**Exercise 4: Interactive phylogenetics using Nextstrain**

EMD 531, Spring 2020

The purpose of this exercise is to learn how to construct an interactive phylogenetics visualization webpage using the Nextstrain pipeline. E.g. <https://nextstrain.org/WNV/NA>

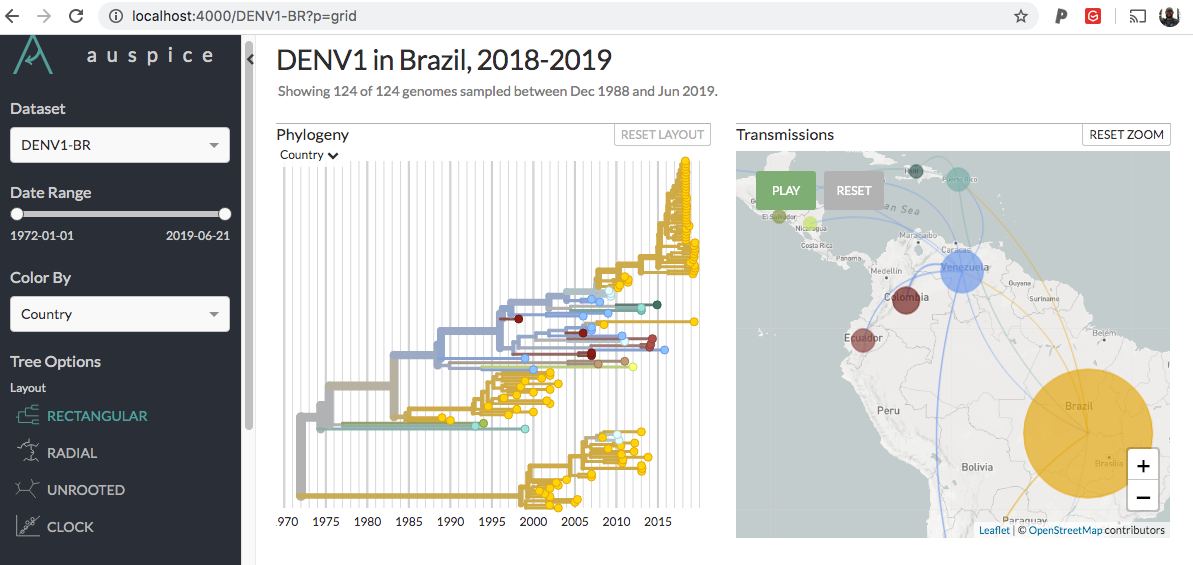
***\*\*Important***: Before class, you must install augur (bioinformatics pipeline) and auspice (visualization tool) so that you can participate. Follow the installation instructions here: <https://github.com/grubaughlab/nextstrain_course>

We will practice in class using our test dataset, but you will need to complete steps 2-4 with your own dataset to complete the exercise.

**Overview:**

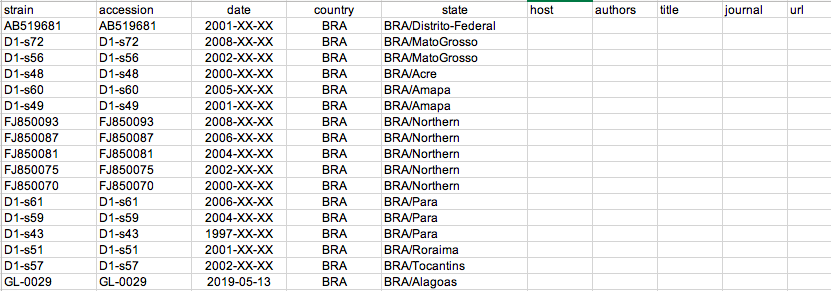
**Step 1: Nextstrain tutorial**

* Follow the instructions listed in “tutorial\_nextstrain.pdf” found here:
  + <https://github.com/grubaughlab/nextstrain_course>
* This will be done in class and will teach you how to:
  + Organize your data
  + Create a reference genome
  + Add your information to the build
  + Run the augur bioinformatic pipeline
  + Visualize your data with auspice
* Using this tutorial, you should create a Nextstrain page that looks like this:



**Step 2: Format your metadata and sequences**

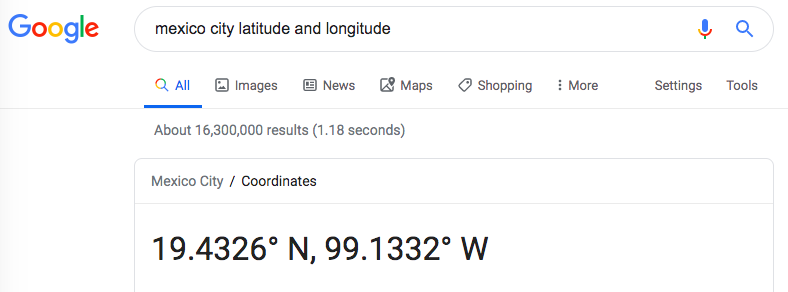
* After completing the tutorial, which ensures that augur and auspice are working correctly, and that you have a general idea of the data format, now it’s time to start working with your data.
* For additional tips on formatting your data, see this resource:
  + <https://neherlab.org/201909_tips_and_tricks.html>
* First, like you did for steps #1-4 in the tutorial, download the files for the course in a new directory
  + <https://github.com/grubaughlab/nextstrain_course>
  + Re-name the directory
* Open the **metadata.xlsx** file
* Add your data using the same format to the **metadata** tab



