Unsupervised cluster analysis of SARS-CoV-2 genomes reflects its geographic progression and identifies distinct genetic subgroups of SARS-CoV-2 virus

Georg Hahn[*], Sanghun Lee[*†], Scott T. Weiss[‡] and Christoph Lange[*]

Abstract

Over 10,000 viral genome sequences of the SARS-CoV-2 virus have been made readily available during the ongoing coronavirus pandemic since the initial genome sequence of the virus was released on the open access Virological website (`http://virological.org/`) early on January 11. We utilize the published data on the single stranded RNAs of $11,132$ SARS-CoV-2 patients in the GISAID (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017) database, which contains fully or partially sequenced SARS-CoV-2 samples from laboratories around the world. Among many important research questions which are currently being investigated, one aspect pertains to the genetic characterization/classification of the virus. Here, we analyze data on the nucleotide sequencing of the virus and geographic information of a subset of $2,540$ SARS-CoV-2 patients without missing entries that are available in the GISAID database. We apply principal component analysis to a similarity matrix that compares all pairs of the $2,540$ SARS-CoV-2 nucleotide sequences at all loci simultaneously, using the Jaccard index (Jaccard, 1901; Tan et al., 2005; Prokopenko et al., 2016; Schlauch et al., 2017). Our analysis results of the SARS-CoV-2 genome data illustrates the geographic progression of the virus, starting from the first cases that were observed in China to the current wave of cases in Europe and North America. We also observe that, based on their sequence data, the SARS-CoV-2 viruses cluster in distinct genetic subgroups. It is the subject of ongoing research to examine whether the genetic subgroup could be related to diseases outcome and its potential implications for vaccine development.

[*]Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA
[†]Department of Medical Consilience, Graduate School, Dankook University, South Korea
[‡]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA 02115

1 Introduction

Now, the outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has created a pandemic as it has spread swiftly from continent to continent, resulting in >2,300,000 infections, with approximately 5% mortality and a devastating effect on public health. Little is known about the rates of mutation in SARS-CoV-2, viral genetic variation is generated on the same timescale as virus transmission which allows us to track the spread of SARS-CoV-2. Phylogenetic and phylogeographic analysis conferred insights into distinct subtypes as the virus spreads through the global population and could be helpful in untangling complex virus transmission dynamics and understanding disease susceptibility (Mousavizadeh and Ghasemi, 2020).

One of the striking features of SARS-CoV-2 infection is the variable clinical features of the disease. In general, older persons seem to have greater severity and mortality (Zhou et al., 2020), as do males (Shi et al., 2020). However some younger subjects have had fatal disease as well (Zhou et al., 2020). Organ involvement also varies with respiratory complications being most frequent but cardiac and renal complications occurring in some other patients (Zhou et al., 2020). This degree of clinical variation may be due to environmental factors such as confinement to a nursing home or chronic care facility (McMichael et al., 2020), or genetics of either the host or the virus. While studies of host genetics are underway, we have chosen here to focus on genetic variation of the virus.

By mid April 2020, the GISAID (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017) database contained samples of the sequenced SARS-CoV-2 genome obtained from thousands of individuals infected with the human coronavirus disease. After routine cleaning and alignment of the data (see the *Methods* section for details), we were left with $n = 2,540$ samples having a trimmed length of $p = 28,988$ nucleotide sequences. In the principal component cluster analysis, we focused on the loci in the SARS-CoV-2 genome where differences to the SARS-CoV-2 reference sequence (in Hamming distance) are observed. The reference sequence is available on GISAID under the accession number EPI_ISL_412026. In particular, denoting with $X \in \mathbb{S}^{n\times p}$ the matrix of nucleotide sequences and letting $r \in \mathbb{S}^p$ be the reference sequence, where $\mathbb{S} = \{A, C, G, T\}$, we compute a Hamming matrix $Y \in \{0, 1\}^{n\times p}$ by defining $Y_{ij} = 1$ if and only if $X_{ij} = z_j$, and $Y_{ij} = 0$ otherwise (see the *Methods* section). The matrix Y indicates the mismatches to the reference sequence, and its row sums are the Hamming distance of each nucleotide sequence to the reference sequence.

