Changes in SARS-CoV-2 before the peak in each country, producing unique variants

Tomokazu Konishi

Graduate School of Bioresource Sciences, Akita Prefectural University

Abstract

The second and third waves of coronavirus disease 2019 (COVID-19) have caused problems worldwide [1]. Those are often thought to have resulted from people's carelessness or people not following restrictions, but in reality, the cause remains unclear. Here, using an objective analytical method, we present the changes in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus causing COVID-19 over time. The virus has mutated in three major directions [2], with three groups remaining to date. The basic structure of the groups was completed by April and shared across all continents. However, the virus continued to mutate independently in each country after the borders were closed. In particular, the virus mutated before the occurrence of a second or third peak. It seems that the mutations conferred higher infectivity to the virus, because of which the virus overcame previously effective protections. Currently, each country may possess such a unique stronger variant, which may cause another peak in other countries. These viruses could also serve as sources of mutations by exchanging parts of the genome. Such mutations could create a variant with superior infectivity.

Introduction

The evolutionary trajectory of an evolved virus is difficult to analyse from nucleotide sequences because the data have complex multivariate structures with numerous dimensions [3]. Phylogenetic

trees have long been used to present relationships in among sequences [4] and many studies on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [5] have employed them [6 7-9], but they have two drawbacks. One is the decisive lack of falsifiability that ruins objectivity [10], and the other is a lack of generality that makes it difficult to integrate with other data sources.

Here, an analysis was performed using a more objective method, i.e., principal component analysis (PCA) [11]. PCA runs a singular value decomposition on a sequence described as a Boolean vector and obtains the principal component (PC) for the sample and base in parallel [12]. PCA finds common directions among a dataset and represents them as independent axes. These axes have an order; PC1 is common to most data, and the lower the contribution to differences in the data, the lower the order. The sample or base is presented on the axis according to its characteristics. As shown in the current study, SARS-CoV-2 data were represented by two axes as a whole. The lower axes represented mutations found in smaller areas. The components were scaled and compared with those of the influenza virus [13] as needed. In addition, the results were integrated with the date and location of collection.

Results and Discussion

Overview of PCA. PCs 1 and 2 formed a triangle comprising four groups that were temporarily termed as 0-3 (Fig. 1A). All groups were found on all continents. In contrast to PC1 and PC2 showing the overall situation, the PC3 axis and the lower-order axes showed differences that relate to only a small number of the samples (Fig. 1B and Extended Data Fig. 1), with smaller contributions to differences in the data (Extended Data Fig. 2); many of them were detected in limited countries after April (comparing Extended Data Fig. 1 and 3). Such mutations occurred since April when countries began to restrict the movement of people across national borders. The

overview of Fig. 1A was completed at the end of March (Fig. 1C). Although group 3 once appeared worldwide, it has been contained to date (Fig. 1D).

These features were similar for both RNA and protein (Extended Data Fig. 1 and 4). The rate of missense mutations, which affects amino acid sequence, was 63% (Extended Data Fig. 5). This is a fairly high rate; rather, null mutations are expected to be 61% [14] even though some differences caused by ignoring damaging mutations have been identified here. In fact, in the case of H1N1 influenza virus from 2001 to 2003 in the United States, the rate was 22%. This suggested that SARS-CoV-2 was under selective pressure to alter its protein structure.

