**Data Management – Bahl Lab**

**For help contact:** Lambodhar –Bahl Lab PhD Student – ld26524@uga.edu

**Programs required:**

Microsoft Excel (Or other datasheet management software)

Tableau (Free for students)

R Studio - <https://www.rstudio.com/>

**Useful Resources:**

gbmunge - <https://github.com/sdwfrost/gbmunge>

Introduction to R : <https://datacarpentry.org/R-genomics/01-intro-to-R.html>

ggplot2 (R package) - <https://ggplot2.tidyverse.org/index.html>

**Computer navigation:**

Command Prompt tutorial (PC users) - <https://www.computerhope.com/issues/chusedos.htm>

Codeacademy for Terminal (Mac/ Linux users)  - <https://www.codecademy.com/learn/learn-the-command-line>

**Flu databases:**

GISAID - <https://www.gisaid.org/>

NCBI Influenza virus database - <https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi>

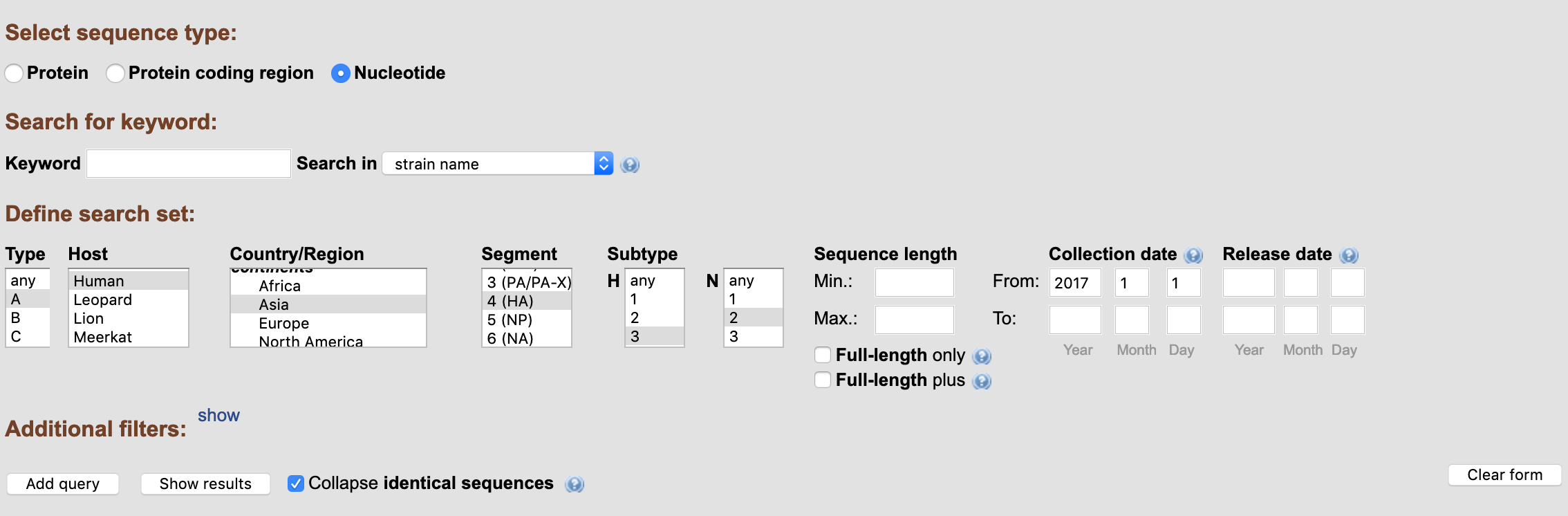
IRD (Influenza Research Database) -  <https://www.fludb.org/brc/home.spg?decorator=influenza>

This tutorial is to help students understand how to collect and manage publically available data. It will describe methods of data collection and methods for the visualization of data.

