# Data-driven whole-genome clustering to detect geospatial, temporal, and functional trends in SARS-CoV-2 evolution


Jean Merlet*
merletjj@ornl.gov
The University of Tennessee
Knoxville, Tennessee, USA

John Lagergren*
lagergrenjh@ornl.gov
Biosciences Division, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

Verónica G. Melesse Vergara
vergaravg@ornl.gov
National Center for Computational Sciences, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

Mikaela Cashman
mcashman@lbl.gov
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory
Berkeley, California, USA

Christopher Bradburne
chris.bradburne@jhuapl.edu
Johns Hopkins University Applied Physics Laboratory
Laurel, Maryland, USA

Raina Plowright
rkp57@cornell.edu
Cornell University
Ithaca, New York, USA

Emily Gurley
egurley1@jhu.edu
Johns Hopkins Bloomberg School of Public Health
Baltimore, Maryland, USA

Wayne Joubert
joubert@ornl.gov
National Center for Computational Sciences, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

Daniel Jacobson
jacobsonda@ornl.gov
Biosciences Division, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA



## ABSTRACT

Current methods for defining SARS-CoV-2 lineages ignore the vast majority of the SARS-CoV-2 genome. We develop and apply an exhaustive vector comparison method that directly compares all known SARS-CoV-2 genome sequences to produce novel lineage classifications. We utilize data-driven models that (i) accurately capture the complex interactions across the set of all known SARS-CoV-2 genomes, (ii) scale to leadership-class computing systems, and (iii) enable tracking how such strains evolve geospatially over time. We show that during the height of the original Omicron surge, countries across Europe, Asia, and the Americas had a spatially asynchronous distribution of Omicron sub-strains. Moreover, neighboring countries were often dominated by either different clusters of the same variant or different variants altogether throughout the pandemic. Analyses of this kind may suggest a different pattern of epidemiological risk than was understood from conventional data, as well as produce actionable insights and transform our ability to prepare for and respond to current and future biological threats.



## CCS CONCEPTS

• **Applied computing** → **Biological networks**; • **Computing methodologies** → **Massively parallel algorithms**.

## KEYWORDS

biological networks, high performance computing, SARS-CoV-2

**ACM Reference Format:**
Jean Merlet, John Lagergren, Verónica G. Melesse Vergara, Mikaela Cashman, Christopher Bradburne, Raina Plowright, Emily Gurley, Wayne Joubert, and Daniel Jacobson. 2023. Data-driven whole-genome clustering to detect geospatial, temporal, and functional trends in SARS-CoV-2 evolution. In *Platform for Advanced Scientific Computing Conference (PASC '23), June 26–28, 2023, Davos, Switzerland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3592979.3593425


## 1 INTRODUCTION

The ongoing COVID-19 pandemic continues to cause considerable mortality worldwide. Mutations in the SARS-CoV-2 genome have been observed throughout the course of the pandemic, and multiple variants (strains) of concern have been identified, some of which exhibit more efficient transmission and higher pathogenicity compared to others. Global efforts in data collection have resulted in more than 15 million geospatially and temporally tagged SARS-CoV-2 genome sequences that have been uploaded to the Global Initiative on Sharing Avian Influenza Data (GISAID) database [8].

Current strategies for classifying genetic lineages of SARS-CoV-2 use the Phylogenetic Assignment of Named Global Outbreak Lineages (PANGO lineage) algorithm [1, 15, 16], which depends on manually curated lineage designations based on a relatively small set of mutations, which may ignore important mechanisms

*Both authors contributed equally to this research.




*PASC '23, Davos, Switzerland,*





ACM ISBN 979-8-4007-0190-0/23/06…$15.00
https://doi.org/10.1145/3592979.3593425

that contribute to infectivity and mortality. Thus, there exists a need to develop a suite of flexible models that leverages the set of all available SARS-CoV-2 sequences to derive whole-genome classifications in a data-driven approach.

To address these challenges, we apply a novel binary encoding to the set of SARS-CoV-2 genome vectors in order to enable efficient all-against-all exhaustive vector comparisons to the population. We leverage the publicly available Combinatorial Metrics (CoMet) codebase, previously used in comparative genomics studies [10, 11], with a modified binary vector similarity metric to compare each SARS-CoV-2 genome vector against all others. We store the results of these comparisons as a sparse network, in which nodes are genome sequences and edges are defined by sequence similarity (i.e., an edge between two genomes exists if their sequences exceed a similarity threshold). These methods scale to leadership-class computing systems and are demonstrated using the Oak Ridge Leadership Computing Facility (OLCF) supercompter, Summit. Then, to discover novel data-driven SARS-CoV-2 lineages, we apply Markov clustering, an unsupervised graph clustering algorithm, to the sequence similarity network to group sets of genomes according to their sequence similarity.

