Rapid whole genome sequence typing reveals multiple waves of SARS-CoV-2 spread

Ahmed M. Moustafa[1], Paul J. Planet[1,2]

1 Children's Hospital of Philadelphia, Philadelphia, PA, USA;2 Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, USA.

Email: moustafaam@email.chop.edu

Background

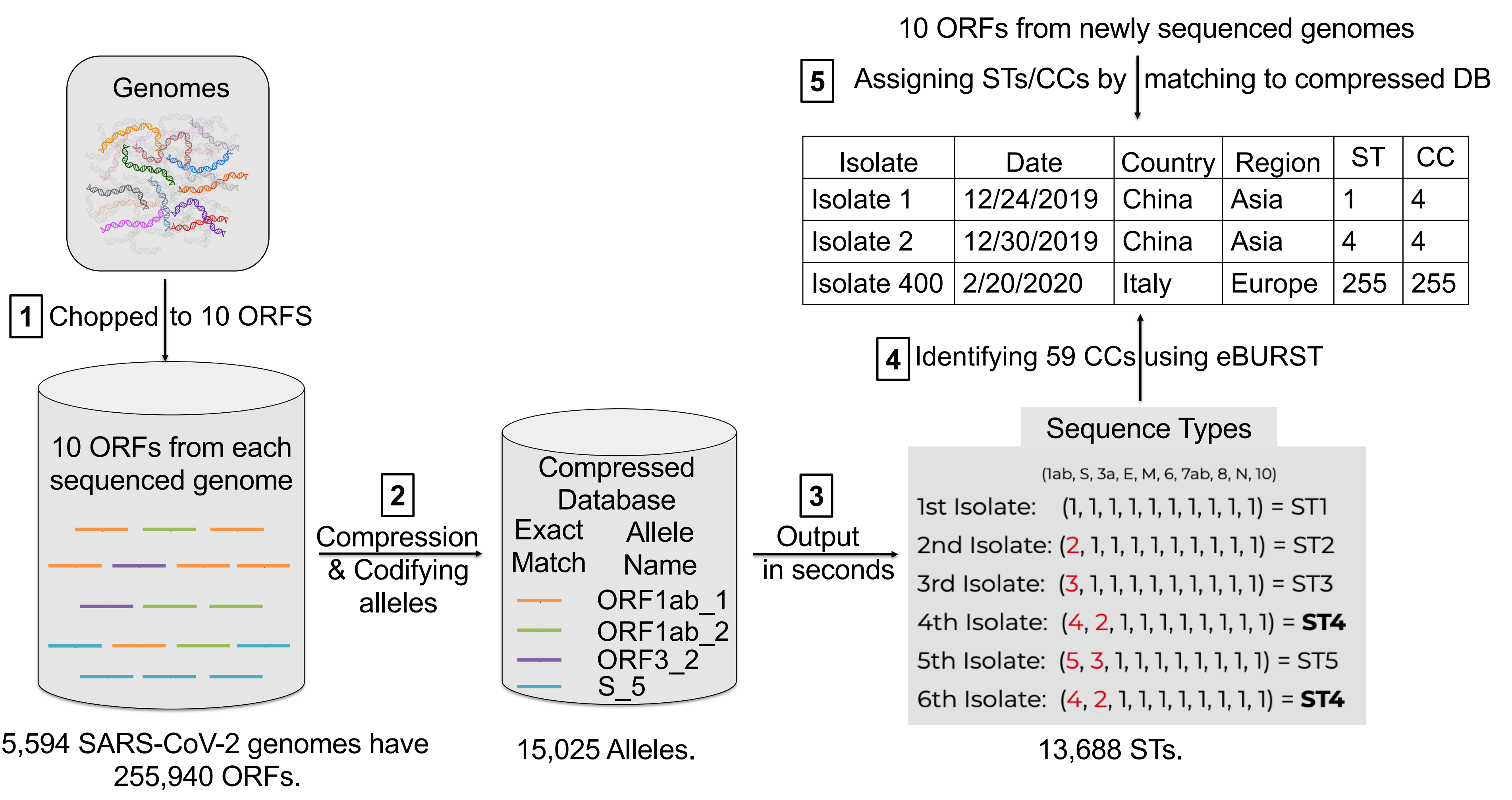
Rapid sequencing of the SARS-CoV-2 pandemic virus has presented an unprecedented opportunity to track the evolution of the virus and to understand the emergence of a new pathogen in near-real time. During its explosive radiation and global spread, the virus has accumulated enough genomic diversity that we are now able to identify distinct lineages and track their spread in distinct geographic locations and over time.

