. The plot contains 75 sequences including 2 sequences from the bat and 4 sequences from pangolin (the orange dots); the rest of the sequences were randomly selected from human virus samples from all over the world (the blue dots)

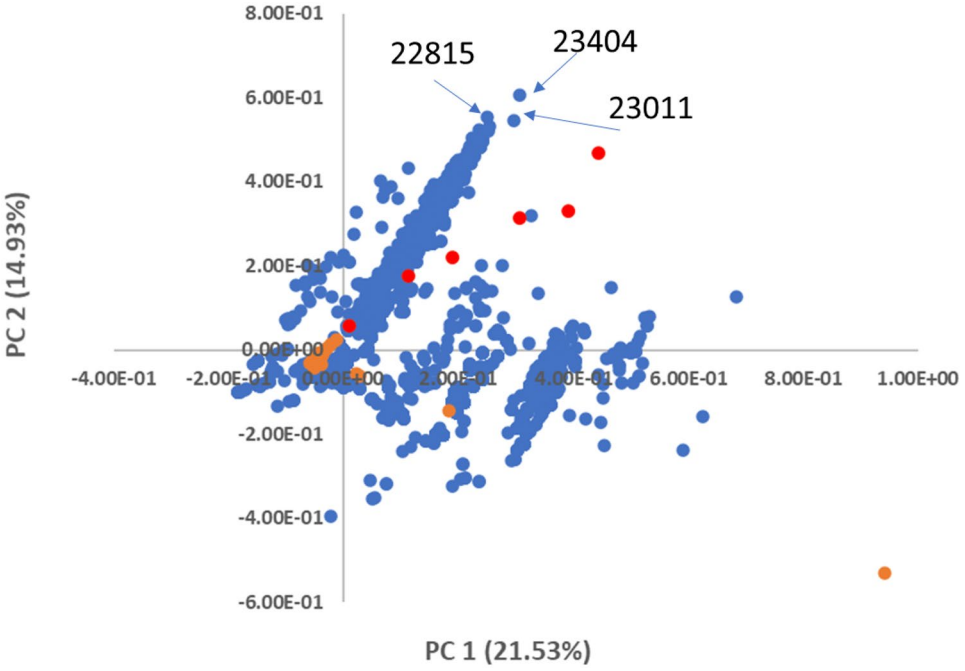


Fig. 2 The score and loading combined plot of the shortlist sequences. The red dots are animal sample scores, orange dots are human sample scores, and the blue dots are loadings. The scores were scaled to the range of the loading plot to allow the plotting in the same figure. The numbers on the figure indicate the representative nucleotide positions and the sequences were listed in Table 1

Briefly, the method converted letters to numerical numbers using the frequency of each letter in their sequence position. The converted numerical matrix was analyzed using the PLS-tool box to do the PCA analysis using the singular value decomposition (SVD) algorithm [29].

Genome Sequence Selections

To test the performance of the method, 75 random sequences were selected from various places of the world samples and 6 animal samples were included for comparison purposes, and the detailed sequence name and date information is in Table 1. After the shortlist of sequences was tested, the total number of more than 20,000 sequences was applied to test the performance of the method in a large number of sample sizes.

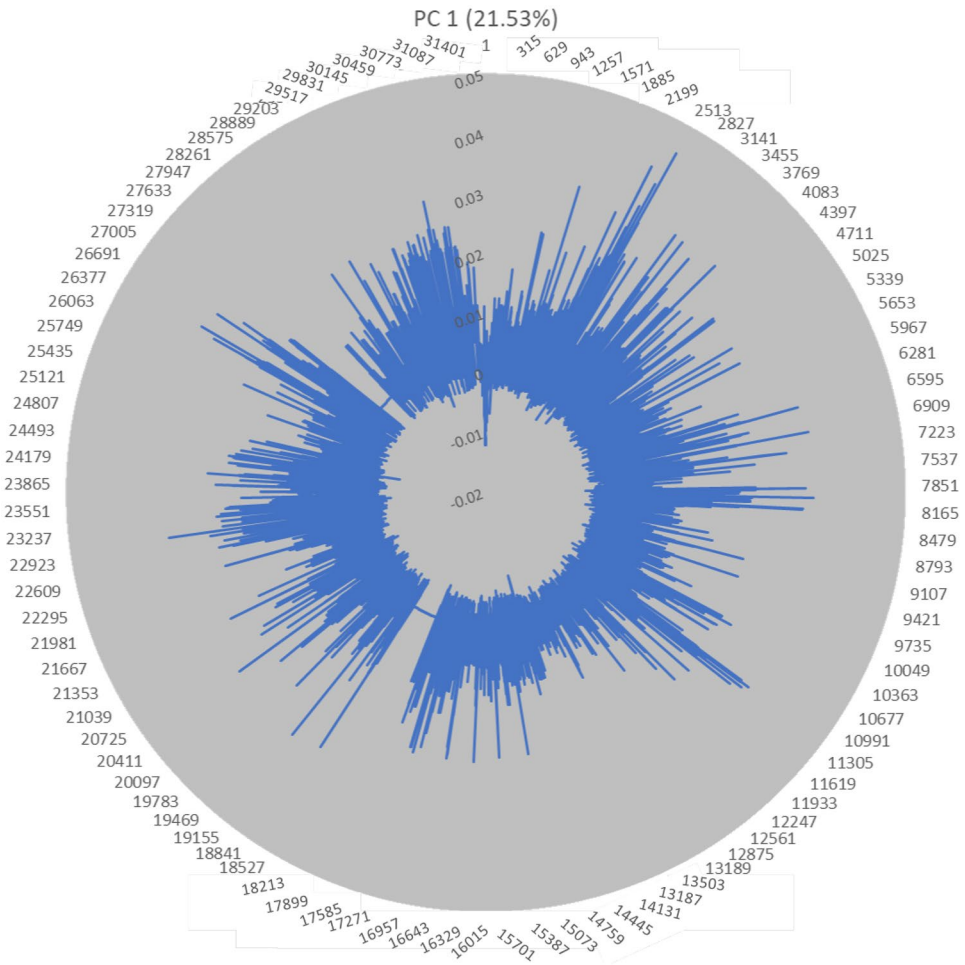
Results

The Sequence Alignment and Conversion

After alignment using MAFFT, a total of 31,690 elements were obtained for the shortlist (75) sequences, and 35,466 elements were obtained when the total sequences were applied to a total number of 21,094 sequences. One example of the aligned sequence is listed in the supporting files. (Dataset S2).