As we are interested in the geographical/temporal change of the SARS-CoV-2 genome, we assessed the similarity between each pair of virus genomes with the Jaccard index (Jaccard, 1901; Tan et al., 2005). By computing the Jaccard index for each pair of samples based on all loci we constructed the Jaccard similarity matrix for the $2,540$ virus genomes (Prokopenko et al., 2016; Schlauch et al., 2017). The Jaccard similarity matrix can efficiently be computed across a large set of genomes and is a powerful tool to detect genetic substructures/clusters in such sets (Hahn et al., 2020b,a).

The article is structured as follows. Section 2 describes the preparation of the dataset including cleaning and alignment, as well as the Jaccard similarity matrices we compute. Clustering results for the SARS-CoV-2 genome samples are presented in Section 3. The article concludes with a discussion in Section 4. The appendix states the anchor nucleotide sequence we employ to align all samples.

2 Methods

Data acquisition All analyses are based on nucleotide sequence data downloaded from GISAID (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017) on 22 April 2020. Only complete sequences (defined as having a length ≥ 29000) were selected, resulting in 10976 sequences.

Data curation We removed sequences with reading errors in the *fasta* file (defined as letters other than A, C, G, T). Carrying out an exhaustive search for a reference sequence found the following string of length 300 which we used as an anchor to align all samples:

AATGTATACATTAAAAATGCAGACATTGTGGAAGAAGCTAAAAAGGTAAA
ACCAACAGTGGTTGTTAATGCAGCCAATGTTTACCTTAAACATGGAGGAG
GTGTTGCAGGAGCCTTAAATAAGGCTACTAACAATGCCATGCAAGTTGAA
TCTGATGATTACATAGCTACTAATGGACCACTTAAAGTGGGTGGTAGTTG
TGTTTTAAGCGGACACAATCTTGCTAAACACTGTCTTCATGTTGTCGGCC
CAAATGTTAACAAAGGTGAAGACATTCAACTTCTTAAGAGTGCTTATGAA

We grouped samples together depending on the location tag in their *fasta* file using the following nine categories: (1) China (all provinces apart from Wuhan), (2) Wuhan, (3) Korea, (4) Europe (all countries apart from Italy), (5) Italy, (6) Australia and New Zealand, (7) Southeast Asia (India, Pakistan, Nepal, Cambodia, Thailand, Vietnam), (8) North America (USA, Canada, Mexico), and (9) Japan. Selecting only samples falling into those regions, having a complete time stamp (year, month, date), and having an exact match of the above reference sequence resulted in $n = 2540$ samples.

Trimming and comparison to reference sequence After alignment of all n samples with the anchor, we trimmed sequences to the left and right in order to establish a sequence window in which all sequences had reads. This window comprised 2811 nucleotides to the left, and 25877 nucleotides to the right of the anchor, resulting in a sequence length of $p = 28988$.

We denote with $X \in \mathbb{S}^{n \times p}$ the matrix of nucleotide sequences. Letting $r \in \mathbb{S}^p$ be the reference sequence published on GISAID (accession number EPI_ISL_412026), where $\mathbb{S} = \{A, C, G, T\}$, we compute a Hamming matrix $Y \in \{0, 1\}^{n \times p}$ by defining $Y_{ij} = 1$ if and only if $X_{ij} = z_j$, and $Y_{ij} = 0$ otherwise. The matrix Y indicates the mismatches to the reference sequence, and its row sums are the Hamming distance of each nucleotide sequence to the reference sequence.

Measuring similarity We employ the locStra (Hahn et al., 2020b,a) R-package to calculate the Jaccard similarity matrix of Y, denoted as $J := Jac(Y) \in \mathbb{R}^{n \times n}$. The Jaccard matrix measures the similarity between all pairs of the n curated samples from GISAID. Computing the first ten eigenvectors of J (for plotting them in pairs) completes the numerical part of the analysis.

