# Spatiotemporal Tracking of SARS-CoV-2 Variants using informative subtype markers and association graphs

Ananya Sen Gupta
*Department of Electrical and Computer Engineering*
*University of Iowa*
Iowa City, IA, USA
ananya-sengupta@uiowa.edu

Zhengqiao Zhao and Gail Rosen
*Department of Electrical and Computer Engineering*
*Drexel University*
Philadelphia, PA, USA
zz374@drexel.edu, glr26@drexel.edu

2020 54th Asilomar Conference on Signals, Systems, and Computers | 978-0-7381-3126-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEEECONF51394.2020.9443496

***Abstract*—Viral subtyping can facilitate visualization and modeling of the geographic distribution and temporal dynamics of disease spread. Understanding the virus's evolution spatiotemporally can help forensic strategies. We have identified mutation variation within SARS-CoV-2 sequences via an entropy measure followed by frequency analysis. These signatures, Informative Subtype Markers (ISMs), define a compact set of nucleotide sites that characterize the most variable (and thus most informative) positions in the viral genomes sequenced from different individuals. Using these ISMs, we show that we can use them for a variety of downstream analyses, such as comparing countries' subtype compositions. We present association graphs as a visualization tool to connect different ISMs based on their co-occurrence across different individuals. In particular, we investigate dominant ISMs for different locations, across different factors such as gender and age.**

***Index Terms*—Bioinformatics, Viral Genomics, Association Graphs, Entropy Measures**

## I. Introduction

The novel coronavirus responsible for COVID-19, SARS-CoV-2, has led to over 50 million confirmed cases worldwide and well on its way to 100 million cases by next year if no intervention is taken. The global nature of the pandemic greatly needs tehcniques to track viral transmission dynamics in real-time. Only a small fraction of the viral samples are sequenced and deposited in the GISAID database,a global science initiative and primary source that provides open-access to genomic data of influenza- and corona-viruses [1]. Since the virus readily mutates, each sequence of an infected individual contains useful information linked to the individual's exposure location and sample date. But, there are over 30,000 bases in the full SARS-CoV-2 genome—so tracking genetic variants on a whole-sequence basis becomes difficult. We use an entropy-based method [2] to produce compact representation, a seventeen base-long compressed label, called an Informative Subtype Marker or "ISM". In this work, we aim to show how regional and temporal distributions of subtypes track the progress of the pandemic. Using the ISMs with association graphs, we provide a quantitative visualization of the relative co-occurrence between different ISM pairs.

## II. Related Work

### A. Tracing SARS-CoV-2's evolving lineages

The Nextstrain group has created a massive phylogenetic tree incorporating sequence data and applied a model of the time-based rate of mutation to create a hypothetical map of viral distribution [3] (available at https://nextstrain.org/ncov). Similarly, the China National Center for Bioinformation has established a "2019 Novel Coronavirus Resource", which includes a clickable world map that links to a listing of sequences along with similarity scores based on alignment (available at https://bigd.big.ac.cn/ncov?lang=en) [4]. Extensive phylogenetic analysis has revealed variations in the genome such as L (70%) and S (30%) clusters as well as A, B, and C clusters [5], [6]. In [6], a promising network graph is used to develop these inferences. However, using the entire genome of the sequence has been shown to have instabilities [7]. Therefore, techniques that are compressed representations that take into account redundancies should be developed to help the forensic tracking of SARS-CoV-2 lineages.

### B. Association Graphs

Association graphs have been traditionally used as a knowledge discovery tool [8] to discover and visualize association rules between variables, known and latent, in high-dimensional and large-scale data analysis. More recently, the idea of using graph-based visualization of associations between microbes in the microbiome was introduced in [9], and a graph-based visualization of chemical contaminants that co-occur in the raw instrument signal was introduced in [10]. In this work, we build upon the visualization in [9], [10] to introduce the idea of quantifying co-occurrence associations between ISMs in a graph setting. The key motivation is to discover ISMs that co-occur across different subject groups to identify dominant ISMs in different locations and time, and then present these ISM associations in a graph-setting for expert interpretation.

