# MEDICAL MICROBIOLOGY AND INFECTIOUS DISEASES CODING WORKSHOP

*Presents*

# *Tidy data: combining and transforming data in R*


INSTRUCTED BY

*Molly Pratt*

*prattm1@myumanitoba.ca*



*Educational series supported by the Department of MMID to promote knowledge translation and dissemination among students and researchers*

# INFORMATION FOR PARTICIPANTS

**All workshops are being recorded and posted to the**
MMID Coding Workshop - YouTube

***Question and Answer period will not be recorded.***

# LAST WEEK...

***Introduction to R (Grace Seo):***

***https://www.youtube.com/watch?v=PhZdW0r0f_8***

***RStudio navigation***

***Installing and loading R packages and libraries***

***Using base R functions and operators***

***R scripts / R markdown***

# LEARNING OBJECTIVES

***Become familiar with tidy data structure***

***Discover R packages for data science (tidyverse)***

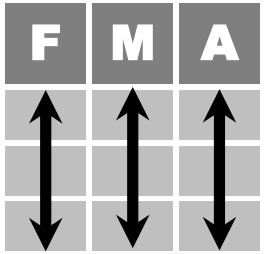
***Use tidyverse functions to reshape data and combine multiple datasets***

***Practice modifying variables and making new variables***

# What is tidy data?

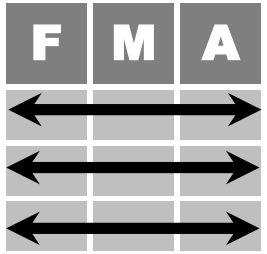
An introduction to the tidyverse

**In a tidy data set there are 3 rules:**



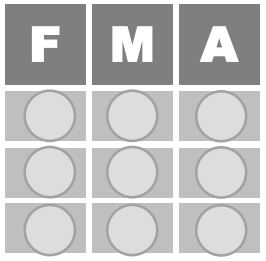

Each **variable** is saved in its own column

&



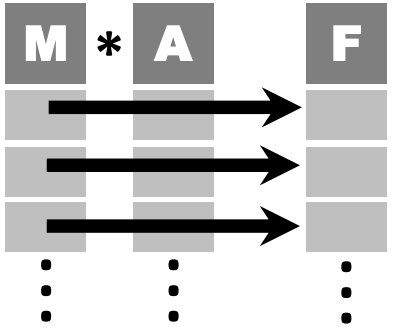

Each **observation** is saved in its own row

&

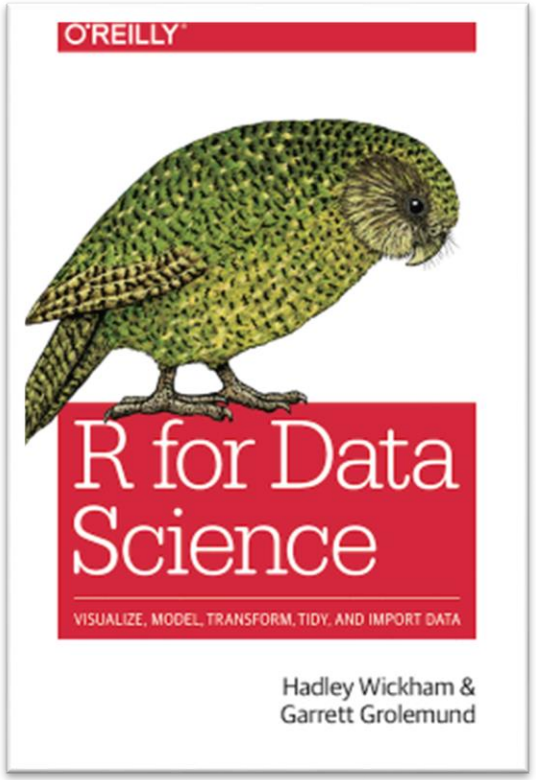



Each **value** has its own cell

Tidy data complements R's **vectorized operations**. R will automatically preserve observations as you manipulate variables. No other format works as intuitively with R.




# R for Data Science Chapter 12: Tidy data




*Free online book*[1]

# R for Data Science
# Chapter 12: Tidy data

Making data **longer:**

| country | year | cases |
|---|---|---|
| Afghanistan | 1999 | 745 |
| Afghanistan | 2000 | 2666 |
| Brazil | 1999 | 37737 |
| Brazil | 2000 | 80488 |
| China | 1999 | 212258 |
| China | 2000 | 213766 |

| country | 1999 | 2000 |
|---|---|---|
| Afghanistan | 745 | 2666 |
| Brazil | 37737 | 80488 |
| China | 212258 | 213766 |

table4

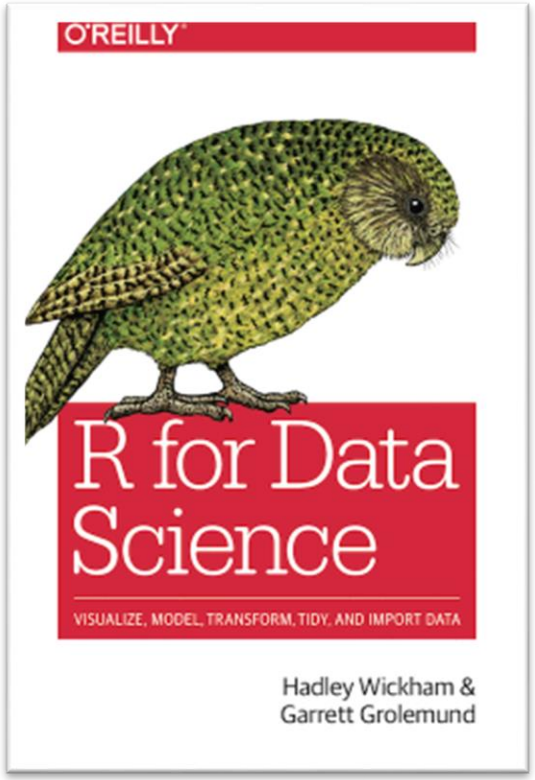
Making data **wider:**

| country | year | key | value |
|---|---|---|---|
| Afghanistan | 1999 | cases | 745 |
| Afghanistan | 1999 | population | 19987071 |
| Afghanistan | 2000 | cases | 2666 |
| Afghanistan | 2000 | population | 20595360 |
| Brazil | 1999 | cases | 37737 |
| Brazil | 1999 | population | 172006362 |
| Brazil | 2000 | cases | 80488 |
| Brazil | 2000 | population | 174504898 |
| China | 1999 | cases | 212258 |
| China | 1999 | population | 1272915272 |
| China | 2000 | cases | 213766 |
| China | 2000 | population | 1280428583 |

table2

| country | year | cases | population |
|---|---|---|---|
| Afghanistan | 1999 | 745 | 19987071 |
| Afghanistan | 2000 | 2666 | 20595360 |
| Brazil | 1999 | 37737 | 172006362 |
| Brazil | 2000 | 80488 | 174504898 |
| China | 1999 | 212258 | 1272915272 |
| China | 2000 | 213766 | 1280428583 |

Test: *Does the content of the <u>cells</u> for a particular <u>column</u> make sense as **values** for that **variable**?*




*Free online book*[1]


# R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

# Introduction to the tidyverse

tidyverse.org

```
>install.packages("tidyverse")
>library(tidyverse)
```

# Introduction to the tidyverse


**Data Import:**
Read and write tabular data / spreadsheets to and from R



**Data Wrangling:**
Reshape, subset, combine, group, and add new variables to your data


**Strings:**
Manipulate character strings

Cheatsheets for working with these packages (highly recommended) can be found in the workshop materials, and are freely available online.

## Subset Observations (Rows)

dplyr::**filter**(iris, Sepal.Length > 7)

Extract rows that meet logical criteria.

dplyr::**distinct**(iris)

Remove duplicate rows.

## Subset Variables (Columns)

dplyr::**select**(iris, Sepal.Width, Petal.Length, Species)

Select columns by name or helper function.

## Make New Variables

dplyr::**mutate**(iris, sepal = Sepal.Length + Sepal. Width)

Compute and append one or more new columns.

