

Title: Continuous mutation of SARS-CoV-2 during migration via three routes

Author: T. Konishi*

Affiliation: Akita Prefectural University

*Correspondence to: konishi@akita-pu.ac.jp

Abstract: Phylogenetic trees are often used to explain evolution, including that estimated from analyses of sequencing data. However, the application of such clustering trees is unsuitable in science as it involves several unverifiable assumptions. Here, we report an objective analysis of SARS-CoV-2 sequences with information on the date of collection and locations. The pandemic started in Wuhan, China and spread along three routes while accumulating mutations at a rate comparable to that of influenza H1N1. Failure to contain the virus can restrict the effect of herd immunity, and newer varieties will repeatedly appear.

One Sentence Summary: The sequence data of SARS-CoV-2, with date of collection and locations shows that the virus has spread from Wuhan along three routes and is rapidly mutating.

Short title: Continuous mutation of SARS-CoV-2

Keywords: COVID-19, genome sequencing, spreading routes, Principal Component Analysis

Introduction

Phylogenetic clustering has been used to observe the relationships among living organisms. It presents the relationships in the form of a tree. However, recent genomic information reveals that this has not been the right way, as evolution does not take the form a tree; in fact, there are many looping branches since genetic information can be transferred horizontally. For example, viruses may exchange a part of their genomes, and this is called a shift. Additionally, estimating the tree shape requires various assumptions that are never verified (1). Such a lack of falsifiability renders the application of the phylogenetic tree unsuitable in science (2).

Recently, Forster et al. found three main clusters in SARS-CoV-2 from 253 sequence samples and estimated the evolution of the virus. It started from one of the clusters, and each of the

clusters appeared to have specificities to human races (3). However, they used phylogenetic trees for analysis.

Sequence data can be regarded as multivariate data in which each base is a variable. Among various methods for multivariate analysis, the principal component analysis (PCA) is acceptable in science as it does not take unverifiable assumptions into account (2, 4). Here, 8,974 samples of SARS-CoV-2 virus were analysed by PCA. The estimated principal components (PCs) were further compared with the corresponding dates of collection and locations, thereby disclosing how this virus emerged, mutated, and spread.

Results and Discussion

The PCs of Sarbecoviruses, including SARS-CoV and SARS-CoV-2, is shown in Fig. 1A. The value \bar{d} is the mean standard deviation at the base, indicating the magnitude of variation in the sequence. The presented variations are due to the increasing number of sequence records in bats that have been reported after the SARS outbreak. Five pangolin and two bat samples were similar to SARS-CoV-2 (Fig. 1A). The variations in SARS and SARS-CoV-2 were not apparent in the PCA of Sarbecoviruses as a whole.

Bats would be the host of many coronaviruses and pangolins were thought to be the intermediate animal (5, 6); however, the reported sequences were away from the human viruses (Fig. 1B). The PCA totally ignored differences among 2,800 human samples; instead, it detected the large differences between human and the other hosts. In contrast, the sample of a tiger was located inside the human samples. It is unnatural to estimate that the bat or pangolin viruses could be the ancestral strain of SARS-CoV-2; rather, they may belong to a subclass that includes SARS-CoV-2 and the undiscovered ancestor. We may know only a part of the variations of bat viruses; sequencing is not easy as the coronavirus has a 30 kB long RNA genome. Storage of virus copies would be helpful and in fact, some of the records describe used lab-hosts. Cultured cells of monkeys may be ideal hosts for this purpose (7). It should however be noted that if a bat virus acclimatized in such cells (8) is leaked, the cell will act as a capable intermediate animal host.

The analysis further focused on SARS-CoV-2 (Fig. 2, Supplementary Information). Till date, the virus has been mutating along three routes (Fig. 2A). The first reports started in Wuhan, China (Fig. 2B), and then diffused horizontally in PC1. A variety that occurred in China transferred to Europe (Fig. 2A, green arrow, Fig. S2) and separated into two routes, divided along the PC2 axis. A new variety would be established when a small number of people carrying a mutant move to a new location and the infection spreads further. The lines that appeared on PCA reflected the trajectory of virus migration; accumulating mutations further elongated along the routes. Since only limited countries reported sequences with varied numbers (Table S2), the width of the lines would become thick or thin in panel A, and show horizontal lines in panel B. The similarity of the varieties found in Australia, England, Taiwan, and the United States (Table S1 and Fig. S3) is probably due to the fact that each country trades with many other countries. It should be noted that the locations are critical for the spreading route and not for the human race.

Reports for the first variety became sparse in April (Fig. 2B), although people may not have gained herd immunity. This may partly reflect the successful infection control in Wuhan; however, the variety has moved to other countries. It is possible that the newer varieties have acclimated to humans, and hence possess better infectivity.

