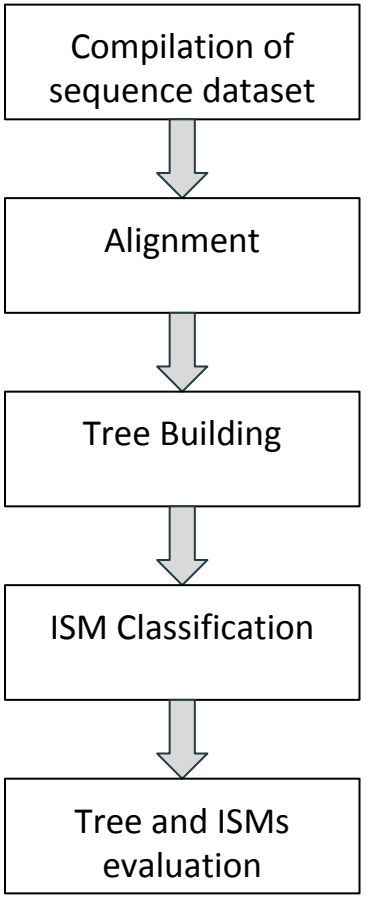
# Examination of COVID-19 Dynamics through Phylogenetic Assembly and ISM Labeling

Presented By:

Anthony Knesis
Dhairav Shah
Jacob Maier
Wenhan Tan
John R Seitz
Alexander Tweed

# Process (4 stages)

1. Data Acquisition
2. Sequence Alignment
3. Phylogenetic Trees
4. ISMs Classification



# Data Acquisition

- SARS-CoV-2 data collected from GISAID's "EpiCoV" database
  - GISAID - Global Initiative on Sharing All Influenza Data, hosted by Germany
  - Expanded to include Covid19 data
  - Currently over 41,600 samples (and counting!)
  - Data must be filtered to reduce number of samples and ensure quality sequences

- Filtered using Nextstrain's augur function
  - Filtering criteria include:
  - Incomplete dates
  - Minimum Length of 25kbp
  - Subsampling of sample locations: 2000/group

```
(nextstrain) john@DESKTOP-T4F4SS0:~/home/ECE5450_Project$ augur filter \
>   --sequences data/sequences.fasta \
>   --metadata data/metadata.tsv \
>   --min-length 25000 \
>   --exclude-where date='2020' date='2020-01-XX' date='2020-02-XX' date='2020-03-XX' date='2020-04-XX' date='2020-05-XX
' date='2020-06-XX' date='2020-01' date='2020-02' date='2020-03' date='2020-04' date='2020-05' date='2020-06' \
--output>   --output results/filtered.fasta \
>   --group-by division year month \
>   --sequences-per-group 2000 \
>

4969 sequences were dropped during filtering
        99 of these were dropped because of 'date=2020'
        7 of these were dropped because of 'date=2020-01-XX'
        14 of these were dropped because of 'date=2020-02-XX'
        1 of these were dropped because of 'date=2020-03-XX'
        0 of these were dropped because of 'date=2020-04-XX'
        0 of these were dropped because of 'date=2020-05-XX'
        0 of these were dropped because of 'date=2020-06-XX'
        2 of these were dropped because of 'date=2020-01'
        3 of these were dropped because of 'date=2020-02'
        135 of these were dropped because of 'date=2020-03'
        15 of these were dropped because of 'date=2020-04'
        0 of these were dropped because of 'date=2020-05'
        0 of these were dropped because of 'date=2020-06'
        129 of these were dropped because they were shorter than minimum length of 25000bp
        4564 of these were dropped because of subsampling criteria
22531 sequences have been written out to results/filtered.fasta
```

# Sequence Alignment



- MUSCLE (Multiple Sequence Comparison by Log-Expectation)
  - uses distance matrices and iterative refinements to finalize alignment
  - up to 500 sequences
- MAFFT on CIPRES (Multiple Alignment using Fast-Fourier Transforms)
  - a web gateway that allows access to phylogenetic tools and the computing resources of the National Science Foundation
  - yields a much higher throughout, allowing alignment of up to 30,000 sequences
- MAFFT must be used since the dataset greatly exceeds MUSCLE’s capacity

