**Data Dictionary for sequencing-results-formatted-cleaned.csv**

*Each row represents one lineage in one sample. Samples in which multiple lineages were detected (most of them) appear in more than one row. Some samples have been sequenced more than once, resulting in the sum of lineage abundances in those samples being > 1. However, this is indicated in the notes column where applicable.*

**Columns are as follows:**

abvec: relative abundance of the given lineage in the given sample, between 0-1.

sample\_id: unique sample identifier. Usually a 5-digit numeric code, but earlier samples have other ID formats.

coverage: percentage of the SARS-CoV-2 reference genome covered by the sample (this is based on all sequences from the sample, not just one variant)

site\_id: 5-character unique identifier for the sample location

wwtp\_name: the name of the sample location (usually a WWTP, but may also be a Cook County Jail ward, manhole id, etc.)

sample\_type: the sample collection method: flow\_weighted\_composite, time\_weighted\_composite, manual\_composite, moore\_swab, or grab.

flow\_rate: the flow rate recorded at sampling time. *Note: the units do not appear to be consistent for all samples.*

sample\_collect\_date: the date the sample was collected (YYYY-MM-DD)

sample\_collect\_time: time the sample was collected (HH:MM:SS, 24-hour scale)

sample\_processed\_date: date of RNA extraction/dPCR analysis

sample\_received\_date: date the sample arrived at the Poretsky lab

sample\_arrival\_temp: temperature recorded when sample arrived at the Poretsky lab (degrees Celsius)

notes: miscellaneous information about the sample

year: the year the sample was collected

week: the week of the year when the sample was collected (2022-01-01 and 2023-01-01 would both be week 1)

full\_lineage\_id: the name of the lineage returned by Freyja

aliases\_removed: merges lineages that have different names, based on this list: <https://github.com/cov-lineages/pango-designation/blob/master/lineage_notes.txt>