The PCA Study of the Shortlist Sequences

The conversion method successfully converted the letter sequences to numerical numbers which are suitable for PCA studies. The PCA study was carried out using mean-centered data, and the first two PCs showed a clear separation between the animal sequences and human sequences (Fig. 1). Most human sequences were separated from the first principal component (PC1) direction, and four sequences were separated from the second principal component (PC2). The loading plot [31] represented the features in the raw sequence which could lead to the separation of the samples in the score plot. In another word, when a potential mutation was observed by the PCA score plot, it is possible to track the changes in the sequence. For example, when the separation was mostly observed in the first PC, it is of interest to study the higher absolute values of PC1. However, the PCA score plot was calculated with a combined contribution of a large number of loadings but not just a single position. In the combined plot of the PCA score and loading plot (Fig. 2), the positions that lead to the separation and two large loadings can be analyzed by studying similar directions, and three representative positions were shown on the top of Fig. 2, and the detailed information was listed in Table 1. Though the score plot was calculated based on the combinations of all the important loadings, the difference between the animal samples and human virus samples can also be observed



in the representative positions (Table 1), and the results showed clear differences between the human samples and animal samples. In addition, the discovered nucleotide position regions with high first PC loadings (Fig. 3) are very similar to previous studies [30]. Position numbers may be different due to the alignment of the different sequences.

Fig. 3 The first PC loading distributions for the shortlist genome sequences which showed the genome site differences. The selected high loadings are included in Table 1, and more details can be obtained in Table 1

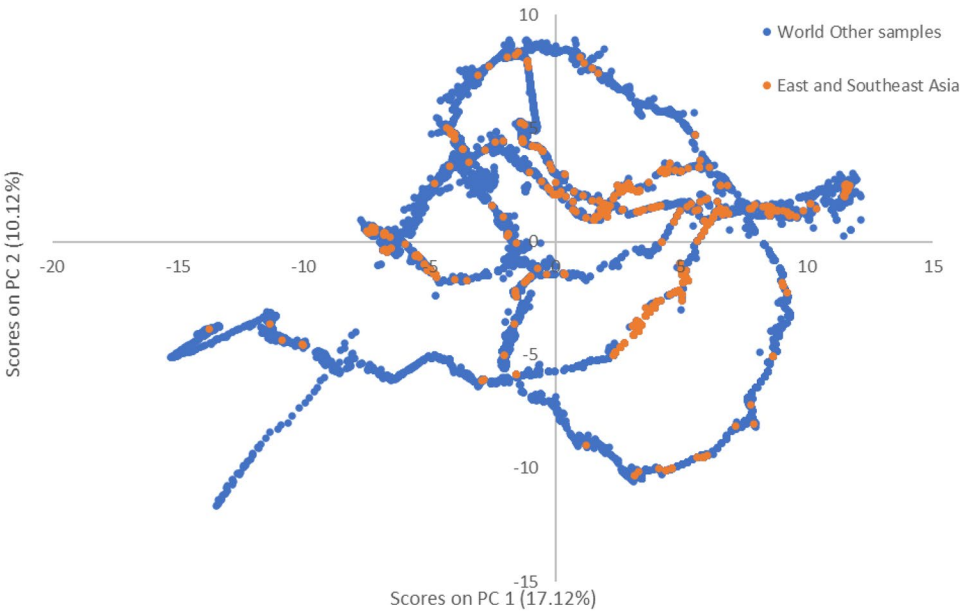


Fig. 4 The PCA score plot of the COVID-19 sequences including 19,697 samples. The orange dots are sequences reported from East and Southeast countries and regions. The blue dots are the samples reported from countries and regions in the other part of the world. The PCA score plot showed the distribution of the genome sequences from all over the world with differences in the first and second PC directions. The other PCs (PC3 to PC10) are listed in Fig. 5