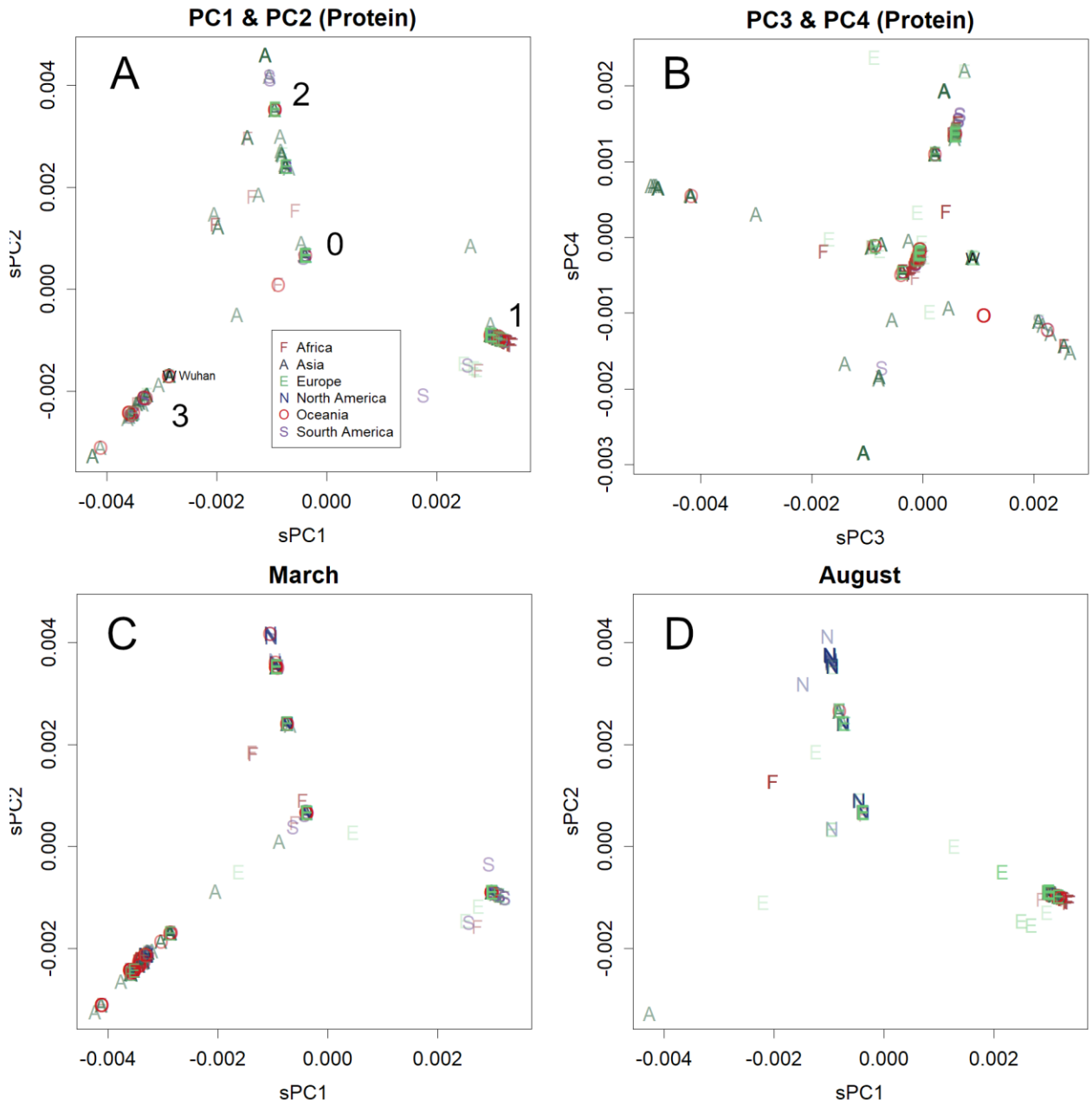


Figure 1. Principal components of SARS-CoV-2 amino acid sequences reported until December 1, 2019. One hundred samples are randomly selected from a continent and presented. **A**. PC1 and PC2. Numbers are given for a temporal explanation of the groups. **B**. PC3 and PC4. The earliest variants found in Wuhan are indicated by the white letter W. The numbers of samples that are separated by the axes are reduced in the lower-order axes (Extended Data Fig. 2 and 3). **C** and **D**. PC1 and PC2 at the end of March and from August 2020. A maximum of 100 samples is selected

from a continent and presented. The components are scaled by the length of sequences; squares of the values are correlated with the expectation of differences per base (Supplementary Information).

The sequences changed prior to the second wave. At the second peak in each country, the virus mutated from the first peak; the changes were observed in the lower PCs, suggesting that the mutations occurred in a region-specific manner, rather than as migrations from other countries (Fig. 2 and Extended Data Fig. 6).

These mutations can be both the result—and the cause—of the second wave. In peak times, there are more patients, which will provide more chances of mutation; this is the result. South Africa could be the case; the variant suddenly appeared, emerged, and peaked. Or, it could have been hidden from national surveillance. However, in Australia, England, France, and the United States, mutations occurred several weeks prior to the peaks (Fig. 2 and Extended Data Fig. 6). This suggested that the mutations were the cause of the peak. In any case, the strains prevalent at those peaks should certainly be more infectious than they were before the mutation, so the variants were replaced. The higher infectivity may have weakened the previously effective protections from the virus. Altered residues between the peaks are shown in Extended Data Table 1; because PCA showed differences among samples and bases in parallel, PCs for bases may help identify those differences. Almost all differences were unique to each country, but this was not surprising because the number of mutations remained too small for a coincidence to be expected. However, group 1 in Australia and group 2 in France were similarly mutated in the spike protein. This mutation would be beneficial regardless of the group because it could function independently of the rest of the virus.

It may be possible to predict the next wave by monitoring the mutations. However, only a limited

number of countries continue sequencing the virus; many have reported only the first peak. However, perhaps the second and subsequent peaks are caused by different variants, as in the cases of countries shown in Figure 2. In many countries and regions, sequences have not been reported at all; it is possible that there are other infectious variants or those that can even repel the acquired immunity in those counties.