3 Results

Figure 1 displays the first two principal components of the Jaccard matrix for the $2,540$ analyzed SARS-CoV-2 genomes, labelled with the country information obtained from GISAID. We observe four distinct branches/clusters for both European and North American samples (Fig.1, left). The top branch contains most cases from the US, the branch below is a mixture of US and European cases and the two lower branches contain predominantly European cases. The predominantly European branch on the x-axis seems to be linked to the cluster at the origin, by genomes from Wuhan, Italy and Europe (Fig. 1, right).

Furthermore, we observe a dense cloud of points around the origin $(0, 0)$, which contains indistinguishable data points for genomes from all countries of the data set. When looking at the

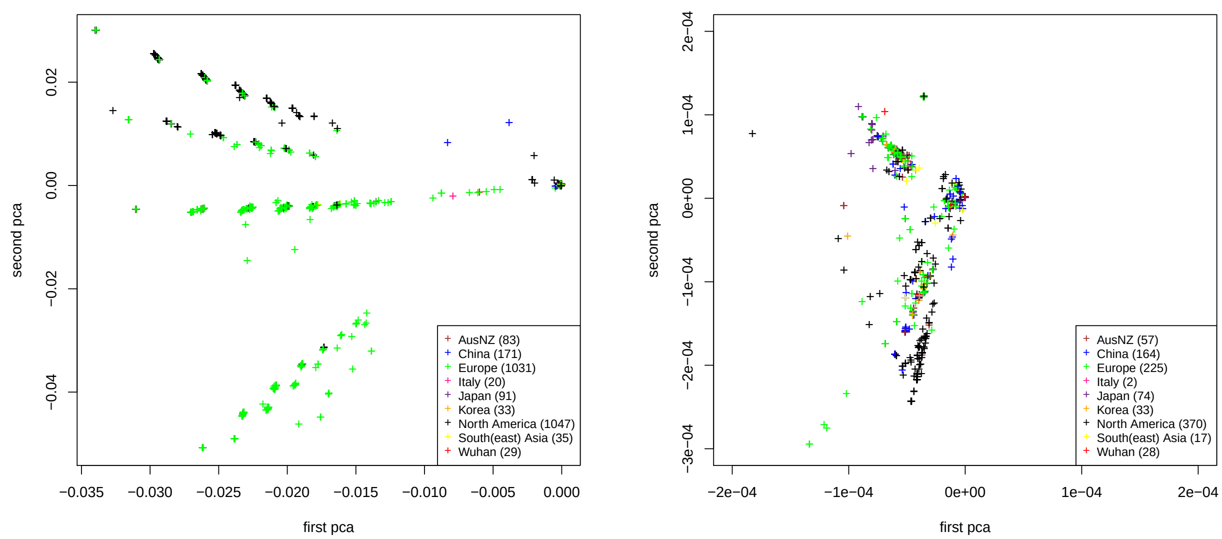

Figure 1: First two principal components of the Jaccard similarity matrix for the $2,540$ SARS-CoV-2 genomes by region/country. Entire dataset (left) and zoomed-in region around the origin $(0,0)$ (right). Here, we denote as "Europe" all genomes that are from cases in the Europe region with the exception of Italy, which is listed separately. Numbers in brackets for each country denote the number of SARS-CoV-2 genomes which are visible in each plot.

zoomed-in region in Fig. 1 (right), we observe that starting from the origin $(0,0)$, which has the highest density of genomes from China and Wuhan, a spatial spread-out is seen in which genomes from geographically close locations (Korea and South/Southeast Asia) cluster in the neighborhood of the origin. This "original" cluster contains also a relatively large number of virus genomes from the US and Europe, which could suggest early transmissions to these geographic regions. Almost all of the data points for the genomes from Wuhan and China are part of this cluster around the origin $(0,0)$, reflecting where SARS-CoV-2 was first reported. Above the cluster at the origin $(0,0)$, a second cluster of genomes is observed, where most of the genomes are from Europe, Japan and Australia/New Zealand.

