COVID-19 Specimen Data Processing: gisaid_code.R

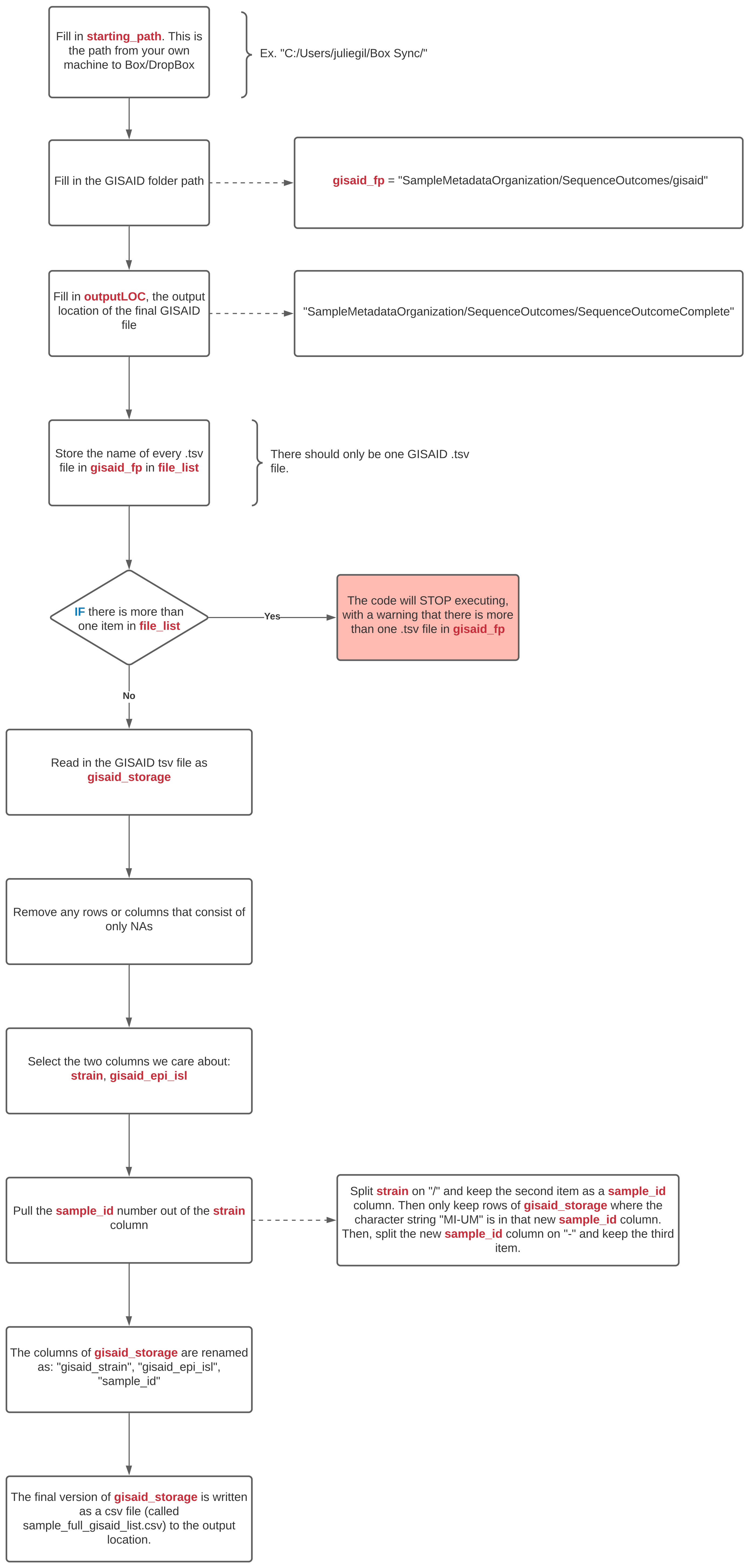
IMPORTANT: The GISAID file name should follow the same format.

metadata_<YYYY-MM-DD_HH-MM>.tsv

YYYY-MM-DD_HH-MM = Date when the GISDAID file was downloaded

Libraries Needed:
library(tidyverse)
library(lubridate)
library(janitor)

Overview: This code file pulls in the single .tsv file with all GISAID info included. The file is processed and then a summary, slimmed down file is output.



Fill in **starting_path**. This is the path from your own machine to Box/DropBox

Ex. "C:/Users/juliegil/Box Sync/"

Fill in the GISAID folder path

gisaid_fp = "SampleMetadataOrganization/SequenceOutcomes/gisaid"

Fill in **outputLOC**, the output location of the final GISAID file

"SampleMetadataOrganization/SequenceOutcomes/SequenceOutcomeComplete"

Store the name of every .tsv file in **gisaid_fp** in **file_list**

There should only be one GISAID .tsv file.

IF there is more than one item in **file_list**

Yes

The code will STOP executing, with a warning that there is more than one .tsv file in **gisaid_fp**

No

Read in the GISAID tsv file as **gisaid_storage**

Remove any rows or columns that consist of only NAs

Select the two columns we care about: **strain**, **gisaid_epi_isl**

Pull the **sample_id** number out of the **strain** column

Split **strain** on "/" and keep the second item as a **sample_id** column. Then only keep rows of **gisaid_storage** where the character string "MI-UM" is in that new **sample_id** column. Then, split the new **sample_id** column on "-" and keep the third item.

The columns of **gisaid_storage** are renamed as: "gisaid_strain", "gisaid_epi_isl", "sample_id"

The final version of **gisaid_storage** is written as a csv file (called sample_full_gisaid_list.csv) to the output location.