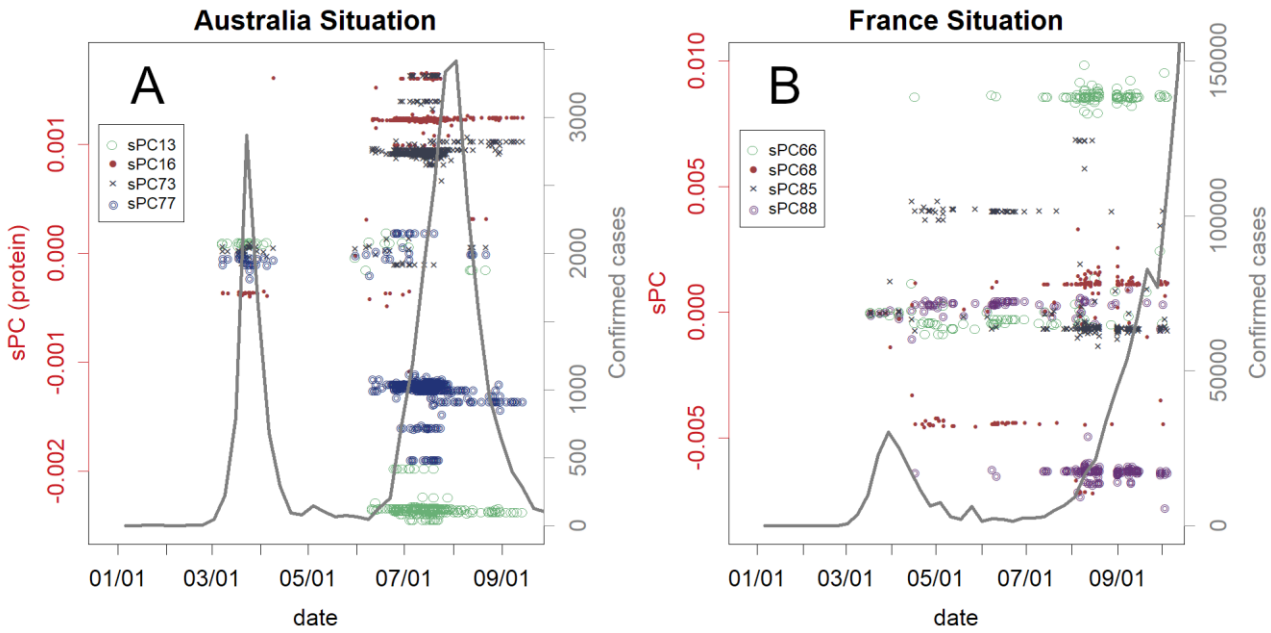


Figure 2. Mutations found on specific principal component (PC) axes and the number of patients. **A.** Australia and **B.** France. The scaled PCs are shown on the left axis, whereas the number of confirmed patients is shown on the right axis. Those for the USA, South Africa, and England are shown in Extended Data Fig. 6.

Magnitude of mutations. Influenza H1N1 is highly infectious in humans [15]. It mutates continuously

but does not appear in the same area during two sequential years. It is prevalent every few years because it takes time to mutate enough to survive against herd immunity in the area. However, the magnitude of changes in the meantime varies (Table 1). Perhaps, the part of the sequence that needs to be modified to escape herd immunity varies from case to case. However, the changes in coronavirus before and after the peaks were much smaller (Table 1). Additionally, at present, the average number of point mutations from the virtual centre is 5.9 bases and 3.7 residues. The farthest groups found in PC1 and PC2 were South Africa (top end of PC1) and South Korea (top of PC2); they differed by 19 residues (Table 1, max [PC1 and PC2]). This difference seems to be smaller than that needed to obtain another peak of seasonal influenza, although this does not guarantee safety from exceptional reinfections [16,17].

A case of pdm09 influenza virus that likely corresponded to SARS-CoV-2 mutations was observed in Thailand from 2009 to 2010 (Extended Data Fig. 7) [18]. The 2009 pandemic was subdued in this tropical country, and three peaks were subsequently identified toward 2010. At the second peak, only the nucleotide sequence changed, and at the third peak, the hemagglutinin protein changed (Table 1) with a high rate of missense mutations (54%). However, these changes were not highly associated with herd immunity because most of the population was not yet infected. These changes might be caused by the adaptation of pdm09—originally a swine strain—to humans.

Influenza H1N1	/1000
USA, 2001-2003	32
New Zealand, 2001-2005	15
(2008-2009)	203
USA, 2009-2014	15
USA, 2014-2016	7.1
Thailand, 2009 - 2010/2/1	0
Thailand, 2010/2/1 - 10/10	2.5
SARS-CoV-2	/1000
Australia (group 1), 4/1 - 8/20	0.5
France (group 2), 6/5 - 9/20	1
England (group 0), 3/30-8/15	1.1
South Africa (group 1), 5/30-8/3	0.8
USA (group 2), 3/30-8/15	0.4
max [PC1, PC2]	2

Table 1. The rate of mutated amino acid residues during the indicated years and location. The numbers show mutations per 1000 amino acid residues. Results of the H1N1 influenza virus hemagglutinin protein observed in the peak seasons, and the whole genome of coronavirus reported in the peaks are shown. (2008–2009) indicates the difference between strains before 2009 and the pdm09 influenza virus. Max (PC1, PC2) is the difference between the most diverged SARS-CoV-2 groups in Fig. 1A.

The influenza virus genome continues to mutate and not only in the part that is directly associated with infection [13]. This property may be the reason it has repeatedly deceived herd immunity for decades, allowing influenza to remain as a pandemic.

In contrast, mutations in SARS-CoV-2 were not uniform; rather, some smaller open reading frames (ORFs) such as E, M, and ORF6 were preserved (Fig. 3). These could have adapted to humans, but they are also conserved in various coronaviruses [19]. If they remain conserved, they could be subjected to herd immunity, extending the period between future pandemics. Evidently, this does not mean that these ORFs do not change forever; the rate of change is just slower.

