**Phylogenetic Network Analysis and Community Discovery of SARS-CoV-2 from US COVID-19 Patients** **Reveal Accumulation of Mutations in the Viral Genome**

**ABSTRACT**

(VC) Add abstract to cover the following points:

* Why did we do this work?
* How did we do it?
* What did we find?

**INTRODUCTION**

(VC) Add Introduction containing the following sections.

Part 1

* Background of COVID-19 pandemic and SARS-CoV-2.

Part 2

* Mutations of SARS-CoV-2 (people usually focus on the mutations in the spike protein).
* SARS-CoV-2 variants detected in UK, Brazil, South Africa, etc.

Part 3

* US COVID-19 pandemic (US so far has the most COVID-19 cases and deaths in the world).
* Mutations of SARS-CoV-2 in the US, especially the mutations occurred in other regions of the viral genome that could enhance the transmission and/or virulence of the virus.

Part 4

* Briefly mention what we are going to present in this paper.

**MATERIALS AND METHODS**

**(TD) Our Pipeline Diagram**

**\*Flowchart\***

*Description*

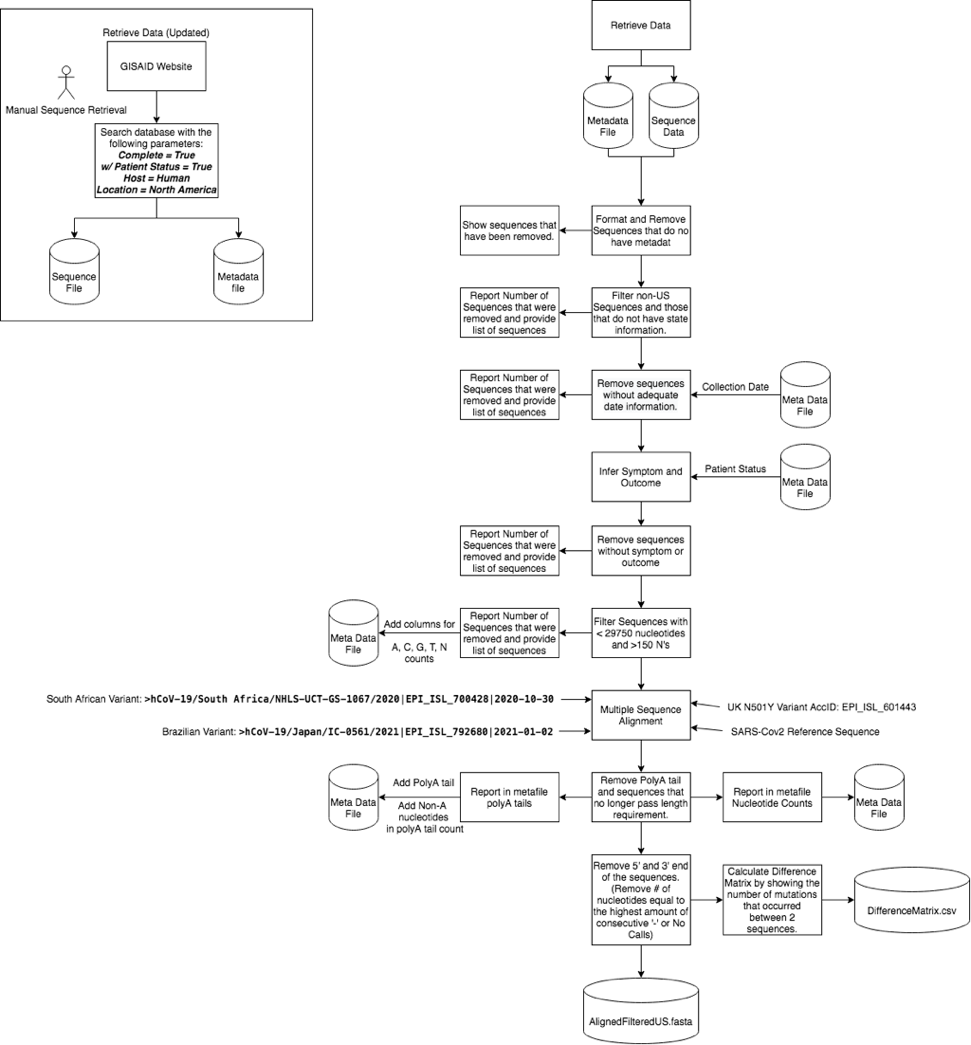
1. **Generation of the master dataset**

**1a. Data retrieval and processing**

(TD) Add: Concise description of how the master dataset was generated

This section should include (but not limited to):

* Retrieval of SARS-CoV-2 genomic sequences of the US COVID-19 cases (Download the SARS-CoV-2 genome sequences from GISAID website).
* Data filtering
  + Eliminate low-quality sequences by applying the following filters.
    - * #nucleotides > 29,750 (including the poly(A) tail (nucleotide count: A, C, G, and T only, but not N or “-“). **Note:** A+C+G+T > 99.5%
    - #N’s < 150. **Note:** N < 5%
    - Collection date has no month or date information.
  + Remove the poly(A) tail from the sequence, i.e. the sequence downstream of the 3' UTR (29675-29870 in the reference sequence). **Note:** The last 25 nucleotides: 5’-AATCTTTAATCAGTGTGTAACATTA-3’
  + Determine the number of remaining cases with the following patient status information.
    - Entry ≠ unknown
    - Entry = unknown
    - Entry = “-“
    - No entry



**Figure 1. Data retrieval and processing pipeline.**

**1b. Standardization of symptoms, outcome and other patient information**

(TD) Add: Concise description how this was done.

**Table 1. Rules used for inferring outcome.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unknown | Survived | Deceased |
| 0 | - | Released | Death |
| 1 | DAMA | Alive | Deceased |
| 2 | EHPAD | Asymptomatic | Deceased |
| 3 | EHPAD\_IRA | Asymptomatic/Released | Deseased |
| 4 | Epidemiology Study | Benigne | Hospitalized, deceased |
| 5 | Hopsitalized | Cured | Hospitalized/Deceased |
| 6 | Hospitaized | Discharged after recovery | deceased |
| 7 | Hospitalised | Home |  |
| 8 | Hospitalized | Hospitalized, released |  |
| 9 | Hospitalized | Hospitalized/Released |  |
| 10 | Hospitalized or to be hospitalized | Hospsitalized, ICU, fully recovered |  |
| 11 | ICU | Live |  |
| 12 | Intensive Care Unit | Live |  |
| 13 | Isolation | Live, acute respiratory infection |  |
| 14 | Naso-pharyngeal swab | Live, physical examination |  |
| 15 | Oro-pharyngeal swab | Mild |  |
| 16 | Physician | Mild / Contact exposure / Asymptomatic |  |
| 17 | Physician network | Mild case |  |
| 18 | Pneumonia (chest X-ray) | Mild symptoms (fever, cardiovascular disorders) |  |
| 19 | Screening | Mild/Contact exposure/Asymptomatic |  |
| 20 | Severe / ICU | Moderate / Outpatient |  |
| 21 | Severe/ICU | Moderate/Outpatient |  |
| 22 | Still hospitalized | Not Hospitalized |  |
| 23 | Suspected Corona | Not hospitalized |  |
| 24 | Symptomatic | Outpatient |  |
| 25 | Symptoms indicative of upper respiratory infection | Pneumonia (chest X-ray), not critical |  |
| 26 | Unknown | Quarantine |  |
| 27 | Unkown | Quarantined |  |
| 28 | Ward | Recovered |  |
| 29 | hospitalized | Recovered and Released |  |
| 30 | hospitalized or to be hospitalized | Recovering |  |
| 31 | inpatient | Released |  |
| 32 | symptomatic | Released |  |
| 33 | uncknown | Released, Live |  |
| 34 | unknow | live |  |
| 35 | unknown | outpatient |  |
| 36 | unkown |  |  |
| 37 | ï»¿unknown |  |  |

**Table 2. Rules used for inferring symptoms.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unknown | Symptomatic (Moderate to Severe Symptoms) | Asymptomatic or Mild Symptoms |
| 0 | - | Released | Asymptomatic |
| 1 | Alive | Cured | Asymptomatic/Released |
| 2 | EHPAD | DAMA | Benigne |
| 3 | EHPAD\_IRA | Death | Home |
| 4 | Epidemiology Study | Deceased | Mild |
| 5 | Isolation | Deceased | Mild / Contact exposure / Asymptomatic |
| 6 | Naso-pharyngeal swab | Deseased | Mild case |
| 7 | Oro-pharyngeal swab | Discharged after recovery | Mild symptoms (fever, cardiovascular disorders) |
| 8 | Physician | Hopsitalized | Mild/Contact exposure/Asymptomatic |
| 9 | Physician network | Hospitaized | Not Hospitalized |
| 10 | Quarantine | Hospitalised | Not hospitalized |
| 11 | Quarantined | Hospitalized | Outpatient |
| 12 | Screening | Hospitalized | outpatient |
| 13 | Suspected Corona | Hospitalized or to be hospitalized | | |
| 14 | Unknown | Hospitalized, deceased | | |
| 15 | Unkown | Hospitalized, released | | |
| 16 | uncknown | Hospitalized/Deceased | | |
| 17 | unknow | Hospitalized/Released | | |
| 18 | unknown | Hospsitalized, ICU, fully recovered | | |
| 19 | unkown | ICU |  |
| 20 | •ÈÀunknown | Intensive Care Unit | | |
| 21 | Live, physical examination | Live, acute respiratory infection | | |
| 22 | Live | Moderate / Outpatient | | |
| 23 | Live | Moderate/Outpatient | | |
| 24 | live | Pneumonia (chest X-ray) | | |
| 25 |  | Pneumonia (chest X-ray), not critical | | |
| 26 |  | Recovered |  |
| 27 |  | Recovered and Released | | |
| 28 |  | Recovering |  |
| 29 |  | Released |  |
| 30 |  | Released |  |
| 31 |  | Released, Live | | |
| 32 |  | Severe / ICU | | |
| 33 |  | Severe/ICU |  |
| 34 |  | Still hospitalized | | |
| 35 |  | Symptomatic | | |
| 36 |  | Symptoms indicative of upper respiratory infection | | |
| 37 |  | deceased |  |
| 38 |  | hospitalized |  |
| 39 |  | hospitalized or to be hospitalized | | |
| 40 |  | inpatient |  |
| 41 |  | symptomatic | | |
| 42 |  | Ward |  |

1. **Phylogenetic analysis**

**2a. Data processing**

(TD) Add: Concise description what data processing steps have been done for the phylogenetic analysis.

This section should include (but not limited to):

* Trimming of the 5’ and 3’ ends, etc.

**2b. Multiple sequence alignment**

(TD) Add: Concise description what data processing steps have been done for the phylogenetic analysis.

* 1. mafft --inputorder --anysymbol --kimura 1 --ep 0.5 --nwildcard --addfragments SequencesToBeAligned.fasta --memsave --6merpair referenceSequence.fasta > output.fasta

**2c. Generation of phylogenetic network**

Construction of phylogenetic networks begins with Using the Network.exe tool (<https://www.fluxus-engineering.com/>). We generated a median joining1 network connection file (network.out). The median joining connection file gives a list of nodes and edges, which were taken and processed through the python library NetworkX. This was to replicate the structure in a python environment to be able to perform the analyses described below. Six network graphs, with nodes indicating the Accession ID, Collection Dates, Locations and Community Identity (see below) of the isolates, as well as the Outcome status and the Symptom status of the patients, were generated.

Using the python package community, community louvain.best\_partition (G, resolution = 2.725) produced the 15 unique sample clusters. Where G is the network graphic and resolution determine the number of communities. Each Community was then separated and colored as mentioned above.

To identify the nucleotide mutation each sample was compared to the reference sequence of SARS-CoV-2 (Accession #: NC\_045512.2) since this is the ancestor of all sequences in the dataset used in this study. Each sample was compared point by pint and all differences were logged. For the amino acid changes a very similar process took place. This conversion took the reference sequence amino acids and compared to the sample amino acids.

1. **Identification of cluster-specific mutational profile**

**3a. Selection of community founder(s)**

The first step in setting up a mutation profile is to select the community founders. Founder were selected based on network position within each cluster.

(VC) Add: Some more details about the rules of selecting the founders (e.g. the definition of a founder).

**3b. Median Vector generation**

1. **(JF)Method 1:**

Using two parent samples (or a parent and a child sample) and the mutational differences from the median vector it was possible to accurately predict the median vector sequence. This was done by producing a copy of the sequence that is closest (Least mutations) to the median vector. This functioned as the base sequence. Then using the second sequence we replaced the mutations in sequence 1 with what the nucleotide should be. This process is repeated until all mutations are replaced.

The program works on the principle of mutational pathway. We have two possibilities with our samples. The first is that both sample A and Sample B are parents of the median vector. The mutations in A will not be the same as B. Letting us remove mutations from A based on non-mutated B and vise-versa. The other scenario is that sample A is a parent and sample B is a child, this can be calculated in a similar way. The only scenario where this does not work is if there are identical mutations, or two child nodes are used since we cannot assume node heritage.

There are checks in place to ensure the sequence is properly generated. The program checks for no change in nucleotide status indicating two children are given or the two sequences have the same mutation. There are also checks in place to insure there are no other user errors in the program.

**ii. (YZ) Method 2:**

To generate all the median vectors, we used a recursive method. This recursive method is based the Method 1 mentioned above. For a median vector node in the network, we will find its two neighbors to generate the sequence for this median vector sequence by Method 1. The two neighbors that are not median vectors will have high priority to be found. If the neighbors are still median vectors, we will call the function that generate the sequence recursively.

Here we use M1 as the function of Method 1 in the pseudocode. Here is the pseudocode for this pipeline:

Def generate\_seq(Node):

If Node is not a median vectoer:

Return sequence\_for\_this\_node

// We can find this information from the sequence database

If node is a median vector:

Neighbor\_1, Neighbor\_2= Find\_two\_neighbors(Node)

// it will return two neighbors for the Node

//We sort the neighbors and the neighbors that are not median vector will //have high priority to be found

Neighbor\_1\_seq=generate\_seq(Neighbor\_1)

Neighbor\_2\_seq=generate\_seq(Neighbor\_2)

Node\_seq=M1(Neighbor\_1\_seq, Neighbor\_2\_seq)

//Here is the recursive part. We call the generate\_seq function and M1 //function recursively

Return Node\_seq

**iii. (JF)Validation of methods:**

Both generated median vector sequences were put into separate text files. The two files were then compared with the Linux diff function. The program showed no differences in the tested median vectors.