# Phylogenetic Trees

RAxML

FastTree 2

- RAxML
  - A slower but more accurate algorithm
  - Incorporates techniques for maximum parsimony and maximum likelihood
- FastTree
  - A clustering technique optimized for speed
  - Assembles a basic topology and refines connections through distance metrics and maximum likelihood models
- IToL
  - a web-based platform for annotating and viewing phylogenetic trees
  - High latency due to large number of leaves

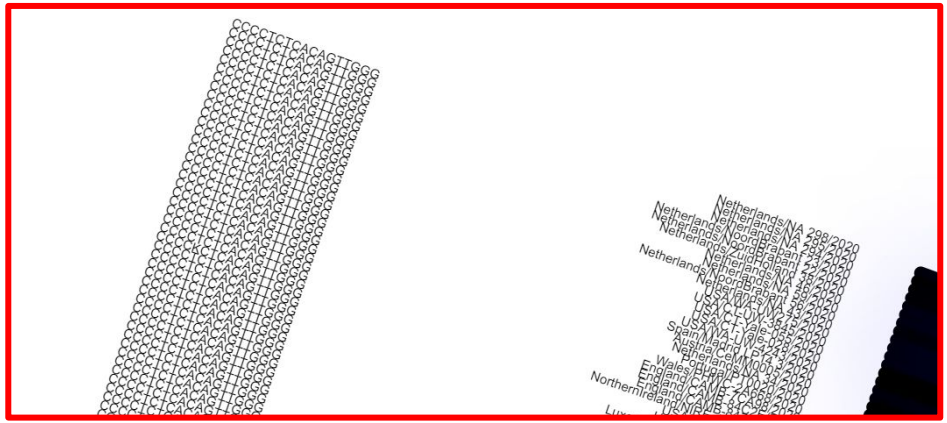
# ISM Classification

- Informative subtype markers
  - Based on entropy calculations which yield highly variable nucleotide sequences
  - Can be extracted as markers from genome sequence
  - Uses a processing pipeline to identify ISMs
  - Has been provided in a research group's GitHub repository (Dr. Rosen & Zhengqiao Zhao)

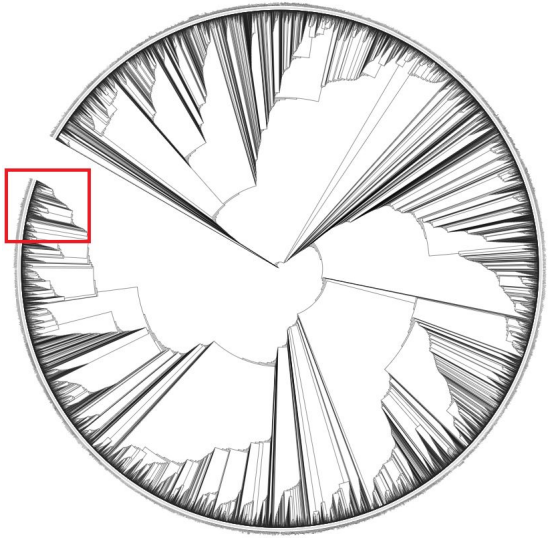
# Strategies/Methods

1. Nextstrain Filtering - subsample unaligned SARS-CoV-2 data
    - Remove incomplete sequences and those with large gaps
    - Decreases data size, while retaining diversity in temporal and geographic samples
2. MAFFT - multiple sequence alignment for condensed data set (approx. 9000 sequences)
3. FastTree - create phylogenetic tree from MSA using CIPRES
4. Proteus - calculate ISMs from alignment with EESI lab pipeline (Python)
5. Tree of Life Viewer - visualize phylogenetic tree with ISMs as node labels
6. Compare FastTree result with Nextstrain tree for reconstruction accuracy
    - e.g. Robinson-Foulds distance (symmetric distance), path difference, subtree pruning and regrafting (SPR) distance, etc.