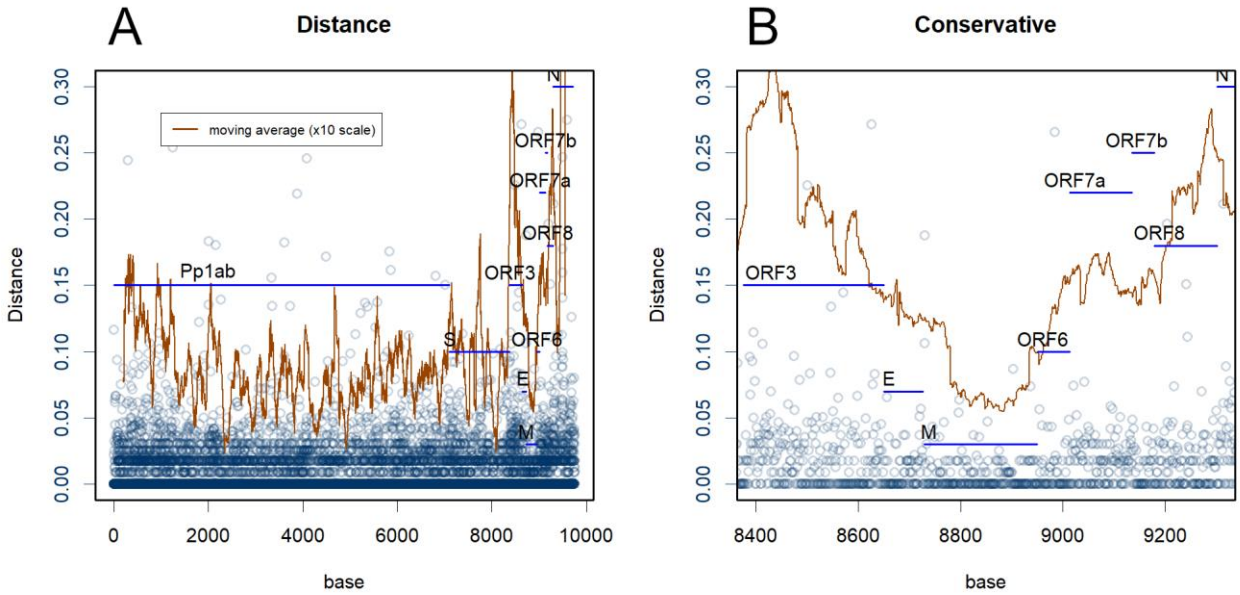


Figure 3 A. Mean distances to the sample mean (protein). The square of the distance reflects the expected mutation occurrence at the residue. The brown line indicates the moving average of 100 residues (×10 scale). **B.** Magnified view of the conserved region.

The strain responsible for the pandemic mutated rapidly. As of March, the bases of PCs 1 and 2 were formed, and each continent harboured the same set of mutated variants (Fig. 1). Since then, human movement has been restricted, and each country accumulated its own variants. In this process, the force that caused the mutations was probably adaptation to humans (Table 1). It caused changes in both codon usage and protein structure (Extended Data Fig. 1 and 4), but the rate of missense mutations was high.

The unique variants in each country would serve as the source of newer variants, which are more infectious than the previous variants; therefore, they replaced the mode (Fig. 2 and Extended Data Fig. 6). If countries had not closed their national borders, many of their variants could have been replaced by other more powerful variants, as in the case of group 3. However, the unique variants survived in a narrower area. International events, such as the Olympic Games, which attract people in large numbers, will serve as an exchange market for these viruses. To make matters worse, coronavirus could acquire shift-type mutations, resulting in a mixture of genomes [19]. When an exported strain is mixed with a domestic strain, a new strain is created. If each of the unique mutations works independently (otherwise, replacement of the older variants with these variants would not be beneficial), the accumulation of the mutations can result in the production of more powerful viruses with cherry-picked the surviving unique mutations.

Materials and Methods

PCA. Nucleotide sequences were obtained from GISAID [20] on October 1, 2020. Only the complete sequences that did not contain N were selected. Sequences were aligned using DECHIPER [21]. The sequences were converted to a Boolean vector and subjected to PCA [3]. Sample PCs and sequence PCs are scaled based on the length of the sequence and the number of samples, respectively [12]. All calculations were performed using R [22]. The ID, acknowledgements, and scaled PCs of samples and bases can be downloaded from Figshare (dx.doi.org/10.6084/m9.figshare.6025748). The newest version of R code and the PCA axes are publicly available (https://github.com/TomokazuKonishi/Fast-calc-for-SARS-CoV-2). Using these methods, any sequence data can be represented on the same axes presented here, and the axes can also be updated with newer sequences. The relationship between PC values and the number of mutated bases is briefly explained in the Supplementary Information. Number of confirmed cases were obtained from WHO [5].

The axes were identified and used as follows. Because PCA axes are sensitive to sample bias, they were determined using 6,092 samples randomly selected from each continent in proportion to their population; 989 from Africa, 3636 from Asia, 610 from Europe, 476 from North America, 347 from South America, and 34 from Oceania; the number of sequences used from Africa was the maximum available; hence, the sampling was not random. For clarity, Fig. 1 presents only 100 samples from each continent, whereas Extended Data figures present approximately 1000 samples. The contributions of the axes are presented in Extended Data Fig. 2.

Distances between samples from different peaks. In influenza, only one strain has emerged over the years, with repeated mutations [13]. Differences between peaks within the same strain were found as

differences between the mean sequences from both peaks. In contrast, in coronavirus, the peak comprises several different variants, especially the first peak. Averaging will cancel the characteristics of these variants. To avoid this problem, the differences were estimated among those that have similar PC1 and PC2 and are in the same group as shown in Table 1, and the differences only appeared in the lower PCs.

The rate of missense mutations. In influenza, differences in both nucleotide and amino acid sequences were presumed as differences between the average peak sequences. For coronavirus, the difference between the mean sequence and each sample was estimated, and the rate was estimated for each sample. The distribution was slightly distorted (Extended Data Fig. 5), and hence, the representative value was found as the median.

1 WHO. *Coronavirus disease (COVID-19) Pandemic*, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> (2020).

2 Konishi, T. *Continuous mutation of SARS-CoV-2 during migration via three routes*. (2020). doi: 10.6084/m9.figshare.12365777.v1

3 Konishi, T. *et al.* Principal Component Analysis applied directly to Sequence Matrix. *Scientific Reports* **9**, 19297, doi:10.1038/s41598-019-55253-0 (2019).

4 Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-314, doi:10.1038/nrg3186 (2012).