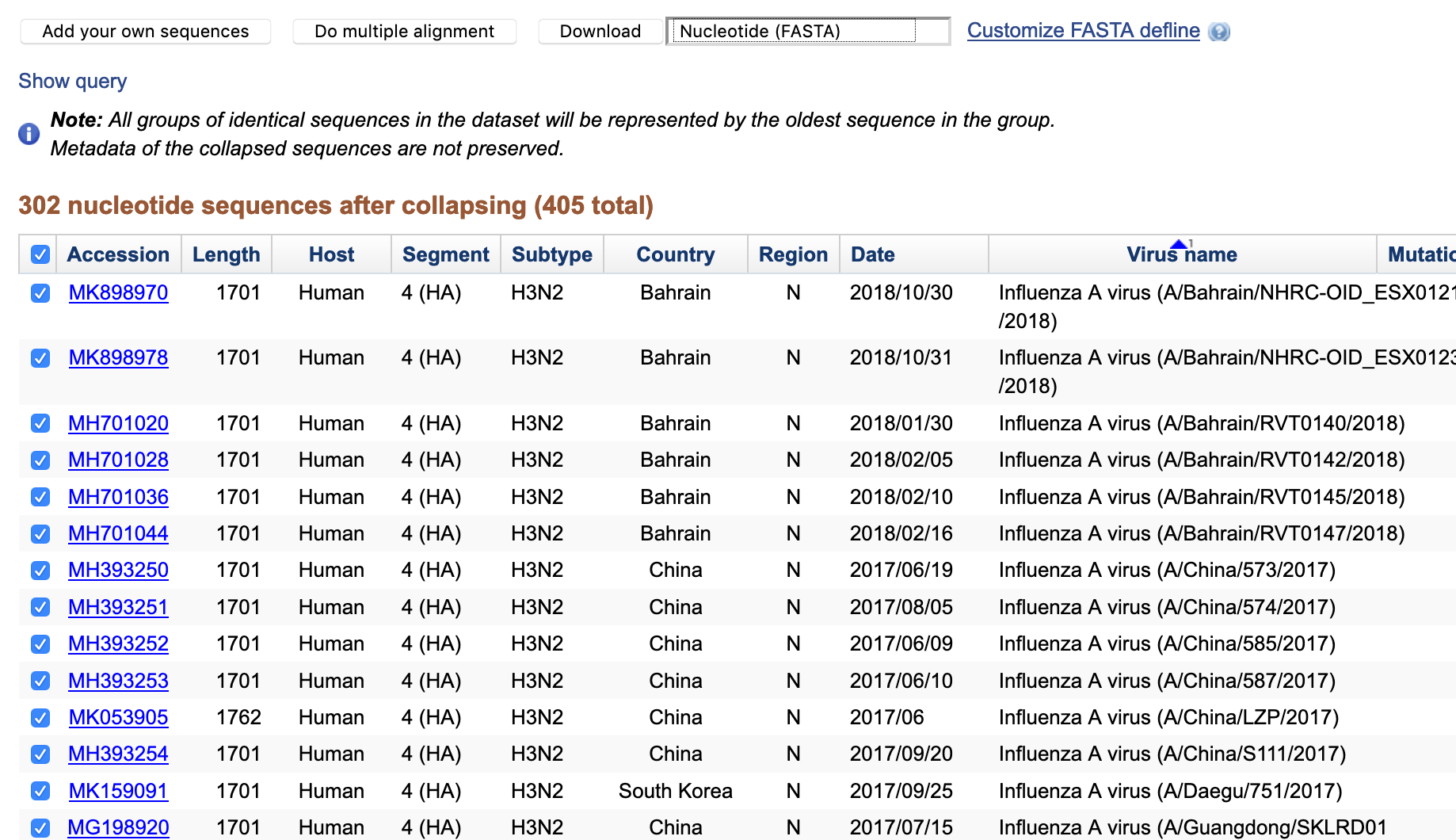
**Data Collection**

There are several different methods of collecting data for studying influenza and other viruses. Major databases such as GISAID (Global Initiative on Sharing All Influenza Data) and NCBI (National Center for Biotechnology Information in the United States) host a large amount of public data available to researchers. In this tutorial, we will learn how to collect data from NCBI for use in molecular epidemiological analyses. This tutorial can be applied to datasets from GISAID and other data sources.

Influenza virus database is a resource that allows users to collect all available public data for a given query. In the example below we are searching for all sequence data for the HA protein of H3N2 Influenza A viruses from humans in Asia from 1/1/2017 to the present.