**3c. Identification of mutations in community founder(s)**

Once the founders were selected each mutation must be identified for each sample and cluster. First each sequence was compared to the reference. Since each sequence is the same length and aligned, it is a one-to-one comparison down the entire sequence. Then when all mutations are identified and logged the DNA sequences are converted into amino acid sequences. The amino acid sequence is compared to the reference again, and all differences are logged. Then an amino level analysis can be performed to assess the differences in samples.

**3d. Selection of viral sequences related to the Alpha, Beta and Gamma variants**

(VC) Add: Concise description of these viral isolates were identified.

1. **Availability of the dataset and computer codes (ALL)**

The dataset and computer codes developed and used in this study have been made publicly available by submitting to the GitHub website (<https://github.com/JonFeige/Covid_project>)

**Result**

1. **Generation of master dataset**

**1a. General information of the master dataset**

The sequences that are shown in the master dataset are those that have passed out strenuous quality control.

(TD) Add: Concise description of the dataset (e.g. the number of sequences available after each step).

1. Manual sequence retrieval (Complete = True / Patient Status = True / Host = Human / Location = North America)
2. Remove sequences with no metadata.
3. Filter non-US sequences and those with no state information.
4. Remove sequences without adequate information on collection date.
5. Remove sequences without adequate information on symptom and outcome.
6. Filter sequences with <29750 ACGT's & >150 N's

* Final number of sequences in the dataset.

**1b. Statistics:**

Table \_. General information of the COVID-19 patients in the dataset.

|  |  |  |
| --- | --- | --- |
| **Patient, Location and Collection Date Information** | | **Percentage** |
| Gender | Male | 48.11% |
| Female | 37.65% |
| Unknown | 14.24% |
| Age | <17 | 1.60% |
| 17 - 37 | 18.75% |
| 38 - 58 | 25.15% |
| 59 - 79 | 25.73% |
| >80 | 12.06% |
| Unknown | 16.72% |
| Symptom Status | Asymptomatic | 6.54% |
| Symptomatic | 60.90% |
| Unknown | 32.56% |
| Outcome | Alive | 57.27% |
| Deceased | 20.49% |
| Unknown | 22.24% |
| Location | California | 47.9651163 |
| Florida | 0.14534884 |
| Louisiana | 5.81395349 |
| Massachusetts | 1.74418605 |
| Montana | 5.81395349 |
| New York | 17.2965116 |
| Puerto Rico | 14.0988372 |
| South Carolina | 3.19767442 |
| Texas | 3.34302326 |
| Utah | 0.14534884 |
| Washington | 0.29069767 |
| Wyoming | 0.14534884 |
| Collection Date | February - March 2020 | 16.2790698 |
| April 2020 | 26.4534884 |
| May - June 2020 | 4.50581395 |
| July 2020 | 8.5755814 |
| August - September 2020 | 7.55813953 |
| October - November 2020 | 5.95930233 |
| December 2020 | 25.7267442 |
| January - February 2021 | 4.94186047 |

The genomic sequences of SARS-CoV-2 reference (NC\_045512.2), Alpha Variant (EPI\_ISL\_601443), Beta Variant (EPI\_ISL\_700428) and Gamma Variant (EPI\_ISL\_792680) were added to the dataset as references.

1. **Phylogenetic analysis and community discovery**

**(TD) 2a. General characteristics** **of the genomic sequences after data processing**

* 1. Describe percentage of A,C,G,T
     1. Percent of A Nucleotides: 29.8316104%
     2. Percent of C Nucleotides: 18.3437434%
     3. Percent of G Nucleotides: 19.6071348%
     4. Percent of T Nucleotides: 32.1188133%
  2. Different statistical measure.
     1. Percent of Ambiguous Calls: 0.00378578%
     2. Percent of Non-A count in Poly-A tail: 0.04776026%

Question: Is there anything else that should be mentioned?

**2b. Multiple sequence alignment**

* + 1. Trimming of the 5’ and 3’ sequences

Trimmed the 5’ end and 3’ end separately. Find the sequence that has the most no calls at the 5’ end of the sequence and eliminate the number of no calls from the beginning of every sequence in order to eliminate noise. After eliminating the poly(A) tail from all sequences, find the sequence that has the most no calls at the 3’ end of the sequence and eliminate the number of no calls from the end of every sequence in order to eliminate noise.

* + 1. mafft --inputorder --anysymbol --kimura 1 --ep 0.5 --nwildcard --addfragments SequencesToBeAligned.fasta --memsave --6merpair referenceSequence.fasta > output.fasta

(TD) Add: Concise description of the result of sequence alignment (e.g. which sections are more conserved, and which sections have more mutations).



Comment: This figure may be shown as supplementary data.

**2c. Phylogenetic network of SARS-CoV-2 isolates**

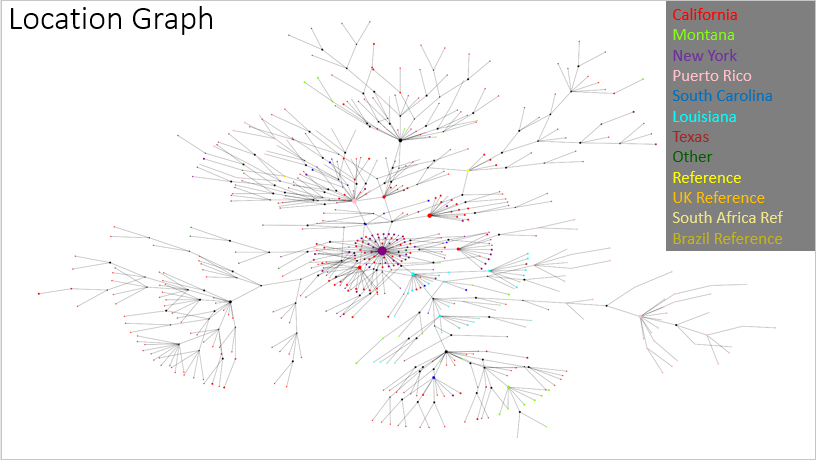


Figure \_. Phylogenetic network graph with nodes showing location information. Color scheme: California - Red; Montana - Lawn green; New York - Purple; Puerto Rico - Pink; South Carolina - Blue; Louisiana - Cyan; Texas - Brown; Other State - Green; Reference - Yellow; Alpha Variant - Gold; Beta Variant - Khaki; and Gamma Variant - khaki.

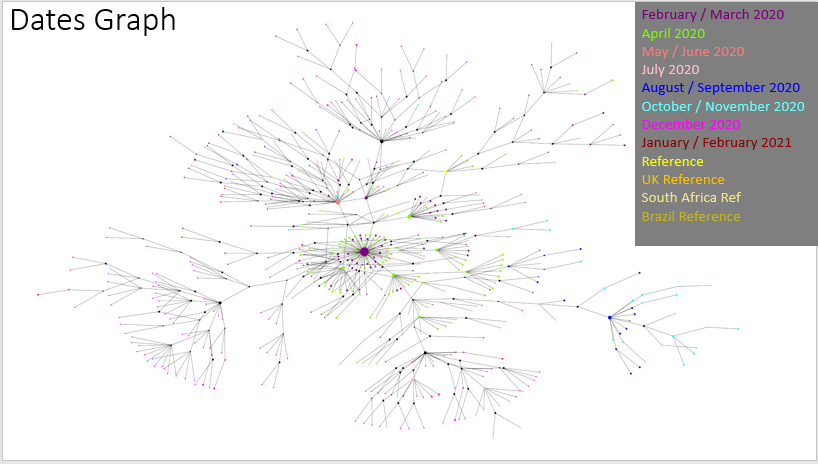


Figure \_. Phylogenetic network graph with nodes showing location information. Color scheme: Reference - Yellow; Alpha Variant - Gold; Beta Variant - Khaki; Gamma Variant – khaki; Feb/Mar 2020 - Purple; Apr 2020 - Lawn green; May/Jun 2020 - Light coral; Jul 2020 - Pink; Aug/Sep 2020 - Blue; Oct/Nov 2020 - Cyan; Dec 2020 - Fuchsia; Jan/Fe 2021 - Dark red; and Other (unknown?) - Green.

**2d. Community discovery - identification of 15 clusters**

(VC) Add: Concise description of the relationship between the clusters.

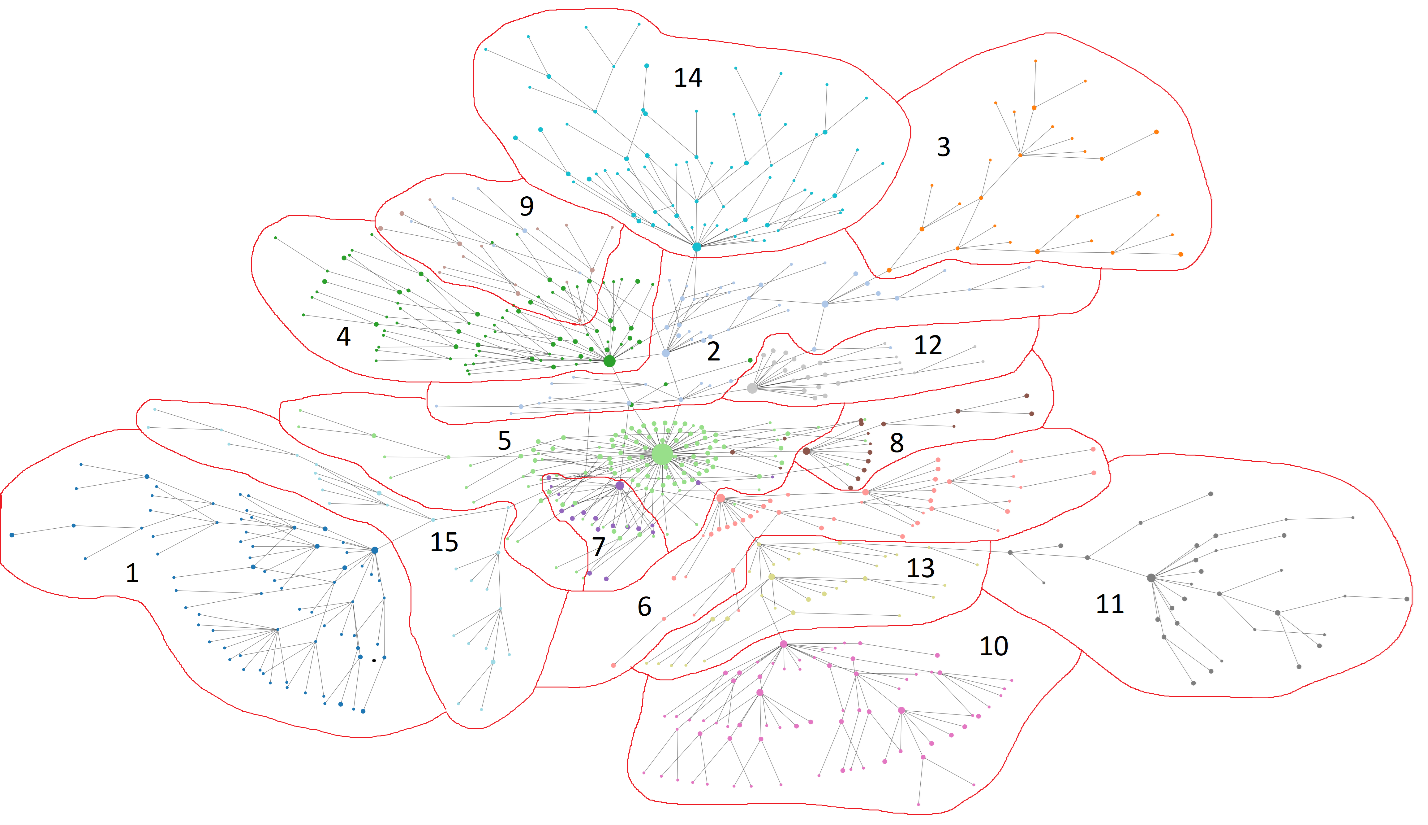


Figure \_. The network diagram showing 15 communities labeled by different colors. Color scheme: Cluster 1 – Dark Blue, Cluster 2 – Sky Blue, Cluster 3 – Orange, Cluster 4 – Dark Green, Cluster 5 – Light Green, Cluster 6 – Pink, Cluster 7 – Purple, Cluster 8 – Brown, Cluster 9 – Tan, Cluster 10 – Violet, Cluster 11 – Dark Gray, Cluster 12 – Gray, Cluster 13 – Gold, Cluster 14 – Cyan, Cluster 15 – Light Blue

(TD) Add: Concise description of the characteristics of each cluster.