## 2 BACKGROUND

### 2.1 SARS-CoV-2 genome sequences

GISAID is a global database for genome sequences established in 2008 that provides open access to genomic data of influenza viruses [8]. In light of the COVID-19 pandemic, it was co-opted to store SARS-CoV-2 sequences and has become the world's largest such repository, holding more than 15.49 million SARS-CoV-2 samples as of May 2023. While GISAID accepts submissions from anywhere, African, Asian, and South American countries generally sequence and deposit an order of magnitude fewer positive COVID-19 samples compared to European, North American, and Oceania countries [12]. Additionally, both testing rates and subsequent sequencing rates per positive test vary considerably from country to country and over time [9]. Most SARS-CoV-2 sequences in the database also contain associated metadata that provides sample collection date and location, as well as a World Health Organization (WHO) variant classification (e.g., Alpha and Delta variants). In this work, we include the 11 million SARS-CoV-2 sequences available through June 2, 2022.

### 2.2 Vector comparison and network analysis

The CoMet application leverages ultra-low precision to exhaustively compare vectors at extreme speed and efficiency, which requires that input vectors (i.e., SARS-CoV-2 genome sequences) are encoded in binary format. There is no standard for the binarization process, and each dataset and application requires its own formulation based on the goals of the vector comparison and downstream analyses. Previous studies have used the CoMet application to conduct exhaustive vector comparisons for genomic and environmental applications [10, 11, 13]. In particular, CoMet has been optimized to run efficiently on the Oak Ridge Leadership Computing Facility's (OLCF) leadership-class systems, including the Summit supercomputer and the OLCF's new exascale supercomputer, Frontier, as well as the JUWELS Booster at Jülich Supercomputing Centre and Perlmutter at the National Energy Research Scientific Computing Center (NERSC). All implementations leverage GPU-accelerated computing and massive parallelism to enable large-scale vector comparisons at extreme scales. CoMet exploits these features to compute similarities between pairs or triplets of binary vectors at record-breaking (exascale) speeds [10, 11, 13], which enables direct comparisons across millions of samples. To enable these accelerations for studying SARS-CoV-2 genomes, we specify our vector binarization scheme in Section 3.2. Here, our all-against-all comparison of SARS-CoV-2 genomes results in approximately $3 \times 10^{13}$ vector comparisons.

The CoMet application supports a number of binary similarity metrics (Duo, CCC, etc.) [5, 6], and the choice of metric and metric parameterization, in tandem with the vector binarization scheme, are important considerations. In this work, we modify the Duo metric to compute the binary set similarity between any two sequences [5]. Duo is a similarity metric that is used to compute the frequency of co-occurring binary elements in each vector pair. Duo is used to group matrix values into high (pairs of 1's), low (pairs of 0's), and anti-correlated (pairs of 10 or 01) categories, which can be modified depending on the domain application. The Duo metric is represented by the equation

$$\text{Duo}_{ij} = mD_{ij}(1 - qf_i)(1 - qf_j) \tag{1}$$

where $i$ and $j$ are binary vectors, $D$ computes the frequency of a specified binary relationship (i.e., 11, 10, 01, or 00) between $i$ and $j$, $f$ computes the frequency of binary elements in vector $i$ or $j$, and $m = 4$ and $q = 2/3$ are scaling constants. In practice, the more elements that exhibit the specified relationship in a pair of vectors, the higher the Duo score. This metric results in four values for each pair of input vectors based on the different bit comparisons: high-high, high-low, low-high, and low-low, with each value computing the frequency of all such comparisons across the two vectors. A similarity threshold is then used to store pairs of vectors only if their Duo score exceeds the set threshold. This threshold is determined by accounting for both storage space constraints of the filesystem of the supercomputer being used as well as post-processing and downstream computational requirements. In large comparisons, the full (non-thresholded) output would require petabytes of storage as well as tens of thousands of node-hours of post-processing time. Thus, thresholds are applied and a sparse network structure emerges, in which nodes are represented by vectors and edges are defined between pairs of nodes if their Duo similarity exceeds the threshold.

To group sets of nodes into mechanistically similar clusters, we apply the high-performance Markov clustering (HipMCL) algorithm for large-scale unsupervised network clustering [2, 7]. Similar to CoMet, the open-source HipMCL application leverages massive parallelism for 1000-fold faster graph clustering compared to the original Markov Clustering algorithm, and has been successfully applied on networks with millions of nodes and billions of edges on the Summit supercomputer [2]. By applying HipMCL to the sequence similarity networks in this work, novel clusters (lineage classifications) of SARS-CoV-2 genomes emerge that leverage whole-genome sequence similarity and provide new insights into the spatiotemporal dynamics and epidemiological implications of the evolving COVID-19 pandemic.

### 2.3 Summit supercomputer

The Summit supercomputer contains 4,608 compute nodes each consisting of 2 IBM POWER9 CPUs, 6 NVIDIA Volta V100 GPUs, and 512GB of DDR4 RAM + 96GB of HBM2 DRAM. The 22-core POWER9 CPU has a 3.2GHz base frequency with 3.8GHz turbo, 90 Watt TDP, 32KB L1 cache, 512KB L2 cache/core, and 10MB L3 cache/core. The NVIDIA Volta V100 GPU has 640 Tensor Cores and 5,120 NVIDIA CUDA cores, 1134 GB/sec memory bandwidth, 8.2 TFLOPS at double precision, 16.4 TFLOPS at single precision, and 130 TFLOPS of Tensor Performance. Each Summit compute node achieves 42 teraflops of performance, totalling 200 petaflops of peak performance.