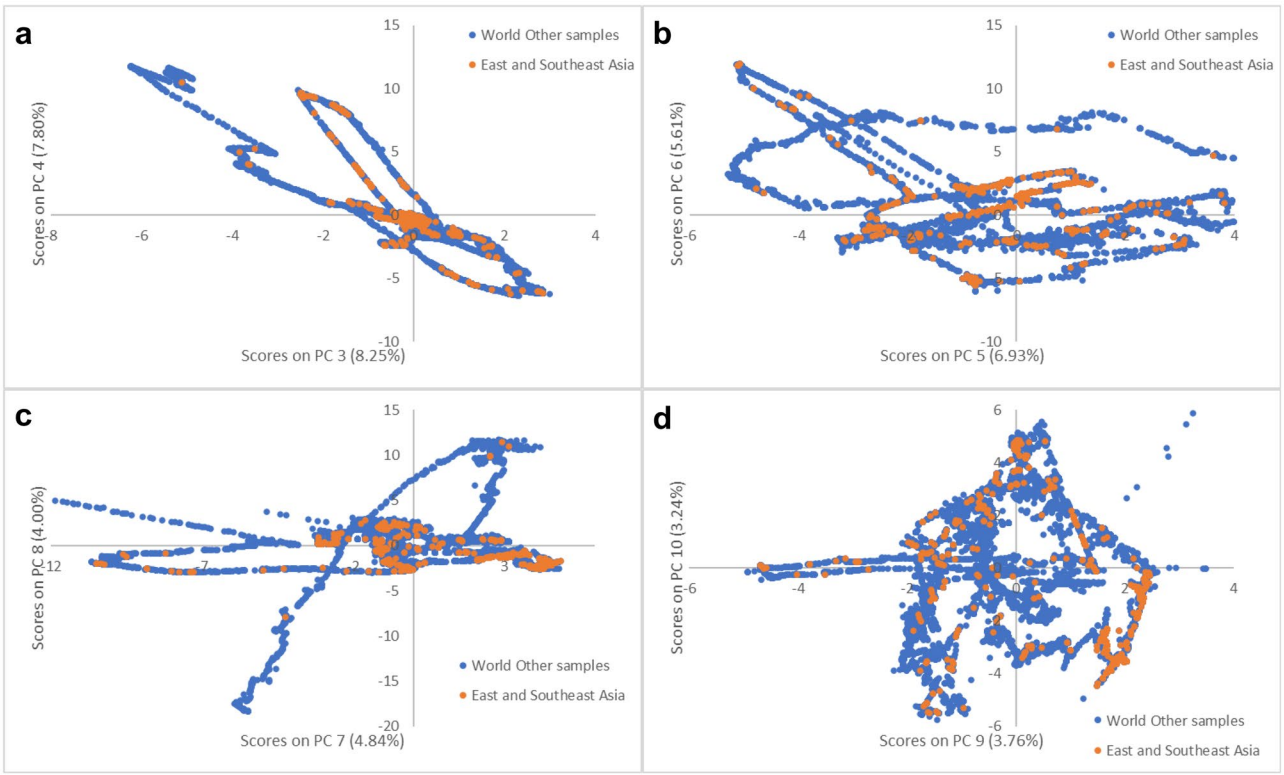


Fig. 5 The PCA score plot of the COVID-19 sequences including 19,697 samples. The orange dots are sequences reported from East and Southeast countries and regions. The blue dots are the samples reported from countries and regions in the other parts of the world. PCA score plots for PCs from 4 to 10. **a** Scores of PC3 vs PC4. **b** Scores of PC5 vs PC6. **c** Scores of PC7 vs PC8. **d** Scores of PC9 vs PC10

The PCA Study of a Large Number of Sequences

A large size dataset with 21,094 genome sequences obtained from GISAID was analyzed using the PCA method. The calculation for about 19,697 sequences cost around 90 min using a personal computer with 12G ram. Due to the complexity of the sequences, a hoteling T2 method was applied to exclude the potential outliers in the PCA score plot to minimize the error caused by potential human errors in sequence upload. The final dataset reserved 19,697 samples from all over the world and a PCA model was built using the mean-centered data in Figs. 4 and 5. The PCA score plot separation is mainly in the PC1 direction with several potential sub-groups. Since the pandemic was firstly reported in China, the East and Southeast Asian samples were highlighted in the score plot. No significant difference was observed in the highlighted samples excepted the left bottom part showed relatively fewer samples. Though the score plot showed that Asian samples tend to appear on the right side of the score plot, a clear difference that can classify virus types like the previous reported phylogenetic network study using 160 samples [3] was not observed. The comparison between Europe and the USA were also highly mixed (Fig. 6). Hence, the data reported time was applied in the following analysis.

Discussion

The Potential Changes with the Sequences Report Time

The mutation of COVID-19 is critical to study the prevention and drug development to fight with the novel coronavirus. When the early reported samples (before Jan 15, 2020) were highlighted in the PCA score plot, they mainly showed on the right side of the score plot (red dots in Figs. 7 and 8). With a slight increase in the time point to the end of January, more sequences were shown to the left side of the plot and the March samples showed more cases to the left side. Though the trend was observed, newer cases were scattered in all the parts of the PCA score plot which is reasonable since the newer reported cases are not necessarily newer mutations. The total of around 1700 samples on the left bottom side

