**Exercise 2: Public repositories and multi-sequence alignments**

EMD 531, Spring 2020

The purpose of this exercise is to learn how to download sequence data (and accompanying metadata) from public repositories and align the sequences for downstream analysis.

***\*\*Important***: We will practice in class using our test dataset, but you will need to complete the following with your own dataset to complete the exercise.

**Download prior to class 6 on Jan 30th.**

UGene: <http://ugene.net/download.html>

**Overview:**

**Step 1:** Select a pathogen and simple epidemiological question to test during exercises 2-5.

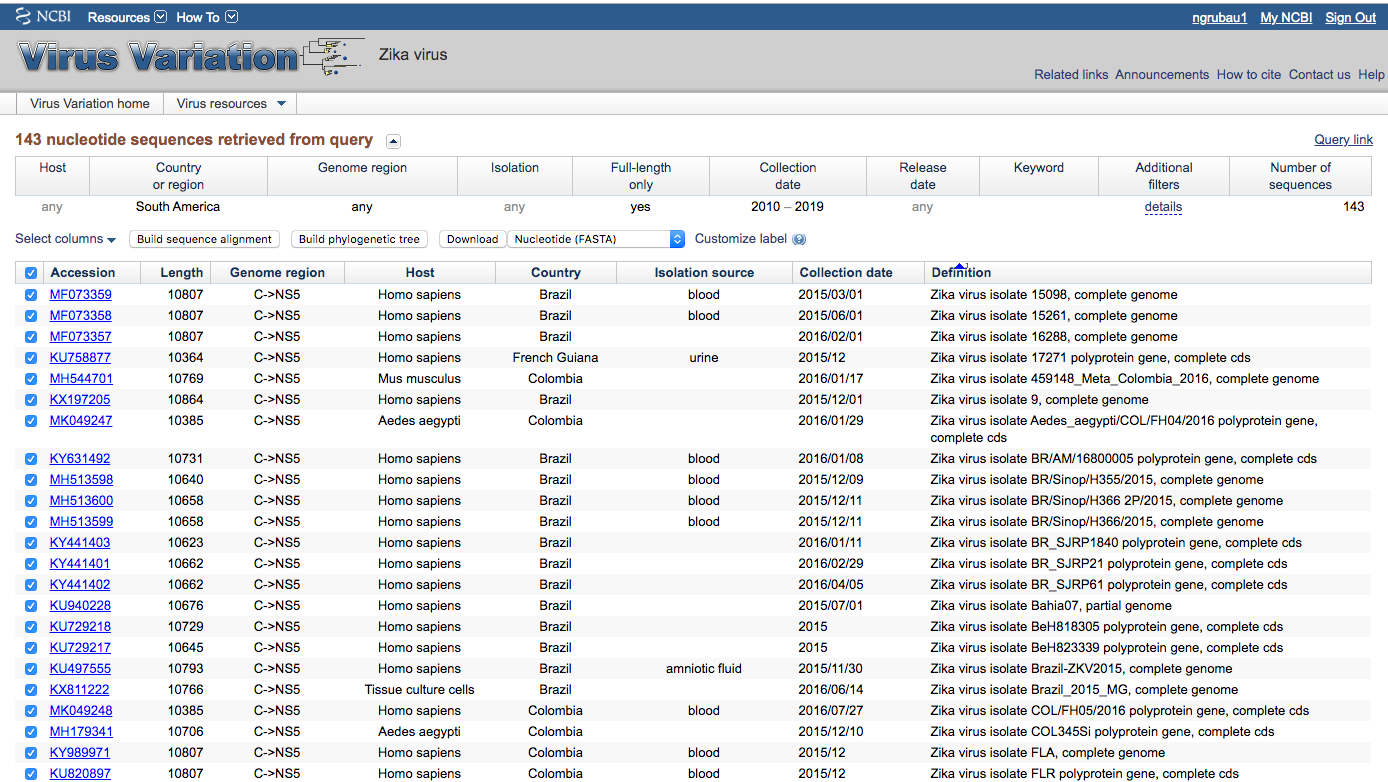
* Pathogen:
  + Can be based on something that you already work with, but please get permission from the instructor first.
  + Prefer selecting a virus as it will make the exercises easier (small genomes, fast rates of evolution).
  + Must get permission from instructor to choose a bacteria or parasite.
* Epi question:
  + Must be answerable with publically available data (can be something that was already answered).
  + Examples:
    - When was pathogen X introduced into region Y?
    - Where was pathogen X introduced from (i.e. what are the origins)?
    - Are all pathogen X outbreaks in region Y related?
    - Are pathogen X outbreaks in region Y and Z related?
    - Class example = When was Zika virus introduced into Brazil?