Accordingly, in Figure 1 we present the schematic diagram of an association graph for ISMs.

We note that each ISM in Figure 1 is denoted by a unique vertex with a numerical index. Two vertices are joined by

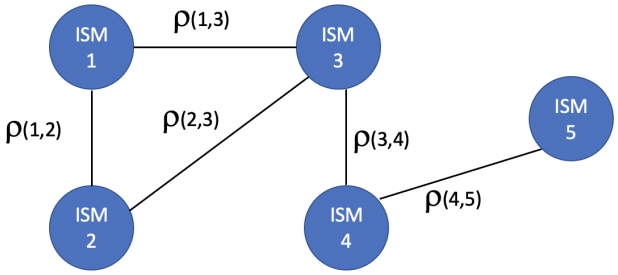


Fig. 1. Schematic diagram of an association graph between six ISMs.

an edge if they co-occur across a given data set. An edge between two connected ISMs with vertex indices $m$ and $n$ has a metric, henceforth called the "association metric", and denoted as $\rho(m, n)$.

The association metric denotes the relative co-occurrence of the ISMs across the data set, and as such this co-occurrence can be measured in multiple ways. In this work, we consider the relative ratio-based metric in Equation (1), adopted with modifications from [11]. In principle, it is similar though not identical in mathematical formulation to relative abundance (i.e., frequency), a popular metric in the bioinformatics literature, also presented for reference in this work. Mathematically, in a dataset of $P$ subjects, if $ISM_m$ is detected $M$ times, and $ISM_n$ is detected $N$ times, then $\rho(m, n)$ is given by:

$$\rho(m, n) = \min\left(\frac{M}{N}, \frac{N}{M}\right) \tag{1}$$

By design, the association metric in Equation (1) is commutative, i.e., $\rho(m, n) = \rho(n, m)$, with highest possible value of $\rho(m, n) = 1$ when $M = N$.

While any two ISMs will always have an association metric, we can filter an association graph to isolate vertices that share high (or low) values of $\rho(\cdot)$.

## III. Methods

### A. Preprocessing step: Forming the ISMs

The ISM procedure is to align the sequences using MAFFT [12]. Then, we calculated the entropy at a given position $i$ by:

$$H(i) = -\sum_{k \in L} p_k(i) * log_2(p_k(i)) \tag{2}$$

where $L$ is a list of unique characters in all sequences and $p_k(i)$ is a probability of observing a character $k$ at position $i$. We estimated $p_k(i)$ from the frequency of characters at that position. The result is that we get 30,000 entropies as seen in Fig. 2. We then select sites that have more than 0.2 entropy and mask out ambiguous bases, which is described more in [2].

### B. Associations between ISMs

Association graphs provide a visual tool to represent the relative co-occurrence of two or more ISM sequences. Figure 1 shows the schematic diagram. We created associations by direct string comparisons between two ISMs.

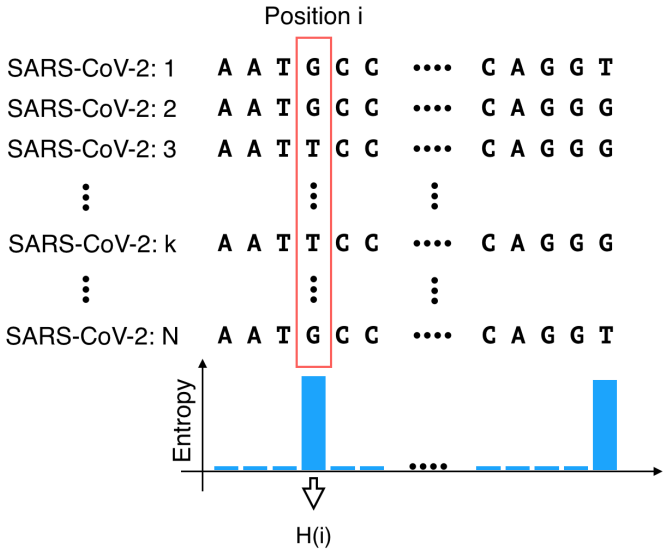


Fig. 2. Variable sites that have a high entropy are chosen to be indicated in the ISM representation.