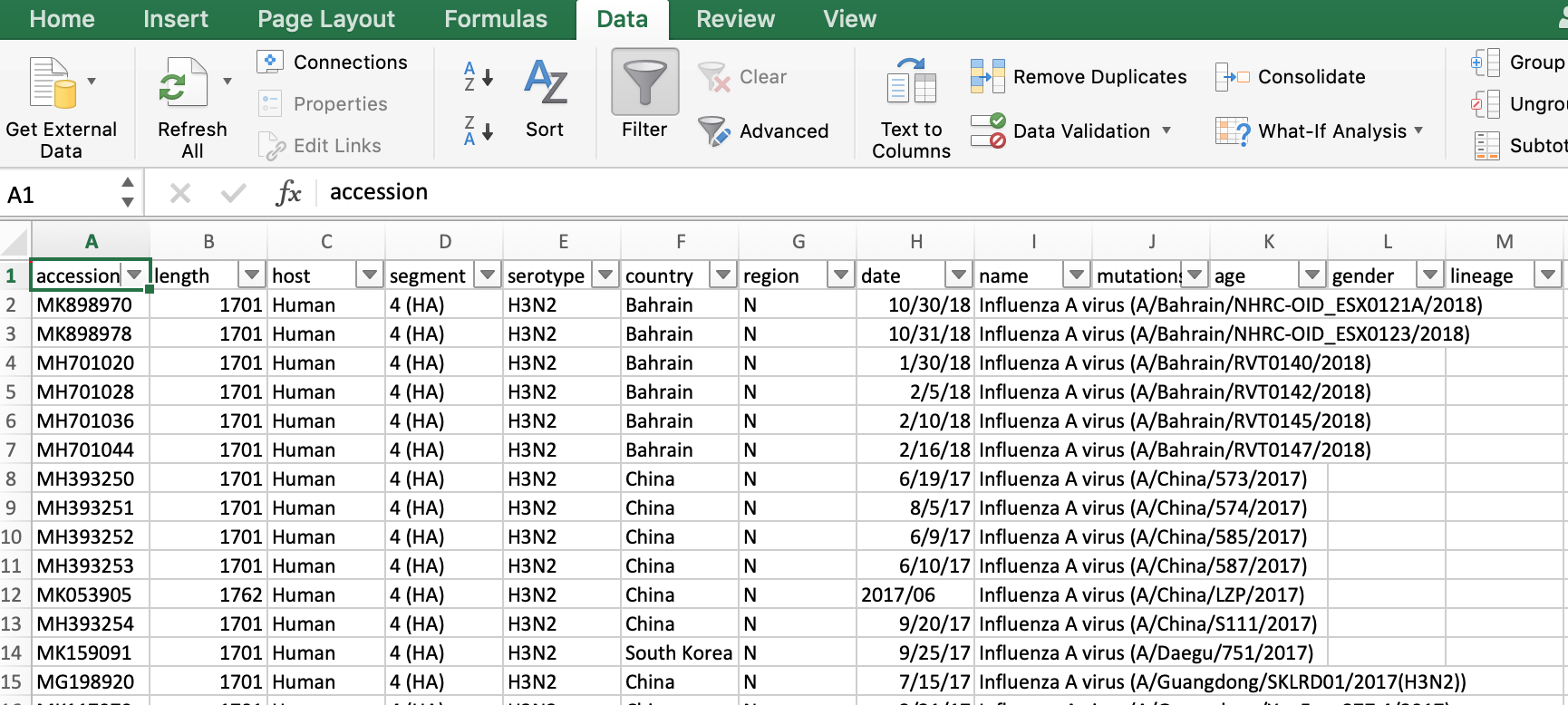


The above query will result in the page seen below:



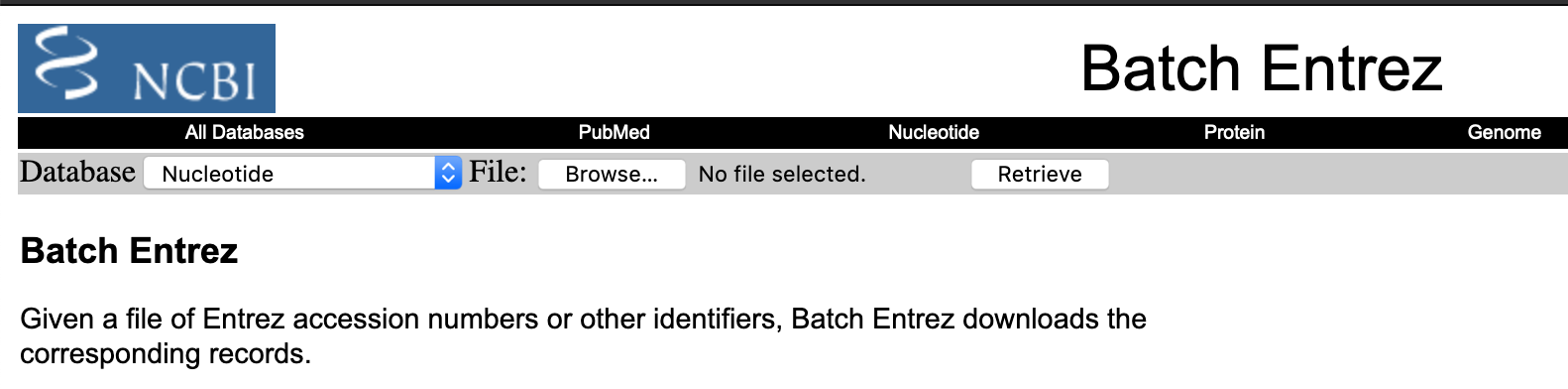
This data can be downloaded in several different ways. The sequence data can be downloaded in fasta format using the download button with the FASTA option. Additionally, the sheet seen above can be downloaded in csv (comma separated value) format. The csv file can be open in excel and is a way to store data. The accession list which is a text file that contains a list of all the accessions can be used in the NCBI tool “Batch Entrez” to retrieve these records as well.

In the example below the data that was retrieved can be filtered based on a column value using the filter option in the data tab of excel.

****

Below is the portal for Batch Entrez, this webpage is a tool that allows you to upload a list of accessions and retrieve the NCBI Genbank records.

**----**

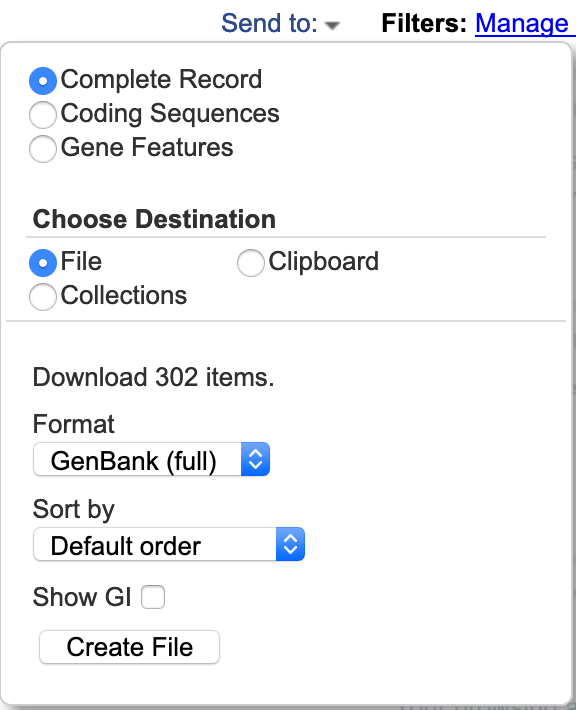


**-----**

This is an example of Genbank nucleotide records. Specific search criteria and filters can be used to edit your search for more specific results.

****

The data can be downloaded as a Genbank file by clicking the Send to table and choosing the option for “Complete Record” and a destination of “File”. See below for example:



Below is an example of a Genbank file format. The Genbank file format is a text based file which takes the Genbank record from NCBI and separates each record with given information. Each record has a given accession number that allows the user to identify that specific sequence. In the file below different data fields such as country or collection\_date are valuable for studying the phylogeography of the organism.



The program gbmunge can take Genbank files (.gb) and extract the sequence data into a fasta format and metadata into a csv file as well as combine both into a csv file. This is very useful if directly downloading sequence data from NCBI’s Genbank.

