(Under materials and methods)

*Genomics data collection and processing*

SARS-CoV-2 sequencing is very rapidly conducted and made available to the public through the National Center for Biotechnology Information (NCBI) site. The reference genome used in this study is the accession: NC\_045512, one of the first complete genomes for SARS-CoV-2 sequenced from Wuhan, China in December 2019 (Baranov et al. 2019). Additionally, another 435 SARS-CoV-2 genomes sequenced from various geographical locations were downloaded through the NCBI Virus database, which are the available genomes sequenced as of 1 April 2020. Using all 436 genomes, multiple alignment sequencing was performed using MAFFT (Multiple Alignment Fourier Fast Transform) software (Katoh 2013) with the NC\_045512 as the reference genome. The sequences were filtered according to its ‘quality’ and usefulness for our study by excluding sequences with excessive amount of non-base character. Specifically, we excluded 72 accessions that has more non-base characters than base characters. The final list of accessions is available in supplementary materials.