**Step 2:** Select a region of the genome to analyze and repository to collect sequence data.

* Region of genome:
  + Viruses:
    - If non-segmented, entire coding sequence (CDS; also called open reading frame [ORF])
    - If segmented, chose one segment that is the most relevant to your question (e.g. HA gene of influenza virus)
  + Bacteria:
    - Focus on single gene related to your question (virulence or resistance factor)
    - Try to keep region for analysis <20,000 nucleotides to avoid very long run times.
* Repository:
  + Examples for viruses:
    - https://www.ncbi.nlm.nih.gov/genome/viruses/variation/
    - https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/find-data/virus
    - https://www.viprbrc.org/brc/home.spg?decorator=vipr
  + Examples for bacteria:
    - https://www.ncbi.nlm.nih.gov/genome/microbes/
    - https://bacteria.ensembl.org/index.html
    - <http://mbgd.genome.ad.jp/>

**Step 3**: Develop selection criteria for sequences to download to address epi question.

* Develop a specific search for pathogen, locations, times of collection, and genome region. For example:



* Only select sequences that include important metadata
  + Date (year-month-day)
  + Location (Country, at least, but lower level classification may be needed for your question)
* Keep it relatively simple, do not download more than 50 sequences.

**Step 4**: Download all data as a **FASTA** file and metadata as a **CSV**.

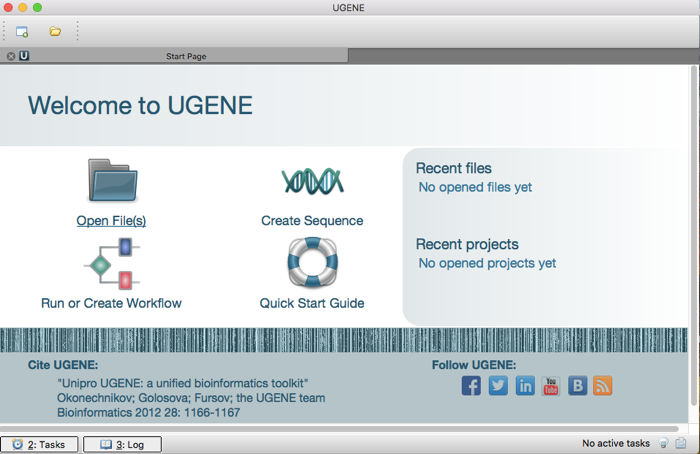
* Select all of the sequences that you want to download, and there is usually a download < FASTA function.
  + Note: In Virus Variation, you can make customized labels to help making remaining headers (below) easier.
* Suggested step: If individual FASTA files, merge all headers and sequences into one file (easier for alignment step, but we can work with individual files too)
  + Open file in a text editor (not Word!), copy and paste headers and sequences into one file. Should look like this where sequences are separated by headers.