# Results



- ISMs linked with GISAID viral identifiers
  - Allows correlation with nodes of phylogenetic tree
- Preliminary labelling annotated in iToL web viewer
  - Tree very large and difficult to visualize effectively
- Color coding ISMs not straightforward
  - Approx. 250 unique ISM strings



# Phylogenetic Trees

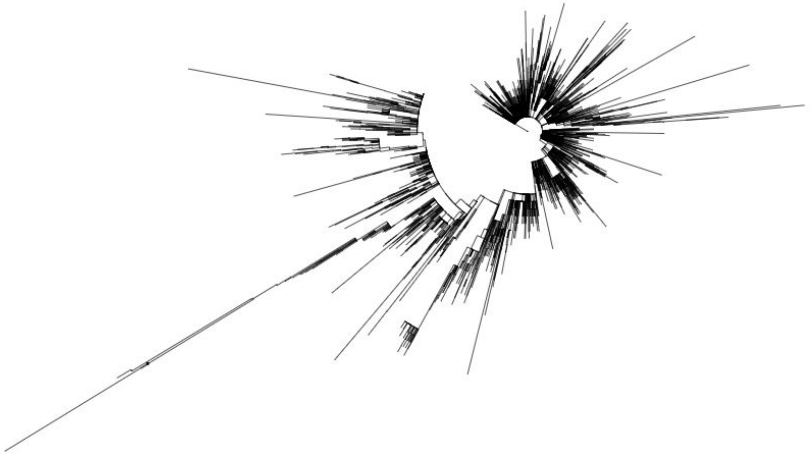


*Figure 1. Phylogenetic tree all US sequences before 5/17/2020*

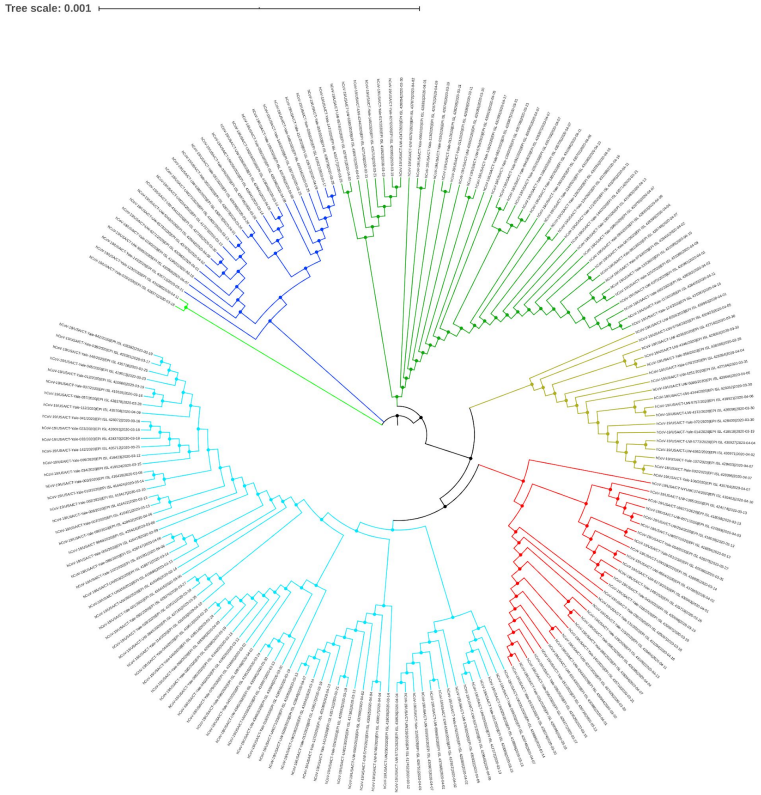



*Figure 2. Color-codes bootstrapping values for 185 sequence Covid-19 dataset*

# ISM Classification

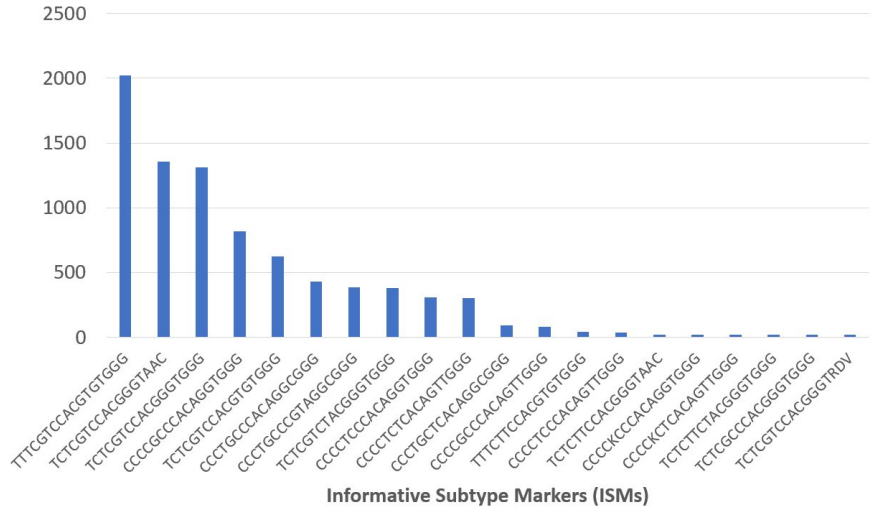



*Figure 3. Number of sequences containing the 20 most abundant ISMs within the total data set.*

| Site | Nucleotide Position | Entropy | Annotation |
|---|---|---|---|
| 1 | 241 | 0.89902 | Non-coding Region |
| 2 | 1059 | 0.81468 | ORF1ab polyprotein |
| 3 | 3037 | 0.90423 | ORF1ab polyprotein |
| 4 | 8782 | 0.51462 | ORF1ab polyprotein |
| 5 | 11083 | 0.47241 | ORF1ab polyprotein |
| 6 | 14408 | 0.90323 | ORF1ab polyprotein |
| 7 | 14805 | 0.30689 | ORF1ab polyprotein |
| 8 | 15324 | 0.28137 | ORF1ab polyprotein |
| 9 | 17858 | 0.26862 | ORF1ab polyprotein |
| 10 | 18060 | 0.2758 | ORF1ab polyprotein |