For a better understanding of the distinct clustering of the European and US samples into the four observed subbranches (Fig. 1, left), we examine the genomic locations where the four subgroups differ. In particular, we are interested in how SARS-CoV-2 samples belonging to those four branches compare to the SARS-CoV-2 reference sequence. For the reference genome subdivided into 50 consecutive bins, Fig. 2 shows the normalized numbers of mismatches with respect to the trimmed reference sequence among all samples from the four branches of European and American

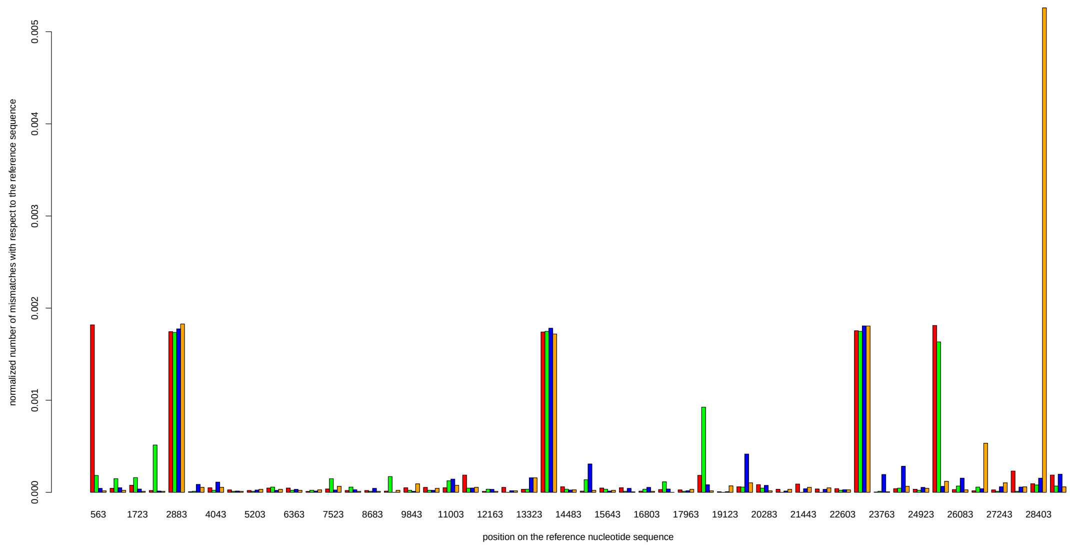

Figure 2: The x-axis shows the roughly 29,000 nucleotides of the trimmed SARS-CoV-2 reference sequence in 50 bins. The y-axis shows per bin the normalized number of mismatches (with respect to the reference sequence) among the samples in the European and North American population, stratified into samples from the top branch (red), the second branch (green), the middle branch (blue), and the bottom branch (orange) visible in the left panel of Fig. 1. The normalization is done with respect to both the bin size and the number of samples in each of the four branches.

samples. The number of mismatches is normalized by both the bin size and the number of samples in each branch. We observe that, besides of the (outlier) mismatches at the beginning and at the end of the virus genome, the samples of the four distinct branches differ from the reference genome at the same four genomic regions, and by roughly similar frequencies. The only exception is the fourth branch at approximately position 25000, where the first two branches are similar, and the third and fourth branches are similar.

4 Discussion

Our analysis demonstrates that the genome of the virus varies by geographic region, with different viral sequences present in Asia vs Europe and the US. The results also suggest that there are four distinct genetic subgroups in Europe and the US.

The viral sequences that were initially observed in Asia seem to be more homogeneous. The genetic sequences of viruses from Europe and the US are more diverse when compared to the sequences from countries in Asia.