# Functions

Syntax

**package::function(x, …)**

or

**function(x, …)**

Where x is a data frame

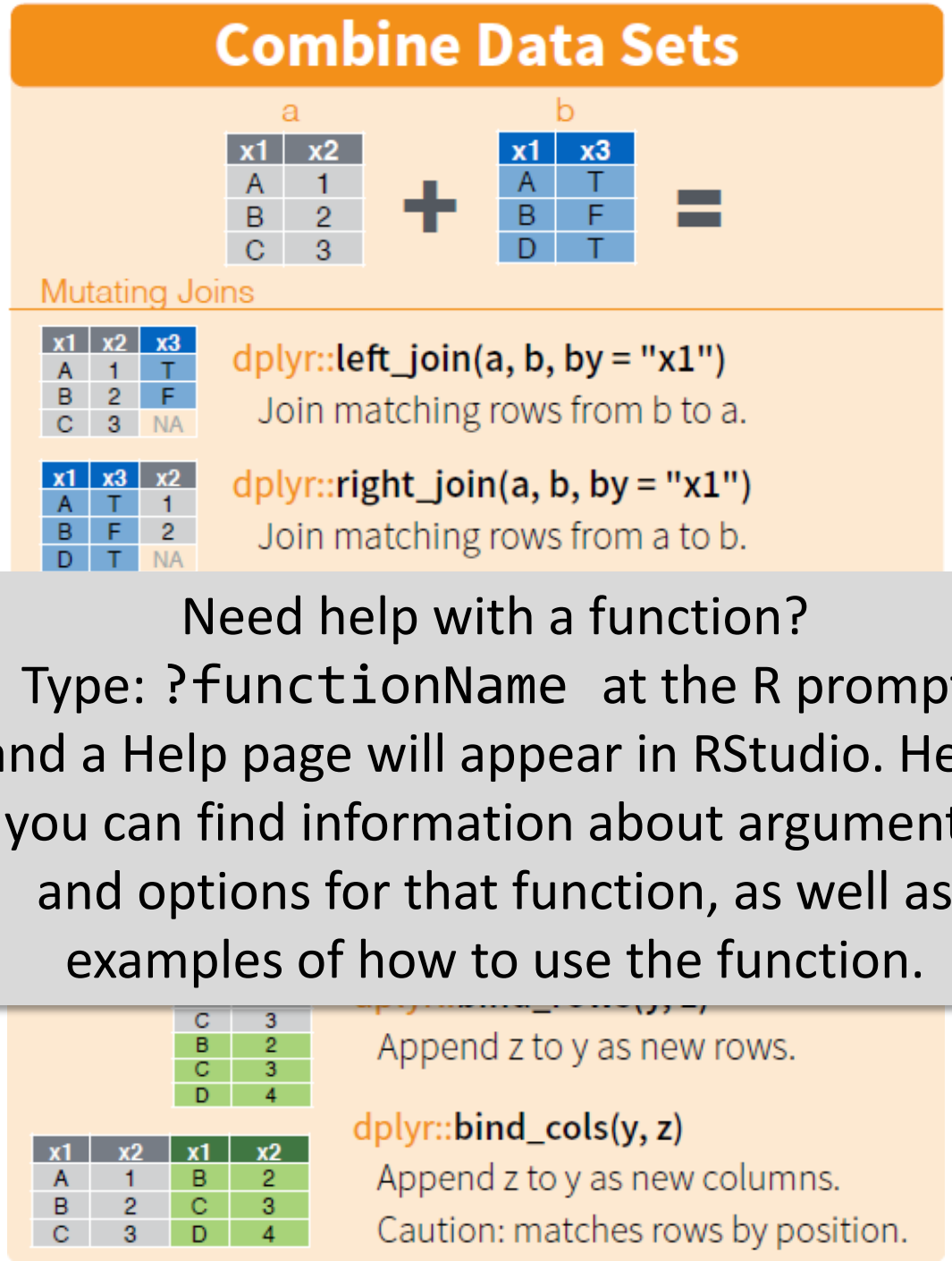
"…" represents additional input

Pipes (%>%)

**x %>% function(…)**

`x %>% f(y)` *is the same as* `f(x,y)`

"Piping" makes R code easier to read and organize. Whatever is before the %>% is passed to the first argument of the function.




# Functions

Syntax

**package::function(x, …)**

or

**function(x, …)**

Where x is a data frame

"…" represents additional input

Pipes (%>%)

**x %>% function(…)**

`x %>% f(y)` *is the same as* `f(x,y)`

"Piping" makes R code easier to read and organize. Whatever is before the %>% is passed to the first argument of the function.

# Assigning objects in R

```
> sum(1, 2, 3, 4, 5)
[1] 15
> b <- sum(1, 2, 3, 4, 5)
> b
[1] 15
> b <- b + 10
> b
[1] 25
```

## Saving Outputs

`function(x,…)`

Print output to the console (with some exceptions)

`newObject <- function(x,…)`

Assign output to newObject

Now typing `newObject` into R will print its contents into the console

# Assigning objects in R

```
> b_10 <- b + 10
> b
[1] 15
> b_10
[1] 25
```

When working with data sets, try to **avoid overwriting** your intermediate outputs until you are very comfortable with the process. You can always remove them later once you've achieved your final product.

## Saving Outputs

`function(x,…)`

Print output to the console (with some exceptions)

`newObject <- function(x,…)`

Assign output to newObject

Now typing `newObject` into R will print its contents into the console

# Data Transformation

Combining Example Datasets

# Workshop materials

If you haven't already, please **download** and **save** the workshop materials in your RStudio working directory.

Example:

`C:/Users/USERNAME/Desktop/MMID-Coding/`**`workshop_materials_Feb9`**
`C:/Users/USERNAME/Desktop/MMID-Coding/`**`MMID-coding.Rproj`**

The `tidy_data_MMID.R` script can be opened and run in Rstudio. Due to time constraints, it is highly recommended you work through the script on your own time.
`# comments` in the script are provided to guide you.

- data/
  - **gisaid_metadata.xls**
  - **line_list_galaxy.csv**
  - example-metadata-tidy.csv
- R_code/
  - **tidy_data_MMID.R**
- cheatsheets/
  - data-import-readr.pdf
  - data-visualization-ggplot2.pdf
  - **data-wrangling-dplyr-tidyr.pdf**
  - strings-stringr.pdf