| 11 | 23403 | 0.90693 | S surface glycoprotein |
| 12 | 25563 | 0.91373 | ORF3a protein |
| 13 | 26144 | 0.31148 | ORF3a protein |
| 14 | 28144 | 0.50899 | ORF8 protein |
| 15 | 28881 | 0.66009 | nucleocapsid phosphoprotein |
| 16 | 28882 | 0.65704 | nucleocapsid phosphoprotein |
| 17 | 28883 | 0.65992 | nucleocapsid phosphoprotein |

*Table 1. Mapping ISM sites to the reference viral genome.*

# Challenges

- Alignment takes a long time when the number of sequences and length of sequences increase
- Large phylogenetic trees with a lot of branches are difficult to work with (figure 1)
  - This applies to visualization and ISM labeling
- ISM’s required further processing which increases computational resources
- Quality of data was questionable at times and was changing rapidly

# References


1. "*Characterizing geographical and temporal dynamics of novel coronavirus SARS-CoV-2 using informative subtype markers*", Zhengqiao Zhao, Bahrad A. Sokhansanj, Gail L. Rosen, bioRxiv 2020.04.07.030759; doi: https://doi.org/10.1101/2020.04.07.030759

2. "iTOL: Interactive Tree of Life", https://itol.embl.de/

3. "Cipres: Cyber Infrastructure for Phylogenetic Research", http://www.phylo.org/

4. "Nextstrain:Augur Bioinformatics Toolkit", https://nextstrain.org/