Table \_. Patient information of the SARS-CoV-2 isolates in the dataset. Values are percentage (%).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Gender | | | Age | | | | | |
| Male | Female | Unknown | <17 | 17 - 37 | 38 - 58 | 59 - 79 | >80 | Unknown |
| ALL | 48.11 | 37.65 | 14.24 | 1.60 | 18.75 | 25.15 | 25.73 | 12.06 | 16.72 |
| Community 1 | 57.89 | 42.11 | 0.00 | 3.51 | 19.30 | 36.84 | 24.56 | 15.79 | 0.00 |
| Community 2 | 48.65 | 48.65 | 2.70 | 0.00 | 13.51 | 21.62 | 48.65 | 8.11 | 8.11 |
| Community 3 | 57.14 | 38.10 | 4.76 | 0.00 | 19.05 | 23.81 | 28.57 | 9.52 | 19.05 |
| Community 4 | 53.13 | 29.69 | 17.19 | 0.00 | 26.56 | 18.75 | 23.44 | 12.50 | 18.75 |
| Community 5 | 62.26 | 34.91 | 2.83 | 0.94 | 6.60 | 29.25 | 36.79 | 21.70 | 4.72 |
| Community 6 | 34.15 | 39.02 | 26.83 | 0.00 | 24.39 | 17.07 | 17.07 | 17.07 | 39.02 |
| Community 7 | 55.56 | 44.44 | 0.00 | 0.00 | 22.22 | 44.44 | 16.67 | 16.67 | 0.00 |
| Community 8 | 52.94 | 23.53 | 23.53 | 0.00 | 17.65 | 5.88 | 35.29 | 17.65 | 23.53 |
| Community 9 | 8.33 | 0.00 | 91.67 | 0.00 | 8.33 | 0.00 | 0.00 | 0.00 | 91.67 |
| Community 10 | 50.00 | 50.00 | 0.00 | 3.70 | 29.63 | 27.78 | 24.07 | 14.81 | 0.00 |
| Community 11 | 0.00 | 0.00 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100.00 |
| Community 12 | 61.90 | 33.33 | 4.76 | 9.52 | 4.76 | 28.57 | 28.57 | 23.81 | 4.76 |
| Community 13 | 52.17 | 43.48 | 4.35 | 0.00 | 13.04 | 47.83 | 34.78 | 0.00 | 4.35 |
| Community 14 | 47.06 | 41.18 | 11.76 | 1.96 | 29.41 | 19.61 | 27.45 | 9.80 | 11.76 |
| Community 15 | 29.41 | 70.59 | 0.00 | 0.00 | 29.41 | 35.29 | 23.53 | 23.53 | 0.00 |

Table \_. Symptom status and outcome of patients in the dataset. Values are percentage (%).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Symptom Status | | | Outcome | | |
| Asymptomatic | Symptomatic | Unknown | Alive | Deceased | Unknown |
| ALL | 6.54% | 60.90% | 32.56% | 57.27% | 20.49% | 22.24% |
| Community 1 | 1.75 | 36.84 | 61.40 | 87.72 | 10.53 | 1.75 |
| Community 2 | 18.92 | 56.76 | 24.32 | 70.27 | 18.92 | 10.81 |
| Community 3 | 23.81 | 66.67 | 9.52 | 71.43 | 9.52 | 19.05 |
| Community 4 | 9.38 | 65.63 | 25.00 | 48.44 | 31.25 | 20.31 |
| Community 5 | 3.77 | 89.62 | 6.60 | 36.79 | 49.06 | 14.15 |
| Community 6 | 4.88 | 56.10 | 39.02 | 48.78 | 0.00 | 51.22 |
| Community 7 | 27.78 | 66.67 | 5.56 | 77.78 | 16.67 | 5.56 |
| Community 8 | 0.00 | 88.24 | 11.76 | 23.53 | 35.29 | 41.18 |
| Community 9 | 0.00 | 91.67 | 8.33 | 8.33 | 0.00 | 91.67 |
| Community 10 | 1.85 | 14.81 | 83.33 | 94.44 | 3.70 | 1.85 |
| Community 11 | 0.00 | 100.00 | 0.00 | 0.00 | 0.00 | 100.00 |
| Community 12 | 0.00 | 80.95 | 19.05 | 80.95 | 14.29 | 4.76 |
| Community 13 | 0.00 | 43.48 | 56.52 | 60.87 | 13.04 | 26.09 |
| Community 14 | 0.00 | 43.14 | 56.86 | 68.63 | 9.80 | 21.57 |
| Community 15 | 0.00 | 41.18 | 58.82 | 82.35 | 11.76 | 5.88 |

Table \_. Collection location of the SARS-CoV-2 isolates in the dataset. Values are percentage (%).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | California | Florida | Louisiana | Massachusetts | Montana | New York | Puerto Rico | South Carolina | Texas | Utah | Washington | Wyoming |
| ALL | 47.9651163 | 0.14534884 | 5.81395349 | 1.74418605 | 5.81395349 | 17.2965116 | 14.0988372 | 3.19767442 | 3.34302326 | 0.14534884 | 0.29069767 | 0.14534884 |
| Community 1 | 98.245614 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.75438596 | 0 | 0 | 0 |
| Community 2 | 70.2702703 | 0 | 0 | 5.40540541 | 2.7027027 | 13.5135135 | 2.7027027 | 0 | 5.40540541 | 0 | 0 | 0 |
| Community 3 | 61.9047619 | 0 | 0 | 0 | 4.76190476 | 14.2857143 | 0 | 0 | 14.2857143 | 0 | 4.76190476 | 0 |
| Community 4 | 37.5 | 0 | 0 | 0 | 1.5625 | 28.125 | 17.1875 | 9.375 | 4.6875 | 1.5625 | 0 | 0 |
| Community 5 | 40.5660377 | 0.94339623 | 0 | 6.60377358 | 0 | 45.2830189 | 3.77358491 | 1.88679245 | 0.94339623 | 0 | 0 | 0 |
| Community 6 | 2.43902439 | 0 | 36.5853659 | 0 | 17.0731707 | 0 | 26.8292683 | 4.87804878 | 12.195122 | 0 | 0 | 0 |
| Community 7 | 94.4444444 | 0 | 0 | 0 | 5.55555556 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Community 8 | 11.7647059 | 0 | 0 | 0 | 0 | 58.8235294 | 23.5294118 | 5.88235294 | 0 | 0 | 0 | 0 |
| Community 9 | 0 | 0 | 0 | 0 | 0 | 0 | 91.6666667 | 8.33333333 | 0 | 0 | 0 | 0 |
| Community 10 | 70.3703704 | 0 | 0 | 0 | 20.3703704 | 1.85185185 | 0 | 1.85185185 | 3.7037037 | 0 | 0 | 1.85185185 |
| Community 11 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| Community 12 | 80.952381 | 0 | 0 | 0 | 0 | 14.2857143 | 4.76190476 | 0 | 0 | 0 | 0 | 0 |
| Community 13 | 13.0434783 | 0 | 56.5217391 | 0 | 13.0434783 | 8.69565217 | 4.34782609 | 4.34782609 | 0 | 0 | 0 | 0 |
| Community 14 | 66.6666667 | 0 | 0 | 0 | 13.7254902 | 7.84313725 | 9.80392157 | 0 | 0 | 0 | 1.96078431 | 0 |
| Community 15 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table \_. Collection date of the SARS-CoV-2 isolates in the dataset. Values are percentage (%).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Feb-Mar 2020 | Apr 20 | May-Jun 2020 | Jul 20 | Aug-Sep 2020 | Oct-Nov 2020 | Dec 20 | Jan-Feb 2021 |
| ALL | 16.2790698 | 26.4534884 | 4.50581395 | 8.5755814 | 7.55813953 | 5.95930233 | 25.7267442 | 4.94186047 |
| Community 1 | 0 | 0 | 0 | 0 | 0 | 1.75438596 | 96.4912281 | 1.75438596 |
| Community 2 | 13.5135135 | 48.6486486 | 5.40540541 | 2.7027027 | 0 | 5.40540541 | 21.6216216 | 2.7027027 |
| Community 3 | 76.1904762 | 23.8095238 | 0 | 0 | 0 | 0 | 0 | 0 |
| Community 4 | 10.9375 | 17.1875 | 4.6875 | 14.0625 | 7.8125 | 6.25 | 17.1875 | 21.875 |
| Community 5 | 31.1320755 | 47.1698113 | 4.71698113 | 2.83018868 | 3.77358491 | 0 | 4.71698113 | 5.66037736 |
| Community 6 | 7.31707317 | 43.902439 | 4.87804878 | 17.0731707 | 24.3902439 | 2.43902439 | 0 | 0 |
| Community 7 | 27.7777778 | 66.6666667 | 5.55555556 | 0 | 0 | 0 | 0 | 0 |
| Community 8 | 11.7647059 | 47.0588235 | 11.7647059 | 5.88235294 | 5.88235294 | 17.6470588 | 0 | 0 |
| Community 9 | 0 | 0 | 0 | 50 | 50 | 0 | 0 | 0 |
| Community 10 | 0 | 0 | 1.85185185 | 20.3703704 | 0 | 5.55555556 | 68.5185185 | 3.7037037 |
| Community 11 | 0 | 0 | 0 | 6.4516129 | 35.483871 | 58.0645161 | 0 | 0 |
| Community 12 | 52.3809524 | 19.047619 | 0 | 4.76190476 | 0 | 0 | 19.047619 | 4.76190476 |
| Community 13 | 0 | 56.5217391 | 8.69565217 | 13.0434783 | 0 | 0 | 13.0434783 | 8.69565217 |
| Community 14 | 1.96078431 | 0 | 11.7647059 | 1.96078431 | 7.84313725 | 9.80392157 | 56.8627451 | 9.80392157 |
| Community 15 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 |

The Accession ID and the metadata of the members of each cluster are shown in Supplementary Data Table \_.

1. **Mutation profiles of SARS-CoV-2 isolates in USA**

**(Dr. Chan)** Add: Concise description of the selection of community founders

* What is a founder?
* How are the founders selected?

A found is a node within the community that holds a strong centrality to the other nodes within the community. Typically this is the parent node or the direct children of the parent node.

A median vector (MV) node is a point between three samples that expresses a genetic split from the parent to the two child nodes. The purpose is to show the point in which the two children nodes split with mutations.)

The founder(s) of each cluster are listed in the table below.

Table \_. The founders of all 15 clusters are listed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **Founder** | | | | |
| 1 | MV19 | MV33 | MV23 |  |  |
| 2 | EPI\_ISL\_475633 | MV1 |  |  |  |
| 3 | MV137 | MV58 | MV52 | MV72 | MV5 |
| 4 | EPI\_ISL\_940900 | MV79 | EPI\_ISL\_475632 |  |  |
| 5 | EPI\_ISL\_677673 | EPI\_ISL\_593478 | EPI\_ISL\_475639 | EPI\_ISL\_631724 | MV2 |
| 6 | EPI\_ISL\_436946 | EPI\_ISL\_463295 | EPI\_ISL\_436944 | EPI\_ISL\_940958 |  |
| 7 | EPI\_ISL\_475674 |  |  |  |  |
| 8 | EPI\_ISL\_525658 | EPI\_ISL\_475590 | MV145 |  |  |
| 9 | MV32 |  |  |  |  |
| 10 | MV25 | EPI\_ISL\_516771 | MV41 |  |  |
| 11 | EPI\_ISL\_940918 | EPI\_ISL\_942009 |  |  |  |
| 12 | EPI\_ISL\_475585 |  |  |  |  |
| 13 | MV16 | EPI\_ISL\_463286 |  |  |  |
| 14 | MV24 | MV88 | MV141 | MV92 | MV130 |
| 15 | MV101 | MV86 | MV49 |  |  |

(VC) Add: Description of the characteristics of the founders

Selected founders – either viral isolates sequenced or median vector (MV) nodes.

(VC) Add: Comparison of the founders between different clusters

* Five clusters (1, 3, 9, 14 & 15) – all founders are MV nodes.
* Four clusters (6. 7. 11 & 12) – all founders are viral sequences.
* Six clusters (2, 4, 5, 8, 10 &13) – both viral sequence and MV founders

**3a. Identification of nucleotide and amino acid changes in the founder(s) of each cluster**

To identify the nucleotide mutation each sample was compared to the reference sequence of SARS-CoV-2 (Accession #: NC\_045512.2) since this is the ancestor of all sequences in the dataset used in this study. Each sample was compared point by pint and all differences were logged. For the amino acid changes a very similar process took place. This conversion took the reference sequence amino acids and compared to the sample amino acids.

Based on the mutation profiles, the clusters can be divided into three groups (Table \_).

Cluster 3, the only member of Group 1, appears to be quite different from all other clusters.

Some nodes in Cluster 2 function as the ancestors of the founders of other clusters in Group 2. For example, MV1 is the ancestor of EPI\_ISL\_475585 and EPI\_ISL\_677673, the founders of Cluster 12 and Cluster 5, respectively.

Some members of Cluster 5 have acquired a large number of mutations. Consequently, they became the ancestor of the founders of several clusters, which formed a separate group, Group 3. In fact, this group can be further divided into 3 subgroups, Group 3a (Clusters 7 and 8), Group 3b (1 and 15), and Group 3c (Clusters 6, 10, 11 and 13).

Table \_. Mutation-based grouping of SARS-CoV-2 clusters.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Group** | **Member** | | | | | |
| Group 1 | Cluster 3 |  |  |  |  |  |
| Group 2 | Cluster 2\* | Cluster 12 | Cluster 5 | Cluster 14 | Cluster 4 | Cluster 9 |
| Group 3a | Cluster 5\* | Cluster 7 | Cluster 8 |  |  |  |
| Group 3b | Cluster 5\* | Cluster 15 | Cluster 1 |  |  |  |
| Group 3c | Cluster 5\* | Cluster 6 | Cluster 13 | Cluster 10 | Cluster 11 |  |

**3b. Group 1 clusters and mutations**

As shown in the table below, MV52/MV5 and MV72 represent two independent lines, with different mutation profiles. This result is consistent with the pattern shown in Figure \_, which showed that viral sequences of these two lineages have some degree of location specificity.

Table \_. Group 1 mutations identified in the founders of Cluster 3.