## 3 METHODS

### 3.1 Data pre-processing

All sequences available on the GISAID platform through June 2nd, 2022 were acquired using GISAID's '`high coverage`' (< 1% unassigned nucleotides) and '`collection date compl`' (has a sample collection date) filters, resulting in approximately 11 million SARS-CoV-2 sequences. Further, GISAID's metadata was used to exclude sequences with non-human hosts and any sequences with missing collection dates and locations. Each sequence was then aligned to the Wuhan reference SARS-CoV-2 genome (NCBI RefSeq `NC_045512.2`) with the Multiple Sequence Alignment using Fast Fourier Transform (MAFFT) software package (version 7) with settings `–auto` (selects an appropriate strategy from L-INS-i, FFT-NS-i, or FFT-NS-2 algorithms, according to data size), `–addfragments` (adds unaligned fragmentary sequence(s) into an existing alignment), and `–keeplength` (keeps alignment length unchanged).

After alignment and to account for missing data and sequencing errors, any genome sequences that contained invalid nucleotide codes, greater than 1000 deletions, or greater than 1% unassigned nucleotides (nucleotide code 'N') were removed from consideration. Additionally, nucleotides to the left of the first Artic primer overlap position and to the right of the last Artic primer overlap position were deleted from all genome vectors, as they contained a disproportionately high amount of Ns compared to the rest of the genomes. Following this, the total number of mutations to each of 'A', 'C', 'G', 'T', and '-' (i.e., deletion) were summed across all sequences at each nucleotide position in the genome. To remove non-mutating nucleotides, any nucleotide position that did not have at least one of the five mutation counts greater than 100 was omitted from all sequences. Additionally, all nucleotide codes other than 'A', 'C', 'G', 'T', and '-' (e.g., 'R') were treated as missing values by converting the nucleotide to 'N'. Finally, any remaining nucleotides that did not pass the mutation filter were assigned the reference genome value at the corresponding position. In summary, these filtering steps resulted in 7,722,980 SARS-CoV-2 sequence vectors with 21,433 nucleotides each. All pre-processing steps were implemented using Python (version 3.9.2), in scripts that are publicly available at https://github.com/jeanmerlet/covid_comet.

### 3.2 Binary vector comparison

To prepare each sequence for exhaustive vector comparison with ultralow precision, each vector is converted to binary using the following strategy. Nucleotides corresponding to 'A', 'C', 'G', 'T', and '-' are replaced with one-hot vectors with five bits. For example, 'A' corresponds to the vector '10000', 'C' corresponds to '01000', and '-' to '00001'. Missing values (i.e., 'N') are replaced by vectors of all zero: '00000'. This process increases the genome vector length from 21,433 nucleotides to $21,433 \times 5 = 107,165$ bits.

To compare binary vectors, we modify the Duo metric described in Equation 1 to account for the novel binarization scheme. In this case, two genome sequences are similar if they share nucleotides in corresponding positions (e.g., A's that line up with A's, C's that line up with C's, etc.). In the case of one-hot representations, two nucleotides are the same if they share a 1 in the same position. Thus, we modify Equation 1 by (i) setting $q = 0$ in order to disable the frequency terms, $f_i$ and $f_j$, and (ii) to only consider the 'high-high' correlation. In other words, this reformulated Duo metric measures the frequency of co-occurrences of 1's between any two SARS-CoV-2 binary vectors, thereby measuring the co-occurrences of similar nucleotides. Thus, this score increases with the number of shared nucleotides between vectors $i$ and $j$, and decreases with the number of mutations.

Exhaustive vector comparisons were computed using CoMet compiled with software versions GCC 6.4.0, Spectrum MPI 10.3.1.2-20200121, and CUDA 10.1.243. The full computational run was launched on the Summit supercomputer with IBM's Job Step Manager (JSM) software, which provides the 'jsrun' tool using six MPI ranks per node, with each MPI rank allocated 1 NVIDIA V100 GPU and 7 POWER9 CPU cores mapped via OpenMP threads. The following environment variables were set prior to the CoMet run:

- `OMP_NUM_THREADS=7`
- `PAMI_IBV_ENABLE_DCT=1`
- `PAMI_ENABLE_STRIPING=1`
- `PAMI_IBV_ADAPTER_AFFINITY=0`
- `PAMI_IBV_QP_SERVICE_LEVEL=8`
- `PAMI_IBV_ENABLE_OOO_AR=1`

The CoMet vector comparisons were thresholded at 0.19991 (i.e., ∼99.9% of the maximum theoretical score, 0.2) in order to retain the top ∼0.1% of most similar vector pairs. This process resulted in a sparse sequence-to-sequence similarity network containing 5,290,386 sequences (nodes) and 777,928,464,646 sequence pairs (edges). Thus, on average, each sequence shares an edge with ∼147,000 other nodes in the network based on vector comparisons that utilize the entire SARS-CoV-2 genome.

