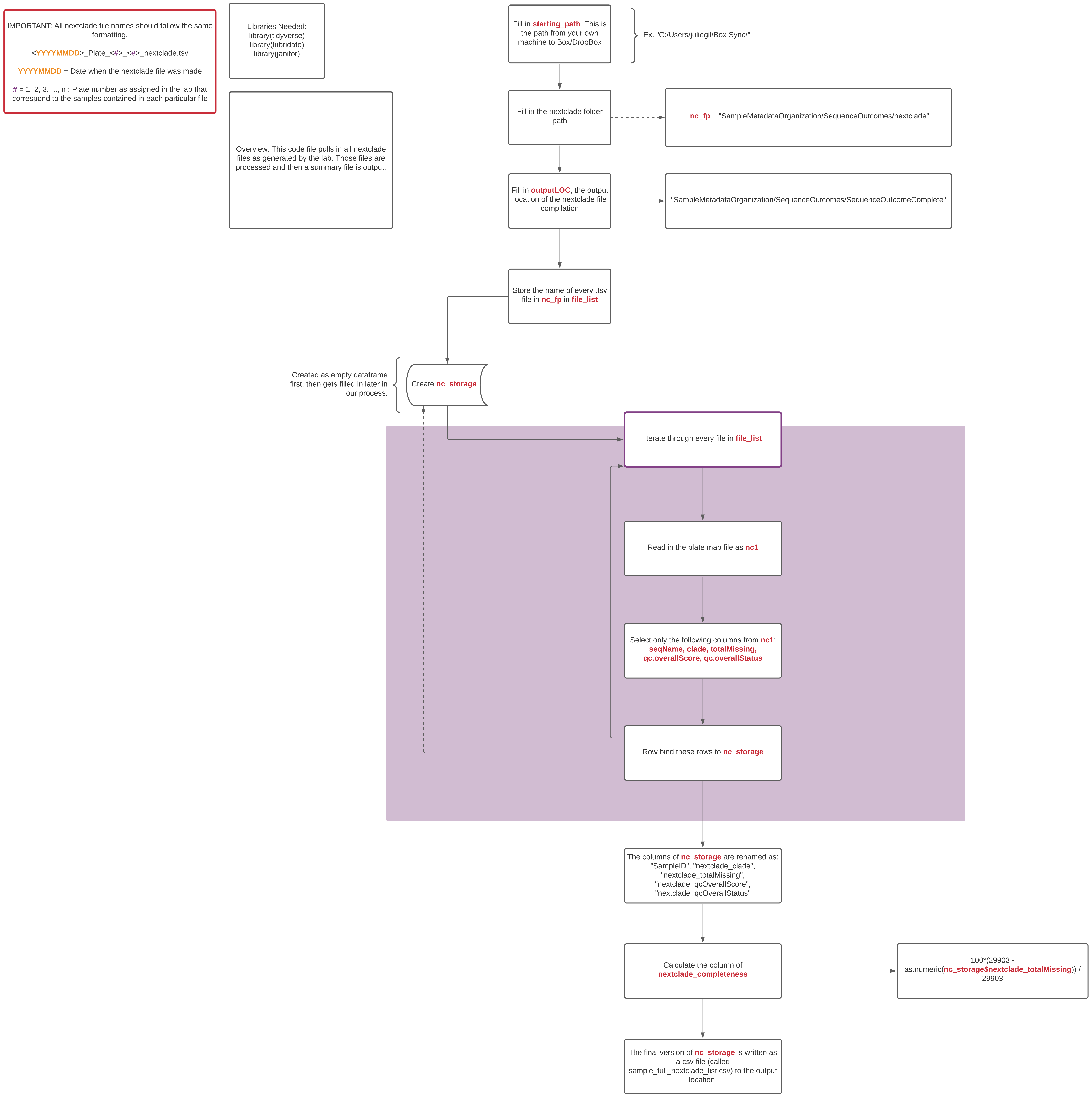
COVID-19 Specimen Data Processing: nextclade_code.R



IMPORTANT: All nextclade file names should follow the same formatting.

<YYYYMMDD>_Plate_<#>_<#>_nextclade.tsv

YYYYMMDD = Date when the nextclade file was made

# = 1, 2, 3, ..., n ; Plate number as assigned in the lab that correspond to the samples contained in each particular file

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

Overview: This code file pulls in all nextclade files as generated by the lab. Those files are processed and then a summary 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 nextclade folder path

nc_fp = "SampleMetadataOrganization/SequenceOutcomes/nextclade"

Fill in **outputLOC**, the output location of the nextclade file compilation

"SampleMetadataOrganization/SequenceOutcomes/SequenceOutcomeComplete"

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

Create **nc_storage**

Created as empty dataframe first, then gets filled in later in our process.

Iterate through every file in **file_list**

Read in the plate map file as **nc1**

Select only the following columns from **nc1**: **seqName, clade, totalMissing, qc.overallScore, qc.overallStatus**

Row bind these rows to **nc_storage**

The columns of **nc_storage** are renamed as: "SampleID", "nextclade_clade", "nextclade_totalMissing", "nextclade_qcOverallScore", "nextclade_qcOverallStatus"

Calculate the column of **nextclade_completeness**

100*(29903 - as.numeric(**nc_storage$nextclade_totalMissing**)) / 29903

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