*Red: missense mutations; black: synonymous mutations; blue: mutations in the untranslated region (UTR).*

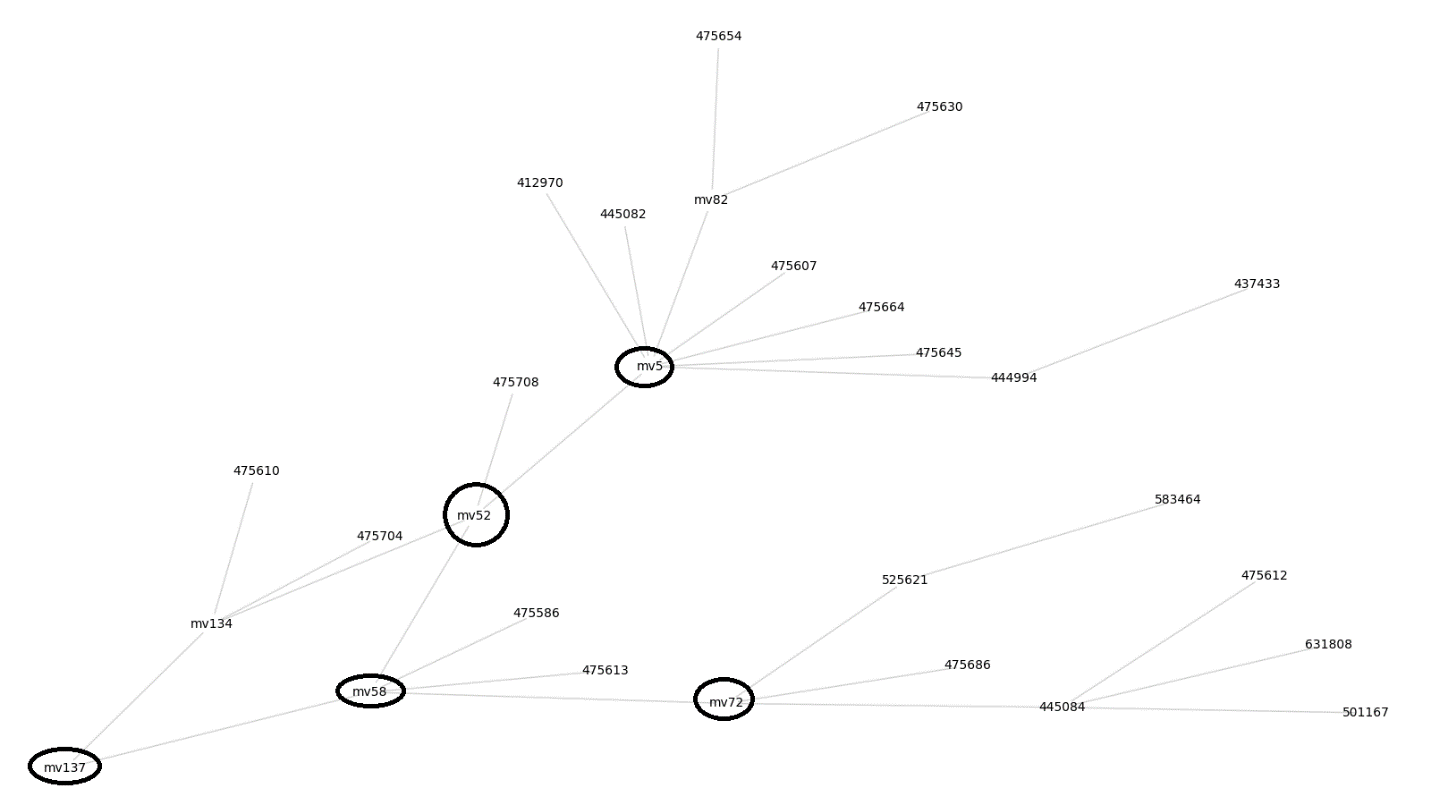


Figure \_. Network graph of Cluster 3. There are five founders (circled) in this cluster, MV137, MV58, MV52, MV72, and MV5.

**3c. Group 2 clusters and mutations**

There are six clusters in Group 2. Two nodes in Cluster 2, EPI\_ISL\_475633 and MV1, appeared to be the ancestors of the founders of the other five member clusters (Clusters 4, 5, 9, 12 and 14). Each of these clusters has acquired a specific combination of mutations. For example, the founders of Cluster 5 (EPI\_ISL\_677673, EPI\_ISL\_475639, EPI\_ISL\_631724, MV2 and EPI\_ISL\_593478) have two unique mutations Nsp2 T85I (PP1a/PP1ab T265I) and in Protein 3a Q57H. Similarly, the founders of Cluster 4, EPI\_ISL\_940900, EPI\_ISL\_475632 and MV79 have acquired the RG203-204KR mutations in the nucleocapsid, due to the GGG>AAC (28881-28883) mutation in the viral genome. The mutation profiles of the Group 2 clusters are shown in Table \_. The network graphs of the clusters in Group 2 are shown below.

Table \_. Group 2 mutations identified in Clusters 2, 4, 5, 9, 12 and 14.



*Red: missense mutations; black: synonymous mutations; blue: mutations in the untranslated region (UTR).*

i. Cluster 2

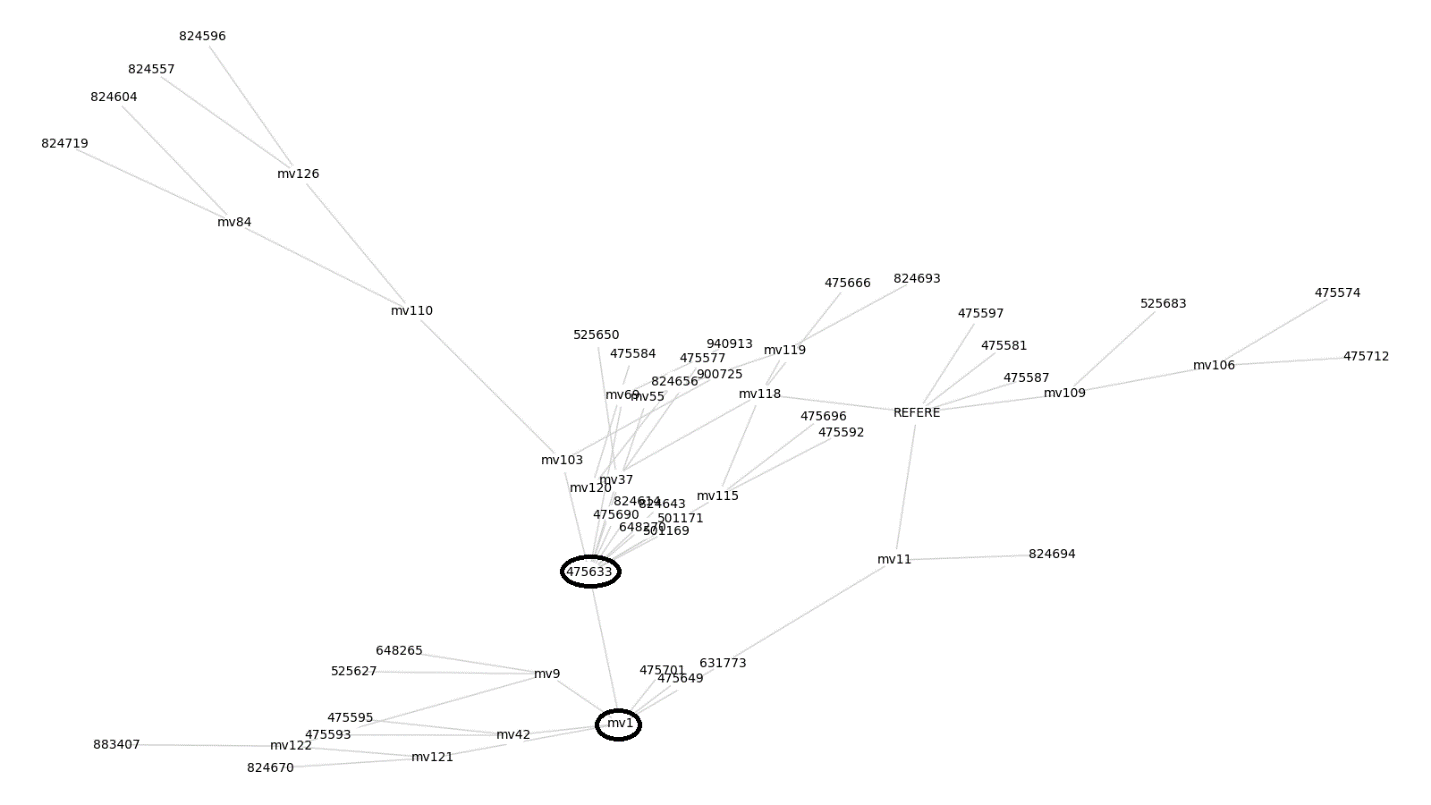


Figure \_. Network graph of Cluster 2. There are two founders (circled) in this cluster, EPI\_ISL\_475633 and MV1.

ii. Cluster 12

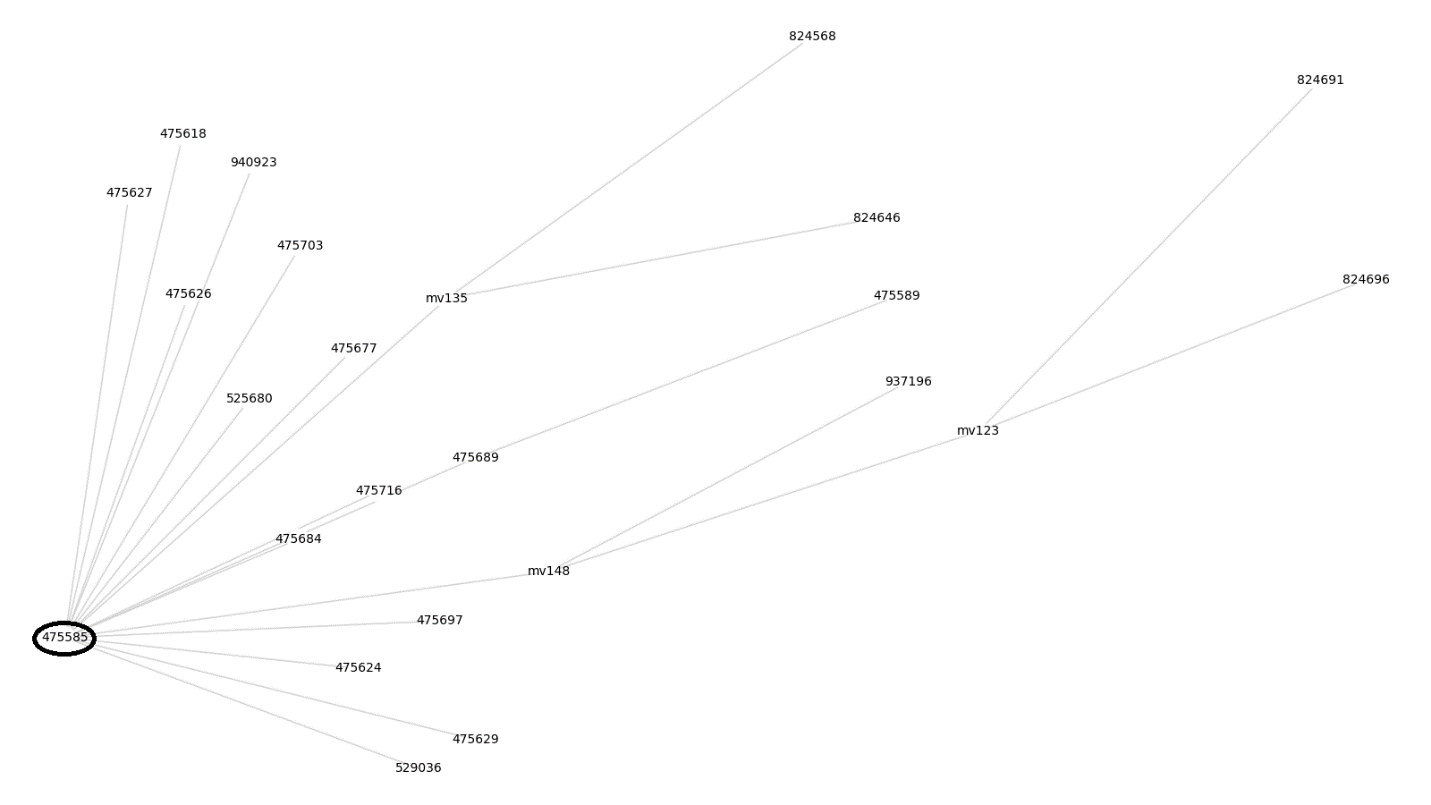


Figure \_. Network graph of Cluster 12. There is only one founder (circled) in this cluster, EPI\_ISL\_475585.

* 1. Cluster 5

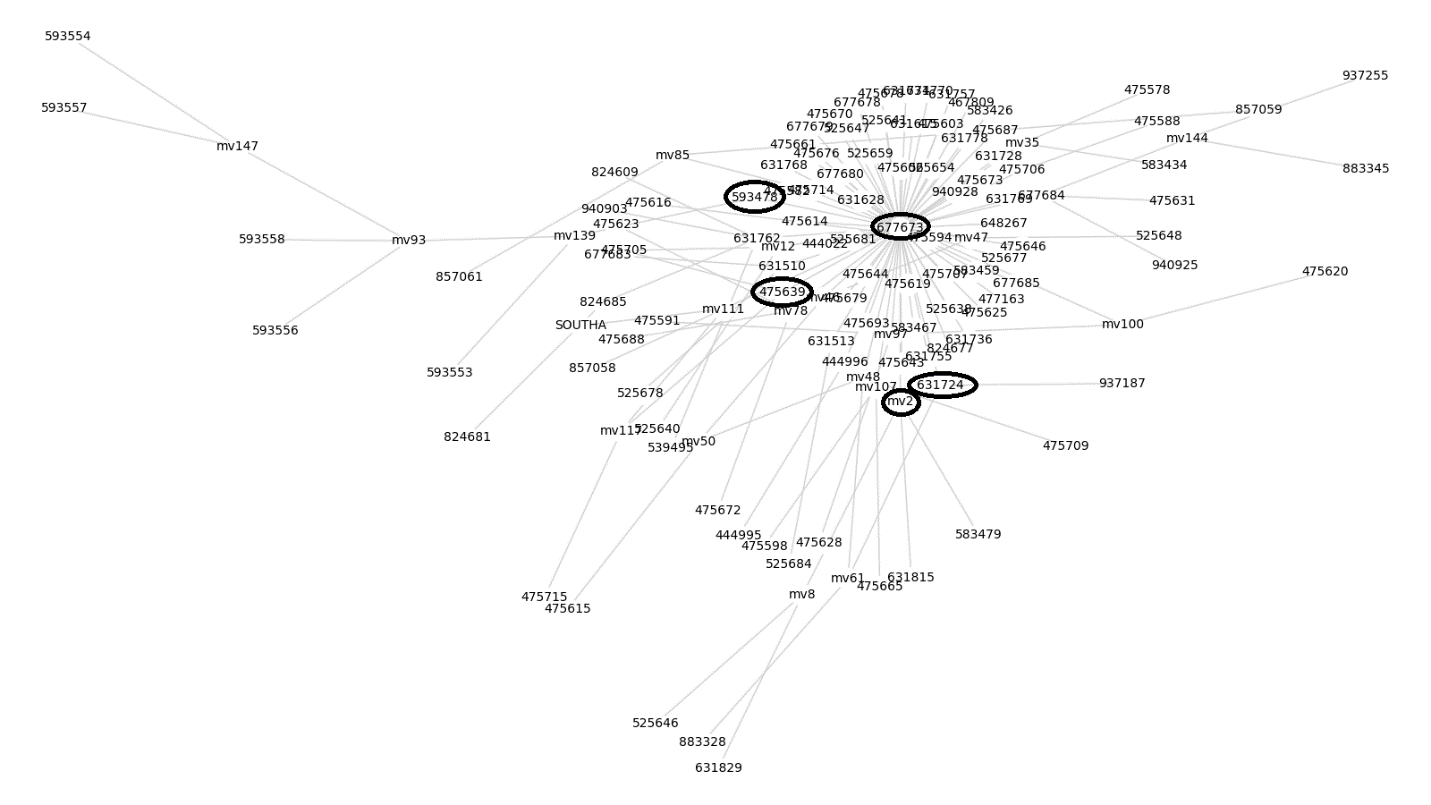


Figure \_. Network graph of Cluster 5. There are five founders (circled) in this cluster, EPI\_ISL\_677673, EPI\_ISL\_593478, EPI\_ISL\_475639, EPI\_ISL\_631724, and MV2.