## IV. Results

We demonstrate that ISMs can be used to show spatiotemporal trends in SARS-CoV-2 data. By simple frequency analysis, we are preliminarily able to observe that countries in similar parts of the world tend to have similar subtypes. We also observe that subtype frequencies in a given location come in waves. However, this analysis does not tell the entire story. We wish to investigate if one subtype was "transmitted" from one country to another and how this progressed over time. In this regard, association graphs allow a simple graph-based visualization of relative co-occurrence between ISM sequences and allow us to track the vertices that cluster together across time, location and other variables.

### A. Preliminary Plots of Geographic and Temporal Relative Abundances

At the country/region level, we assess the geographic distribution of SARS-CoV-2 subtypes, and, in turn, we count the frequency of unique ISMs per location. The ISM pipeline creates pie charts for different locations to show the geographical distribution of subtypes. Fig. 3 show the distributions of ISMs per country. Each subtype is also labeled with the earliest date associated with sequences from a given location in the dataset. ISMs with less than 5% abundance are plotted as "OTHER".

To study the progression of SARS-CoV-2 viral subtypes in the time domain, we group all sequences in a given location that were obtained no later than a certain date (as provided in the sequence metadata) together and compute the relative abundance (i.e., frequency) of corresponding subtypes. Any subtypes with a relative abundance that never goes above 2.5% for any date are collapsed into "OTHER" category per location. Fig. 4 shows the ISM relative abundance over time in the USA.

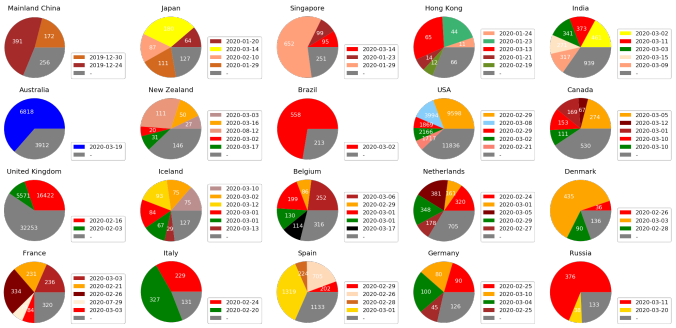


Fig. 3. Distribution of ISM subtypes in various countries as of mid-October 2020. exact ISM sequences can be seen at [2]. We can see that East Asia has a different composition from North American and Central Europe.

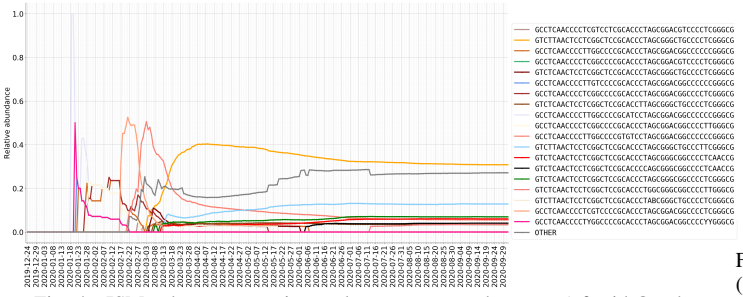


Fig. 4. ISM subtypes over time, where waves can be seen. (of mid-October 2020). At the beginning, newly introduced ISMs became highly abundant as previous outbreaks were squashed. However, now two subtypes seem to persist.