Reconstructing a robust phylogeny of the existing virus variants is an intuitive approach. However, one of the main problems of this approach is scalability. Building phylogenetic trees will be harder as more genomes are added every day. Building a phylogeny of a selected subset genomes necessarily loses information and explanatory power.

Because of this roadblock, our goal was to come up with a rapid way to categorize genomes that scales readily and leads to as little information loss as possible. We saw an opportunity to combine our allele identifying tool “WhatsGNU”[1] with the Multilocus Sequence Typing (MLST)[2] approach that has been widely used in bacterial categorization, tracking the emergence of new lineages, and associating specific Sequence Types/Clonal Complexes (STs/CCs) with certain diseases.

Here we developed the GNU-based Virus IDentification (GNUVID) tool[3] that implements a wgMLST scheme composed of all ten ORFs in the SARS-CoV-2 genome. It rapidly assign an allele number to each gene nucleotide sequence in the virus’s genome creating a ST which is codified as the sequence of allele numbers for each of the ten genes in the viral genome. The STs are then linked into clearly defined CCs that are consistent with phylogeny. We show that assessment of STs and CCs agrees with multiple introductions of the virus in certain geographical locations. In addition, we use temporal assessment of STs/CCs to uncover waves of expansion and decline, and the apparent replacement of certain STs with emerging lineages in specific geographical locations.

GNUVID Workflow



10 ORFs from newly sequenced genomes

5 Assigning STs/CCs by matching to compressed DB

Isolate	Date	Country	Region	ST	CC
Isolate 1	12/24/2019	China	Asia	1	4
Isolate 2	12/30/2019	China	Asia	4	4
Isolate 400	2/20/2020	Italy	Europe	255	255

Genomes

1 Chopped to 10 ORFS

10 ORFs from each sequenced genome

25,594 SARS-CoV-2 genomes have 255,940 ORFs.

2 Compression & Codifying alleles

Compressed Database

Exact Match	Allele Name
—	ORF1ab_1
—	ORF1ab_2
—	ORF3_2
—	S_5

15,025 Alleles.

3 Output in seconds

Sequence Types

(1ab, S, 3a, E, M, 6, 7ab, 8, N, 10)

1st Isolate: (1, 1, 1, 1, 1, 1, 1, 1, 1, 1) = ST1
2nd Isolate: (2, 1, 1, 1, 1, 1, 1, 1, 1, 1) = ST2
3rd Isolate: (3, 1, 1, 1, 1, 1, 1, 1, 1, 1) = ST3
4th Isolate: (4, 2, 1, 1, 1, 1, 1, 1, 1, 1) = **ST4**
5th Isolate: (5, 3, 1, 1, 1, 1, 1, 1, 1, 1) = ST5
6th Isolate: (4, 2, 1, 1, 1, 1, 1, 1, 1, 1) = **ST4**

13,688 STs.

4 Identifying 59 CCs using eBURST

- Complete and high coverage SARS-CoV-2 genomes (n=43,406) were downloaded from GISAID[4] on July 17th 2020. The 10 ORFs were identified in the genomes using Blastn[5].
- 25,594 genomes passed our quality check (No ambiguous or degenerate bases in the 10 ORFs).
- The 25,594 genomes were fed to the GNUVID tool, which eliminated copies of redundant sequences while retaining the metadata and assigned a ST profile to each genome/isolate.
- Global optimum eBURST[6] minimum spanning tree (MST) implemented in PHYLOViZ[7] tool was then used to cluster the STs in clonal complexes (CCs) at the double locus variant (DLV) level.
- Single CPU & 16 GB of RAM on Standard Desktop.
- GNUVID is available so that new WG sequences can be rapidly assigned to an ST/CC (https://github.com/ahmedmagds/GNUVID).

References

1- Moustafa AM, Planet PJ. WhatsGNU: a tool for identifying proteomic novelty. Genome Biology 2020; 21(1): 58.

2- Maiden MC, Bygraves JA, Feil E, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. PNAS 1998; 95(6): 3140-5.

3- Moustafa AM, Planet PJ. Rapid whole genome sequence typing reveals multiple waves of SARS-CoV-2 spread. bioRxiv 2020.06.08.139055.

4- Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill 2017; 22(13).

5- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990; 215(3): 403-10.