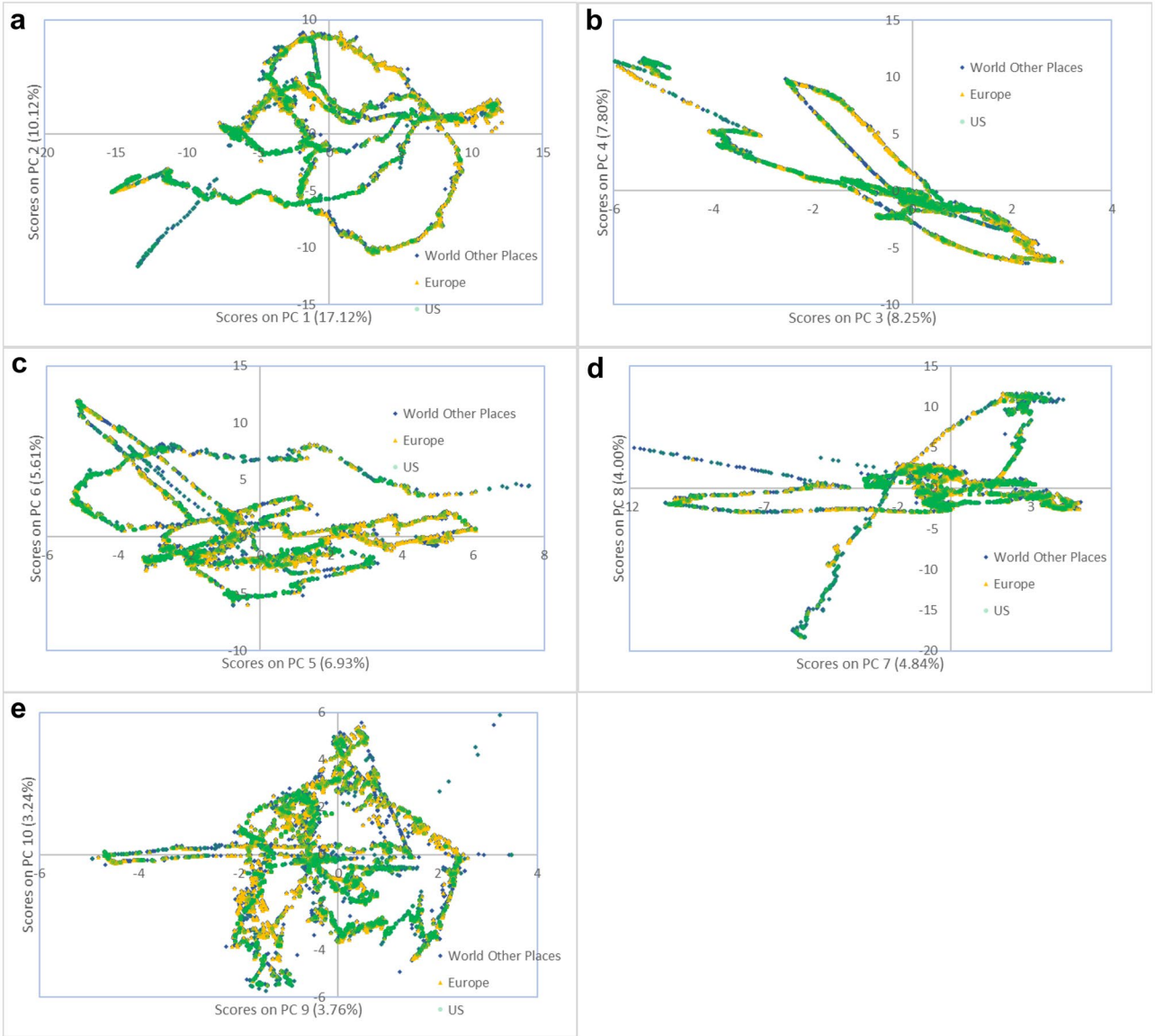


Fig. 6 The PCA score plot with the comparison of Europe and the US. **a** Scores of PC1 vs PC2. **b** Scores of PC3 vs PC4. **c** Scores of PC5 vs PC6. **d** Scores of PC7 vs PC8. **e** Scores of PC9 vs PC10

of the score plot are mainly from Europe and North America (around 95%) which may draw some attention for further mutation studies. The geographic clustering and time section were analyzed after the unsupervised PCA study without pre-settings. The nonlinear PCA method [32] could help to reduce errors judging by the short test data (Fig. 9); however, the methods consume hundreds of times processing time which could be used only for a small number of sequences.

PCA Loading Plot Information

The score plot was calculated with a combined contribution of a large number of loadings but not just a single position, so the loading plot provides the protentional important positions. For example, if there are group differences in the score plot, the loadings in the same direction will be more important in the differences. In the shortlist sequences, the beginning and end positions usually have

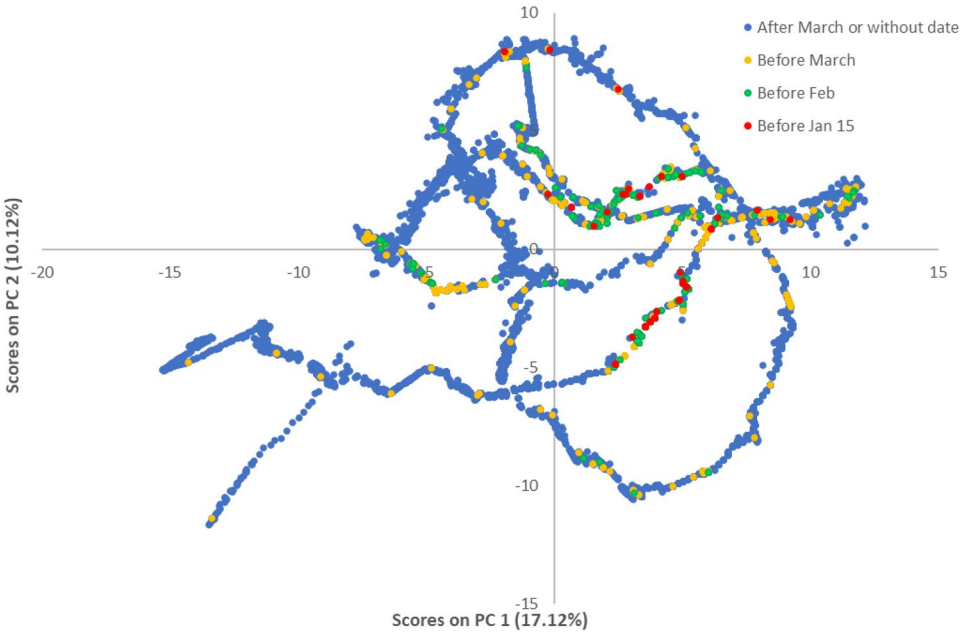


Fig. 7 The PCA score plot of the COVID-19 sequences including 19,697 samples. The red dots, green dots, and yellow dots are sequences reported before Jan 15, 2020, before Feb 1, 2020, and before March 1, 2020, representatively. The blue dots are the samples reported after March 1, 2020. The other PCs (PC3 to PC10) are listed in Fig. 8