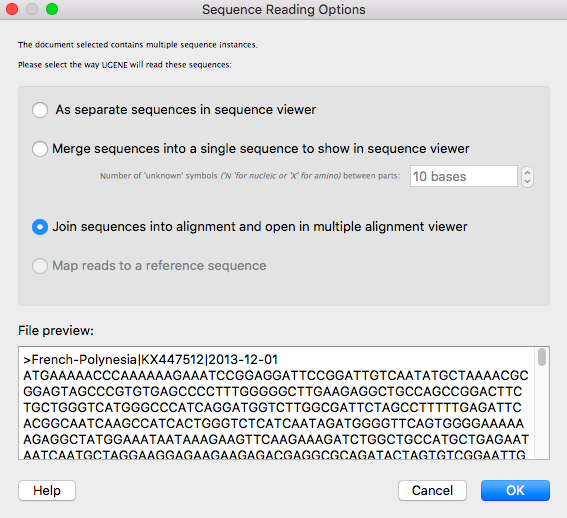
* Rename headers using a text editor using the following format (and see above)
  + > Repository-Identifier|Location|Date (*note, above uses location first, but ID should be first*)
  + Notes:
    - No spaces, use dashes (-) instead
    - Date format: YYYY-MM-DD
* Also, download all of the header information as a CSV, or copy the headers into a CSV. This will be used for Exercise 4.

**Step 5**: Upload data into UGene for multi-sequence alignment.

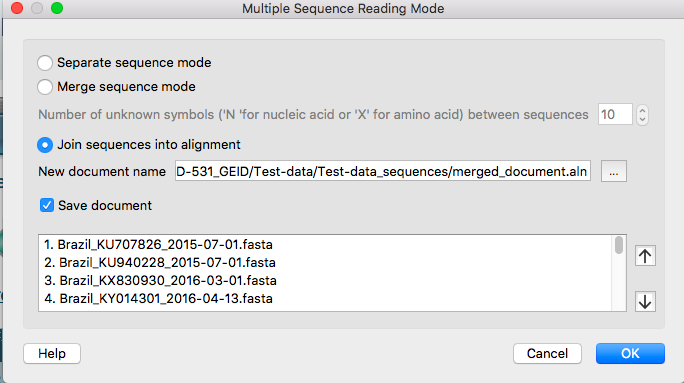
* Open UGene and open your fasta file containing all of your sequences



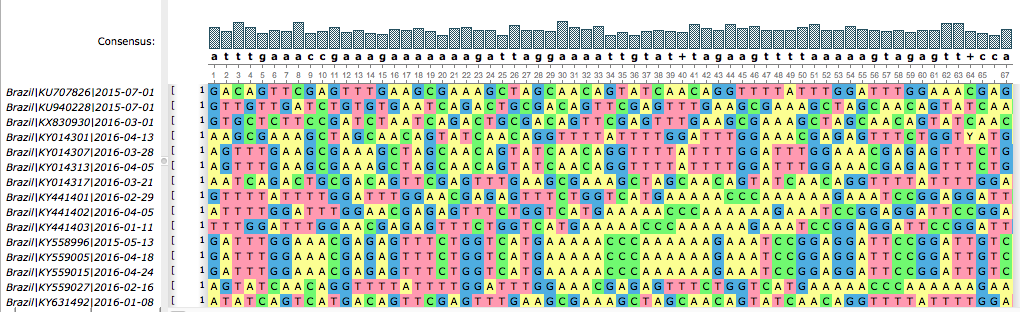
* Select **Join sequences into alignment and open in multiple alignment viewer**



* \*\*Alternatively, if you have individual FASTA files for each sequence, you can select all to open, the select **Join sequences into alignment**

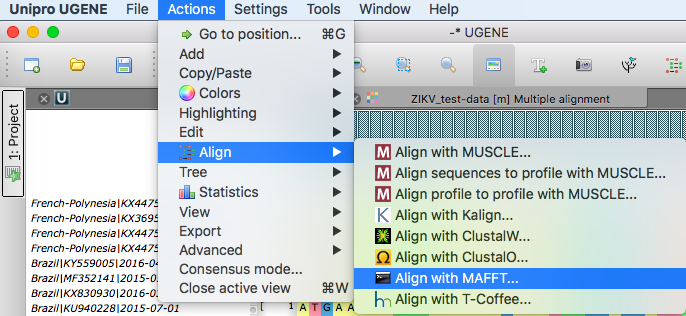


* At this point, the sequences are not aligned.