### 3.3 Unsupervised network clustering

The network representation described above is useful for multiple downstream scientific tasks. In this work, we applied Markov clustering to group sets of SARS-CoV-2 sequences into novel, data-driven lineages. Clustering was performed using HipMCL [2], which iterates between two matrix operations: expansion and inflation. The expansion step simulates a random walk and effectively spreads flow across the similarity network, while the inflation step contracts the flow [17]. Oscillating between these two steps increases edge weights between nodes of the same cluster and decreases edge weights to nodes of different clusters. Importantly, unlike alternative clustering methods (e.g., $k$-means), the number of genome clusters is not pre-defined, but rather emerges based on the graph

topology and parameterization of the MCL algorithm. The 'inflation rate' is the primary hyperparameter used to control the granularity of the emergent lineages, in which a lower inflation rate results in fewer, larger clusters, and higher values produce multiple, smaller clusters. Note that inflation rate can only control cluster granularity to a point, while the underlying graph topology has a strong influence on the outcome. For example, a fully-connected graph will produce one cluster regardless of inflation rate. In this work, we considered a range of inflation values: 1.2, 1.5, 2.0, 4.0, and 8.0, which produced up to 80,000 novel lineages. We chose to focus on inflation 1.2 (which produced the smallest number of clusters) for further downstream analyses. Markov clustering at this inflation yielded 22,040 clusters, with the top 20 largest clusters containing 26% of all edges in the underlying sequence-to-sequence similarity network.

The resulting clusters were visualized using a number of geospatial and temporal views. Using the GISAID metadata for each sequence, we assigned sequences to dates and locations. For dates, any sequence with only the month and year of collection was automatically assigned the 15th day of the month. For locations, we extracted both the country and continent of collection. Any sequences missing such date or location information were ignored in the visualization. Further, we used the PANGO lineage labels of each sequence to map them to the corresponding World Health Organization (WHO) variant, or "Other" if there was no corresponding WHO variant (e.g., "AY.1" was mapped to "Delta"). Then, for each cluster, we assigned a WHO variant label if there was a clear majority variant (e.g., $\geq$ 90% of the sequences within that cluster) and "Other" otherwise. While sequence metadata for the clusters mapped to 1,712 different PANGO lineages, each of the 1,000 largest clusters was composed of 99.9% sequences from a single WHO variant after mapping PANGO lineages to WHO variants. As there are only a handful of WHO variants, but more than 20,000 clusters, many clusters were classified to the same WHO variant. Importantly, although multiple clusters may belong to the same WHO variant, these clusters consisted of different sets of mutations, and hence were kept separate in our analysis. We then chose the top four largest clusters belonging to each of the three largest strains through June 2022: "Alpha", "Delta", "Omicron", and "Other", resulting in a total of 16 data-driven lineages (i.e., 4 clusters $\times$ 4 WHO variants). The WHO clusters with majority variants not included in visualizations (e.g. Mu) comprised a small subset of all sequences (2.7% of sequences) and will be considered in future work. To construct global spatiotemporal views, each country is chosen to display a unique color corresponding to the cluster from the set of top 16 clusters with the most sequences at that time point. While the light blue cluster (variant) in "Other" contains Wuhan-1, it also contains other SARS-CoV-2 sub-strains of Wuhan-1 that are close to but not completely identical to Wuhan-1. Finally, we applied a 7-day moving average to each time series in order to smooth the noisy signal and improve the quality of the visualizations. Using these methods, we identified novel subgroupings of the larger WHO variants and subsequently tracked the evolution and spread of these subgroups over time and space.

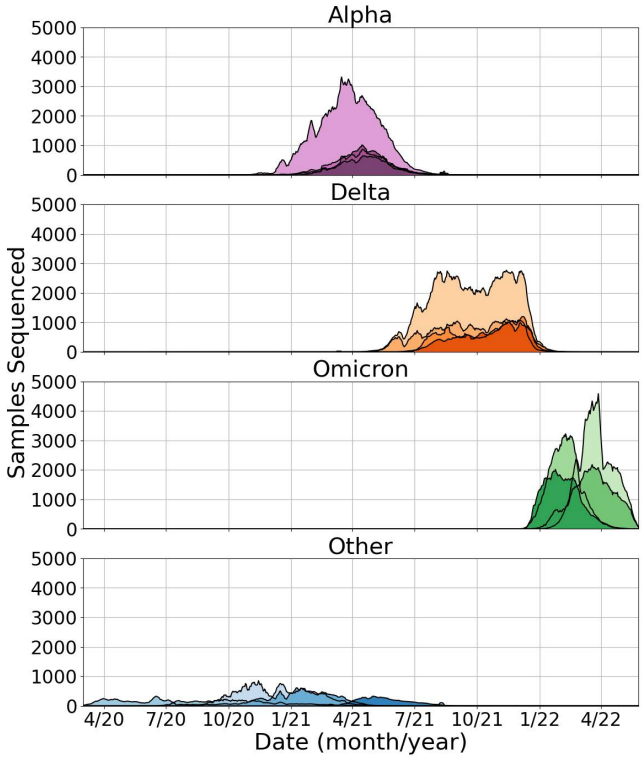



**Figure 1: Longitudinal view of the top four SARS-CoV-2 clusters for WHO variants, Alpha, Delta, Omicron, and "Other" over the first ~2.5 years of the Covid-19 pandemic. Each color and subplot refers to a different WHO variant. The different hues of each color represent the four largest data-driven clusters (i.e., SARS-CoV-2 lineages) corresponding to the WHO variant.**