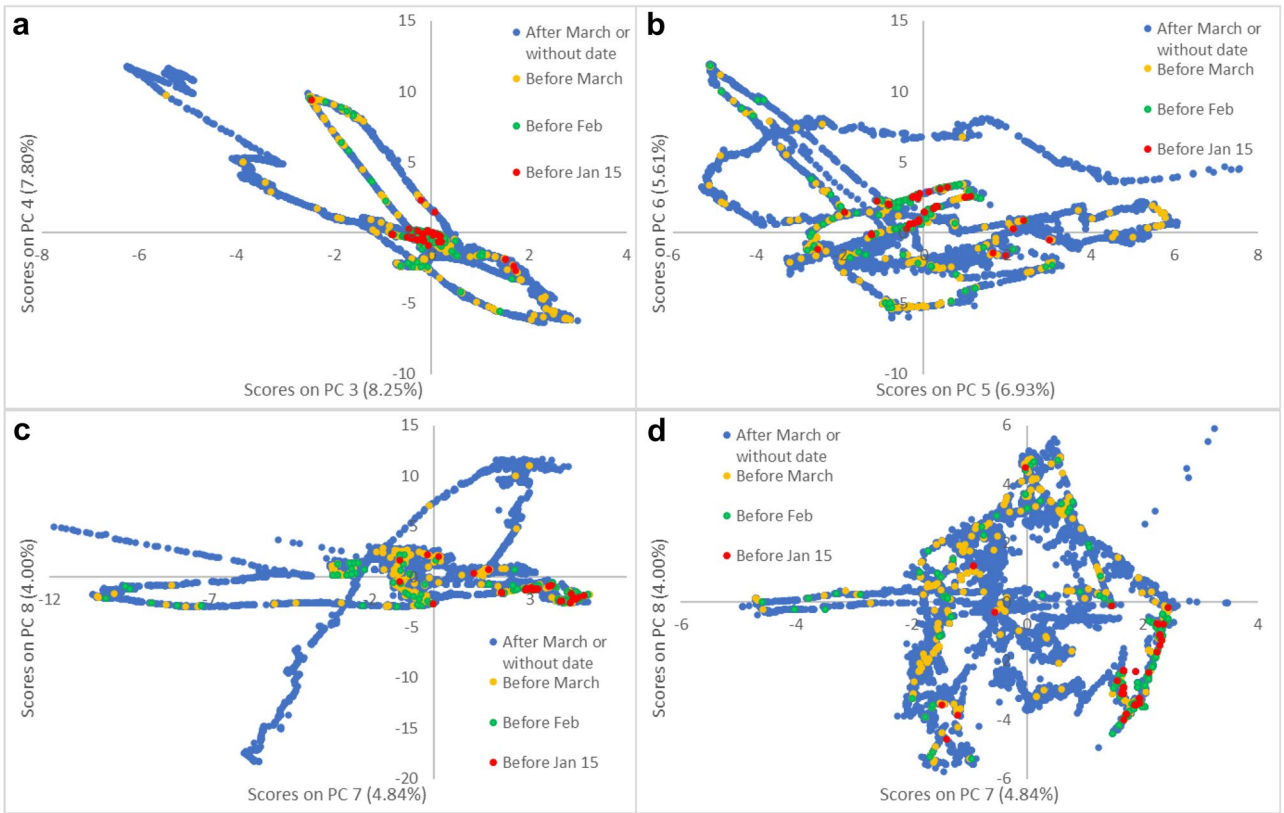


Fig. 8 The PCA score plot of the COVID-19 sequences including 19,697 samples. The red dots, green dots, and yellow dots are sequences reported before Jan 15, 2020, before Feb 1, 2020, and before March 1, 2020, representatively. The blue dots are the samples reported after March 1, 2020. **a** Scores of PC3 vs PC4. **b** Scores of PC5 vs PC6. **c** Scores of PC7 vs PC8. **d** Scores of PC9 vs PC10

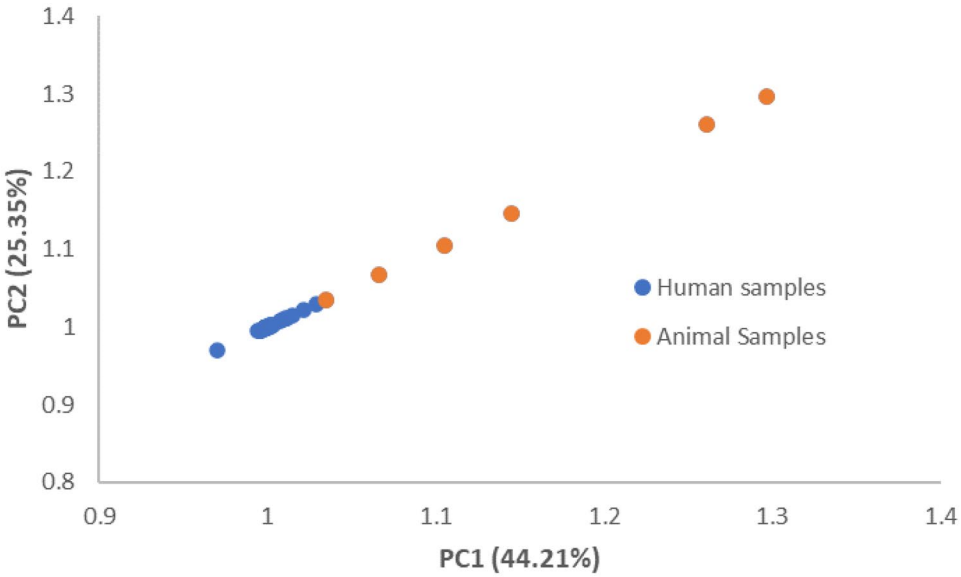


Fig. 9 The PCA score plot for non-linear PCA for the shortlist sample in Fig. 1 (processing time is about 2 h using a personal computer with 12 G ram)