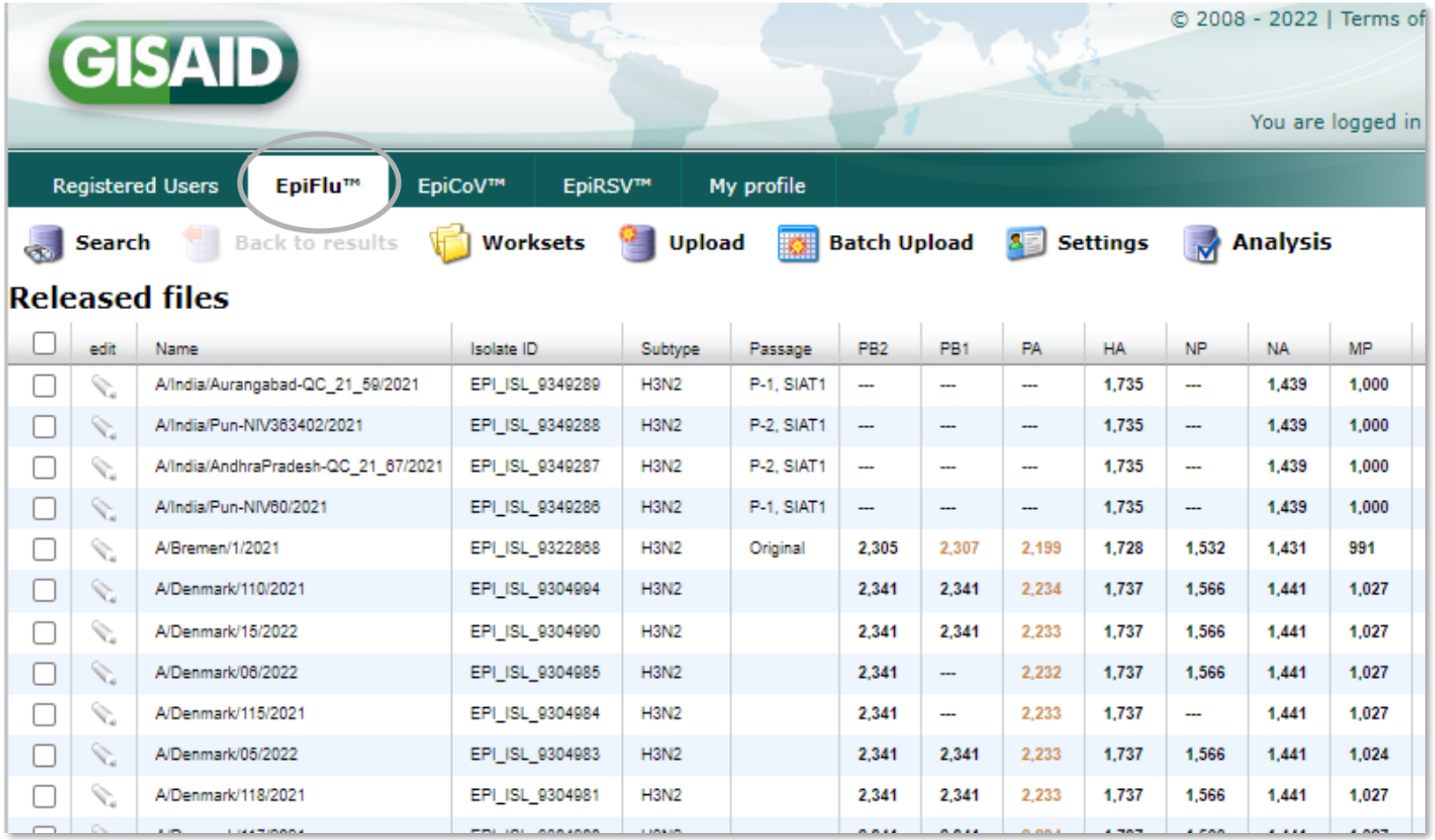
*Example*

**The data**:

**The metadata**:

- author
- publisher
- year
- keywords
- genre

etc…




| edit | Name | Isolate ID | Subtype | Passage | PB2 | PB1 | PA | HA | NP | NA | MP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A/India/Aurangabad-QC_21_59/2021 | EPI_ISL_9349289 | H3N2 | P-1, SIAT1 | --- | --- | --- | 1,735 | --- | 1,439 | 1,000 |
| | A/India/Pun-NIV383402/2021 | EPI_ISL_9349288 | H3N2 | P-2, SIAT1 | --- | --- | --- | 1,735 | --- | 1,439 | 1,000 |
| | A/India/AndhraPradesh-QC_21_67/2021 | EPI_ISL_9349287 | H3N2 | P-2, SIAT1 | --- | --- | --- | 1,735 | --- | 1,439 | 1,000 |
| | A/India/Pun-NIV60/2021 | EPI_ISL_9349286 | H3N2 | P-1, SIAT1 | --- | --- | --- | 1,735 | --- | 1,439 | 1,000 |
| | A/Bremen/1/2021 | EPI_ISL_9322868 | H3N2 | Original | 2,305 | 2,307 | 2,199 | 1,728 | 1,532 | 1,431 | 991 |
| | A/Denmark/110/2021 | EPI_ISL_9304994 | H3N2 | | 2,341 | 2,341 | 2,234 | 1,737 | 1,566 | 1,441 | 1,027 |
| | A/Denmark/15/2022 | EPI_ISL_9304990 | H3N2 | | 2,341 | 2,341 | 2,233 | 1,737 | 1,566 | 1,441 | 1,027 |
| | A/Denmark/06/2022 | EPI_ISL_9304985 | H3N2 | | 2,341 | --- | 2,232 | 1,737 | 1,566 | 1,441 | 1,027 |
| | A/Denmark/115/2021 | EPI_ISL_9304984 | H3N2 | | 2,341 | --- | 2,233 | 1,737 | --- | 1,441 | 1,027 |
| | A/Denmark/05/2022 | EPI_ISL_9304983 | H3N2 | | 2,341 | 2,341 | 2,233 | 1,737 | 1,566 | 1,441 | 1,024 |
| | A/Denmark/118/2021 | EPI_ISL_9304981 | H3N2 | | 2,341 | 2,341 | 2,233 | 1,737 | 1,566 | 1,441 | 1,027 |

**The data**: ATGCTCATGGAC…

**The metadata**:

- subtype
- host
- location
- sample date
- submitting lab

etc…

# Example Data – Influenza metadata from GISAID EpiFlu database[2]

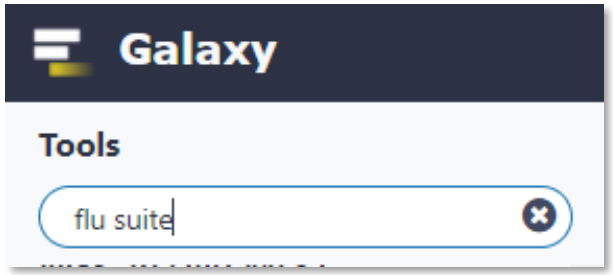
gisaid_metadata.xls

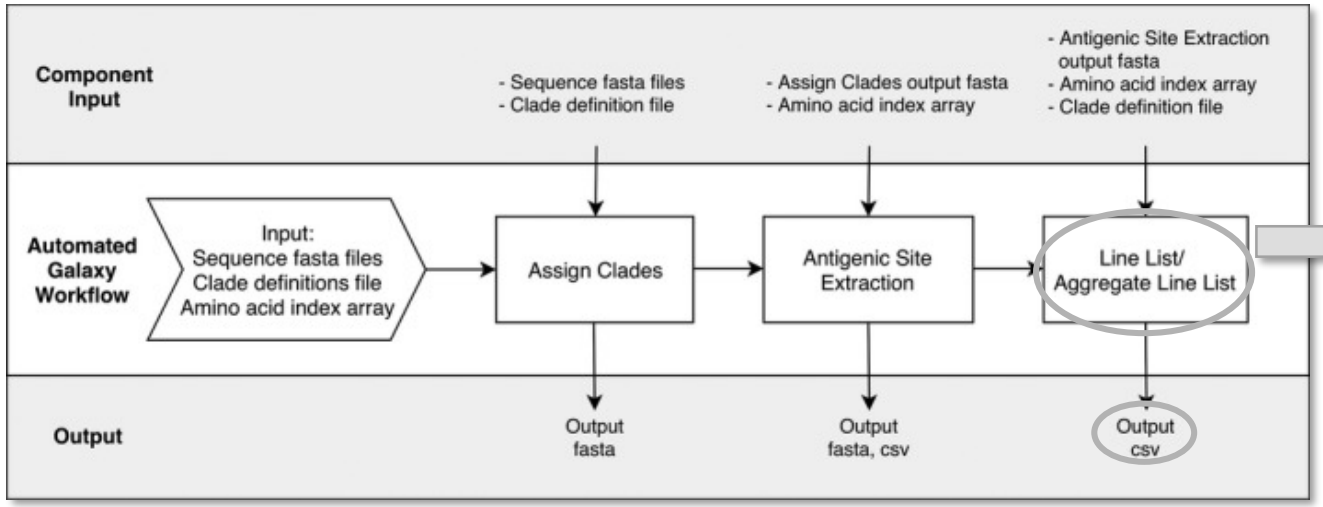
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Isolate_Id | PB2 Segment_Id | PB1 Segmen | PA Segment_ | HA Segment_Id | NP Segment_ | NA Segment_ | MP Segment | NS Segment_ | HE Segment_ | P3 Segment_ | Isolate_Name | Subtype | Lineage |
| 2 | EPI_ISL_505940 | | | | EPI1789728\|A/swir | | EPI1789729\|A | | | | | A/swine/South Dakota/A02245350/2019 | A / H3N2 | |
| 3 | EPI_ISL_505938 | | | | EPI1789724\|A/swir | | EPI1789725\|A | | | | | A/swine/Minnesota/A02245351/2019 | A / H3N2 | |
| 4 | EPI_ISL_505915 | | | | EPI1789660\|A/swir | | EPI1789661\|A | | | | | A/swine/North Carolina/A02478865/201 | A / H3N2 | |
| 5 | EPI_ISL_502566 | | | | EPI1774147\|A/swir | | EPI1774148\|A | | | | | A/swine/Iowa/A02479185/2020 | A / H3N2 | |
| 6 | EPI_ISL_499123 | EPI1766439\|A/swine/ | EPI1766403\|A | EPI1766323\|A | EPI1766067\|A/swir | EPI1766293\|A | EPI1766161\|A | EPI1766237\|A | EPI1766496\|A | | | A/swine/Spain/090/2018 | A / H3N2 | |
| 7 | EPI_ISL_487158 | EPI1755388\|LNS9440 | EPI1755389\|L | EPI1755387\|L | EPI1755391\|LNS9 | EPI1755384\|L | EPI1755390\|L | EPI1755386\|L | EPI1755385\|L | | | A/Luxembourg/LNS9440568/2020 | A / H3N2 | |
| 8 | EPI_ISL_513958 | EPI1796774\|A_ENG_ | EPI1796775\|A | EPI1796773\|A | EPI1796777\|A_EN | EPI1796770\|A | EPI1796776\|A | EPI1796772\|A | EPI1796771\|A | | | A/England/190480558/2019 | A / H3N2 | |
| 9 | EPI_ISL_486700 | | | | EPI1754046\|52-A_ | | EPI1754047\| | | | | | A/Finland/181/2020 | A / H3N2 | |
| 10 | EPI_ISL_486766 | | | | EPI1754177\|76-A_ | | EPI1754178\| | | | | | A/Switzerland/6206/2020 | A / H3N2 | |
| 11 | EPI_ISL_498294 | | | | EPI1763581\|82-A_ | | EPI1763582\| | | | | | A/Bosnia and Herzegovina/211/2020 | A / H3N2 | |
| 12 | EPI_ISL_498293 | | | | EPI1763579\|77-A_ | | EPI1763580\| | | | | | A/Beirut/331/2020 | A / H3N2 | |
| 13 | EPI_ISL_498357 | | | | EPI1763706\|20-A_ | | EPI1763707\| | | | | | A/Lyon/1826/2020 | A / H3N2 | |
| 14 | EPI_ISL_482811 | | | | EPI1751828\|HA_2 | | | | | | | A/Berlin/29/2020 | A / H3N2 | |
| 15 | EPI_ISL_506028 | | | | EPI1790230\|A/Indi | | | | | | | A/India/Pun-1922052/2019 | A / H3N2 | |
| 16 | EPI_ISL_506017 | | | | EPI1790219\|A/Indi | | | | | | | A/India/Pun-1923665/2019 | A / H3N2 | |
| 17 | EPI_ISL_505458 | EPI1788188\|A/Wuhar | EPI1788181\|A | EPI1788210\|A | EPI1788167\|A/Wul | | EPI1788174\|A | EPI1788194\|A | EPI1788201\|A | | | A/Wuhan/345/2019 | A / H3N2 | |
| 18 | EPI_ISL_505051 | | | | EPI1785567\|A/Beij | | | | | | | A/Beijing/PUMCH22/2017 | A / H3N2 | |
| 19 | EPI_ISL_505049 | | | | EPI1785575\|A/Beij | | EPI1785564\|A | | | | | A/Beijing/PUMCH05/2017 | A / H3N2 | |
| 20 | EPI_ISL_503101 | EPI1776906\|A/Chile/J | EPI1776900\|A | EPI1776904\|A | EPI1776901\|A/Chil | EPI1776899\|A | EPI1776902\|A | EPI1776903\|A | EPI1776905\|A | | | A/Chile/JM-R6138/2001 | A / H3N2 | |
| 21 | EPI_ISL_503075 | EPI1776680\|A/Santia | EPI1776683\|A | EPI1776679\|A | EPI1776681\|A/San | EPI1776678\|A | EPI1776677\|A | EPI1776682\|A | EPI1776684\|A | | | A/Santiago/p004d1/2017 | A / H3N2 | |
| 22 | EPI_ISL_502005 | EPI1770811\|A/Arizon | | EPI1770797\|A | EPI1770741\|A/Ariz | EPI1770776\|A | EPI1770765\|A | EPI1770753\|A | EPI1770786\|A | | | A/Arizona/9775/2019 | A / H3N2 | |
| 23 | EPI_ISL_501997 | EPI1770691\|A/Washi | EPI1770690\|A | EPI1770689\|A | EPI1770684\|A/Wa | EPI1770687\|A | EPI1770686\|A | EPI1770685\|A | EPI1770688\|A | | | A/Washington/9306/2019 | A / H3N2 | |
| 24 | EPI_ISL_501981 | EPI1770564\|A/Spain/ | EPI1770563\|A | EPI1770562\|A | EPI1770557\|A/Spa | EPI1770560\|A | EPI1770559\|A | EPI1770558\|A | EPI1770561\|A | | | A/Spain/9287/2019 | A / H3N2 | |
| 25 | EPI_ISL_501980 | EPI1770556\|A/South | EPI1770555\|A | EPI1770554\|A | EPI1770549\|A/Sou | EPI1770552\|A | EPI1770551\|A | EPI1770550\|A | EPI1770553\|A | | | A/South Korea/9286/2019 | A / H3N2 | |
| 26 | EPI_ISL_501931 | EPI1770172\|A/Germa | EPI1770171\|A | EPI1770170\|A | EPI1770165\|A/Ger | EPI1770168\|A | EPI1770167\|A | EPI1770166\|A | EPI1770169\|A | | | A/Germany/9209/2019 | A / H3N2 | |
| 27 | EPI_ISL_501873 | | EPI1769886\|A | EPI1769871\|A | EPI1769647\|A/New | EPI1769789\|A | EPI1769743\|A | EPI1769695\|A | EPI1769835\|A | | | A/New York/9251/2019 | A / H3N2 | |
| 28 | EPI_ISL_501862 | | EPI1769883\|A | | EPI1769636\|A/Italy | EPI1769778\|A | EPI1769732\|A | EPI1769684\|A | EPI1769824\|A | | | A/Italy/9229/2019 | A / H3N2 | |
| 29 | EPI_ISL_501539 | | | | EPI1768143\|A/Chir | | | | | | | A/China/31/2017 | A / H3N2 | |
| 30 | EPI_ISL_501529 | | | | EPI1768133\|A/Chir | | | | | | | A/China/21/2017 | A / H3N2 | |
| 31 | EPI_ISL_514718 | | | | EPI1797926\|N1003 | | EPI1797925\|N | EPI1797924\|N | | | | A/Malaysia/RP0701/2019 | A / H3N2 | |
| 32 | EPI_ISL_510011 | | | | EPI1795434\|N1003 | | EPI1795433\|N | EPI1795432\|N | | | | A/South Africa/9178/2019 | A / H3N2 | |
| 33 | EPI_ISL_514710 | | | | EPI1797902\|N1003 | | EPI1797901\|N | EPI1797900\|N | | | | A/Singapore/KK0001/2020 | A / H3N2 | |

