

COVID-19 Specimen Data
Processing:
plate_map_code.R

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

<YYYYMMDD>_<Platform>_Plate_<#>.xlsx

YYYYMMDD = When the plate map file was made (theoretically, when the plate was run)

Platform = The system the plate was run on (Illumina, Nanopore, etc.)

= 1, 2, 3, ..., n; Plate number as assigned in the lab

IMPORTANT: All platemap contents should follow the same formatting.

Seven columns: A) Processing_Plate, B) Slot, C) Sample ACCN, D) Sample MRN, E) Sample Order#, F) Barcode, G) Source

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

Overview: This code file pulls in all plate map files as generated by the lab.

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

Ex: "C:/Users/julegll/Box Sync"

Fill in the plate map folder path (there should be a single folder for all plate maps, regardless of platform)

platemap_fp = "SampleMetadataOrganization/PlateMaps"

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

"SampleMetadataOrganization/PlateMaps/PlateMapsComplete"

Store the name of every .xlsx file in **platemap_fp** in **file_list**

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

Create **platemap_storage**

Iterate through every file in **file_list**

Read in the plate map file as **plate1**

Select all rows but only columns 1-7

Row bind these rows to **platemap_storage**

The **Processing_Plate** column is split up into 3 different columns

The first 8 characters are re-formatted into a YYYY-MM-DD column = **Plate_Date**

Processing_Plate is split on "." characters and the second item is kept as **Plate_Platform**

Processing_Plate is split on "-" characters and the fourth item is kept as **Plate_Number**

The **Source** column is split on "-" and the second item is kept, and that piece is split on "-" and formatted into <third item>-<first item>-<second item> to form **Source_Date**

The **Source** column is split on "-" and the first item is kept to create **Source_Location**. In instances where **Source_Date** is NA (so the second item couldn't be successfully turned into a date), the entire **Source** column is used to fill in **Source_Location**

Only the following columns are kept, in this order: **Processing_Plate**, **Slot**, **Sample_ACCN**, **Barcode**, **Source_Date**, **Source_Location**, **Plate_Date**, **Plate_Platform**, **Plate_Number**

Those columns are renamed as:
"PlateName", "PlatePosition", "SampleID",
"SampleBarcode", "SampleSourceDate",
"SampleSourceLocation", "PlateDate",
"PlatePlatform", "PlateNumber"

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