5 WHO. *WHO Coronavirus Disease (COVID-19) Dashboard*, <https://covid19.who.int/> (2020).

6 Alm, E. *et al.* Geographical and temporal distribution of SARS-CoV-2 clades in the WHO

European Region, January to June 2020. *Euro Surveill* **25**, 2001410, doi:10.2807/1560-7917.ES.2020.25.32.2001410 (2020).

7 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).

8 Lam, T. T.-Y. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* **583**, 282-285, doi:10.1038/s41586-020-2169-0 (2020).

9 Murakami, S. *et al.* Detection and Characterization of Bat Sarbecovirus Phylogenetically Related to SARS-CoV-2. *Emerging Infectious Diseases* **26**, 3025-3029 (2020).

10 Ellis, G. & Silk, J. Scientific method: Defend the integrity of physics. *Nature* **516**, 321-323, doi:10.1038/516321a (2014).

11 Jolliffe, I. T. *Principal Component Analysis*. (Springer-Verlag 2002).

12 Konishi, T. Principal component analysis for designed experiments. *BMC Bioinformatics* **16 Suppl 18**, S7, doi:10.1186/1471-2105-16-S18-S7 (2015).

13 Konishi, T. Re-evaluation of the evolution of influenza H1 viruses using direct PCA. *Scientific Reports* **9**, 19287, doi:10.1038/s41598-019-55254-z (2019).

14 Wang, T. *et al.* Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database. *Nature Communications* **9**, 441, doi:10.1038/s41467-017-02806-4 (2018).

15 CDC. *Similarities and Differences between Flu and COVID-19*, <https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm> (2020).

16 Gupta, V. *et al.* Asymptomatic Reinfection in 2 Healthcare Workers From India With Genetically Distinct Severe Acute Respiratory Syndrome Coronavirus 2. *Clinical Infectious Diseases*, doi:10.1093/cid/ciaa1451 (2020).

17 Edridge, A. W. D. *et al.* Seasonal coronavirus protective immunity is short-lasting. *Nature*

Medicine **26**, 1691-1693, doi:10.1038/s41591-020-1083-1 (2020).

18 WHO. *Influenza surveillance outputs*, <https://www.who.int/influenza/resources/charts/en/> (2020).

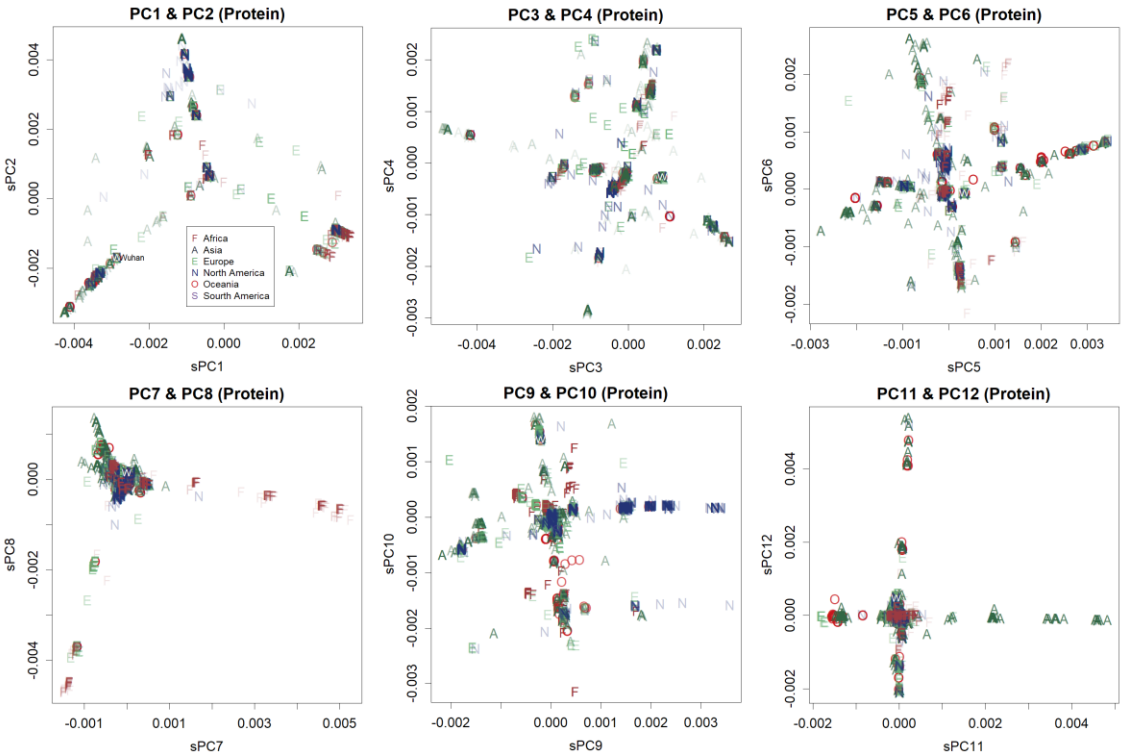
19 Konishi, T. Principal component analysis of coronaviruses reveals their diversity and seasonal and pandemic potential. *PLoS ONE* **15**, e0242954, doi:https://doi.org/10.1371/journal.pone.0242954 (2020).

20 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global challenges (Hoboken, NJ)* **1**, 33-46, doi:10.1002/gch2.1018 (2017).