# Example Data – Influenza metadata from GISAID EpiFlu database[2]

**41** Influenza A / H3N2 isolates x **63** columns of data

.xls file




Taxonomic clade assignments
+
Comparison with reference sequence

**= metadata**

# Line list – Analysis with Influenza Classification Suite[3]

line_list_galaxy.csv

Galaxy PROJECT

| | A | B | C | D | E | | DY | DZ | EA | EB | EC | ED | EE | EF | EG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | 44 | | 305 | 307 | 308 | 309 | 310 | 311 | 312 | | |
| 2 | Clade_3C.2a_A/Hong_Kong/4801/2014_X-263B_EGG | | | | Q | N | | R | Y | V | K | H | S | | |
| 3 | Sequence Name | N | Clade | Extra Substitutions | | | | | | | | | | Number of Amino Acid Substitutions | % Identity of Antigenic Site Residues |
| 4 | A/Arizona/9775/2019 | | 3C.3a | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 5 | A/Beijing/PUMCH05/2017 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 12 | 0.908396947 |
| 6 | A/Beijing/PUMCH22/2017 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 12 | 0.908396947 |
| 7 | A/Beirut/331/2020 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 17 | 0.870229008 |
| 8 | A/Berlin/29/2020 | | 3C.3a | | . | . | | . | . | . | . | Q | . | 12 | 0.908396947 |
| 9 | A/Bosnia_and_Herzegovina | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 10 | A/Bretagne/963/2020 | | No_Match | | . | . | | . | . | . | . | Q | . | 15 | 0.885496183 |
| 11 | A/Chile/JM-R6138/2001 | | No_Match | | . | S | | . | . | . | . | Q | N | 25 | 0.809160305 |
| 12 | A/China/21/2017 | | 3C.2a_+_N121K_+_S144K | | . | . | | . | . | . | . | . | . | 9 | 0.93129771 |
| 13 | A/China/31/2017 | | 3C.2a_+_N121K_+_S144K | | . | . | | . | . | . | . | . | . | 9 | 0.93129771 |
| 14 | A/England/190480558/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 12 | 0.908396947 |
| 15 | A/Finland/181/2020 | | 3C.3a | | . | . | | . | . | . | . | Q | . | 11 | 0.916030534 |
| 16 | A/Germany/9209/2019 | | 3C.3a | | . | . | | . | . | . | . | Q | . | 11 | 0.916030534 |
| 17 | A/Haiti/394/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 18 | A/Hawaii/28/2020 | | No_Match | | . | S | | K | . | . | . | Q | K | 45 | 0.65648855 |
| 19 | A/India/Pun-1922052/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 20 | A/India/Pun-1923665/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 21 | A/Italy/9229/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 22 | A/Luxembourg/LNS9440568 | | No_Match | | . | . | | . | . | . | . | Q | . | 12 | 0.908396947 |
| 23 | A/Lyon/1826/2020 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 17 | 0.870229008 |
| 24 | A/Macedonia/364/2020 | | 3C.3a | | . | . | | . | . | . | . | Q | . | 11 | 0.916030534 |
| 25 | A/Malaysia/RP0701/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 17 | 0.870229008 |
| 26 | A/New_York/9251/2019 | | 3C.3a | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 27 | A/Niger/7221/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 17 | 0.870229008 |
| 28 | A/Perth/20/2020 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 17 | 0.870229008 |
| 29 | A/Santiago/p004d1/2017 | | 3C.2a1_+_R142G | | . | . | | . | . | . | . | . | . | 6 | 0.954198473 |
| 30 | A/Singapore/KK0001/2020 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 16 | 0.877862595 |
| 31 | A/South_Africa/9178/2019 | | No_Match | | . | . | | . | . | . | . | Q | . | 14 | 0.893129771 |
| 32 | A/South_Korea/9286/2019 | | No_Match | | . | . | | . | . | . | . | Q | . | 13 | 0.900763359 |
| 33 | A/Spain/9287/2019 | | 3C.3a | | . | . | | . | . | . | . | Q | . | 11 | 0.916030534 |
| 34 | A/Switzerland/6206/2020 | | No_Match | | . | . | | . | . | . | . | Q | . | 17 | 0.870229008 |
| 35 | A/Virginia/03/2020 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 17 | 0.870229008 |
| 36 | A/Washington/9306/2019 | | 3C.2a1_+_N121K_+_K92R_+_H311Q | | . | . | | . | . | . | . | Q | . | 12 | 0.908396947 |