* 1. Cluster 14

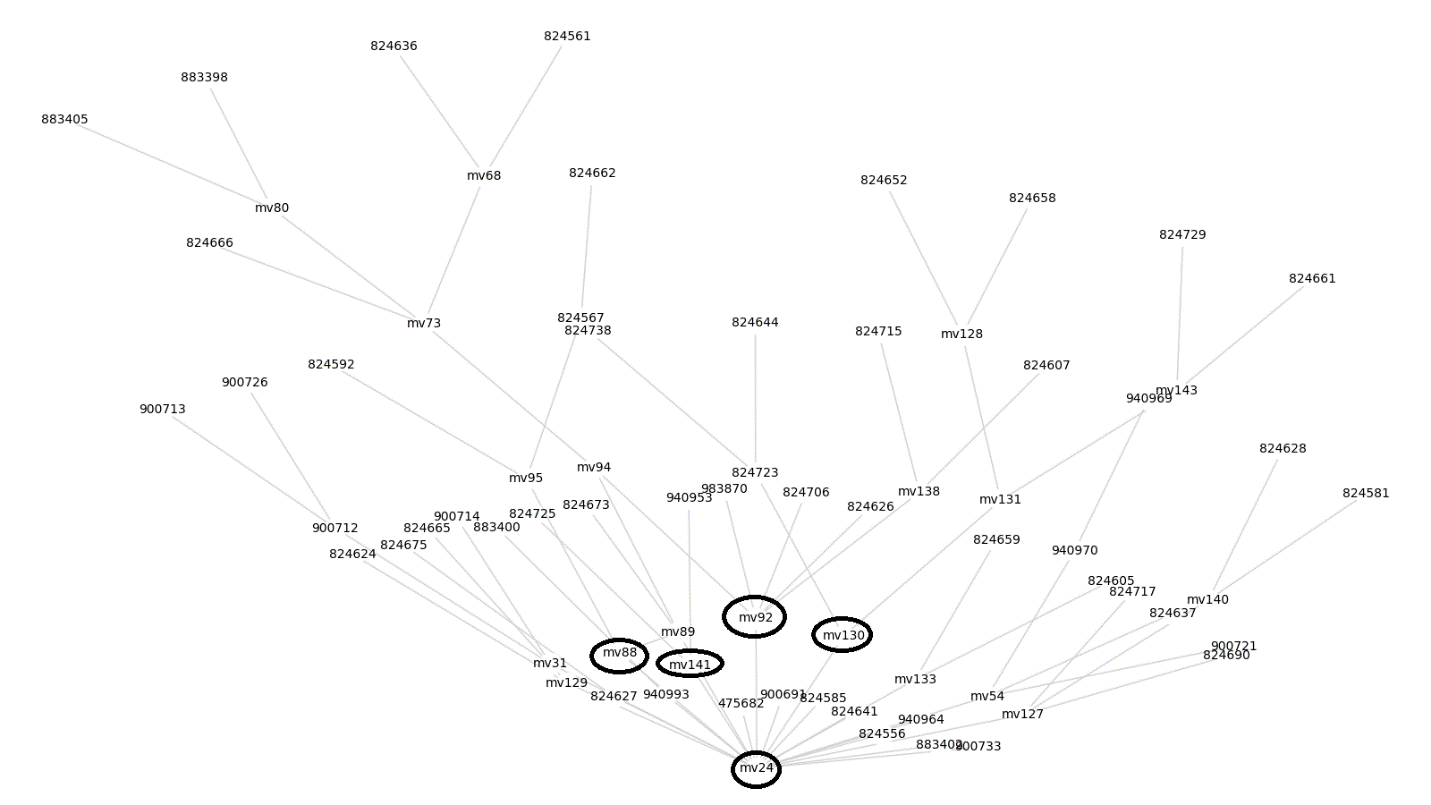


Figure \_. Network graph of Cluster 14. There are five founders (circled) in this cluster, MV24, MV88, MV141, MV92, and MV130.

v. Cluster 4

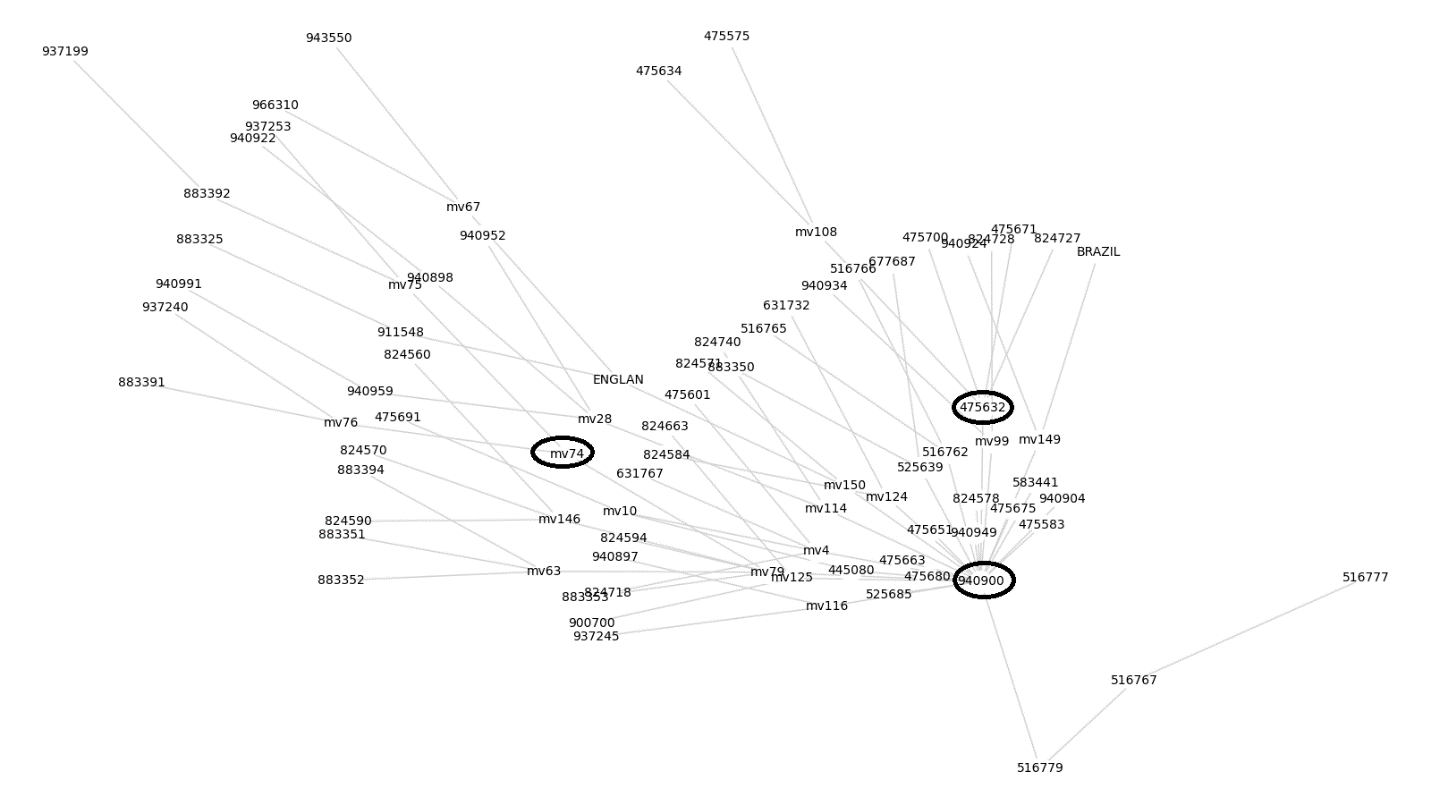


Figure \_. Network graph of Cluster 4. There are three founders (circled) in this cluster, EPI\_ISL\_940900, MV79, and EPI\_ISL\_475632.

vi. Cluster 9

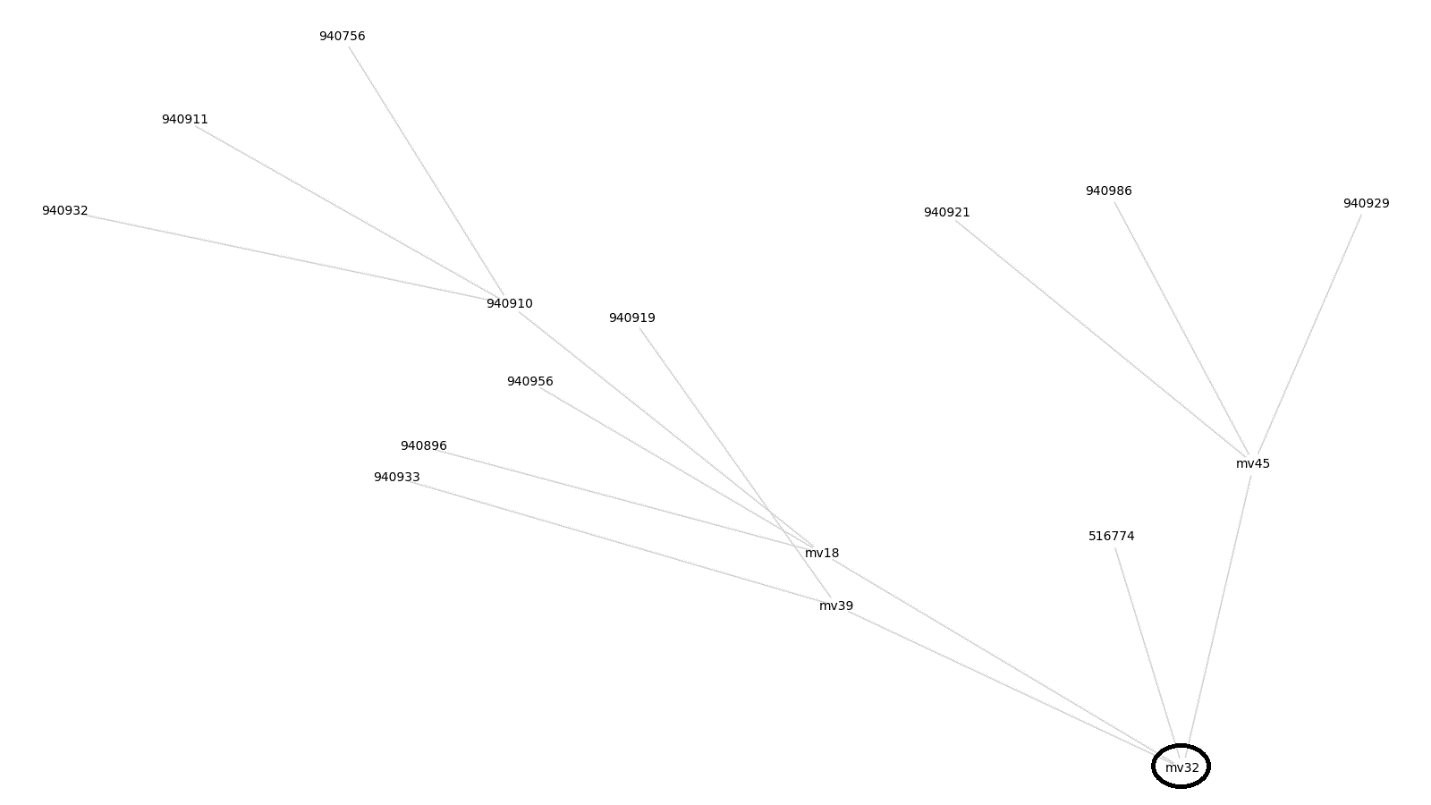


Figure \_. Network graph of Cluster 9. There is only one founder (circled) in this cluster, MV32.

**3d. Group 3 clusters and mutations**

As shown in Figure \_, the main founder of Cluster 5, EPI\_ISL\_677673 is highly connected. Based on the mutation profiles, some of these connected nodes became the founders of Clusters 6, 7, 8 and 15. These clusters, together with their descendants (Clusters 1, 10, 11 and 13), as well as Cluster 5, formed a separate group (i.e. Group 3) that contains different mutation profiles from that of Groups 1 and 2. Further analysis revealed that Group 3 can be divided into 3 subgroups. Group 3a contain two members, Cluster 7 and Cluster 8, in addition to Cluster 5.

1. **Group 3a**

EPI\_ISL\_475674, the founder of Cluster 7, has a synonymous mutation in V38 of Nsp1 (PP1a/PP1ab) , while the founders of Cluster 8 (EPI\_ISL\_525658, EPI\_ISL\_475590 and MV145) have two missense mutations, Nsp7 S25L (PP1a/PP1ab S3884L) and Nsp14 A320V (PP1ab A6245V) and a G>A at position 29540 (UTR 10). The mutation profiles of Clusters 7 and 8 are shown in Table \_. The network graphs of the Clusters 7 and 8 are shown below.

Table \_. Group 3a mutations identified in Clusters 5, 7 and 8.