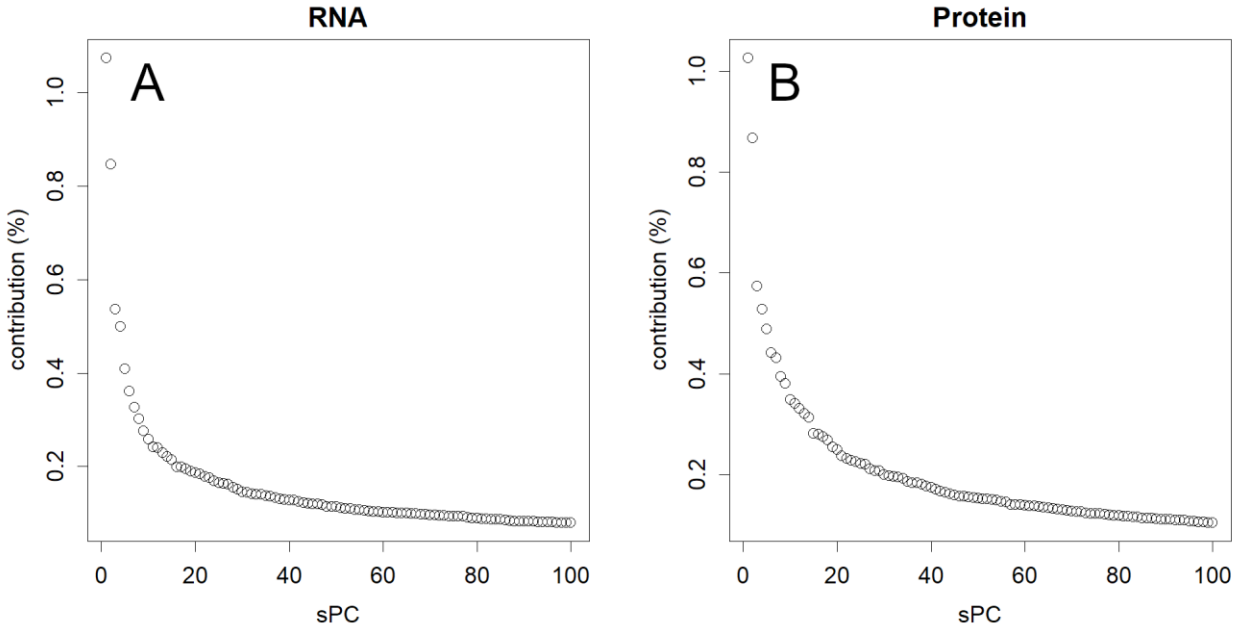
21 Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16**, 322, doi:10.1186/s12859-015-0749-z (2015).

22 R_Core_Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2020).

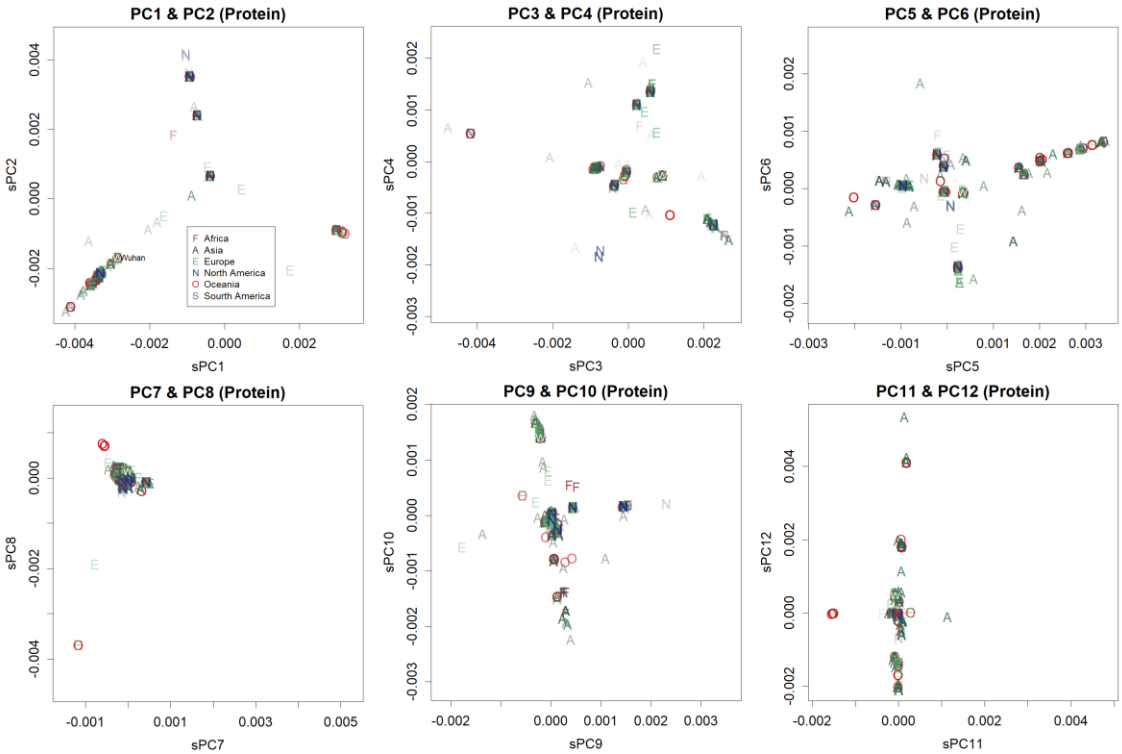
Extended Data for Changes in SARS-CoV-2 before the peak in each country, producing unique variants