## B. Associations

In this preliminary work, we introduce the concept of association graphs, and given practical limitations, we only provide the bar-graphs of the relative distributions of dominant ISMs for different groups of subjects taken from the GISAID database. Figures 5-8 provide the results over diverse subject groups across a range of age, location and gender from the GISAID database. From these types of graphs, we can see that there is more diversity in England of viral subtypes (from the uniform spread) than in North America.

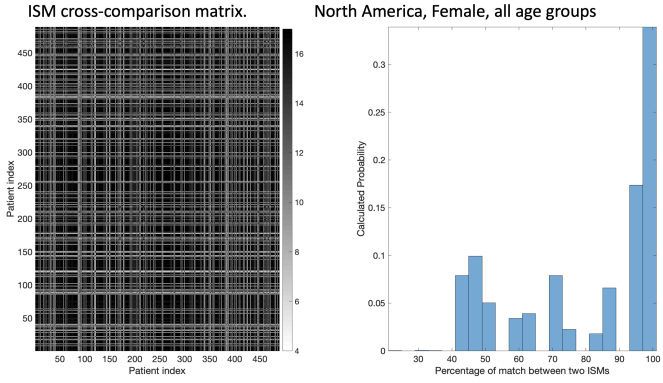


Fig. 5. Measuring associations between ISM sequences for North American females, all age groups

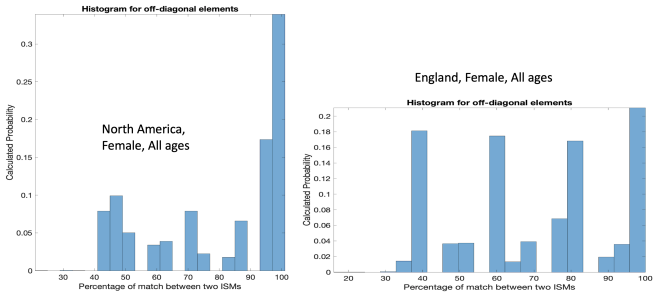


Fig. 6. Association profile variability between female subjects, across all age groups, between North American and England.

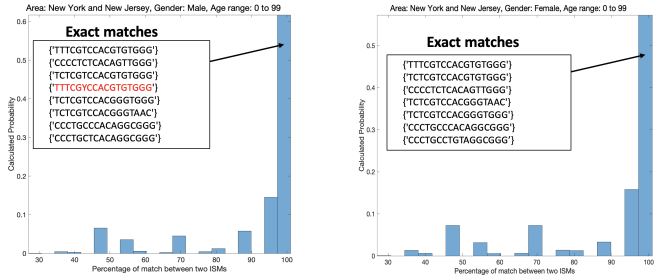


Fig. 7. Association profile variability based on gender, for a particular region (New York and New Jersey), for all age groups. ISM sequences marked in red are distinct between male and female, i.e, their association metrics are either extremely low or zero.

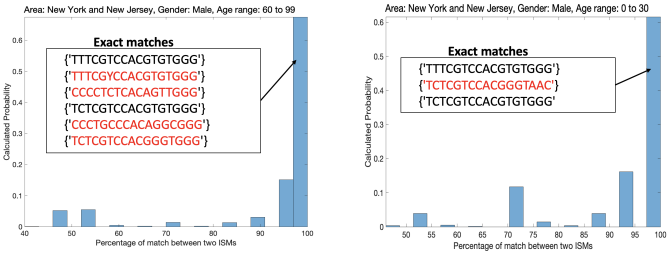


Fig. 8. Association profile variability based on age group, for all male subjects within a given region (New York and New Jersey). ISM sequences marked in red are distinct between the two age classes, i.e, their association metrics are either extremely low or zero.

## V. Conclusion

We use an entropy-based method [2] to produce a seventeen base-long compressed label, called an Informative Subtype Marker or "ISM", to create a compact representation. Based on GISAID database, we demonstrate how regional and temporal distributions of subtypes track the progress of the pandemic. Using the ISMs with association graphs and related association variability charts, we provide a quantitative visualization of the relative co-occurrence between different ISM pairs.