*Red: missense mutations; black: synonymous mutations; blue: mutations in the untranslated region (UTR).*

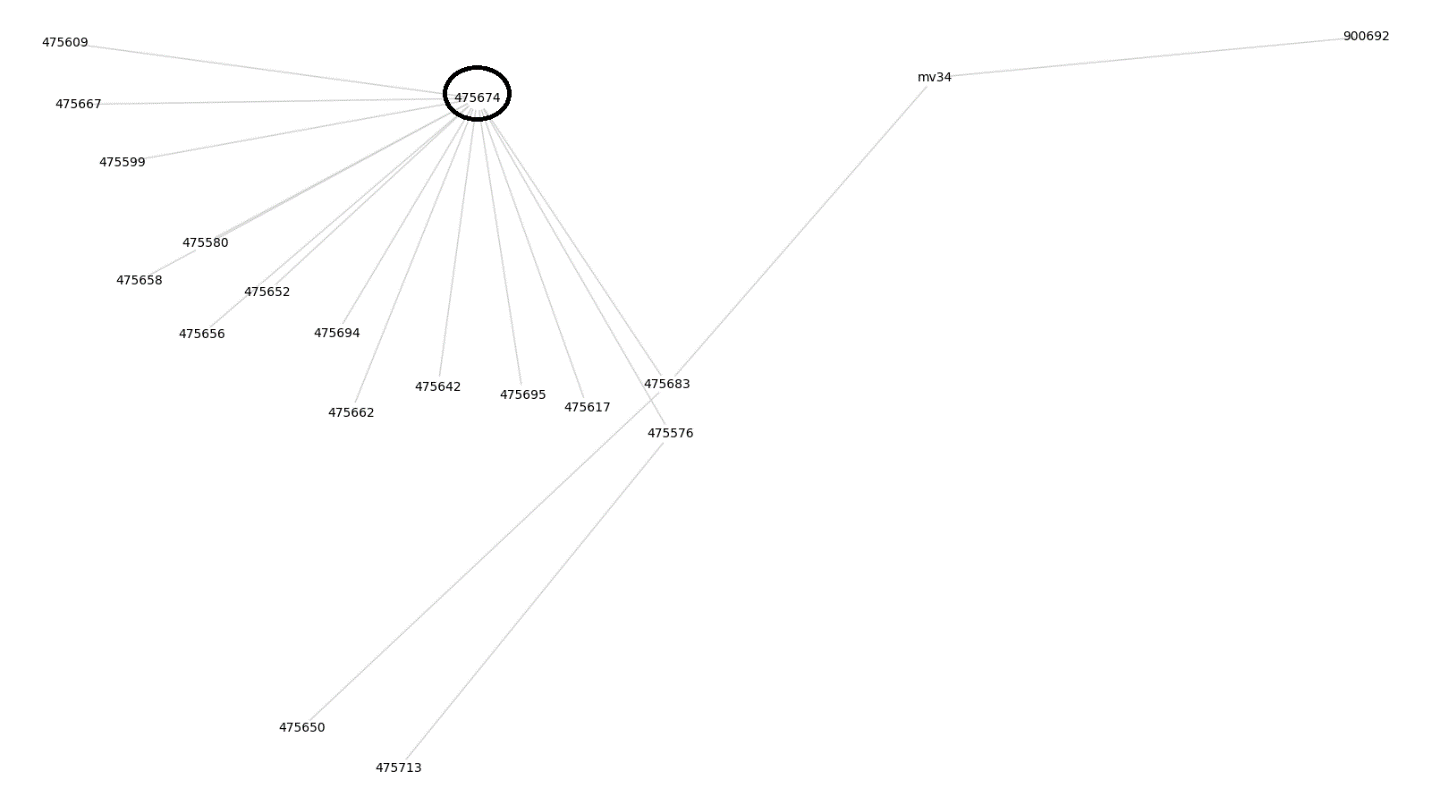


Figure \_. Network graph of Cluster 7. There is only one founder (circled) in this cluster, EPI\_ISL\_475674.

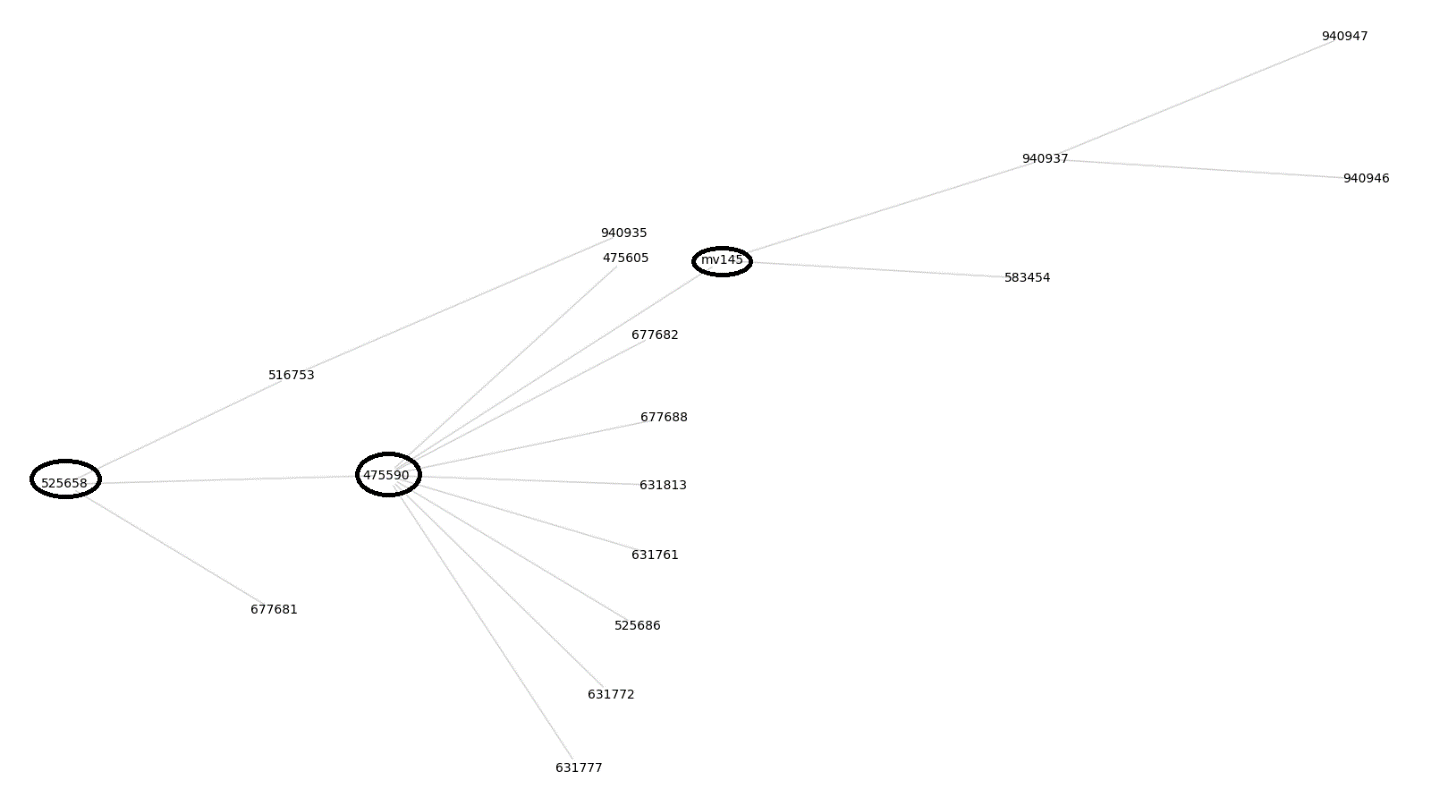


Figure \_. Network graph of Cluster 8. There are three founders (circled) in this cluster, EPI\_ISL\_525658, EPI\_ISL\_475590, and MV145.

1. **Group 3b**

Cluster 5 also gave rise of Cluster 15, which subsequently gave rise to Cluster 1. Together, these clusters form the Group 3b, with a unique mutation profiles different from that of other groups. For example, the founders of Cluster 15, MV101, MV49 and MV86 contain eight novel mutations, in addition to the mutations occurred in Cluster 5. These include five missense mutations, (Nsp13 D260Y (PP1ab D5584Y), Spike protein S13I, W152C and L452R, and nucleocapsid T205I) and three synonymous mutations in the membrane protein (F53), the nucleocapsid (F363) and UTR 9 (28272 A>T). MV86 also contains 5 novel mutations that are not shared by MV101 or MV49, but are also detectable in MV19, MV33 and MV23, the founders of Cluster 1, confirming that MV86 is the ancestor of Cluster 1, as observed in the phylogenetic network. Interestingly, all these are synonymous mutations except one, Nsp9 I65V (PP1a / PP1ab I4205V). Table \_ shows the mutation profiles of Group 3b clusters. The network graphs of the Clusters 15 and 1 are shown below.

Table \_. Group 3b mutations identified in Clusters 1, 5 and 15.



*Red: missense mutations; black: synonymous mutations; blue: mutations in the untranslated region (UTR).*

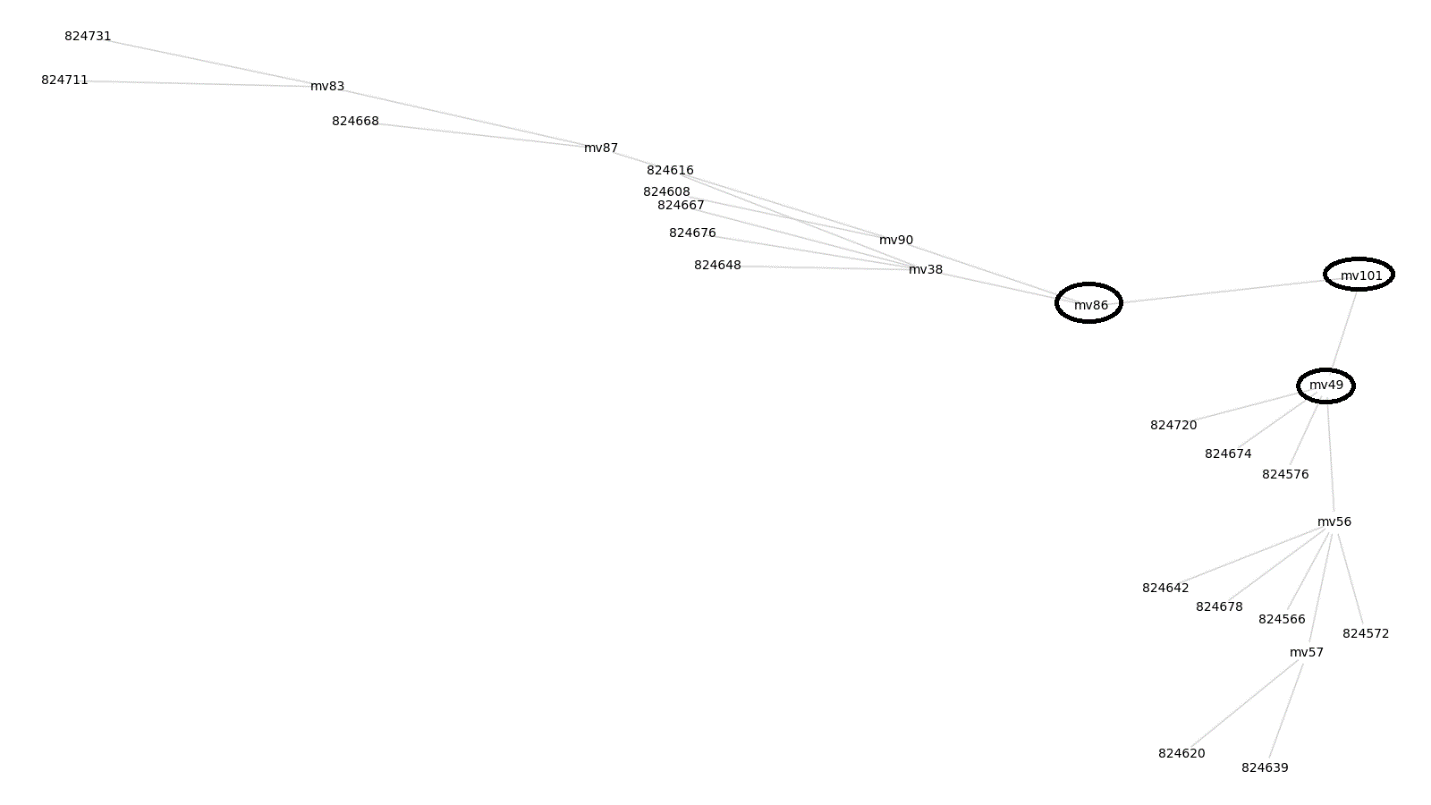


Figure \_. Network graph of Cluster 15. There are three founders (circled) in this cluster, MV101, MV86, and MV49.