# Line list – Analysis with Influenza Classification Suite[3]

**41** Influenza A / H3N2 isolates x **137** columns of data!

.csv file

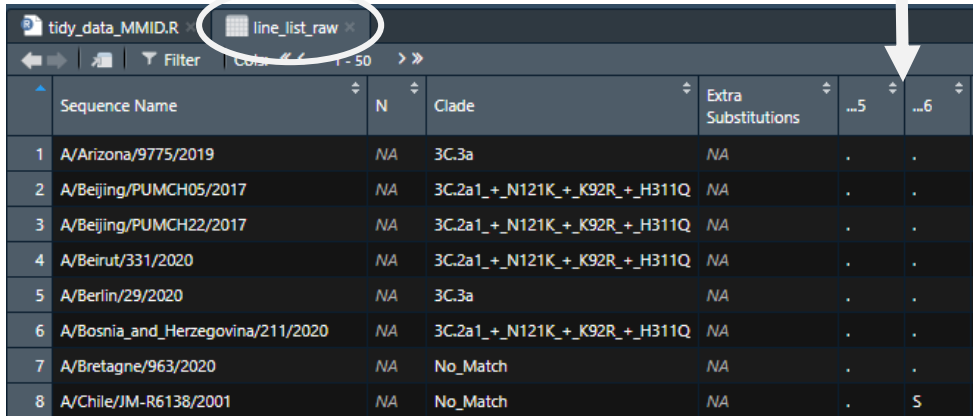
# Code

```
library(readr)

line_list_raw <-
read_csv("workshop_materials_Feb9/data/line_list_galaxy.csv",
skip = 2,
show_col_types = FALSE)

View(line_list_raw)
```

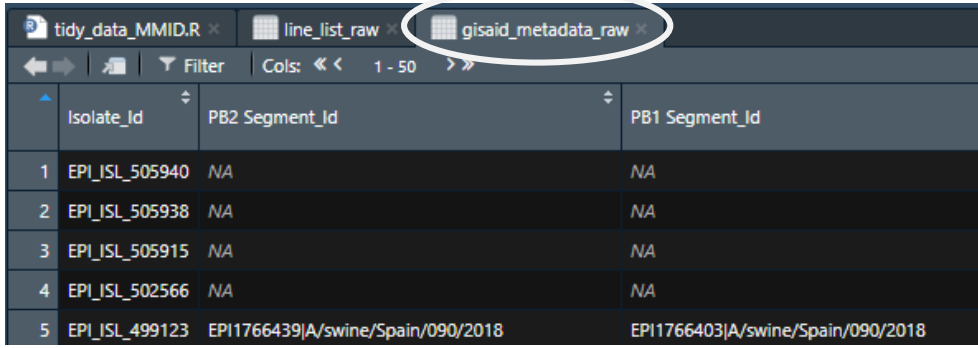
```
library(readxl)

gisaid_metadata_raw <-
read_excel("workshop_materials_Feb9/data/gisaid_metadata.xls")

View(gisaid_metadata_raw)
```

Rstudio

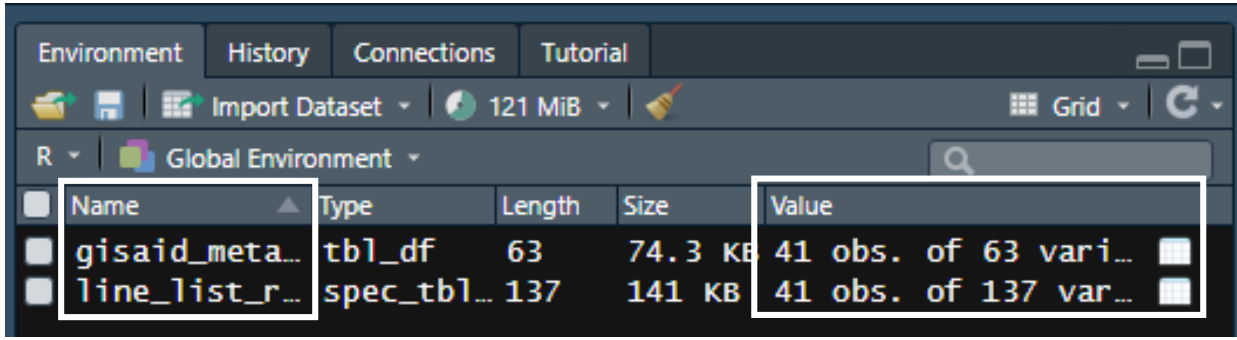
"...#"




| | Sequence Name | N | Clade | Extra Substitutions | ...5 | ...6 |
|---|---|---|---|---|---|---|
| 1 | A/Arizona/9775/2019 | NA | 3C.3a | NA | . | . |
| 2 | A/Beijing/PUMCH05/2017 | NA | 3C.2a1_+_N121K_+_K92R_+_H311Q | NA | . | . |
| 3 | A/Beijing/PUMCH22/2017 | NA | 3C.2a1_+_N121K_+_K92R_+_H311Q | NA | . | . |
| 4 | A/Beirut/331/2020 | NA | 3C.2a1_+_N121K_+_K92R_+_H311Q | NA | . | . |
| 5 | A/Berlin/29/2020 | NA | 3C.3a | NA | . | . |
| 6 | A/Bosnia_and_Herzegovina/211/2020 | NA | 3C.2a1_+_N121K_+_K92R_+_H311Q | NA | . | . |
| 7 | A/Bretagne/963/2020 | NA | No_Match | NA | . | . |
| 8 | A/Chile/JM-R6138/2001 | NA | No_Match | NA | . | S |




| | Isolate_Id | PB2 Segment_Id | PB1 Segment_Id |
|---|---|---|---|
| 1 | EPI_ISL_505940 | NA | NA |
| 2 | EPI_ISL_505938 | NA | NA |
| 3 | EPI_ISL_505915 | NA | NA |
| 4 | EPI_ISL_502566 | NA | NA |
| 5 | EPI_ISL_499123 | EPI1766439\|A/swine/Spain/090/2018 | EPI1766403\|A/swine/Spain/090/2018 |



# Remove unwanted columns:

Rstudio

```
library(dplyr)

line_list_keep <- line_list_raw %>%
        select("Sequence Name",
               "Clade",
               "Number of Amino Acid Substitutions in
                       Antigenic Sites",
               "% Identity of Antigenic Site Residues") %>%

        rename(Isolate_Name = "Sequence Name",
               Num.AA.Sub = "Number of Amino Acid
                       Substitutions in Antigenic Sites",
               percent.id = "% Identity of Antigenic
                       Site Residues")


View(line_list_keep)

metadata_keep <- gisaid_metadata_raw %>%
        select(12, 1, 13, 16, 17, 19, 25, 28, 31:43)

View(metadata_keep)
```

tidy_data_MMID.R | metadata_keep | line_list_raw | gisaid_metadata_raw

Filter

| Isolate_Name | Isolate_Id | Subtype | Location |
|---|---|---|---|
| A/swine/South Dakota/A02245350/2019 | EPI_ISL_505940 | A / H3N2 | North America / United Sta |
| A/swine/Minnesota/A02245351/2019 | EPI_ISL_505938 | A / H3N2 | North America / United Sta |
| A/swine/North Carolina/A02478865/2019 | EPI_ISL_505915 | A / H3N2 | North America / United Sta |
| A/swine/Iowa/A02479185/2020 | EPI_ISL_502566 | A / H3N2 | North America / United Sta |
| A/swine/Spain/090/2018 | EPI_ISL_499123 | A / H3N2 | Europe / Spain |
| A/Luxembourg/LNS9440568/2020 | EPI_ISL_487158 | A / H3N2 | Europe / Luxembourg |
| A/England/190480558/2019 | EPI_ISL_513958 | A / H3N2 | Europe / United Kingdom / |
| A/Finland/181/2020 | EPI_ISL_486700 | A / H3N2 | Europe / Finland |
| A/Switzerland/6206/2020 | EPI_ISL_486766 | A / H3N2 | Europe / Switzerland |
| A/Bosnia and Herzegovina/211/2020 | EPI_ISL_498294 | A / H3N2 | Europe / Bosnia and Herze |
| A/Beirut/331/2020 | EPI_ISL_498293 | A / H3N2 | Asia / Lebanon |
| A/Lyon/1826/2020 | EPI_ISL_498357 | A / H3N2 | Europe / France |
| A/Berlin/29/2020 | EPI_ISL_482811 | A / H3N2 | Europe / Germany |
| A/India/Pun-1922052/2019 | EPI_ISL_506028 | A / H3N2 | Asia / India |
| A/India/Pun-1923665/2019 | EPI_ISL_506017 | A / H3N2 | Asia / India |
| A/Wuhan/345/2019 | EPI_ISL_505458 | A / H3N2 | Asia / China |
| A/Beijing/PUMCH22/2017 | EPI_ISL_505051 | A / H3N2 | Asia / China |
| A/Beijing/PUMCH05/2017 | EPI_ISL_505049 | A / H3N2 | Asia / China |
| A/Chile/JM-R6138/2001 | EPI_ISL_503101 | A / H3N2 | South America / Chile |
| A/Santiago/p004d1/2017 | EPI_ISL_503075 | A / H3N2 | South America / Chile |
| A/Arizona/9775/2019 | EPI_ISL_502005 | A / H3N2 | North America / United Sta |
| A/Washington/9306/2019 | EPI_ISL_501997 | A / H3N2 | North America / United Sta |