## 4 RESULTS AND DISCUSSION

Using the set of 22,040 clusters based on exhaustive whole-genome sequence comparisons discussed in Section 3.3, we show spatial and temporal views of the top 16 novel data-driven lineage across countries and continents. First, we visualize the top 16 data-driven clusters (i.e., novel SARS-CoV-2 lineages) over time in a split view by WHO variant in Figure 1.

This view enables tracking the emergence and disappearance of various subclusters over time, and demonstrates that each WHO variant generally appears and disappears as a group, apart from the miscellaneous "Other" category, which appears consistently over the time course in the form of various subvariants. Note that the time between sample collection and availability on GISAID varies from days to several weeks, by which we observe a sharp decrease in the number of sampled sequences at the end of the time series (i.e., when the GISAID data was accessed). This decrease therefore does not indicate a drop in the number of genomes sequenced, but rather that the sequences had not all been uploaded to GISAID. The SARS-CoV-2 genome sequences in these clusters included metadata tags for 1,712 unique PANGO lineages, which were mapped to named WHO variants or "Other", based on the WHO variant definitions.

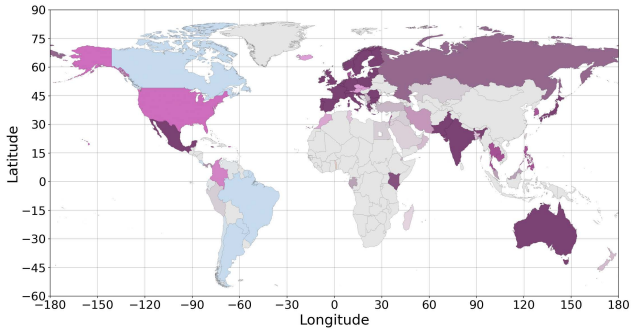
While we are only displaying the 4 largest clusters for each of the major WHO variants, there are 20,856 clusters mapping to these 4 WHO variants (Alpha, Delta, Omicron, and Other), and each of these clusters consists of a separate set of mutations from the others. Thus our methods result in significantly higher resolution of variant classification than WHO variants or PANGO lineages. Fewer than 5% of sequences (i.e., 1,184 clusters) corresponded to a WHO variant other than the four considered here, whereas Alpha, Delta, Omicron, and Other each accounted for 2,022, 11,301, 2,396, and 5,137 of the clusters, respectively. Note that the vertical axis in these figures corresponds to the number of daily samples sequenced, and does not necessarily correspond to the number of cases of any particular variant at any particular time.

We adopt an alternative spatiotemporal view in Figure 2, which considers the dynamics of the top clusters across country, continent, and sample collection date. The full animated view of these snapshots is available at https://github.com/pasc-2023-anonymous/pasc_2023_anonymous/blob/main/maps.gif.

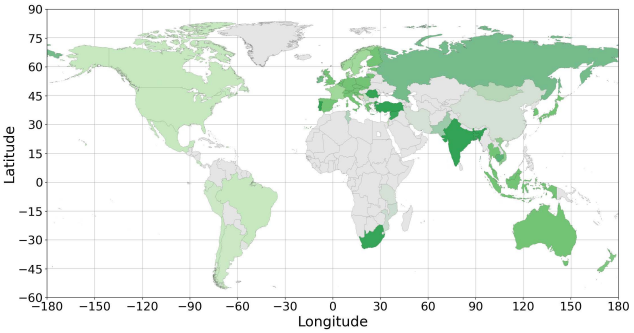
Note that the other continents (i.e., South America, Africa, Asia, and Oceania) are excluded from Figure 2c due to very small numbers of sampled sequences. The geospatial views in Figures 2a and 2b show that different subclusters of the larger WHO variants affected different countries and continents at different times. The continental time series view in Figure 2c confirms this by demonstrating that different subclusters dominate the samples being sequenced in North America and Europe at different times.