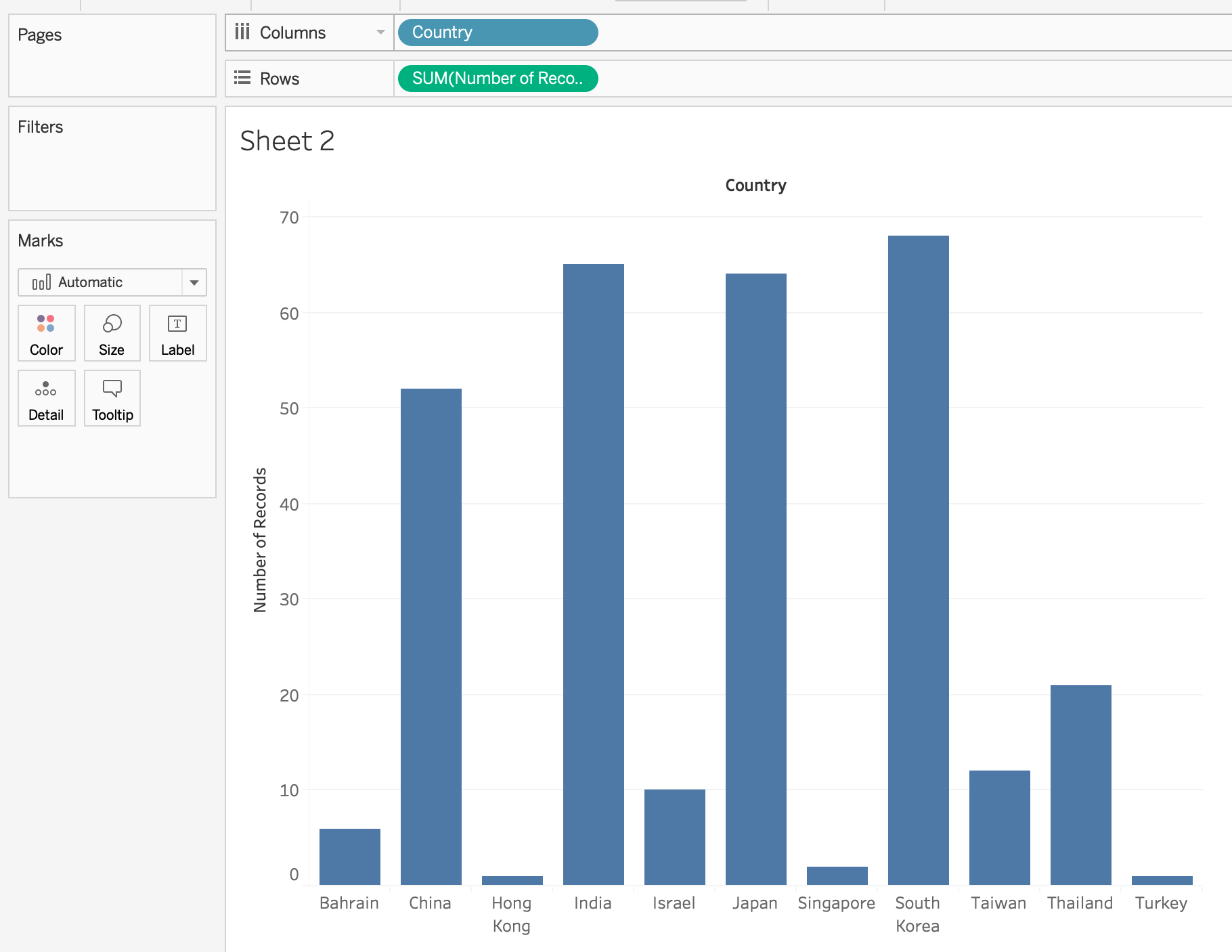
*/Users/ldamodar/gbmunge/src/gbmunge -i sequence.gb -f sequence.fas -o metadata.csv –t -s*

Additionally, when extracting data from resources such as VIPR or the IRD, there are options to download relevant metadata in csv format.