**Step 6**: Perform multi-sequence alignment.

* To start the multi-sequence alignment, go to:
  + **Actions > Align >** [choose an alignment program]
    - *\*\*Read Chapter 3 in The Phylogenetic Handbook and be prepared to defend your choice in the question below and during our discussion in class.*
  + Do not select any advanced options

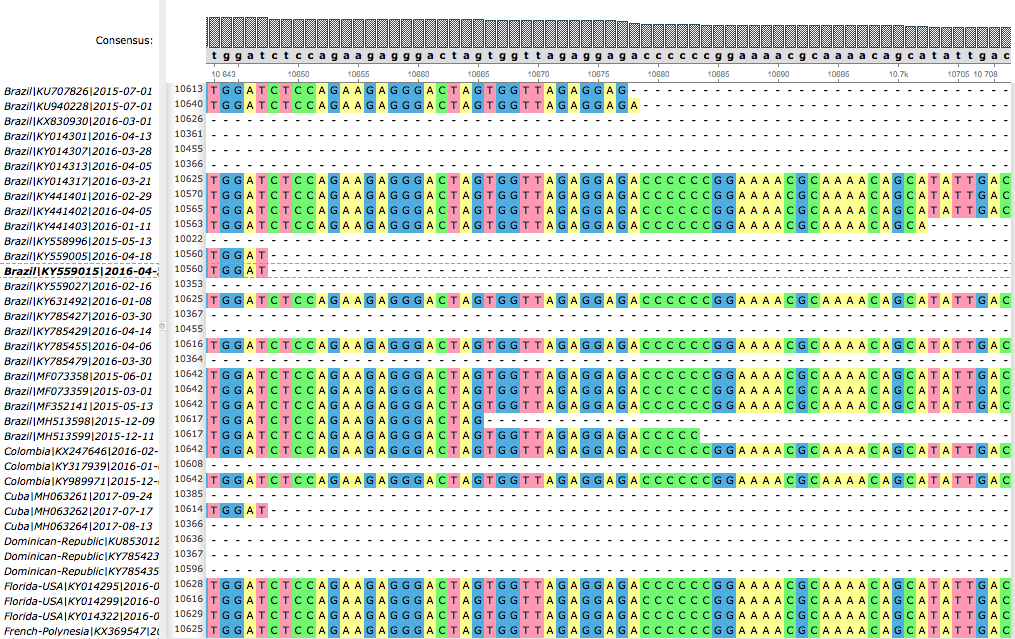


* Now the sequences are aligned and look like this:

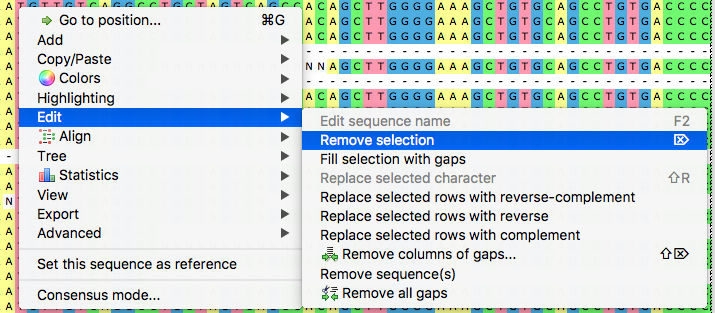


**Step 7**: Trim untranslated regions and/or other genome regions not needed for analysis.

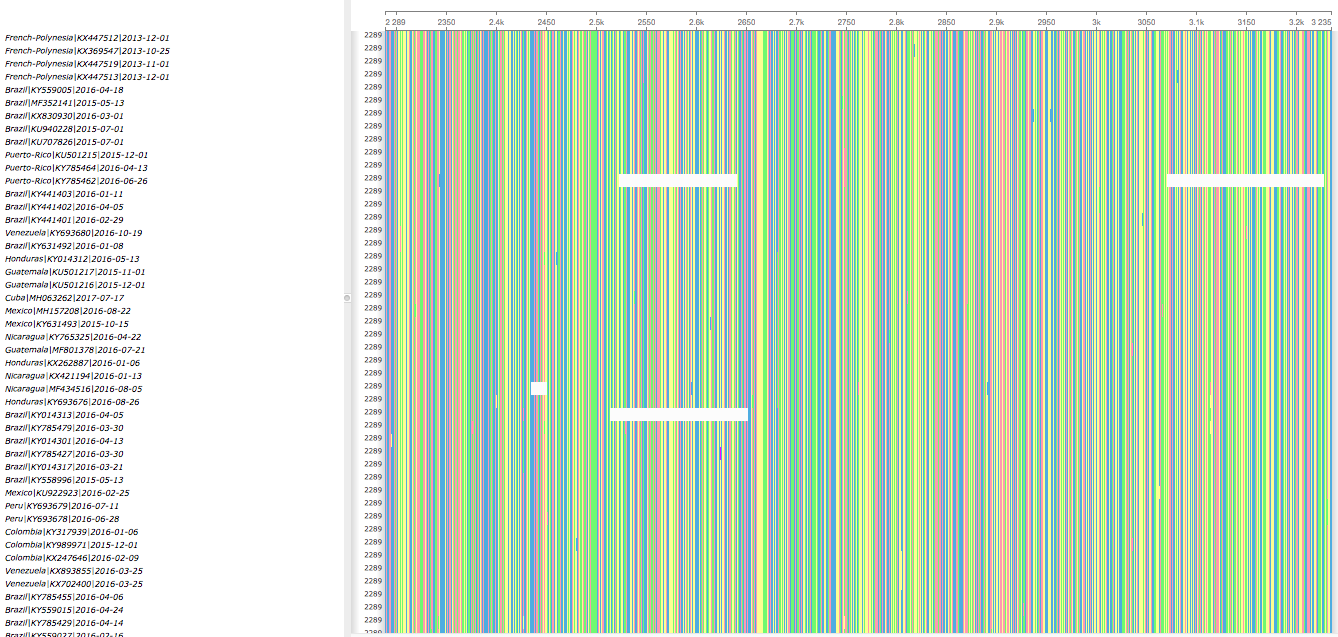
* If your sequences contain any regions that you don’t want to analyze, such as untranslated regions, you’ll want to remove those prior to your phylogenetic analyses (they evolve differently than protein-coding sequences). Also, your sequences may not all be the same length:



* Find the coordinates and sequence motifs for the region of interest. You can find this information by going to GenBank and search for the Reference Sequence for your pathogen. E.g.: <https://www.ncbi.nlm.nih.gov/nuccore/NC_001563.2>
* Start with the 3’ end (right side, as shown above), find the end of your CDS or ORF, verify the sequence motif (numbering may be off from the reference genome),
* To delete parts of the sequences, highlight the columns to remove, and press [backspace].
  + or right click, select: **Edit > Remove selection**.



* + Note that it appears difficult to delete columns that are off screen, so you may need to repeat a few times to remove entire unwanted regions.
* Repeat for 5’ region, and fix any potential gaps.
* Save the file, zoom out as far as you can (magnifying glass symbols), and take a screenshot like below so that I know that you made an alignment (paste as answer to **E3-3**).
  + Note that the white spaces are “NNNNs” – basically missing data. Pathogen sequencing can be messy and incomplete.



**Questions:**

*Combine answers with all exercises,* ***due before class 16 on March 5th****. Be prepared to discuss your answers during class 16.*

**Rerun the steps about using your data, not the ZIKV test. Then answer the following questions.**

**E2-1:** What is your epidemiological question? Briefly describe the pathogen and how phylogenetics can help to answer. (Max 300 words; 5 points)

**E2-2:** Describe how you selected genomic data. What repository did you use, and why? What part of the genome did you target, any why? What were your selection criteria for sequences (e.g. location, time, host, disease)? (Max 300 words; 5 points)

**E2-3**: What are the general criteria for selecting a multiple sequence alignment program? Which program did you choose and why? (Max 200 words; 3 points)

**E3-4**: Paste a screen-shot of the trimmed alignment from your sequences (i.e. not the ZIKV test data). (7 points)