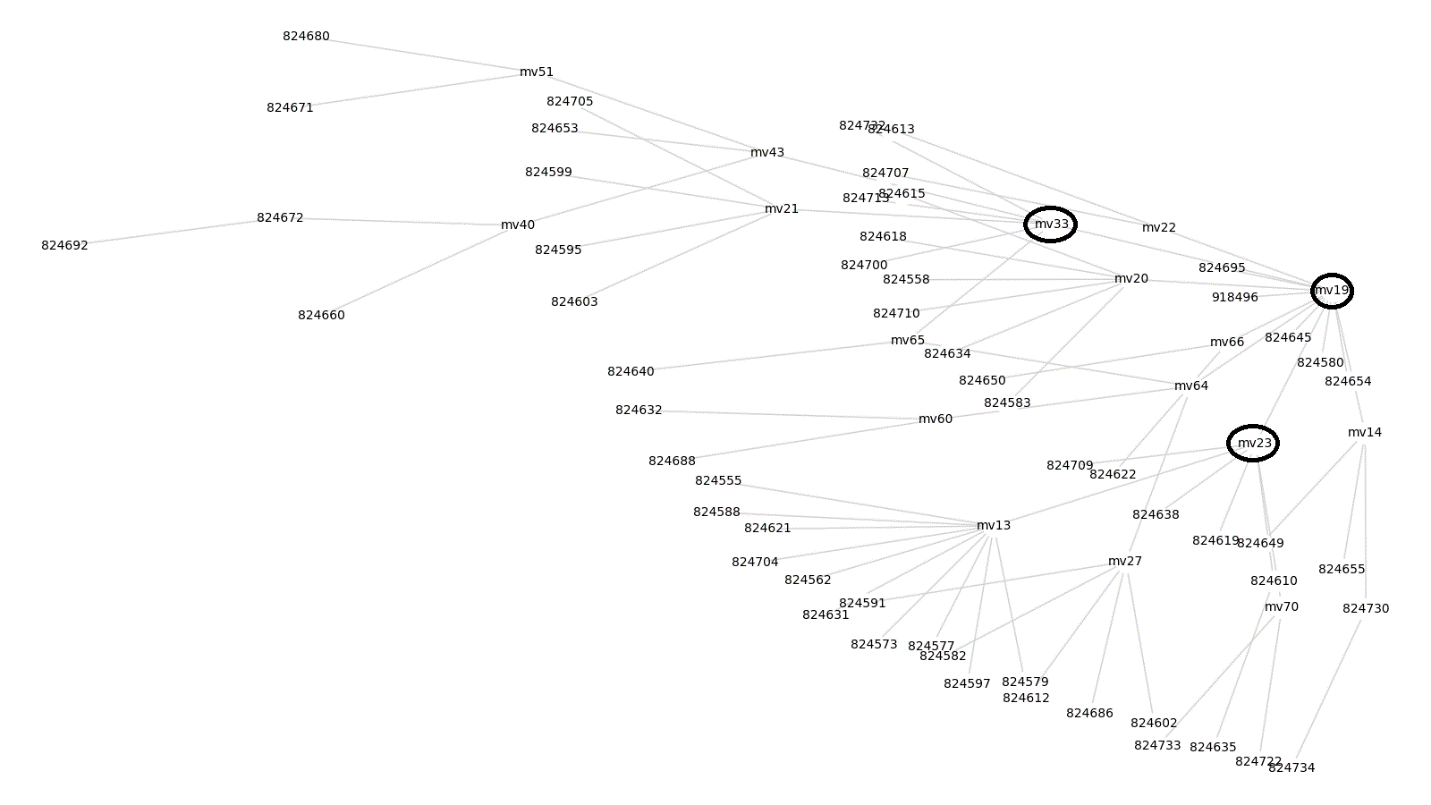


Figure \_. Network graph of Cluster 1. There are three founders (circled) in this cluster, MV19, MV33, and MV23.

1. **Group 3c**

EPI\_ISL\_677673, the main founder of Cluster 5, is also the ancestor of the clusters in Group 3c (that include Cluster 6, 10, 11 and 13). The founder of Cluster 6, EPI\_ISL\_436946, is a descendant of EPI\_ISL\_677673. This isolate contains a Protein 8 S24L mutation, in addition to the mutations detected in Cluster 5. While the emergence of Cluster 13 is associated with the Nsp5 L89F (PP1a / PP1ab L3352F) mutation, this cluster subsequently gave rise to Cluster 10 via the acquisition of five missense mutations (Nsp14 N129D (PP1ab N6054D), Nsp16 R216C (PP1ab R7014C), Protein 3a G172V, and nucleocapsid P67S and P199L). Cluster 11, another lineage derived from Cluster 13, on the other hand, has acquisition five mutations, including four synonymous mutations (Nsp3 K1111 (PP1a / PP1ab K1929), Nsp15 V84 (PP1ab V6536), and Nucleocapsid S327 and D401), one missense mutation (Protein 3a D155Y) and a 29764 G>T mutation that is located in the 3' UTR pseudoknot stem-loop 2. The mutation profiles of Group 3c clusters are shown in Table \_. The network graphs of the clusters in Group 3c are shown below.

Table \_. Group 3c mutations identified in Clusters 5, 6, 10, 11 and 13.



*Red: missense mutations; black: synonymous mutations; blue: mutations in the untranslated region (UTR).*

* Cluster 6

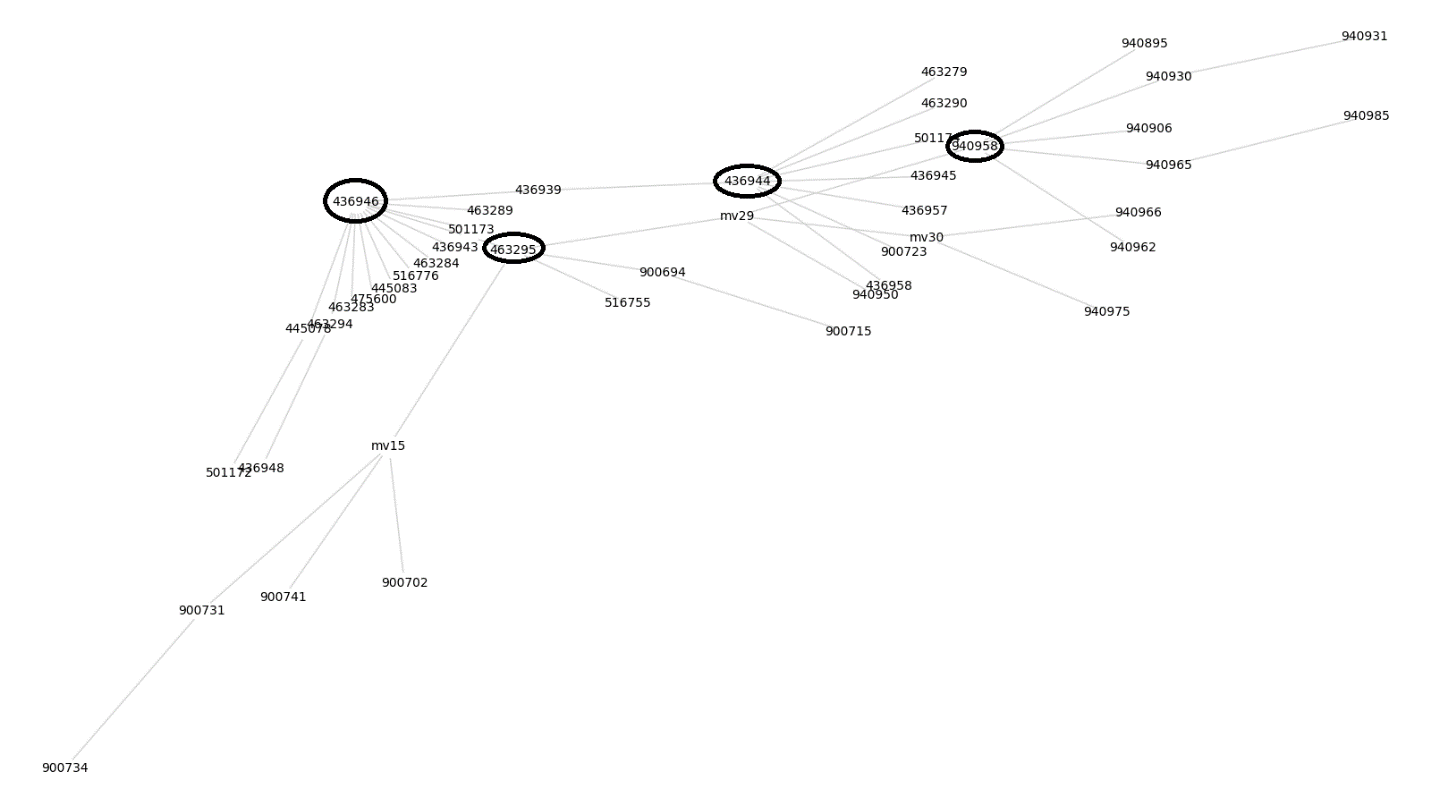


Figure \_. Network graph of Cluster 6. There are four founders (circled) in this cluster, EPI\_ISL\_436946, EPI\_ISL\_463295, EPI\_ISL\_436944, and EPI\_ISL\_940958.

* Cluster 13

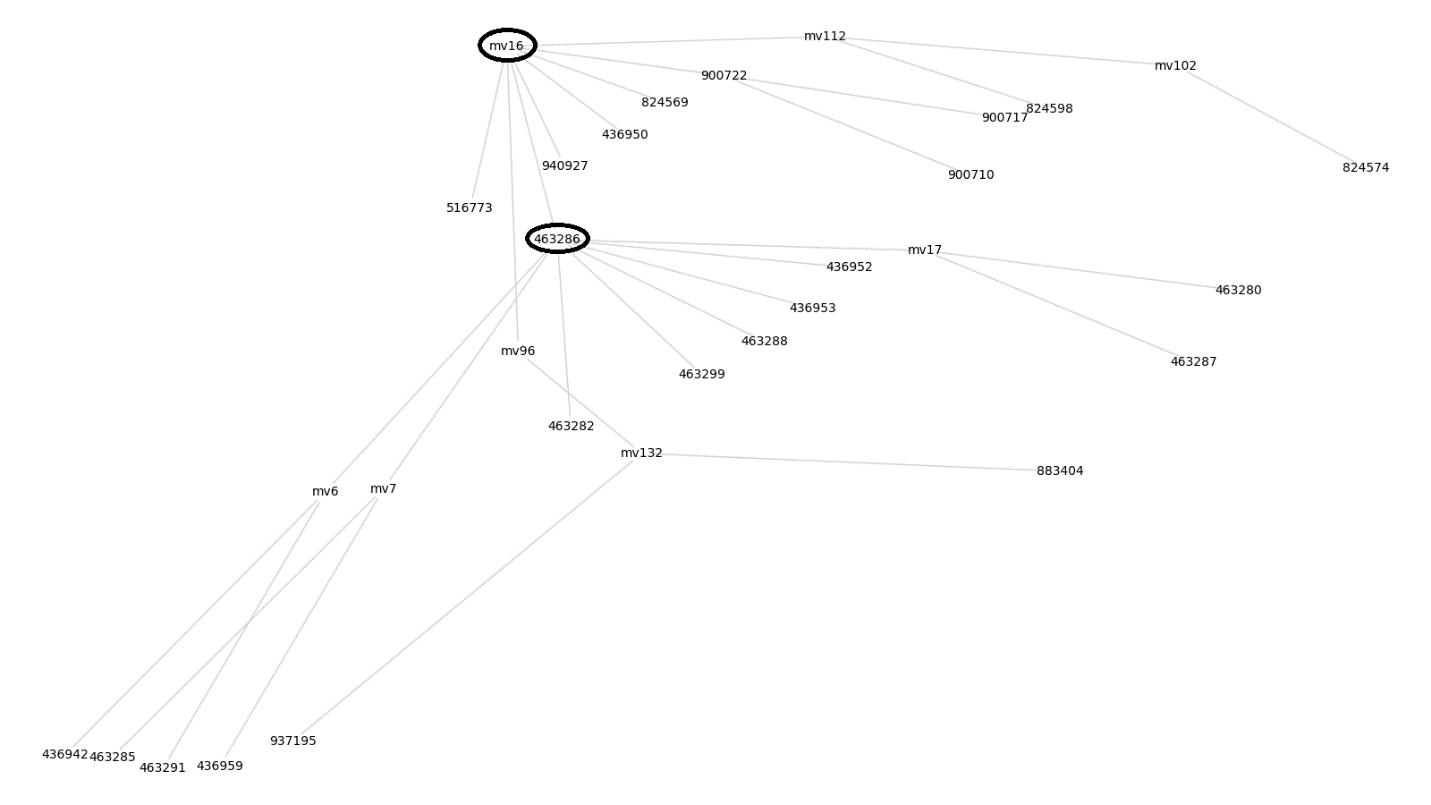


Figure \_. Network graph of Cluster 13. There are two founders (circled) in this cluster, MV16 and EPI\_ISL\_463286.

* Cluster 10

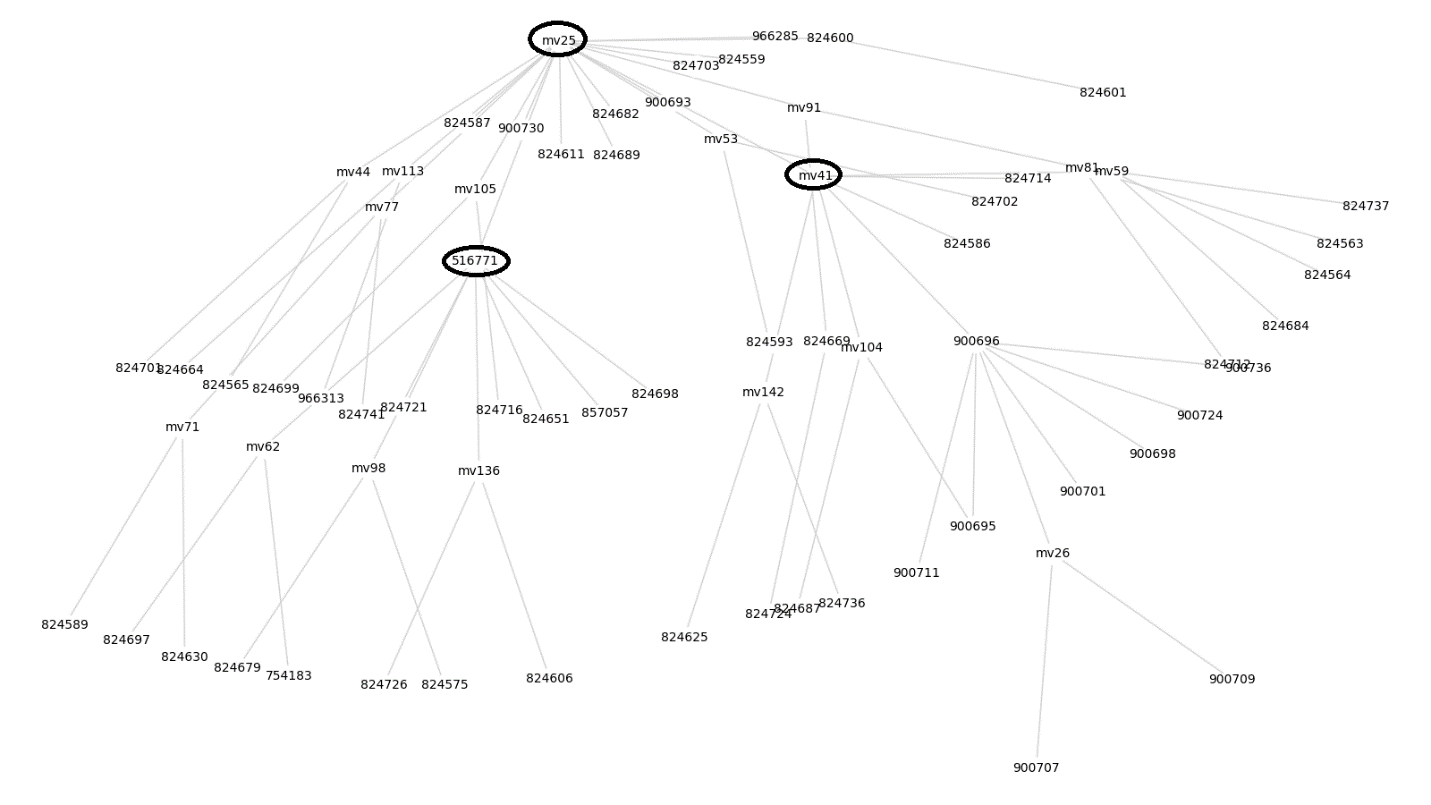


Figure \_. Network graph of Cluster 10. There are two founders (circled) in this cluster, MV25, EPI\_ISL\_516771, and MV41.