**Data Visualization and Descriptive Stats**

**Tableau**

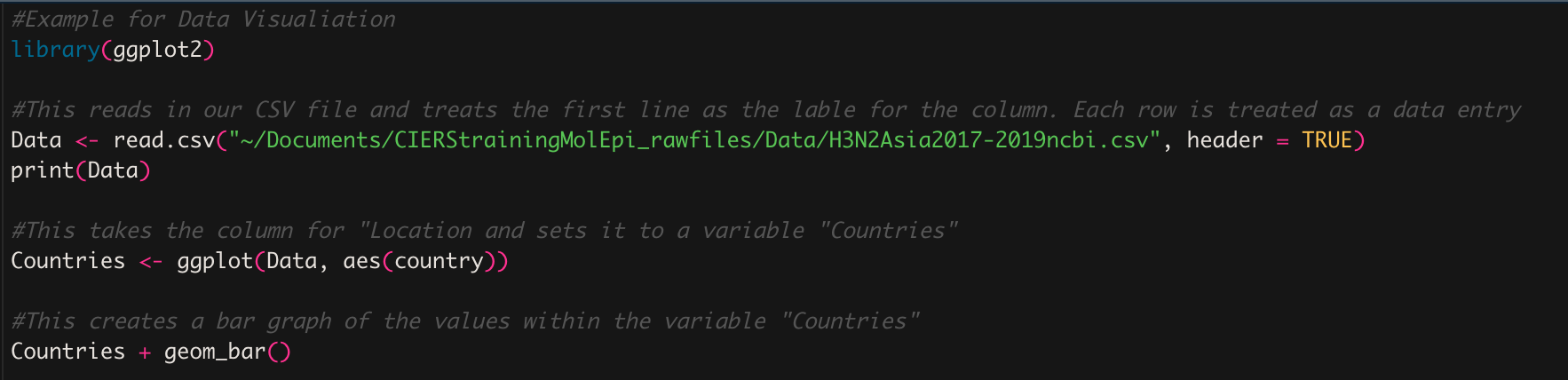
Tableau is a data management software that is offered free to students. It allows you to important data in several different formats and create a wide range of different figures from that data. If we want to get a good idea of what information we have for our sequences we can visualize it through tableau using the steps seen below.

****

The column and row values can be manipulated to provide different orientations to your data and pair information, for example data and country of collection can be paired to see the distribution of samples from a country within a given time period.

**R Studio**

R studio is software that provides a programming environment for coding the computer language R. This programming language its very useful for data manipulation and visualization, additionally it is very useful for statistical analysis of data. For infectious disease researchers, it is a very useful language to learn to help analyze and represent raw data. The tutorial linked in the beginning of this document is a good introduction to the language. Below is an example of how to create an R script that takes data in the form of csv file and creates a bar chart with the counts of entries from one column of the data file. The code below uses an R library called ggplot2 which is a popular library for making information graphics and to organize data.



The result of the above code is the plot below. This plot can be manipulated using different commands and different labels can be given.

Data/Rplot.pdf

Visualization of data allows you to make important decisions about how to conduct your analysis and what biases there might be within your dataset. In the example above we can see that there is a large amount of sequence data from China, India, Japan, and South Korea, this would mean that a researcher should consider a subsampling scheme in order prevent any geographic signals in the data from being obscured by oversampling.

When understanding your data descriptive stats allow for reasonable study design. The number of records/sequences, where sequences come from, how many sequences from each place, the range of dates of collection, are all examples of important descriptive stats that can inform any potential problems/biases that you might face when doing your work. By identifying the biases and normalizing data, by using random sampling for example, you create a dataset that is less prone to yielding deceptive or inaccurate results.

It is important to note that the results from the Influenza virus database should scrutinized and examined carefully. The records that are reported are those that include the metadata that would allow it to show up for specific search criteria, if researchers who submitted sequences did not provide enough metadata or accurate metadata that can result in data that is not shown in your search. If you are doing a search like the one we have done above for Asian sequences, you must check specifically for data from countries that sequence data was not reported from such as Vietnam or Malaysia, two very large countries. You can search directly in NCBI’s Genbank or using the specific country criteria in the Influenza virus database to check for other sequence data.

Once all data has been collected and you have done your descriptive statistics, you can move forward with the resulting dataset after your subsampling or removal of undesired records. In the second handout, you can see how to randomly sample your dataset and add metadata to sequence data labeled for use in phylodynamic analysis.