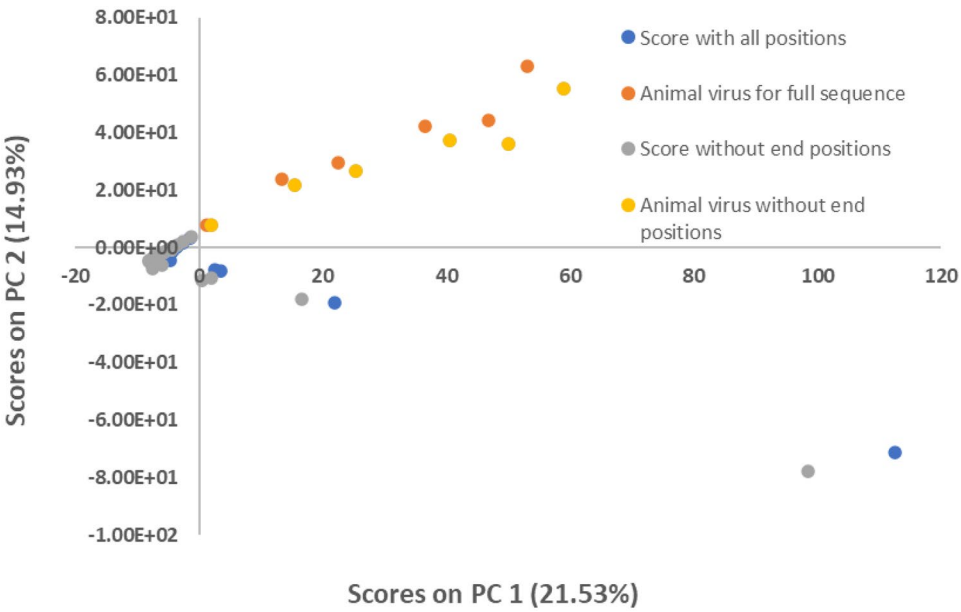


Fig. 10 The comparison of the PCA score plot between the original one and the one after removing the first and last 1000 positions. The blue and orange dots are the full sequences for human virus and animal virus, representatively. The grey dots and yellow dots are the sequences without end positions for the human and animal virus, representatively

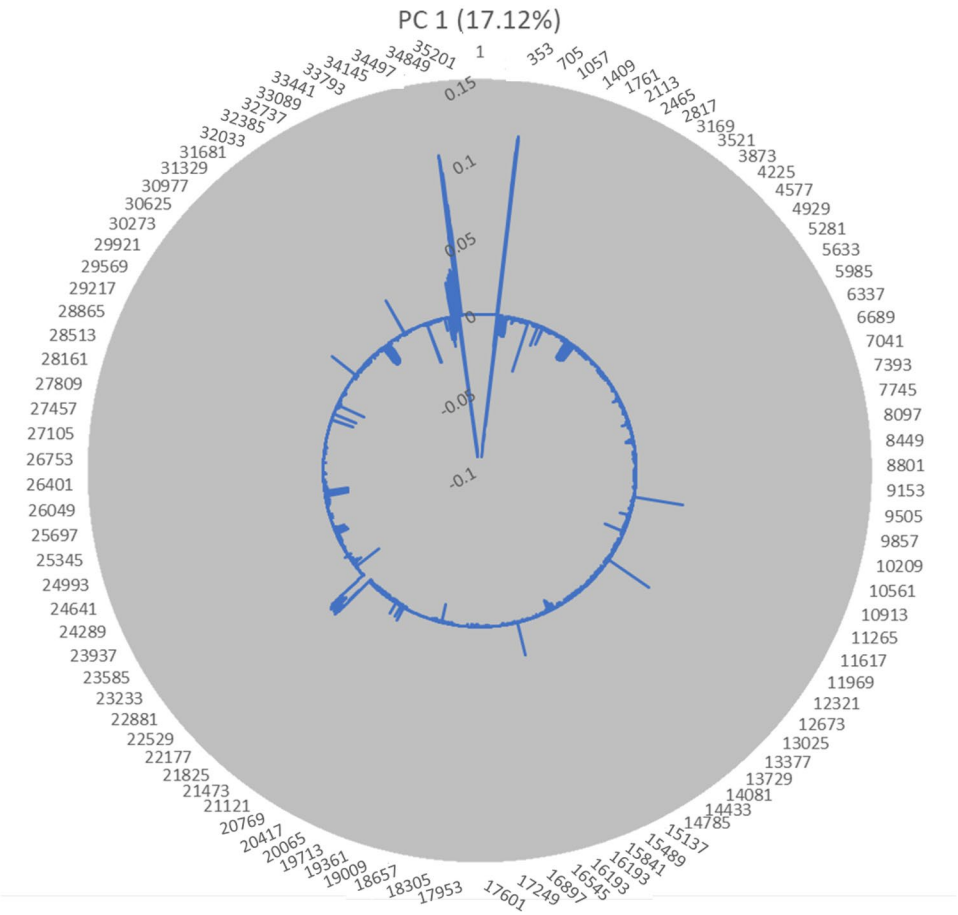


Fig. 11 The first PC loading distributions for the total 19,697 genome sequences which potentially indicated the mutation directions from time. The large loadings are 1782 3584, 9804, 11126, 12276, 16358, 20593, 22182, 22220, 30306, and 32631

differences due to the low coverages or the undetermined nucleotide; however, when the first and last 1000 positions were excluded from the calculation, almost no significant difference could be observed (Fig. 10). This showed a great advantage in a large amount of data interpretation when potential errors could exit with a hard cleanup process. Figure 11 showed that the first PC loadings lead to the separation of Fig. 7 which indicates the potential mutation after time. The important loadings are listed under Fig. 11, and a large area with multiple important loadings lies in position 22,182 which is the spike protein gene site [33]. The predicted gene site positions may contribute to the mutation studies in COVID-19. In addition, though the loading plot can only provide the general trend of the virus sequence difference, it can still be powerful in filtering important changes.