* Cluster 11

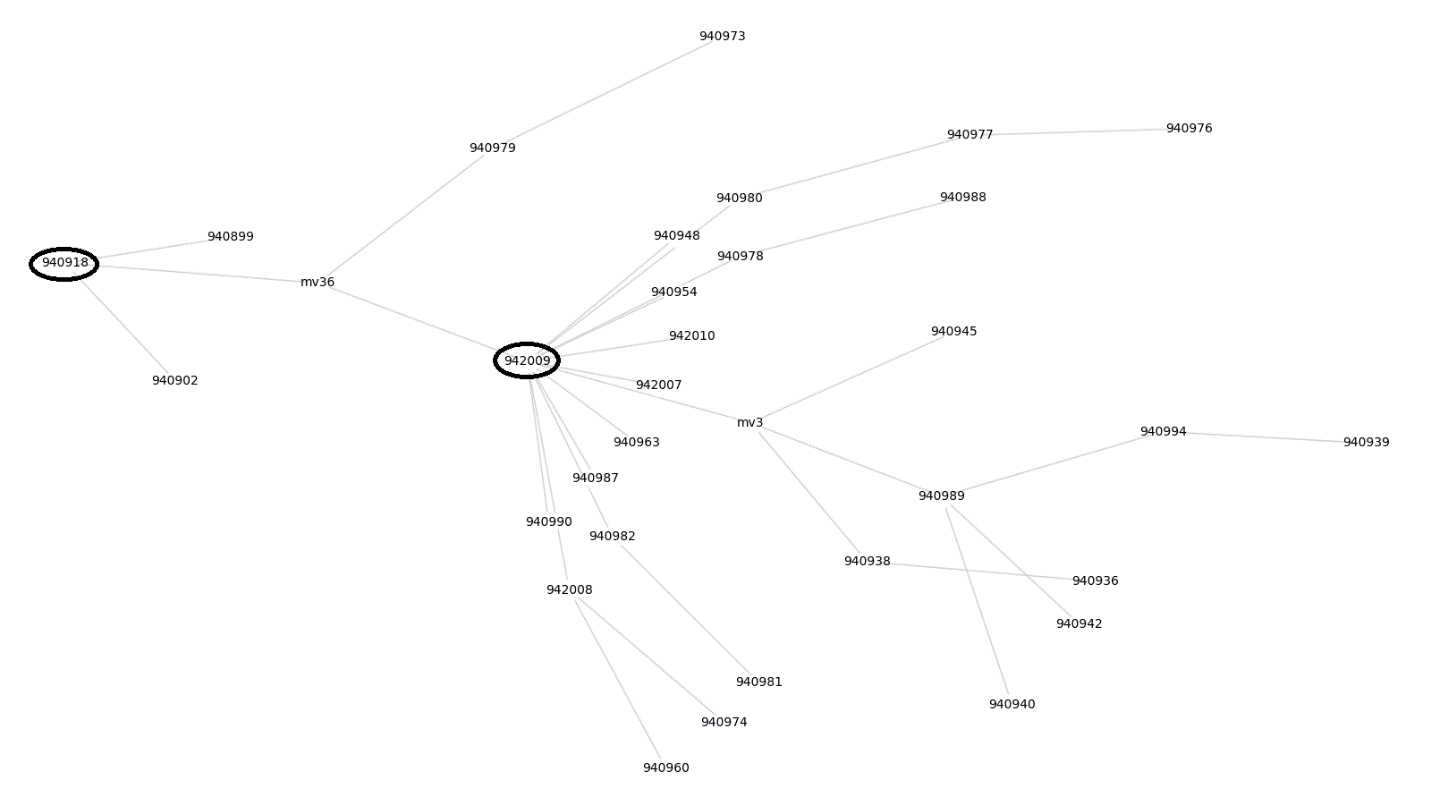


Figure \_. Network graph of Cluster 11. There are two founders (circled) in this cluster, EPI\_ISL\_940918 and EPI\_ISL\_942009.

The mutation profiles of the founders of these 15 clusters and all the mutations detected in the viral isolates and median vector nodes of this study are listed in the Supplementary Data Tables 2 and 3.

(View Supplementary tables (Too big to put in: Mutdiffs and prodiffs))

1. **Emergence of viral isolates related to variants reported in other countries**

Four viral sequences are highly similar to the Alpha variant that was first detected in UK. In fact, they contain all the mutations of the Alpha variant (Tables \_). Interestingly, they have acquired additional novel mutations that were not detected in the Alpha variant. As shown in Table \_, EPI\_ISL\_911548 and EPI\_ISL\_883325 share five of eight novel mutations, while EPI\_ISL\_943550 and EPI\_ISL\_966310 share six of another ten novel mutations. Thus, it is likely that the Alpha variant gave rise to two independent lineages that can be detected in the US.

Table \_. Mutations identified in Alpha variant and related isolates.



*Red: missense mutations; black: synonymous mutations; blue: mutations in the untranslated region (UTR).*

Table \_. Additional mutations identified in Alpha-related isolates.



*Red: missense mutations; black: synonymous mutations; blue: mutations in the untranslated region (UTR).*

On the other hand, viral sequences that are highly related to the Beta or Gamma variants were not detected in our dataset (data not shown).

The mutations associated with the transition from Clusters 2 and 5 to their respective descendent clusters can be summarized in the figure below.

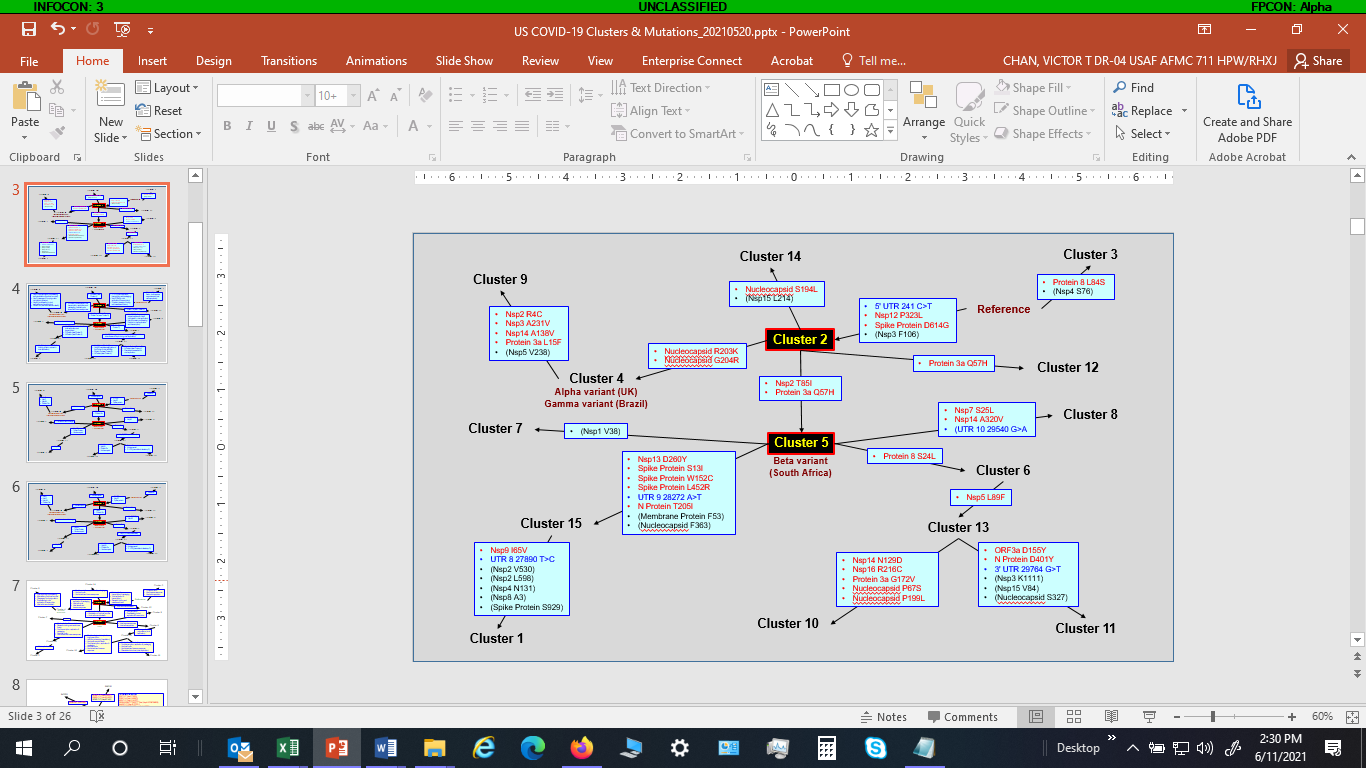


Figure \_. Mutations associated with the transition from the ancestor cluster to the descendant cluster(s) in the US, from February 2020 to February 2021.

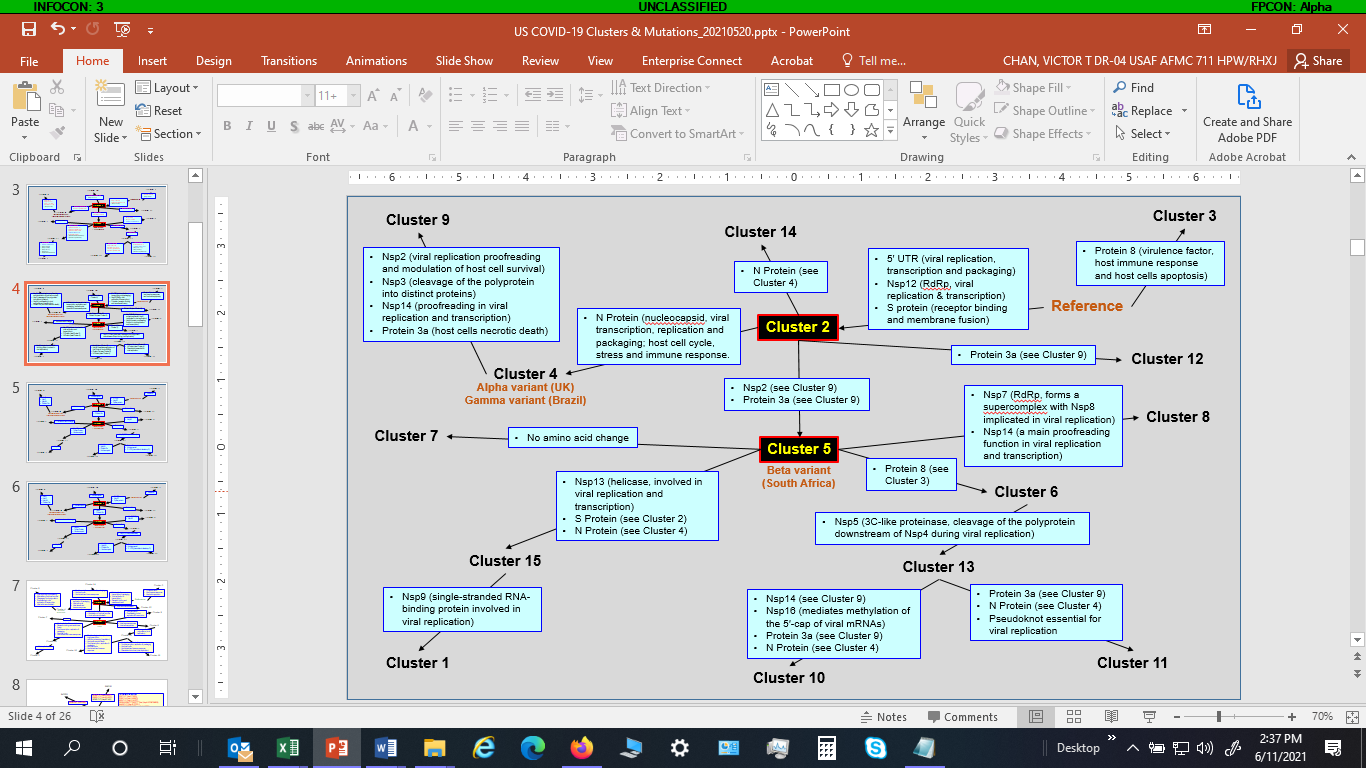
**Discussion (Dr. Chan)**

US so far has the most COVID-19 cases and deaths in the world.

Mutations of SARS-CoV-2 (e.g., variants detected in UK, Brazil, South Africa, India, etc.) will have significant implications in combating the pandemic.

* Effectiveness of vaccine and/or therapeutic agents
* Transmission
* Virulence…

Understanding the evolution of this virus will facilitate the development of more effective strategies against this virus.



Not all mutations will confer selective advantages.

It is important to identify mutations that confer selective advantages.

We took this approach to identify the accumulation of mutations during the pandemic in the US

Potential effects of these mutations, especially the effects of the combination of these mutations.

Major limitation of the study - we are limited by the sequences that have been sequenced and submitted to the GISAID.

Compared to other countries (e.g., UK), the sequencing effort in the US is relatively low.

A large portion of the SARS-CoV-2, including many mutants were not detected 🡪 would not be in our dataset.

The median vector nodes might represent some mutants that play a role in the route during the accumulation of mutations described in this study.

**(Dr. Welch)** We provide a software pipeline that allows the analysis of the contents of the public GISAID database. The two main contributions are:

1. A public code repository that enables other researchers to perform similar analyses, e.g., for different geographic regions and/or time periods.
2. A graph-based approach, which enables the discovery of more complex phylogenetic relationships than is possible with trees.

**Reference**

1. Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48