Miscellaneous COVID-19 lineages dominate the beginning of the pandemic (January 2020), but almost entirely disappear with the arrival of Alpha in December 2020 (1). We did not find that the geographic proximity of two countries guaranteed the same cluster, PANGO lineage, or WHO variant. Instead, neighboring countries were often occupied by different clusters of the same variant or entirely different WHO variants. At the peak of the number of sequenced Alpha SARS-CoV-2 samples uploaded to GISAID on April 17, 2021, Canada and much of South America were dominated by the Alpha variant, while the USA and Mexico had majorities of different clusters of Alpha sequences (see Figure 2a). Even after the Omicron variant had almost entirely taken over globally, e.g., during the height of the original Omicron surge on February 19, 2022, Europe and Asia reported sequences of four different substrains (clusters) of Omicron. On the other hand, the United States and South America were reporting the same Omicron cluster, yet a different one compared to the rest of the world (see Figure 2b). Note that sequencing enrichment optimization may play a role in differences between observed geographical variants. The instance from December to February of different Omicron subvariants in Europe and North America, and a stark gap between Delta and Omicron sequences in Europe (see Figure 2b), may have been exacerbated by short successive changes in the oligo primer pool at that time. In particular, the ARTIC primer sets transitioned from V3 to V4 in January of 2022, but V4 was plagued by amplicon drop offs, and was not optimized until later that February to V4.1 [14].

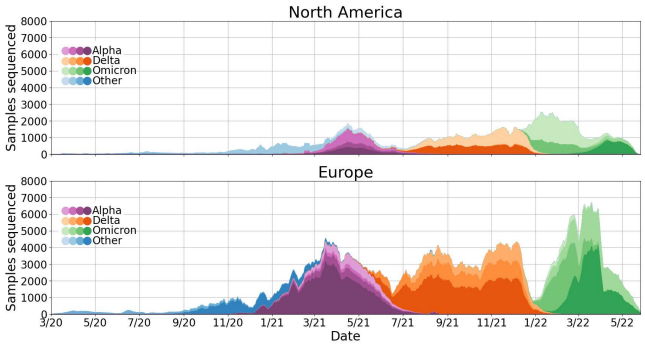
Our novel lineage classification method demonstrates that an all-against-all, whole-genome vector comparison applied to millions of SARS-CoV-2 sequences can yield tens-of-thousands of novel data-driven clusters, each composed of closely genetically-related




**(a) Top SARS-CoV-2 strains on April 17, 2021.**




**(b) Top SARS-CoV-2 strains on Feb. 19, 2022.**




**(c) Cluster prevalence time series comprising named WHO variants.**

**Figure 2: Longitudinal views of the top SARS-CoV-2 strains by continent during the first ~2.5 years of the COVID-19 pandemic. Each color refers to a different WHO variant, and the different hues of each color represent the four largest data-driven clusters (i.e., SARS-CoV-2 lineages) corresponding to the WHO variant. (a) shows a snapshot of the distribution of clusters on April 17, 2021, during the peak of the Alpha variant. (b) shows a snapshot of the distribution of clusters on February 19, 2022, during the initial peak of the Omicron variant. (c) shows a similar time series view compared to Figure 1, but split by continent rather than WHO variant. The $y$-axis represents the cumulative sum of sequenced samples at daily resolution.**

sequences. The PANGO lineage designation system relies on phylogenetic trees (which, by definition, exclude recombination) and hand-curation to classify the several million sequence samples considered in this work into 1,380 different lineages, while the WHO variant naming scheme classifies them into a small handful of named variants. Both of these classification systems focus on a subset of hand-picked mutations to make their assignments. Importantly, unlike previous classification paradigms that rely on such heuristics, our methodology leverages the entirety of the SARS-CoV-2 genome in a network context to produce novel lineages in an unbiased data-driven way, which results in 22,040 distinct clusters (SARS-CoV-2 lineages). Further, the use of unsupervised Markov clustering does not assume a set number of lineages *a priori*, but rather produces clusters based on the topology of the emergent sequence-to-sequence similarity network, which is only possible to derive using extremely efficient exhaustive vector comparisons on leadership-class HPC systems.

## 4.1 Limitations

While the scale of the available SARS-CoV-2 sequences on GISAID is unprecedented, the available data come with several caveats. Metadata is available for every uploaded sequence, but it is often incomplete. For example, of the 11M sequences, 7M mark gender as unknown, with the rest composed of slightly more women than men (2.13M women and 1.98M men). Other notable categories with significant amounts of missing or unknown data are patient age, geographical origin of the sample, and sample collection date. As a result of this missing metadata, any sequence analyses requiring associated metadata (e.g. sample collection date and location for spatiotemporal analyses) must necessarily exclude these sequences, reducing the number of overall usable sequences.

Another major consideration when interpreting results stems from the disparity between infection rates and sequencing rates throughout countries globally. These differences can arise for various reasons, such as income disparity (78% of high-income but only 42% of low-income countries sequenced $\geq$ 0.5% of COVID-19 cases) [3] or unwillingness to share sequencing data (37% of countries submitted less than half of their sequences to public databases when sequencing variants of concern) [4]. Moreover, sequencing rates overall are vastly lower than infection rates, with only 6.8% of countries sequencing $\geq$ 5% of their total confirmed cases [3]. Lastly, our current spatiotemporal visualizations do not account for population density. One should keep these geospatial and socioeconomic sequencing disparities in mind when interpreting results involving comparisons such as differences between countries or populations.