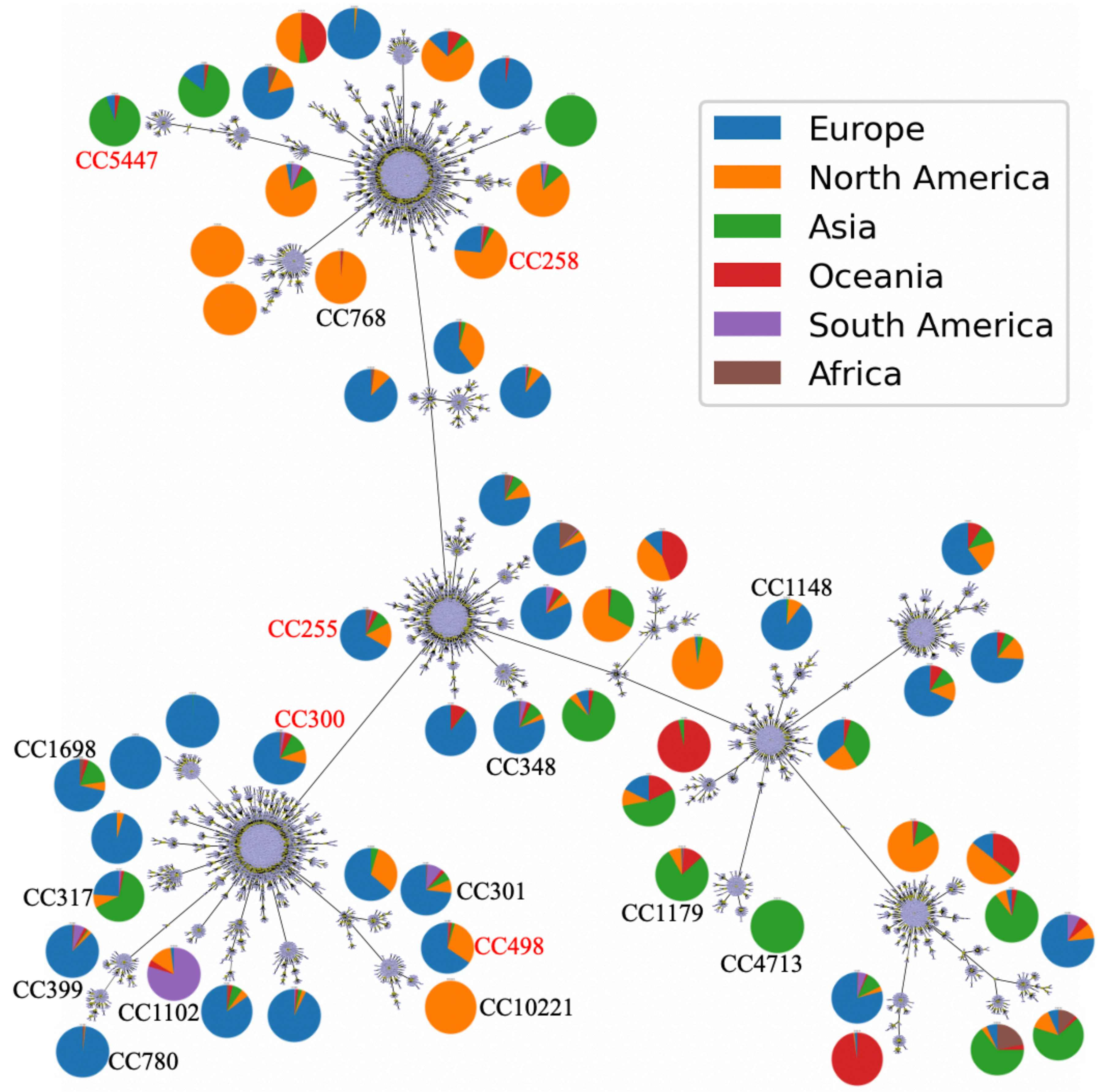
6- Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data. Journal of Bacteriology 2004; 186(5): 1518.

7- Nascimento M, Sousa A, Ramirez M, Francisco AP, Carrico JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics 2017; 33(1): 128-9.

