**Analysis of evolution and adaptation of SARS-CoV-2 in humans**

**Generation of data sets**

1. ***Retrieve high-quality SARS-CoV-2 genome sequences.***
2. Download the first 100,000 SARS-CoV-2 genome sequences (based on submission date) from GISAID website.
3. 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’
4. Eliminate low-quality sequences by applying the following filters.
5. A+C+G+T<29725.
6. Collection date has no month or date information.
7. Determine the number of remaining cases with the following patient status information.
8. Entry ≠ unknown
9. Entry = unknown
10. Entry = “-“
11. No entry
12. ***Standardization of patient status information.***

The description of patient status is quite variable (many different terms were used). However, they can be classified into two categories, symptom and outcome. Therefore, patient status can be standardized according to these two categories. **Note:** in some cases, patient status information can be found under “additional host information” or even “additional location information”.

1. Symptom
2. Asymptomatic/Mild (original entries include asymptomatic, mild case, mild clinical signs without hospitalization, outpatient, etc.).
3. Symptomatic (original entries include symptomatic, bronchitis, pneumonia, hospitalized, etc.).
4. NA (original entries include unknown, “–“, and blank).
5. Outcome
6. Recovered (original entries include alive, live, recovered, released, stable, cured, discharged, etc.).
7. Deceased (original entry = deceased).
8. NA (original entries include: unknown, “–“, and blank).
9. ***Generation of master data set***
10. Generate the master data set containing high-quality SARS-CoV-2 genome sequences with standardized patient status information and other information. The metadata will include:
11. Collection date (with year, month and date)
12. Gender
13. Patient age
14. Symptom
15. Outcome
16. Additional host information
17. Location
18. Additional location information
19. Virus name
20. Specimen
21. Passage
22. Lineage
23. Clade
24. Host
25. Perform the following statistics on the master data set
26. Number of cases for each day since the first collection date
27. World
28. Continent
29. Country
30. Gender (male and female)
31. Age (<35, 35 – 50, 50 – 65 and >65)
32. Symptom (asymptomatic/mild, symptomatic and NA)
33. Outcome (recovered, deceased and NA)
34. Lineage
35. Clade
36. ***Generation of top five (5) country-specific data sets***
37. Generate subsets genome sequences from the top five countries that have the most COVID-19 cases.
38. Perform the following statistics on the country-specificdata set
39. Number of cases for each day since the first collection date
40. Gender (male and female)
41. Age (<35, 35 – 50, 50 – 65 and >65)
42. Symptom (asymptomatic/mild, symptomatic and NA)
43. Outcome (recovered, deceased and NA)
44. Lineage
45. Clade