The origin of the PCA, (0, 0), is the data average. According to the central limiting theorem, if a strain continuously mutates, it should appear near the data average among the formed variations. The first samples were located in close proximity to the center. They showed slightly negative values in PC1, but this bias may be attributed to a large number of reports from Europe.

Within a few months, the coronavirus has changed by 0.01 in PC1, a speed comparable to or even faster than that of R-type influenza viruses, which change in a single direction (Fig. S1). In European countries such as Iceland, several varieties have been found (Fig. 2A). Accumulation of these mutations demand novel immunities for virus control and hence, weaken the effect of herd immunity. The virus is bound to repeat seasonal epidemics such as influenza.

The results presented here are quite different from those reported by Forster et al. (3). One of the reasons is in the decision that the bat sample is the direct origin of SARS-CoV-2; however, the

appropriateness of this decision was not verified. This relates to the most basal unverifiable assumption of the clustering methods, whether the samples should be directly connected or not. Additionally, the previous study did not take the passage of time into account. Another disadvantage of a phylogenetic tree is that it is more difficult to reconcile with other information.

In contrast, PCA generates a spreadsheet with the same dimension as the sequence matrix, facilitating data integration. Actually, the estimation of the origin of the outbreak in this report has two distinct roots of evidence: the close proximity to the mean data and the start of the records.

Seeking the parent virus for SARS-CoV-2 would be difficult. Tigers may not be an intermediate animal because the sample was far from the mean data (Fig. 2A) and the collection date was too recent (Table S1). Even if we find a similar virus in bats, it may be transferred or contaminated by humans. There would be a subclass that includes humans, pangolins, and bats; however, sequencing a newer variety would require new sets of primers. Hence, it is necessary to store copies of the virus in the lab, but this involves the risk of accidental leaks without sufficient benefit. Some of the reports used new generation sequencers, which will produce short, unassembled reads (5, 6) and even may mix up different varieties. Therefore, although the approach is quite interesting, it is not suitable for identifying a novel strain.

References

1. Ellis G & Silk J (2014) Scientific method: Defend the integrity of physics. *Nature* 516:321-323.
2. Konishi T, *et al.* (2019) Principal Component Analysis applied directly to Sequence Matrix. *Scientific Reports* 9(1):19297.
3. Forster P, Forster L, Renfrew C, & Forster M (2020) Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences*:202004999.
4. Konishi T (2015) Principal component analysis for designed experiments. *BMC Bioinformatics* 16 Suppl 18:S7.
5. Lam TT-Y, *et al.* (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*.
6. Zhou H, *et al.* (2020) A novel bat coronavirus closely related to SARS-CoV-2 contains

natural insertions at the S1/S2 cleavage site of the spike protein. *Current Biology*.

7. Matsuyama S, *et al.* (2020) Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. *Proceedings of the National Academy of Sciences* 117(13):7001-7003.

8. Graham RL & Baric RS (2010) Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *Journal of Virology* 84(7):3134-3146.

Acknowledgments: the author would like to thank Editage (www.editage.com) for English language editing.

Competing interests: the author has no competing interests.

Data and materials availability: all data is available in the Supplementary Information.