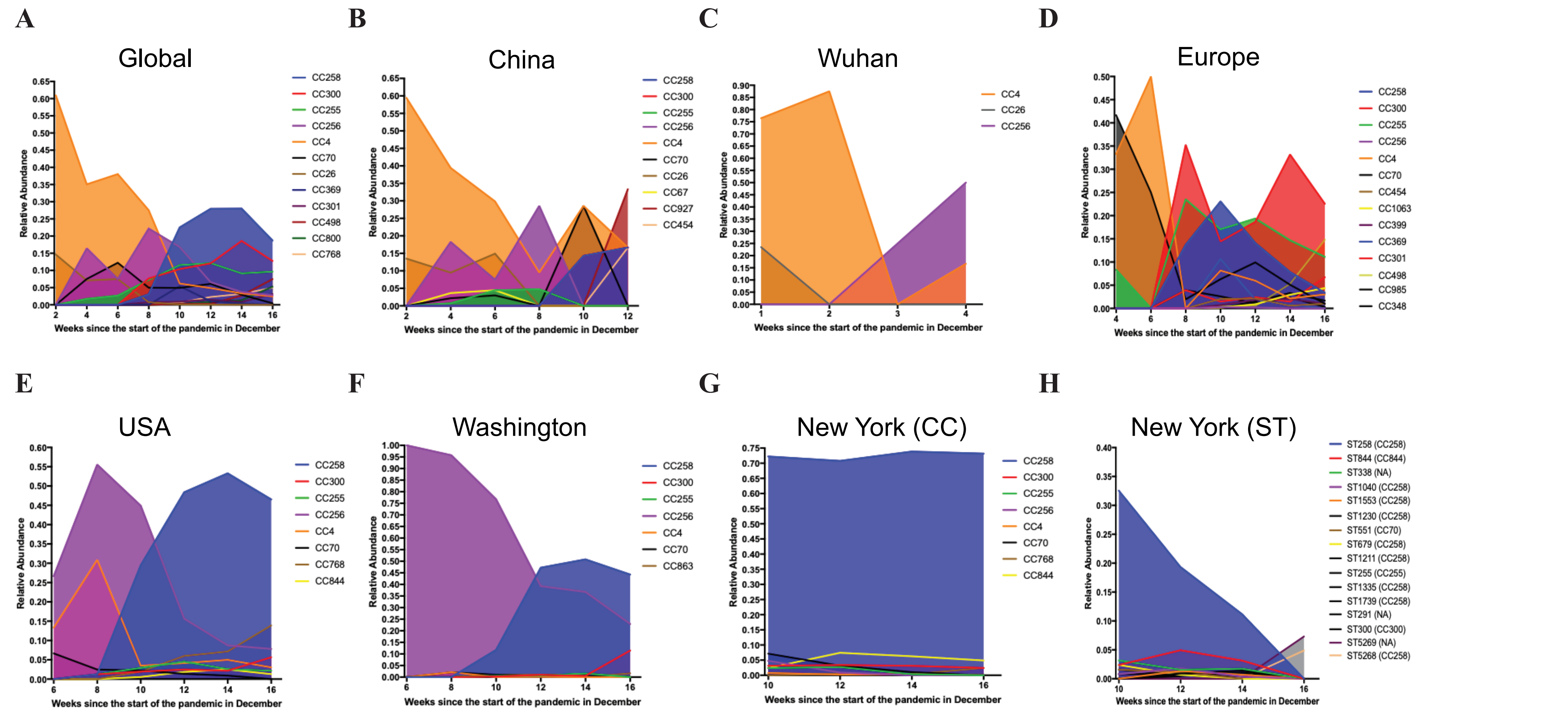
* We would like to thank Elizabeth Lugones for helping with the GNUVID workflow figure.

Results

Minimum spanning tree of the 13688 STs showing the 59 CCs. The 5 Active CCs are in red and the 12 Quiet CCs are in black. The pie charts show the percentage distribution of genomes from the different geographic regions in each CC. The analysis uncovered strong associations of ST/CCs with certain geographical regions but also dynamic local changes in ST/CC prevalence.



Temporal Plots of circulating STs/CCs at different geographical locations (Global, China, Wuhan, Europe, USA, Washington, NY (CC) and NY (ST)) until 05/17. The plots show waves of expansion and replacement of SARS-CoV-2 STs and CCs in different geographical locations.



Conclusion

The genomic epidemiology of the 25,594 SARS-CoV-2 isolates studied here show six predominant CCs circulated/circulating globally. Our tool (GNUVID) allows for fast sequence typing and clustering of whole genome sequences in a rapidly changing pandemic. This be used to temporally track emerging clones or identify the likely origin of viruses. With stored metadata for each sequence on date of isolation, geography, and clinical presentation, new genomes could be matched almost instantaneously to their likely origins and potentially related clinical outcomes.