Methods:

1-Sequencing of two Australian isolates: they are sequenced by MISEQ platform

RNA was purified from each isolate using (ZYMO research)

purified RNA was reverse transcribed using (applied bio system)

with random octamers linked to a specific primer sequence followed by a second strand synthesis DNA polymerase and complementary DNA was further amplified with (Roche)

resulting DNA was purified using DNA clean and ZYMO research

fragmentation and dual index library preparation using alumina and denatured using a 300- cycle

sequencing reads were trimmed for a quality and mapped to the published reference sequence

2-gisaid preprocessing and alignment: all available viral sequences were downloaded from GISAID

Filtering

complete sequences of human origin (187genomes in total) low quality sequences were filtered out leaving 178 strains we also reported viral sequences from the European virus archive global

This data set of 181 sequences was aligned against each other using muscle

Sequencing artifacts and errors preventing the full viral genome from being sequenced

Once trimmed and converting all (u) to (t) we identified identical sequences to limit the effect of the duplicates we collapsed it in one entry

3- phylogenetic trees: the max likelihood was generated from the above alignments using RAXML-ng

The evolutionary model used was a general time reversible with GAMA and distributed rate hetero o genetic and invariant sites (GTR+G+L) and this is the most general model and produces equal trees

The tree was visualized using ITOL and midpoint rooted tree and show the likely evolutionary relationships between the sampled strains

4- k-MER method: every organism can have a unique signature based on composition of the genomic sequence to quantify this signature we will determine the k MER frequency then counting all possible strings of length k in the sequence of the virus acts as an alternative to phylogenetic. the conceptual isolated can be visualized by running PCA over all genomic signatures to reduce this high dimensional k MER frequency vector into a 2-dimensional space

Custom scripts were used to calculate the k-MER frequency

k-MERS containing ambiguous bases were removed then we calculate the relative proportion of each k-MER resulting in frequency vector then using PCA implementation to reduce the genomic signatures containing proportions into a vector containing 2 principal components

custom scripts were used to compare the genomic signatures for all COV- 2 sequences

Results

1- phylogeny reveals 3 clusters: the evolutionary structure of the 181 isolates (by phylogenetic tree) reveals the 3 major clusters (c1_c3)

C1 represents early stages but c2 and c3 represent the later isolates

The three clusters are separated by distinct mutations

There may be 3 additional clusters emerging (c4_c6)

C4 capturing the suspected community spread from Lombardy

This finding is different who postulate 2 clusters (s and l) however the etiology is not fully demonstrated especially with the intermediate vector is unknown

Positions are relative to trimmed alignments

High number of mutations due to sequencing errors

Irrespective of the root placement both trees allow the assessment of individual isolates

The full impact of these genomic variations can only be confirmed through functional genomics experiments

Also having a methodology able to take deletions into consideration when calculating genomic distance is desirable

2- alignment free phylogeny captures evolutionary distances

Aiming to overcome the limitations of phylogenetic tree approach can be used to understand SARS cov2 and how it changes

We calculate 10 MERS across all viral genome followed by PCA to reduce k-MER vector into 2-dimensional image

It will be separated into 3 SARS-COV- 2, SARS and MERS and all distances between strains are very small we isolate MERS into 2 strains we re-ran the PCA on SARS-COV-2 we find out that isolates are likely to be closed while strains separated by time are far apart

While the k- MER is not as suggestive as phylogenetic trees it reflects the fluidity of changes and capture recombination events

Phylogenetic analysis is based on the presence of shared mutations

3- How repetitive are the currently chosen isolates for preclinical models: more isolates are chosen while next strain is powerful in visualizing and it relies in phylogeny

More fluid distance measure of alignment free methods

The dominant driver placing isolates away from the center is missing bases