Conclusions

In this study, we adapted the frequency method-based PCA approaches originally designed for protein sequence to COVID-19 genome sequences which have more than 30,000 positions after alignment. The study was first demonstrated using a randomly selected shortlist of sequences with 75 samples with animal samples from bat and pangolin for testing the method, and the PCA showed as a powerful tool to separate the human samples from the animal samples using in the PCA score plot. The study also indicated that the end sequence uncertainty and gap positions have very limited influence on the score plot clustering which is suitable for the COVID-19 data which were submitted by users from all over the world and may have potential errors. The method was applied to a total of more than 20,000 sequences which showed the potential direction of the virus mutation with the time. The PCA method is expected to provide a fast analysis method with limited requirements for data cleaning but may have limitations in classifications. The authors suggest researchers use the tool in combination with other methods when a classification result is expected. When the COVID-19 genome sequence database becomes larger and larger, the method provides a great opportunity to study the mutation of the coronavirus and may also be served as a pre-processing for other methods

such as the tree building-related approaches. The PCA score plot interpretation also discovered that potential virus evolution directions may worth further study to investigate COVID-19 potential mutations.

Acknowledgments We gratefully acknowledge all the authors who submitted the sequences to the GISAID database. The detailed information of the contributors is listed in Supplementary Dataset S1.

Data Availability Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s12559-020-09790-w .

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix

Table 1 The detailed sequences names reported locations and dates with the three nucleotides positions listed. The animal samples are listed at the bottom of the table. The first 3 positions were selected by using the 2D loading plot and the rest positions were selected using the PC1 plot. The difference could be clearly observed from the last few rows for the animal samples

Sequence names	23011	22815	23404	2576	2534	11141	11201	19245	18748	26368	26437
Japan	C	A	A	A	A	A	A	C	C	A	A
Japan	C	A	A	A	A	A	A	C	C	A	A
Beijing	C	A	A	A	A	A	A	C	C	A	A
Beijing	C	A	A	A	A	A	A	C	C	A	A
Belgium	C	A	A	A	A	A	A	C	C	A	A
England	C	A	A	A	A	A	A	C	C	A	A
Ireland	C	A	A	A	A	A	A	C	C	A	A
Ireland	C	A	A	A	A	A	A	C	C	A	A
England	C	A	A	A	A	A	A	C	C	A	A
Algeria	C	A	A	A	A	A	A	C	C	A	A
Algeria	C	A	A	A	A	A	A	C	C	A	A
Austria	C	A	A	A	A	A	A	C	C	A	A
Austria	C	A	A	A	A	A	A	C	C	A	A
Greece	C	A	A	A	A	A	A	C	C	A	A
Brazil	C	A	A	A	A	A	A	C	C	A	A
Germany	C	A	A	A	A	A	A	C	C	A	A
Lebanon	C	A	A	A	A	A	A	C	C	A	A
Russia	C	A	A	A	A	A	A	C	C	A	A
Russia	C	A	A	A	A	A	A	C	C	A	A
Switzerland	C	A	A	A	A	A	A	C	C	A	A
France	C	A	A	A	A	A	A	C	C	A	A
Latvia	C	A	A	A	A	A	A	C	C	A	A
Switzerland	C	A	A	A	A	A	A	C	C	A	A
Iceland	C	A	A	A	A	A	A	C	C	A	A
Iceland	C	A	A	A	A	A	A	C	C	A	A
Sweden	C	A	A	A	A	A	A	C	C	A	A
Denmark	C	A	A	A	A	A	A	C	C	A	A
Germany	C	A	A	A	A	A	A	C	C	A	A
Lebanon	C	A	A	A	A	A	A	C	C	A	A
Slovakia	C	A	A	A	A	A	A	C	C	A	A
Denmark	C	A	A	A	A	A	A	C	C	A	A
Latvia	C	A	A	A	A	A	A	C	C	A	A
Sweden	C	A	A	A	A	A	A	C	C	A	A
Scotland	C	A	A	A	A	A	A	C	C	A	A
Greece	C	A	A	A	A	A	A	C	N	A	A
France	C	A	A	A	A	A	A	C	C	A	A
Scotland	C	A	A	A	A	A	A	C	C	A	A
Belarus	C	A	A	A	A	A	A	C	C	A	A

Table 1 (continued)

Sequence names	23011	22815	23404	2576	2534	11141	11201	19245	18748	26368	26437
Czech Republic	C	A	A	A	A	A	A	C	C	A	A
Lithuania	C	A	A	A	A	A	A	C	C	A	A
Argentina	C	A	A	A	A	A	A	C	C	A	A
Argentina	C	A	A	A	A	A	A	C	C	A	A
Lithuania	C	A	A	A	A	A	A	C	C	A	A
Slovakia	C	A	A	A	A	A	A	C	C	A	A
Brazil	C	A	A	A	A	A	A	C	C	A	A
Colombia	C	A	A	A	A	A	A	C	C	A	A
Colombia	C	A	A	A	A	A	A	C	C	A	A
Chile	C	A	A	A	A	A	A	C	C	A	A
Netherlands	C	A	A	A	A	A	A	C	C	A	A
Netherlands	C	A	A	A	A	A	A	C	C	A	A
Chile	C	A	A	A	A	A	A	C	C	A	A
Australia	C	A	A	A	A	A	A	C	C	A	A
Costa Rica	C	A	A	A	A	A	A	C	C	A	A
Croatia	C	A	A	A	A	A	A	C	C	A	A
Belgium	C	A	A	A	A	A	A	C	C	A	A
Australia	C	A	A	A	A	A	A	C	C	A	A
USA	C	A	A	A	A	A	A	C	C	A	A
USA	C	A	A	A	A	A	A	C	C	A	A
USA	C	A	A	A	A	A	A	C	C	A	A
Belarus	C	A	A	A	A	A	A	C	C	A	A
Wales	C	A	A	A	A	A	A	C	C	A	A
Wales	C	A	A	A	A	A	A	C	C	A	A
USA	C	A	A	A	A	A	A	C	C	A	A
Costa Rica	C	A	A	A	A	A	A	C	C	A	A
Singapore	C	A	A	A	A	A	A	C	C	A	A
Ghana	C	A	A	A	A	A	A	C	C	A	A
Czech Republic	C	A	A	A	A	A	A	C	C	A	A
Singapore	C	A	A	A	A	A	A	C	C	A	A
Ghana	C	A	A	A	A	A	A	C	C	A	A
bat	C	A	A	A	A	A	G	T	T	A	A
pangolin	A	G	C	-	-	-	-	N	T	-	-
pangolin	A	G	C	G	G	G	G	T	T	C	C
pangolin	N	N	N	N	N	N	N	Y	N	N	C
pangolin	T	T	G	T	T	T	G	A	A	T	T
bat	G	C	T	C	C	G	G	C	N	N	N