## 4.2 Performance comparison

CoMet has been leveraged for multiple large-scale exhaustive vector comparisons with applications ranging from genomics to environmental modeling [10, 11, 13]. However, these applications differ significantly from our use case, including different data modalities, binarization scheme, choice of similarity metric, number of vectors, and vector length. As such, it can be difficult to draw meaningful performance comparisons between these use cases. In an earlier version of this work, we reported on the set of 4 million available SARS-CoV-2 sequences that were available through September, 2021. In this work, we have completed our analysis on an expanded set of 11 million SARS-CoV-2 sequences, and we now present a performance comparison between these two instances. The first and second runs are labeled Sept. 2021 and June 2022, respectively, in reference to the latest month of sequence collection time (see Table 1).

**Table 1: SARS-CoV-2 CoMet run comparison.**

| | # Vectors | # Elements | # Nodes | Node hours |
|---|---|---|---|---|
| Sept. 2021 | 2,978,754 | 73,965 | 150 | 136.0 |
| June 2022 | 7,722,980 | 107,165 | 840 | 631.2 |

Our expanded set of SARS-CoV-2 vectors contained approximately 3 times as many sequences as an earlier version of this work that included 4 million genome sequences. After pre-processing and thresholding, there were 2.65 times more input vectors in the second run, although the vectors were longer. The increased vector length is due to our removal of any nucleotide position that did not have at least one of the five mutation counts greater than 100 across all the sequences (see Section 3.1), which increased when we increased the number of sequences from 4 to 11 million. Note that the number of sequences in Table 2 is lower than the number of vectors in Table 1, as we thresholded the CoMet outputs to a subset of highly significant edges (top ~5% of comparisons) in order to generate a manageable volume of data for downstream analyses. Despite this thresholding, the September and June runs generated 269GB and 12TB of output data, respectively.

**Table 2: SARS-CoV-2 HipMCL run comparison.**

| | # Sequences | # Edges (top %) | # Nodes |
|---|---|---|---|
| Sept. 2021 | 2.07M | 112.47B (5.20%) | 49 |
| June 2022 | 5.29M | 777.92B (5.56%) | 40 |

The updated CoMet run required 4.63 times more total hours of computational time. However, while compute time scaled similarly to the increased number of vectors, the time increase was mainly due to the write time required to output the much larger set of CoMet results. Since we chose to save the top ~5% most variable edges for both runs, the number of edges in the second run was proportionally much higher than in the original.

## 4.3 Conclusion

In this work, we performed an all-against-all whole-genome vector comparison of every SARS-CoV-2 sequence available on GISAID through September 2022. Although we are only visualizing a small number of the most prevalent strains in this manuscript, this workflow has produced the genome-driven identification of more than 22,000 strains or haplotypes, for each of which we have geospatial and temporal trajectories spanning the COVID-19 pandemic. Our identification of the trend for dominant SARS-CoV-2 strains to be spatially asynchronous suggests a possibly different pattern of epidemiological risk than has been understood thus far. These

will be fascinating resources for future work that focuses on the evolution of the virus over the course of the pandemic and to aid in detecting possible recombination events. Furthermore, this work will help in future analyses intended to identify epistatic relationships in strains that have evolved over the course of the pandemic. These geospatially and temporally resolved haplotypes will also enable the use of explainable-AI approaches to find whether there are associations between sets of mutations and local environmental variables (e.g. climatic and socio-demographic features) and/or whether they affect mortality rates.

## ACKNOWLEDGMENTS

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work is supported as part of the Genomic Sciences Program DOE Systems Biology Knowledgebase (KBase) funded by the Office of Biological and Environmental Research's Genomic Science program within the US Department of Energy Office of Science under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05–00OR22725, and DE-AC02-98CH10886. This work was also supported by the U.S. National Science Foundation (EF-2133763). We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. This manuscript has been co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

## REFERENCES

[1] [n. d.]. Pango Network. https://www.pango.network/the-pango-nomenclature-system/statement-of-nomenclature-rules.

[2] Ariful Azad, Georgios A Pavlopoulos, Christos A Ouzounis, Nikos C Kyrpides, and Aydin Buluç. 2018. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Research* 46, 6 (Jan. 2018), e33–e33. https://doi.org/10.1093/nar/gkx1313

[3] Anderson F. Brito, Elizaveta Semenova, Gytis Dudas, Gabriel W. Hassler, Chaney C. Kalinich, Moritz U.G. Kraemer, Joses Ho, Houriiyah Tegally, George Githinji, Charles N. Agoti, Lucy E. Matkin, Charles Whittaker, Benjamin P Howden, Vitali Sintchenko, Neta S. Zuckerman, Orna Mor, Heather M Blankenship, Tulio de Oliveira, Raymond T. P. Lin, Marilda Mendonça Siqueira, Paola Cristina Resende, Ana Tereza R. Vasconcelos, Fernando R. Spilki, Renato Santana Aguiar, Ivailo Alexiev, Ivan N. Ivanov, Ivva Philipova, Christine V. F. Carrington, Nikita S. D. Sahadeo, Céline Gurry, Sebastian Maurer-Stroh, Dhamari Naidoo, Karin J von Eije, Mark D. Perkins, Maria van Kerkhove, Sarah C. Hill, Ester C. Sabino, Oliver G. Pybus, Christopher Dye, Samir Bhatt, Seth Flaxman, Marc A. Suchard, Nathan D. Grubaugh, Guy Baele, Nuno R. Faria, , , and and. 2021. Global disparities in SARS-CoV-2 genomic surveillance. (Aug. 2021). https://doi.org/10.1101/2021.08.21.21262393