## Acknowledgments

We downloaded all SARS-Cov-2 sequences available from and acknowledge the contributions of the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database, which has made accessible novel coronavirus sequencing data, including from the NIH Genbank resource [1]. We would also

like to acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu Database on which this research is based, as well as all future SARS-CoV-2 sequence contributors in GISAID's EpiFlu Database. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) [13], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). This work was supported by NSF awards #1919691 and #1936791.

## REFERENCES

[1] Y. Shu and J. McCauley, "Gisaid: Global initiative on sharing all influenza data – from vision to reality," *Eurosurveillance*, vol. 22, no. 13, 2017. [Online]. Available: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.13.30494

[2] Z. Zhao, B. A. Sokhansanj, C. Malhotra, K. Zheng, and G. L. Rosen, "Genetic grouping of sars-cov-2 coronavirus sequences using informative subtype markers for pandemic spread visualization," *PLOS Computational Biology*, vol. 16, no. 9, pp. 1–32, 09 2020. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1008269

[3] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121–4123, 05 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty407

[4] W. Zhao, S. Song, M. Chen, D. Zou, L. Ma, Y.-K. Ma, R. Li, L. Hao, C. Li, D. Tian, B. Tang, Y.-Q. Wang, J. Zhu, H. Chen, Z. Zhang, Y. Xue, and Y. Bào, "The 2019 novel coronavirus resource." *Yi chuan = Hereditas*, vol. 42 2, pp. 212–221, 2020.

[5] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian, J. Cui, and J. Lu, "On the origin and continuing evolution of SARS-CoV-2," *National Science Review*, vol. 7, no. 6, pp. 1012–1023, 03 2020. [Online]. Available: https://doi.org/10.1093/nsr/nwaa036

[6] P. Forster, L. Forster, C. Renfrew, and M. Forster, "Phylogenetic network analysis of sars-cov-2 genomes," *Proceedings of the National Academy of Sciences*, vol. 117, no. 17, pp. 9241–9243, 2020. [Online]. Available: https://www.pnas.org/content/117/17/9241

[7] Y. Turakhia, N. De Maio, B. Thornlow, L. Gozashti, R. Lanfear, C. R. Walker, A. S. Hinrichs, J. D. Fernandes, R. Borges, G. Slodkowicz, L. Weilguny, D. Haussler, N. Goldman, and R. Corbett-Detig, "Stability of sars-cov-2 phylogenies," *PLOS Genetics*, vol. 16, no. 11, pp. 1–34, 11 2020. [Online]. Available: https://doi.org/10.1371/journal.pgen.1009175

[8] Show-Jane Yen and A. L. P. Chen, "A graph-based approach for discovering various types of association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 5, pp. 839–845, 2001.

[9] C. M. Cullen, K. K. Aneja, S. Beyhan, C. E. Cho, S. Woloszynek, M. Convertino, S. J. McCoy, Y. Zhang, M. Z. Anderson, D. Alvarez-Ponce *et al.*, "Emerging priorities for microbiome research," *Frontiers in Microbiology*, vol. 11, p. 136, 2020.

[10] R. A. McCarthy, A. S. Gupta, B. Kubicek, A. M. Awad, A. Martinez, R. F. Marek, and K. C. Hornbuckle, "Signal processing methods to interpret polychlorinated biphenyls in airborne samples," *IEEE access*, vol. 8, pp. 147 738–147 755, 2020.

[11] H. G. Damavandi, A. S. Gupta, R. K. Nelson, and C. M. Reddy, "Interpreting comprehensive two-dimensional gas chromatography using peak topography maps with application to petroleum forensics," *Chemistry Central Journal*, vol. 10, no. 1, p. 75, 2016.

[12] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 01 2013. [Online]. Available: https://doi.org/10.1093/molbev/mst010

[13] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, "Xsede: Accelerating scientific discovery," *Computing in Science Engineering*, vol. 16, no. 5, pp. 62–74, Sep. 2014.