It is important to consider the methodological aspects of our analysis to describe the viral genetic diversity of SARS-CoV-2. We used a direct, unsupervised approach to compare the entire viral genomes, i.e. principle components analysis based on Jaccard similarity matrices. We did not include information about the geographic origin of the samples in the analysis nor did we attempt to directly model the evolutionary relationship of the different SARS-CoV-2 genomes, e.g. via phylogenetic analysis. Nevertheless, our analysis results reflect clearly the chronological spread of SARS-CoV-2 around the globe. The advantages of our methodology compared to existing network analyses are twofold: First, our unsupervised approach is capable of recovering geographic subgroups and geographic progression without any (spacial) information, and it takes all loci into account simultaneously. Second, it is particularly simple as it only involves the Hamming matrix and its Jaccard representation, both of which can be efficiently computed using binary operations only (Hahn et al., 2020a).

Our analysis results have two potentially important implications. It is the subject of ongoing research to understand whether the four genetic subgroups of the SARS-CoV-2 viruses have different clinical implications in terms of disease progression and disease outcome. Furthermore, our results suggest that vaccine creation will have to take into account the genetic variability in viral sequence described here. As the virus has spread over time, its genetic diversity has increased and is likely to increase even further. To combat the virus successfully, it will be fundamental to understand the consequences of this development on the features of the virus.

Acknowledgements

The authors gratefully acknowledge the contributors, originating and submitting laboratories of the sequences from GISAID's EpiCoV™ Database (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017) on which this research is based. A detailed list of contributors is available in the Supplementary Information.

Data Availability Statement

Sequence data that support the findings of this study are deposited in the GISAID database with accession numbers in the range of EPI_ISL_402119 to EPI_ISL_426289 (`https://www.gisaid.org/`).

Funding

The initial methodology work for this paper was funded by Cure Alzheimer's Fund; Funding for this research was provided through the National Human Genome Research Institute [R01HG008976]; and the National Heart, Lung, and Blood Institute [U01HL089856, U01HL089897, P01HL120839, P01HL132825].

References

Elbe, S. and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(33–46).

Hahn, G., Lutz, S., Hecker, J., Prokopenko, D., Cho, M., Silverman, E., Weiss, S., and Lange, C. (2020a). locstra: Fast analysis of regional/global stratification in whole genome sequencing (wgs) studies. Preprint at bioRxiv:2020.03.06.981050.

Hahn, G., Lutz, S., and Lange, C. (2020b). locStra: Fast Implementation of (Local) Population Stratification Methods (v1.3). `https://cran.r-project.org/web/packages/locStra/index.html`.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaud. Des. Sci. Nat.*, 37:547–579.

McMichael, T., Clark, S., Pogosjans, S., Kay, M., Lewis, J., Baer, A., Kawakami, V., Lukoff, M., Ferro, J., Brostrom-Smith, C., Riedo, F., Russell, D., Hiatt, B., Montgomery, P., Rao, A., Currie, D., Chow, E., Tobolowsky, F., Bardossy, A., Oakley, L., Jacobs, J., Schwartz, N., Stone, N., Reddy, S., Jernigan, J., Honein, M., Clark, T., and Duchin, J. (2020). COVID-19 in a Long-Term Care Facility – King County, Washington, February 27–March 9, 2020. *MMWR Morb Mortal Wkly Rep*, 69(12):339–342.

Mousavizadeh, L. and Ghasemi, S. (2020). Genotype and phenotype of COVID-19: Their roles in pathogenesis. *Journal of Microbiology, Immunology and Infection*, pages 1–5.

Prokopenko, D., Hecker, J., Silverman, E., Pagano, M., Nöthen, M., Dina, C., Lange, C., and Fier, H. (2016). Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics*, 32(9):1366–1372.

Schlauch, D., Fier, H., and Lange, C. (2017). Identification of genetic outliers due to sub-structure and cryptic relationships. *Bioinformatics*, 33(13):1972–1979.

Shi, Y., Yu, X., Zhao, H., Wang, H., Zhao, R., and Sheng, J. (2020). Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. *Crit Care*, 24(1):108.

Shu, Y. and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *EuroSurveillance*, 22(13).

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Pearson; 1st Edition.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., and Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*, 395(10229):1054–1062.