⋮

# Combining data sets:

Rstudio

*A **"key"** variable can be used to uniquely identify each observation (row).*

```
a <- line_list_keep$Isolate_Name %>% sort()
b <- metadata_keep$Isolate_Name %>% sort()
summary(a == b)
```

```
   Mode   FALSE    TRUE
logical       7      34
```

```
> a[29]
[1] "A/South_Africa/9178/2019"
> b[29]
[1] "A/South Africa/9178/2019"
```

```
library(stringr)
metadata_keep$Isolate_Name <- metadata_keep$Isolate_Name %>%
                              str_replace_all(" ", "_")
```

```
   Mode    TRUE
logical      41
```

```
metadata_full <- full_join(metadata_keep, line_list_keep,
                           by = "Isolate_Name")

View(metadata_full)
```

# Cleaning the combined data:

Rstudio

```
library(tidyr)
metadata_clean <- metadata_full %>%
        mutate(Collection_Year = substr(Collection_Date, 1, 4),
                Host_Gender = if_else(Host_Gender %in%
                                c("Male", "M"), "Male",
                                if_else(Host_Gender %in%
                                c("Female", "F"), "Female",
                                if_else(is.na(Host_Gender) ==
                                FALSE, "Other", "NA"))),
                Host_Age_Y = if_else(Host_Age_Unit == "M",
                                Host_Age / 12, Host_Age),
                .keep = "unused") %>%
```

| Collection_Date | Host_Gender | Host_Age | Host_Age_Unit |
|---|---|---|---|
| 2020-02-25 | Female | 50 | Y |
| 2020-03-16 | Male | 54 | Y |
| 2020-02-03 | X | 84 | Y |
| 2019-08-05 | Female | 94 | Y |
| 2019-09-01 | Male | NA | Y |
| 2019-01 | Male | NA | Y |
| 2017-08-07 | Male | NA | Y |
| 2017-08-07 | Female | NA | Y |
| 2001 | M | NA | Y |
| 2017 | Female | 24 | M |
| 2019-05-10 | Male | 48 | M |
| 2019-02-03 | Other | 7 | NA |
| 2019-02-22 | Self-Identify | 14 | NA |
| 2019-03-17 | NA | 24 | NA |
| 2019-03-11 | NA | 25 | NA |
| 2019-03-27 | Female | 33 | NA |
| 2019-02-21 | Female | | |
| 2017-10-23 | Male | | |
| | Female | | |
| | Female | | |
| | F | | |

*Remember `?functionName` if you need help!

# Cleaning the combined data: Rstudio

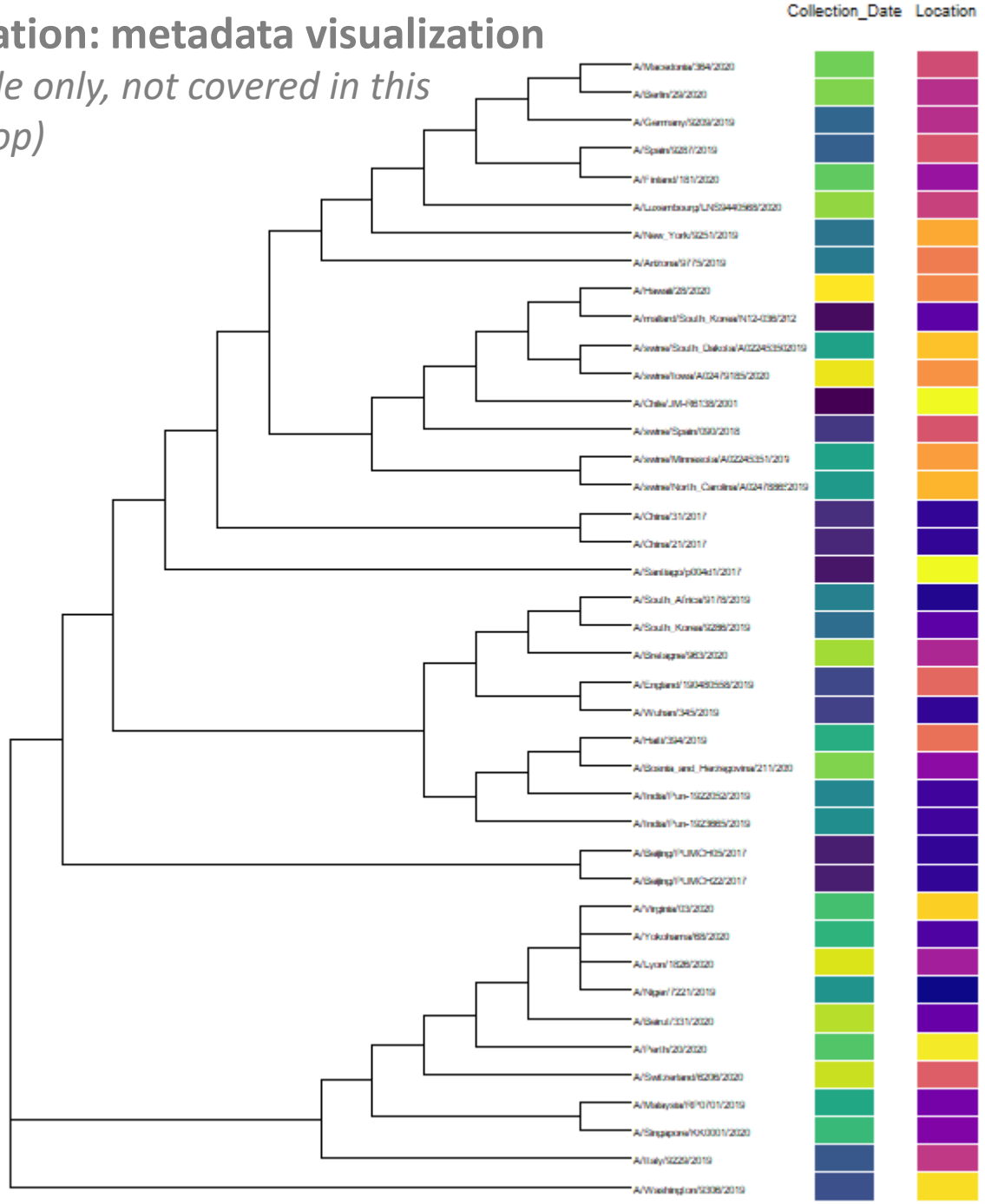
```
library(tidyr)
metadata_clean <- metadata_full %>%
        mutate(Collection_Year = substr(Collection_Date, 1, 4),
                Host_Gender = if_else(Host_Gender %in%
                                c("Male", "M"), "Male",
                               if_else(Host_Gender %in%
                                c("Female", "F"), "Female",
                               if_else(is.na(Host_Gender) ==
                                FALSE, "Other", "NA"))),
                Host_Age_Y = if_else(Host_Age_Unit == "M",
                                Host_Age / 12, Host_Age),
                .keep = "unused") %>%
        separate(col = Location,
                        c("Continent", "Country", "Region"),
                        sep = "/")
View(metadata_clean)
```

| Location |
|---|
| Oceania / Australia |
| North America / United States / Virginia |
| Asia / Japan |
| North America / Haiti |
| Europe / France / Bretagne |
| Asia / Malaysia |
| Asia / Singapore |
| Europe / United Kingdom / LEICESTER |
| North America / United States / Hawaii |
| Europe / Macedonia, the former Yogoslav Republic of |
| Africa / South Africa / Province of the Western Cape |
| Asia / Korea, Republic of |
| Asia / India |
| Asia / India |
| North America / United States / South Dakota |
| North America / United States / Minnesota |
| North America / United States / North Carolina |
| Asia / China |

*Remember `?functionName` if you need help!