References

1. Li H, Liu S-M, Yu X-H, Tang S-L, Tang C-K. Coronavirus disease 2019 (COVID-19): current status and future perspectives. Int J Antimicrob Agents. 2020:105951.
2. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. Int J Antimicrob Agents. 2020;55(3).
3. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci. 2020:202004999.
4. Kim J-M, Chung Y-S, Jo HJ, Lee N-J, Kim MS, Woo SH, et al. Identification of coronavirus isolated from a patient in Korea with COVID-19. Osong Public Health Res Perspect. 2020;11(1):3–7.
5. Wang CT, Liu ZP, Chen ZX, Huang X, Xu MY, He TF, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 8.
6. Giovanetti M, Angeletti S, Benvenuto D, Ciccozzi M. A doubt of multiple introduction of SARS-CoV-2 in Italy: a preliminary overview. J Med Virol.
7. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Res. 2018;46(20):e120.

8. Shafee T, Bacic A, Johnson K. Evolution of sequence-diverse disordered regions in a protein family: order within the chaos. Mol Biol Evol. 2020;37(8):2155–72.
9. Silvin A, Chapuis N, Dunsmore G, Goubet AG, Dubuisson A, Derosa L, et al. Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. Cell. 2020.
10. Xu L, Yuille A. Robust principal component analysis by self-organizing rules based on statistical physics approach. IEEE Trans Neural Netw. 1995;6(1):131–43.
11. Tharwat A. Principal component analysis - a tutorial. Int J Appl Patt Rec. 2016;3(3):197–240.
12. Statheropoulos M, Pappa A, Karamertzanis P, Meuzelaar HLC. Noise reduction of fast, repetitive GC/MS measurements using principal component analysis (PCA). Anal Chim Acta. 1999;401(1–2):35–43.
13. Rymarczyk T, Sikora J. Optimization method and PCA noise suppression application for ultrasound transmission tomography. Przeglad Elektrotechniczny. 2020;96(2):90–3.
14. Yata K, Aoshima M. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. J Multivar Anal. 2012;105(1):193–215.
15. Chawla MPS. PCA and ICA processing methods for removal of artifacts and noise in electrocardiograms: a survey and comparison. Applied Soft Computing. 2011;11(2):2216–26.
16. Reid MK, Spencer KL. Use of principal components analysis (PCA) on estuarine sediment datasets: the effect of data pre-treatment. Environ Pollut. 2009;157(8–9):2275–81.
17. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017;22(13).
18. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Chall. 2017;1(1):33–46.
19. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol. 1995;2(2):171–8.
20. Wallace I, Higgins D. Supervised multivariate analysis of sequence groups to identify specificity determining residues. Bmc Bioinformatics. 2007;8.
21. Shafee T, Anderson MA. A quantitative map of protein sequence space for the cis-defensin superfamily. Bioinformatics. 2019;35(5):743–52.
22. Konishi T, Matsukuma S, Fuji H, Nakamura D, Satou N, Okano K. Principal component analysis applied directly to sequence matrix. Sci Rep. 2019;9(1):19297.
23. Wang B, Kennedy MA. Principal components analysis of protein sequence clusters. J Struct Funct Genomics. 2014;15(1):1–11.
24. Adams E, De Maesschalck R, De Spiegeleer B, Vander Heyden Y, Smeyers-Verbeke J, Massart D. Evaluation of dissolution profiles using principal component analysis. Int J Pharm. 2001;212(1):41–53.
25. Goodpaster A, Kennedy M. Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies. Chemom Intell Lab Syst. 2011;109(2):162–70.
26. Katoh K, Misawa K, Kuma K-i, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
27. Rose R, Golosova O, Sukhomlinov D, Tiunov A, Prosperi M. Flexible design of multiple metagenomics classification pipelines with UGENE. Bioinformatics. 2019;35(11):1963–5.
28. Protsyuk IV, Grekhov GA, Tiunov AV, Fursov MY. Shared bioinformatics databases within the Unipro UGENE platform. J Integr Bioinform. 2015;12(1):11.
29. Vogt F, Tacke M. Fast principal component analysis of large data sets. Chemom Intell Lab Syst. 2001;59(1–2):1–18.
30. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol. 2020;30(8):1578.
31. Yamamoto H, Fujimori T, Sato H, Ishikawa G, Kami K, Ohashi Y. Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. BMC Bioinformatics. 2014;15:9.
32. Karhunen J, Joutsensalo J. Representation and separation of signals using nonlinear PCA type learning. Neural Networks. 1994;7(1):113–27.
33. Wang R, Hozumi Y, Yin C, Wei G-W. Mutations on COVID-19 diagnostic targets. arXiv preprint. 2020.