[4] Zhiyuan Chen, Andrew S. Azman, Xinhua Chen, Junyi Zou, Yuyang Tian, Ruijia Sun, Xiangyanyu Xu, Yani Wu, Wanying Lu, Shijia Ge, Zeyao Zhao, Juan Yang, Daniel T. Leung, Daryl B. Domman, and Hongjie Yu. 2022. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics* 54, 4 (March 2022), 499–507. https://doi.org/10.1038/s41588-022-01033-y

[5] Sharlee Climer, Alan R Templeton, Michael Garvin, Daniel Jacobson, Matthew Lane, Scott Hulver, Brittany Scheid, Zheng Chen, Carlos Cruchaga, and Weixiong Zhang. 2020. Synchronized genetic activities in Alzheimer's brains revealed by heterogeneity-capturing network analysis. *bioRxiv* (2020).

[6] Sharlee Climer, Wei Yang, Lisa de las Fuentes, Victor G. Dávila-Román, and C. Charles Gu. 2014. A Custom Correlation Coefficient (CCC) Approach for Fast Identification of Multi-SNP Association Patterns in Genome-Wide SNPs Data. *Genetic Epidemiology* 38, 7 (Aug. 2014), 610–621. https://doi.org/10.1002/gepi.21833

[7] Stijn Van Dongen. 2008. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* 30, 1 (Jan. 2008), 121–141. https://doi.org/10.1137/040608635

[8] Stefan Elbe and Gemma Buckland-Merrett. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 1, 1 (Jan. 2017), 33–46. https://doi.org/10.1002/gch2.1018

[9] Joe Hasell, Edouard Mathieu, Diana Beltekian, Bobbie Macdonald, Charlie Giattino, Esteban Ortiz-Ospina, Max Roser, and Hannah Ritchie. 2020. A cross-country database of COVID-19 testing. *Scientific Data* 7, 1 (Oct. 2020). https://doi.org/10.1038/s41597-020-00688-8

[10] Wayne Joubert, James Nance, Deborah Weighill, and Daniel Jacobson. 2018. Parallel accelerated vector similarity calculations for genomics applications. *Parallel Comput.* 75 (July 2018), 130–145. https://doi.org/10.1016/j.parco.2018.03.009

[11] Wayne Joubert, Deborah Weighill, David Kainer, Sharlee Climer, Amy Justice, Kjiersten Fagnan, and Daniel Jacobson. 2018. Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction. In *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE. https://doi.org/10.1109/sc.2018.00060

[12] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, GISAID Core Curation Team, and Sebastian Maurer-Stroh. 2021. GISAID's Role in Pandemic Response. *China CDC Weekly* 3, 49 (2021), 1049–1051. https://doi.org/10.46234/ccdcw2021.255

[13] John Lagergren, Mikaela Cashman, Veronica Melesse Vergara, Paul Eller, Joao Gabriel Felipe Machado Gazolla, Hari Chhetri, Jared Streich, Sharlee Climer, Peter Thornton, Wayne Joubert, and Daniel Jacobson. 2022. Climatic clustering and longitudinal analysis with impacts on food, bioenergy, and pandemics. *Phytobiomes Journal* (Sept. 2022). https://doi.org/10.1094/pbiomes-02-22-0007-r

[14] Arnold W. Lambisia, Khadija S. Mohammed, Timothy O. Makori, Leonard Ndwiga, Maureen W. Mburu, John M. Morobe, Edidah O. Moraa, Jennifer Musyoki, Nickson Murunga, Jane N. Mwangi, D. James Nokes, Charles N. Agoti, Lynette Isabella Ochola-Oyier, and George Githinji. 2022. Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome Sequencing Protocol. *Frontiers in Medicine* 9 (Feb. 2022). https://doi.org/10.3389/fmed.2022.836728

[15] Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, Corin Yeats, Louis Du Plessis, Daniel Maloney, Nathan Medd, Stephen W Attwood, David M Aanensen, Edward C Holmes, Oliver G Pybus, and Andrew Rambaut. 2021. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. *Virus Evolution* (July 2021). https://doi.org/10.1093/ve/veab064

[16] Andrew Rambaut, Edward C. Holmes, Áine O'Toole, Verity Hill, John T. McCrone, Christopher Ruis, Louis du Plessis, and Oliver G. Pybus. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* 5, 11 (July 2020), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

[17] Stijn Van Dongen. 2008. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* 30, 1 (2008), 121–141.

Received 23 December 2022; revised 1 February 2023; accepted 6 April 2023