# Application: metadata visualization

*(Example only, not covered in this workshop)*




Location

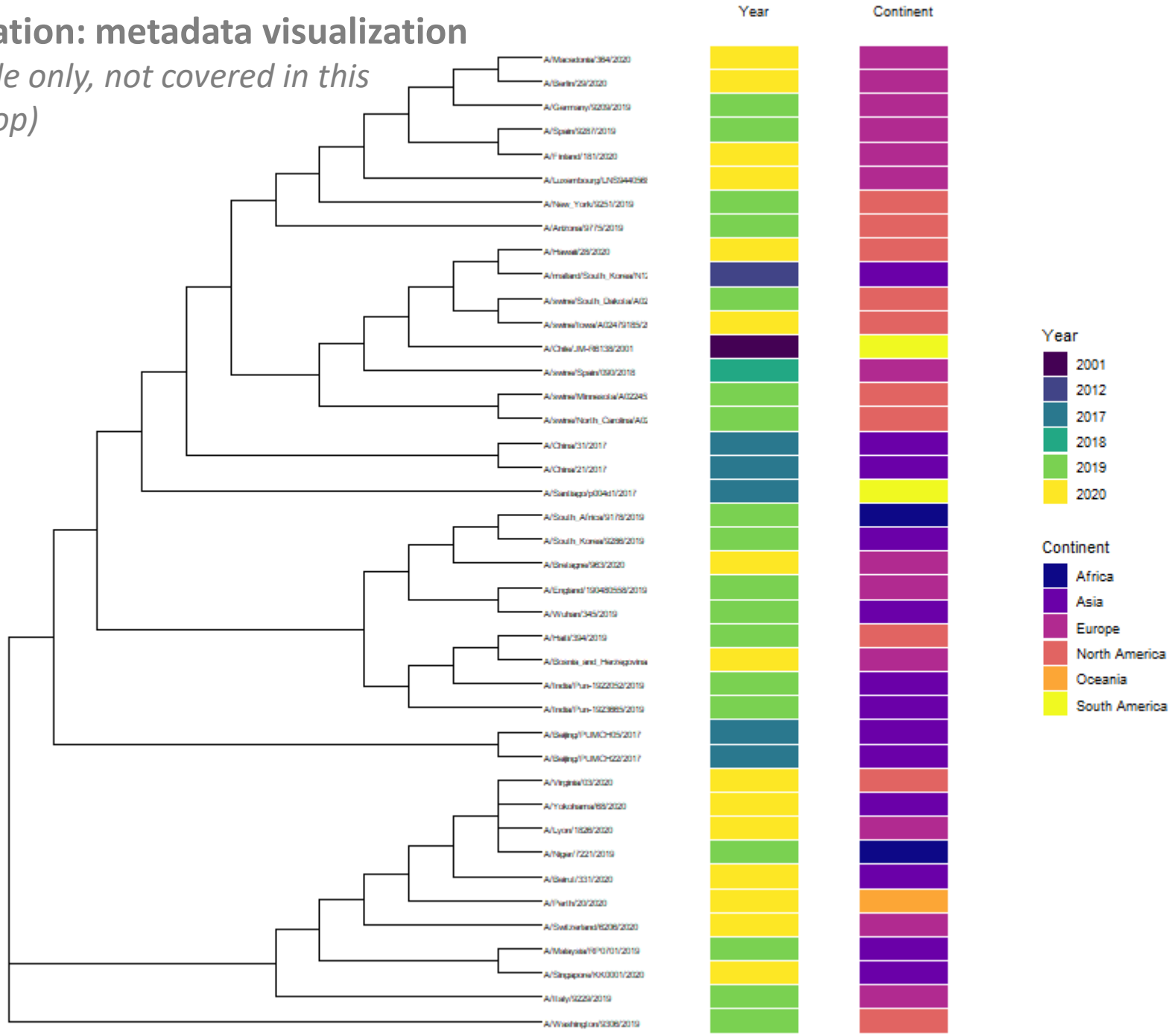
- Africa / Niger
- Africa / South Africa / Province of the Western Cape
- Asia / China
- Asia / India
- Asia / Japan
- Asia / Korea, Republic of
- Asia / Lebanon
- Asia / Malaysia
- Asia / Singapore
- Europe / Bosnia and Herzegovina
- Europe / Finland
- Europe / France
- Europe / France / Bretagne
- Europe / Germany
- Europe / Italy
- Europe / Luxembourg
- Europe / Macedonia, the former Yogoslav Republic of
- Europe / Spain
- Europe / Switzerland
- Europe / United Kingdom / LEICESTER
- North America / Haiti
- North America / United States / Arizona
- North America / United States / Hawaii
- North America / United States / Iowa
- North America / United States / Minnesota
- North America / United States / New York
- North America / United States / North Carolina
- North America / United States / South Dakota
- North America / United States / Virginia
- North America / United States / Washington
- Oceania / Australia
- South America / Chile

Collection_Date

- 2001
- 2012-12-11
- 2017
- 2017-08-07
- 2017-09-14
- 2017-10-23
- 2018-01-23
- 2019-01
- 2019-01-23
- 2019-02-03
- 2019-02-21
- 2019-02-22
- 2019-03-11
- 2019-03-17
- 2019-03-27
- 2019-05-10
- 2019-07-22
- 2019-08-05
- 2019-09-01
- 2019-09-25
- 2019-11-13
- 2019-12-02
- 2019-12-05
- 2019-12-16
- 2020
- 2020-01-02
- 2020-01-23
- 2020-01-29
- 2020-02-01
- 2020-02-02
- 2020-02-03
- 2020-02-07
- 2020-02-10
- 2020-02-25
- 2020-03-09
- 2020-03-16
- 2020-03-27
- 2020-07-01

# Application: metadata visualization

*(Example only, not covered in this workshop)*



| | Isolate_Name | Isolate_Id | Subtype | Continent | Country | Region | Host | Submitting_Lab | Submissio |
|---|---|---|---|---|---|---|---|---|---|
| 1 | A/swine/South_Dakota/A02245350/2019 | EPI_ISL_505940 | A / H3N2 | North America | United States | South Dakota | Swine | NA | 2020-01-1 |
| 2 | A/swine/Minnesota/A02245351/2019 | EPI_ISL_505938 | A / H3N2 | North America | United States | Minnesota | Swine | NA | 2020-01-1 |
| 3 | A/swine/North_Carolina/A02478865/2019 | EPI_ISL_505915 | A / H3N2 | North America | United States | North Carolina | Swine | NA | 2020-01-1 |
| 4 | A/swine/Iowa/A02479185/2020 | EPI_ISL_502566 | A / H3N2 | North America | United States | Iowa | Swine | NA | 2020-04-2 |
| 5 | A/swine/Spain/090/2018 | EPI_ISL_499123 | A / H3N2 | Europe | Spain | NA | Swine | NA | 2020-01-1 |
| 6 | A/Luxembourg/LNS9440568/2020 | EPI_ISL_487158 | A / H3N2 | Europe | Luxembourg | NA | Human | Laboratoire National de Santé | 2020-07-1 |
| 7 | A/England/190480558/2019 | EPI_ISL_513958 | A / H3N2 | Europe | United Kingdom | LEICESTER | Human | Microbiology Services Colindale, Public Health England | 2020-08-1 |
| 8 | A/Finland/181/2020 | EPI_ISL_486700 | A / H3N2 | Europe | Finland | NA | Human | Crick Worldwide Influenza Centre | 2020-07-1 |
| 9 | A/Switzerland/6206/2020 | EPI_ISL_486766 | A / H3N2 | Europe | Switzerland | NA | Human | Crick Worldwide Influenza Centre | 2020-07-1 |
| 10 | A/Bosnia_and_Herzegovina/211/2020 | EPI_ISL_498294 | A / H3N2 | Europe | Bosnia and Herzegovina | NA | Human | Crick Worldwide Influenza Centre | 2020-07-2 |
| 11 | A/Beirut/331/2020 | EPI_ISL_498293 | A / H3N2 | Asia | Lebanon | NA | Human | Crick Worldwide Influenza Centre | 2020-07-2 |
| 12 | A/Lyon/1826/2020 | EPI_ISL_498357 | A / H3N2 | Europe | France | NA | Human | Crick Worldwide Influenza Centre | 2020-07-2 |
| 13 | A/Berlin/29/2020 | EPI_ISL_482811 | A / H3N2 | Europe | Germany | NA | Human | Robert Koch Institute Nationales Referenzzentrum für Influe... | 2020-07-0 |
| 14 | A/India/Pun-1922052/2019 | EPI_ISL_506028 | A / H3N2 | Asia | India | NA | Human | NA | 2020-01-1 |
| 15 | A/India/Pun-1923665/2019 | EPI_ISL_506017 | A / H3N2 | Asia | India | NA | Human | NA | 2020-01-1 |
| 16 | A/Wuhan/345/2019 | EPI_ISL_505458 | A / H3N2 | Asia | China | NA | Human | NA | 2019-10-2 |
| 17 | A/Beijing/PUMCH22/2017 | EPI_ISL_505051 | A / H3N2 | Asia | China | NA | Human | NA | 2018-01-0 |
| 18 | A/Beijing/PUMCH05/2017 | EPI_ISL_505049 | A / H3N2 | Asia | China | NA | Human | NA | 2018-01-0 |
| 19 | A/Chile/JM-R6138/2001 | EPI_ISL_503101 | A / H3N2 | South America | Chile | NA | Human | NA | 2019-06-1 |
| 20 | A/Santiago/p004d1/2017 | EPI_ISL_503075 | A / H3N2 | South America | Chile | NA | Human | NA | 2019-06-1 |
| 21 | A/Arizona/9775/2019 | EPI_ISL_502005 | A / H3N2 | North America | United States | Arizona | Human | NA | 2020-06-1 |
| 22 | A/Washington/9306/2019 | EPI_ISL_501997 | A / H3N2 | North America | United States | Washington | Human | NA | 2020-06-1 |
| 23 | A/Spain/9287/2019 | EPI_ISL_501981 | A / H3N2 | Europe | Spain | NA | Human | NA | 2020-06-1 |
| 24 | A/South_Korea/9286/2019 | EPI_ISL_501980 | A / H3N2 | Asia | Korea, Republic of | NA | Human | NA | 2020-06-1 |
| 25 | A/Germany/9209/2019 | EPI_ISL_501931 | A / H3N2 | Europe | Germany | NA | Human | NA | 2020-06-1 |

```
View(metadata_clean)
```