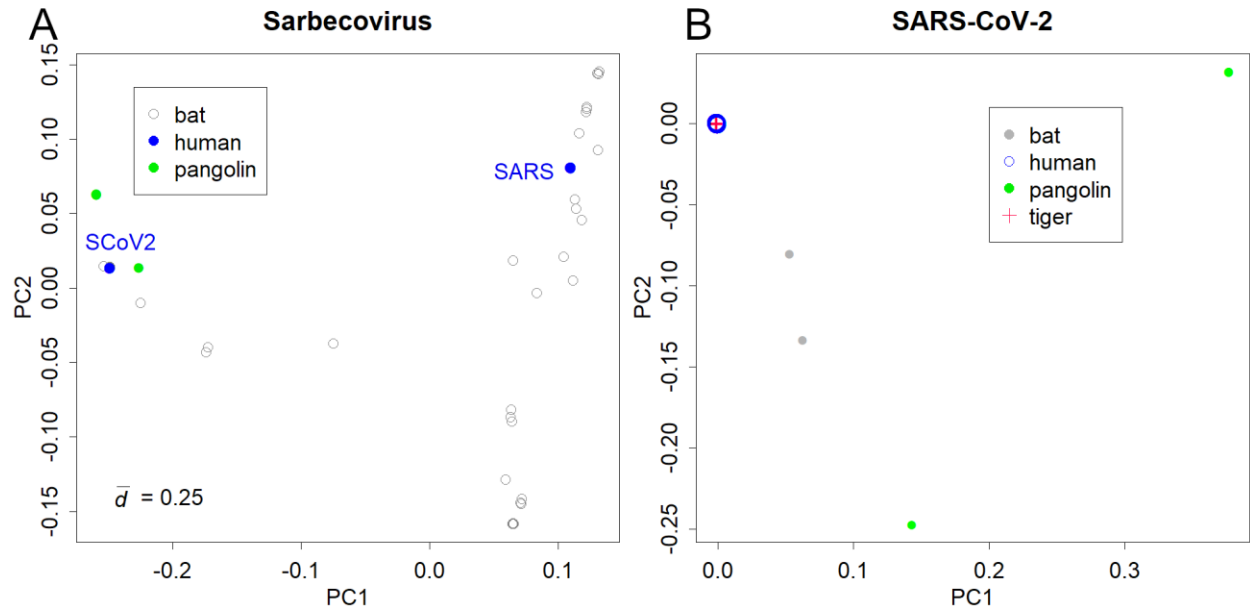


Fig. 1. PCA for coronavirus. **A.** Sarbecovirus. SARS: severe acute respiratory syndrome virus, SCoV2; SARS-CoV-2. The value \bar{d} is the standard deviation at each base that shows the magnitude of variation. **B.** SARS-CoV-2 virus in humans, two bats, five Malayan pangolins, and a tiger.

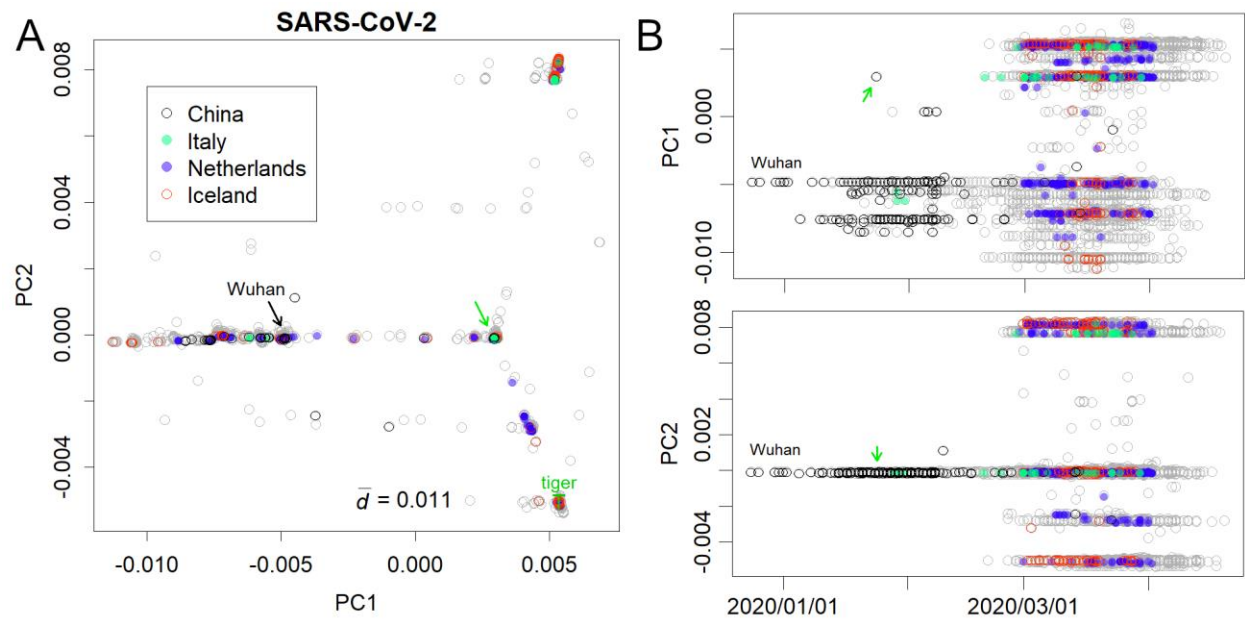


Fig. 2. **A.** Human SARS-CoV-2 virus. **B.** PCs against date of collection. A 3D version is available in Fig. S4.

Materials and Methods

Data and classification

Sequencing data were obtained from the DNA Data Bank of Japan (DDBJ) and GISAID
5 databases. Unfortunately, some of the records were rather preliminary and contained sections of
uncertain reads as well as extra stretches of repetitions. In order to avoid artifacts that are
provoked by the sequencing errors, the corresponding regions were replaced with the average
data in the PCA. This treatment practically removes such uncompleted regions from the analysis.
The sequences were aligned with DECIPHER (8) and manually completed using MEGA (9).
10 They were further processed to observe the relationships among samples by using the direct PCA
method (1), the recent codes of which are presented in GitHub
(<https://github.com/TomokazuKonishi/direct-PCA-for-sequences>).

Estimation of the magnitude of sample variations

15 The scale among sample sequences was estimated by mean distances, scaled by the length of
sequence m , of virus types. This is a type of standard deviation, $\bar{d} = \sqrt{\sum (x - \bar{x})^2 / 2mn}$, where
 x , \bar{x} , and n are the Boolean of each sample sequence, the sample mean, and the number of
samples, respectively. The unit of length is the same as that of the PCA, which will extract the
length towards particular directions.