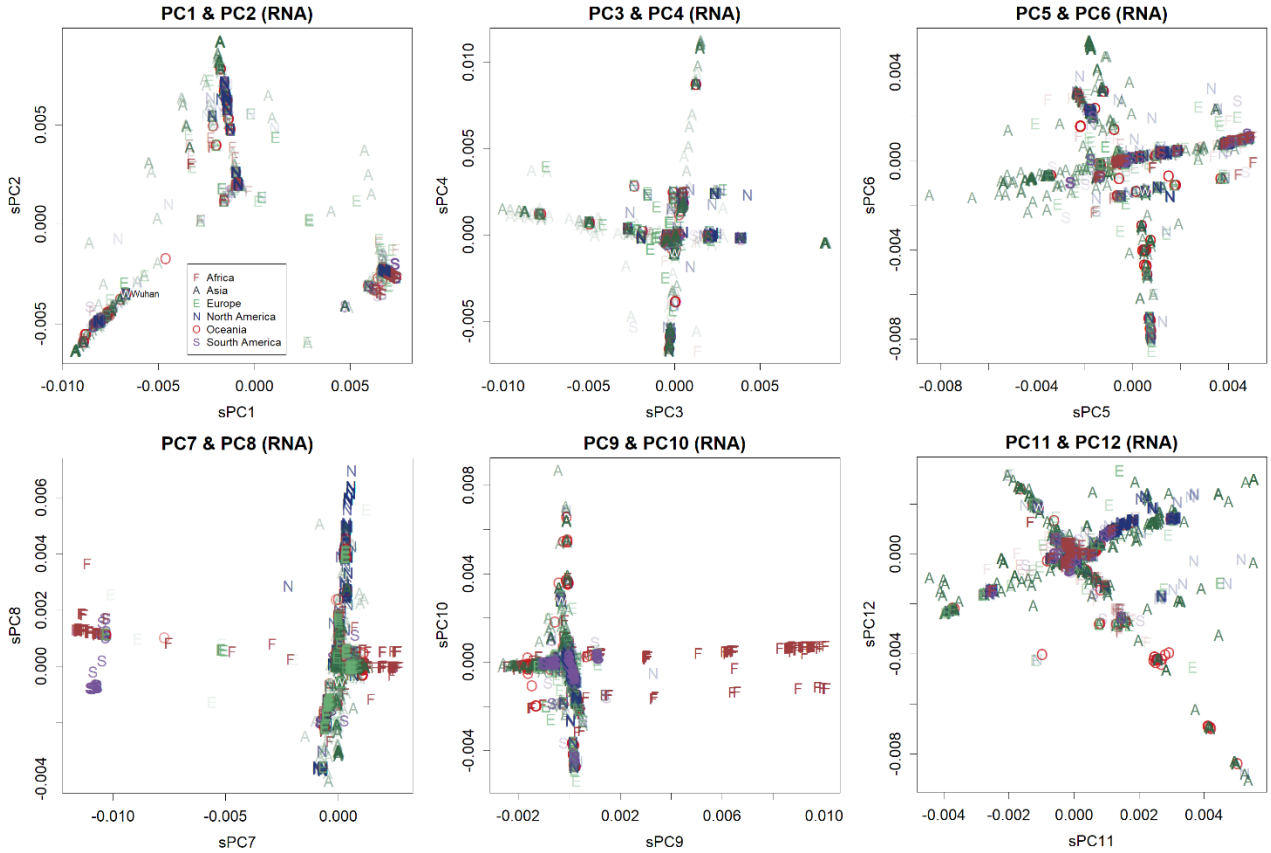
Extended Data Figure 1. Principal components of SARS-CoV-2. A thousand samples are selected and shown from each continent. Data are estimated from amino acid sequences.



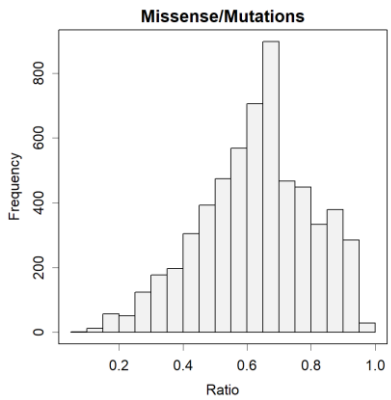
Extended Data Figure 2. Principal component analysis (PCA) contributions. **A.** RNA, **B.** Protein. The superiorities of PC1 and 2, which separate many samples, are clear. In fact, these are very high scores for 6,000 dimensions.



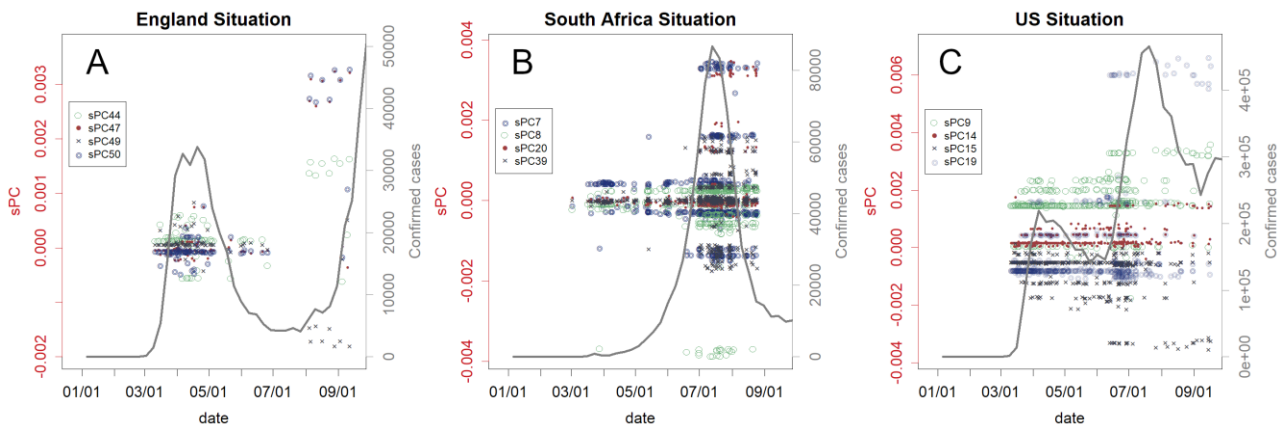
Extended Data Figure 3. Principal component analysis (PCA) of amino acid sequences. Samples collected until the end of March are presented. It is clear that PC1 and PC2 are complete, but the lower axes lack samples that are characteristic of each axis; those have not appeared yet. See Fig. 1 for the full set of amino acid sequences collected until September.