# Saving the final data set:

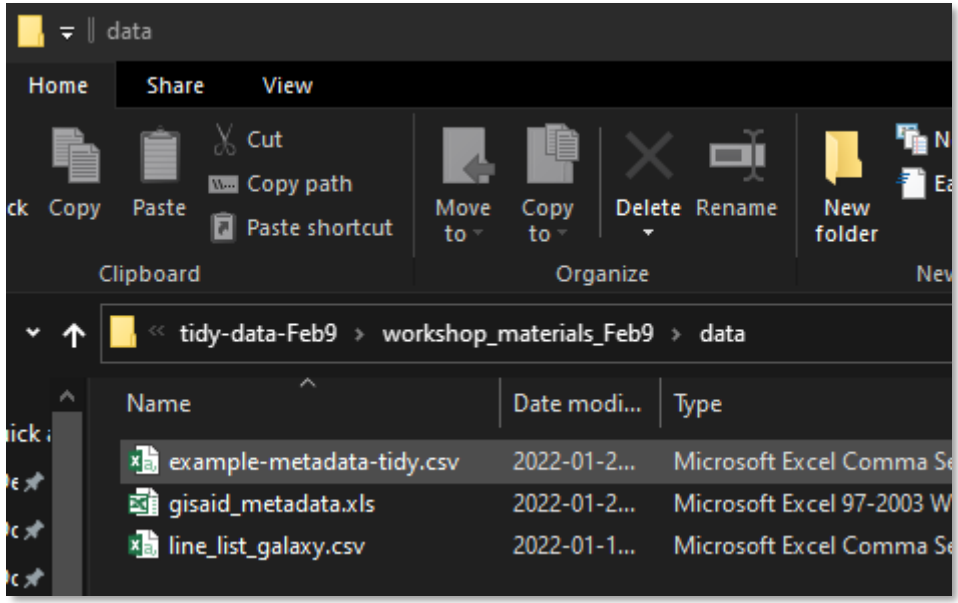
*One final rearrangement before we save:*

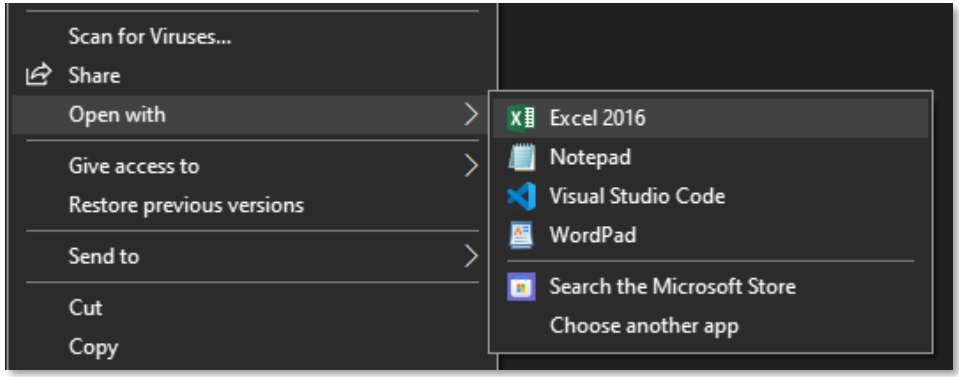
*(put related variables next to each other)*

```
metadata_final <- metadata_clean %>% select(1:3, 21, 23,
7, 20, 25, 4:6, 24, 8:19)

View(metadata_final)

write_csv(metadata_final, "YOUR-PATH/metadata-tidy.csv")
```

path *(exists)* name *(new)*

## File explorer / GUI

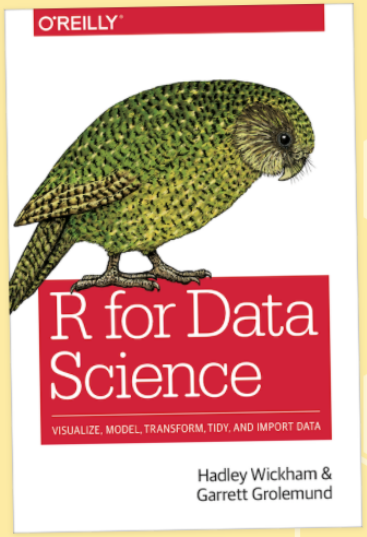


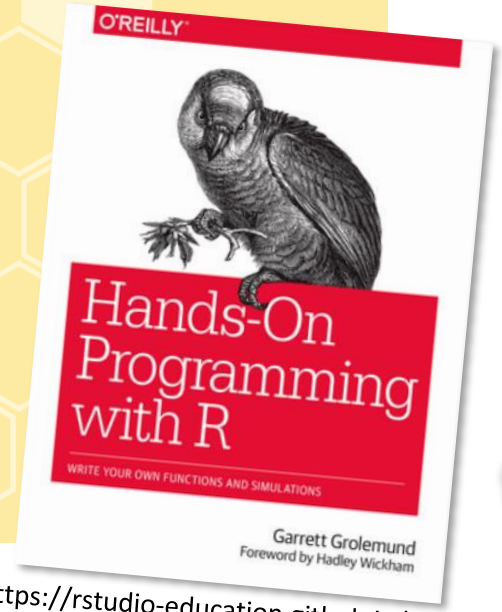



# HELPFUL (and free) RESOURCES

Learn the tidyverse

See how the tidyverse makes data science faster, easier and more fun with "R for Data Science". Read it online, buy the book or try another resource from the community.




https://rstudio-education.github.io/hopr/

SKILL TRACK

datacamp.com

## Tidyverse Fundamentals with R

Experience the whole data science pipeline from importing and tidying data to wrangling and visualizing data to modeling and communicating with data. Gain exposure to each component of this pipeline from a variety of different perspectives in this tidyverse R track.

R | 20 hours | 5 Courses | 1 Project

training.galaxyproject.org

**Data Wrangling**
**with dplyr and tidyr**
Cheat Sheet
RStudio

# REFERENCES


1. Grolemund G & Wickham H. *R for Data Science*. O'Reilly Media, 2017. Available at: https://r4ds.had.co.nz/index.html
2. Shu Y and McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Euro Surveill 22, 30494. 2017. doi:10.2807/1560-7917.ES.2017.22.13.30494.
3. Eisler D, Fornika D, Tindale LC, Chan T, Sabaiduc S, Hickman R, Chambers C, Krajden M, Skowronski DM, Jassem A, Hsiao W. **Influenza Classification Suite: An automated Galaxy workflow for rapid influenza sequence analysis.** Influenza Other Respir Viruses. 2020 May;14(3):358-362. doi: 10.1111/irv.12722. Epub 2020 Feb 16. PMID: 32064792; PMCID: PMC7182599.

# THANK YOU FOR ATTENDING!

# *The Q&A Session will now begin.*

***Please make sure to fill out the Exit Survey***

***We value your feedback!***

***More questions? Please email us at***

***mmid.coding.workshop@gmail.com or post them to the workshop slack channel***

*Educational series supported by the Department of MMID to promote knowledge translation and dissemination among students and researchers*