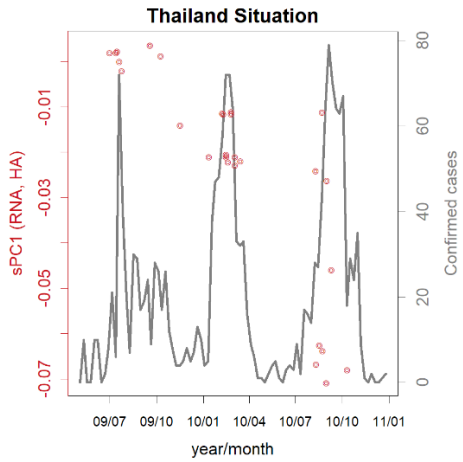
Extended Data Figure 4. Principal components of SARS-CoV-2. A thousand samples are selected and shown from each continent. Data are estimated from nucleotide sequences.



Extended Data Figure 5. The ratio of missense mutations found in SARS-CoV-2 compared with the average sequence. Ratios found in comparison between each sample and the average sequence data are the rate of differences that affect the amino acid sequence.



Extended Data Figure 6. Extension for Fig. 2. **A**. England, **B**. South Africa. **C**. USA.



Extended Data Figure 7. Pdm09 influenza in Thailand. The first peak of cases is much smaller than that in other countries. In the second peak, there are no amino acid changes; however, the rate of missense mutation increases in the third peak. The axes have been identified only within this season of the country; otherwise, these small differences will be neglected among the larger variations in the influenza virus.

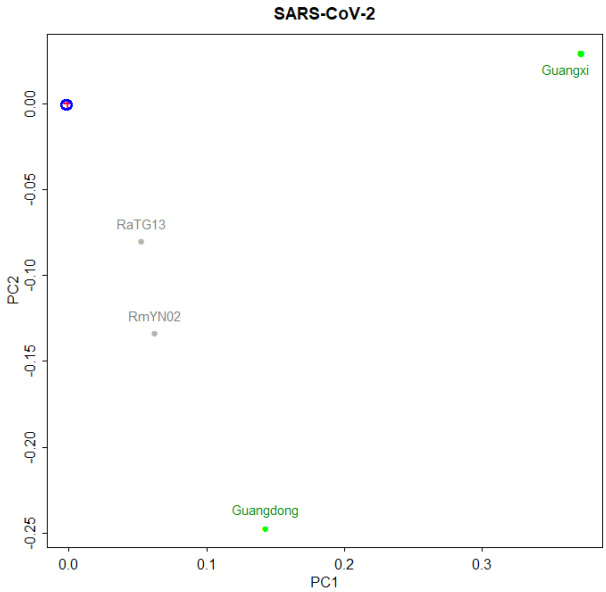
SARS-CoV-2	Example of Mutations
Australia (group 1), 4/1 - 8/20	ORF1ab 300, I/F; S 478, S/N; 1069, V/F
France (group 2), 6/5 - 9/20	ORF1ab 2702, H/Q; 3087, M/I; 4582, A/S; 5173, V/L; 5547, K/R; 5590, E/D; S 478, S/N; N 234, M/I; N 376, A/T
England (group 0), 3/30-8/15	ORF1ab 7325 A/V; N 221, A/V; ORF10 31 V/L
South Africa (group 1), 5/30-8/3	ORF1ab 2166, T/I; ORF3 125, R/S
USA (group 2), 3/30-8/15	ORF1ab 3352, L/D; 6060, N/D; 7020, R/C; S 173, G/V; ORF8 25, S/L; N 67, P/S; 200, P/L

Extended Data Table 1. Example of mutations. These are frequent mutations between the peaks. ORF1ab 2702, H/Q: the 2702th residue of ORF1ab protein that changed from H to Q.

Supplementary Information for Changes in SARS-CoV-2 before the peak in each country, producing unique variants

Relationship between sPC and changed number of bases

The principal component of a sample represents a part of the Euclidean distance from the mean sample, the square root of the sum of squares of the difference, $l = \sqrt{\sum(\text{sample} - \text{mean})^2/2}$, where "sample" is the Boolean vector of a sample and "mean" is the arithmetic mean [1]. The components were further scaled by nucleotide length owing to the generality of the values. Otherwise, because influenza virus and coronavirus are very different in terms of genome length, the results cannot be compared.



The correlation between base changes and the PCs is as follows. For example, in this figure, the Euclidean distance between humans (blue) and bat RaTG13 (grey) is 0.1. The figure shows the square root of the expected difference in each base (because the values are scaled). Therefore, the total difference was 0.1 ^ 2 × length = 280. This was only part of the total difference of 1100 because PCA identified many other axes and the distance was partitioned for other axes as well. If all differences that appear on all axes are summed, the total is 1100. The length of 280 is the difference that appeared in the two directions shown in the PC1 and PC2 axes.