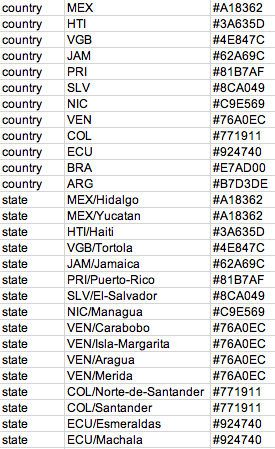
* + When you downloaded data from the public repository, you should have recorded information that you can use here. You must have data in the first 4 columns: **strain**, **accession**, **date**, **country**
    - **strain**: This name must match the headers for your **sequences.fasta** file. It can be whatever you like.
    - **accession**: this should be the database (e.g. genbank) ID number to track the sequence back to its source
    - **date**: make sure that this is formatted as YYYY-MM-DD, and use “XX” if you do not have the information (as shown above). More precise dates are better. Note that excel may change the format. If you are getting an error message when running the snakemake file for augur that sequences are being pruned, check the date formats!
    - **country**: use the full name or an abbreviation
    - **state**: if you have sublocation information, either state, region, county, or city, you can add it here. If you change the name “state” to something else, you must also change it in the **auspice\_config.json** as it will look for that name
    - You may add info to the other columns if you’d like, or leave blank.
* Add your data using the same format to the **lat\_longs** tab



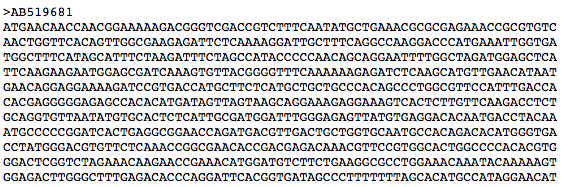
* + For each location listed in **country** and **state** in the metadata tab, there must be a corresponding location in the **lat\_longs** tab (written in the exact same format!). If you left **state** blank, then you do not need to include state locations here.
    - The 3rd and 4th columns are for latitude and longitude, respectively. You must look these up for each of your locations. There are lots of ways to find this data, but good old google will do the trick:



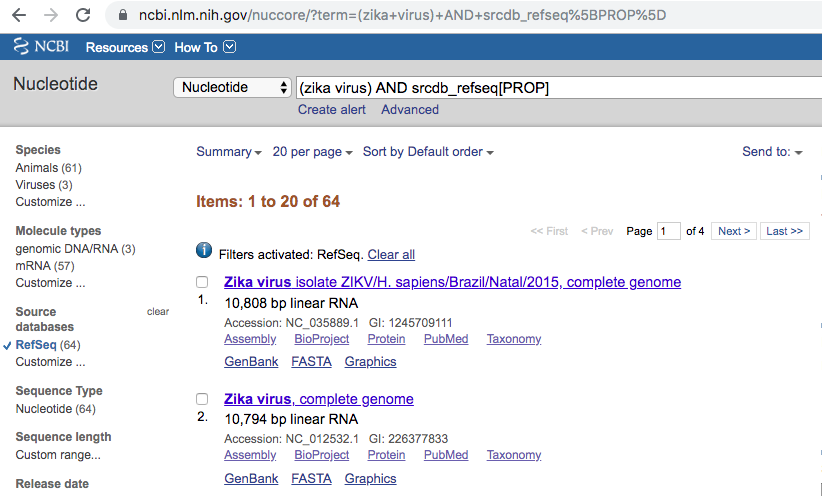
* + - *Just remember to add negatives (-) to South latitudes and West longitudes!*
* Add your data using the same format to the **colors** tab

**

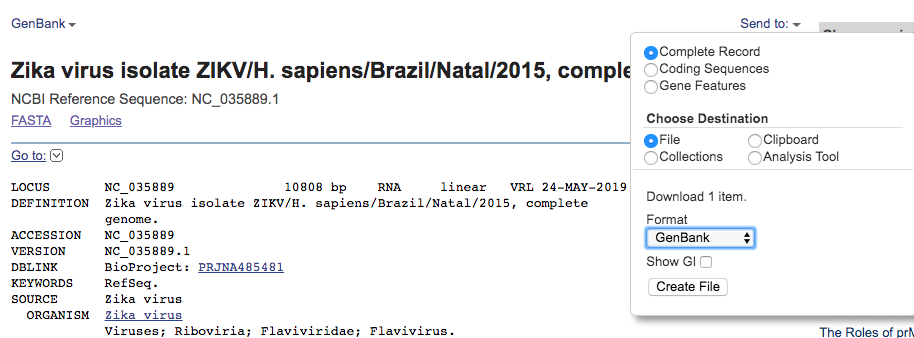
* + The first 2 columns should be exactly the same as the **lat\_longs** tab
  + The 3rd column is the color that auspice will use for each location listed by hex code
    - There are several websites to find hex colors, here is one that I like:
      * <https://htmlcolorcodes.com/color-picker/>
* Just like in steps 5 and 6 of the tutorial, replace the data in your **metadata.xlsx** file by matching the tab names to the corresponding files
  + data > **metadata.tsv**
  + config > **lat\_longs.tsv**
  + config > **colors.tsv**
* Replace the **sequences.fasta** file in the data folder with your aligned and trimmed sequences.
  + Name your file **sequences.fasta** (important for the build)
  + Make sure that the headers match exactly with the **strain** names in column 1 of the metadata (e.g. see below and the metadata image above)



* + *Sequences without matching metadata will cause an error in augur and it will fail to generate a completed build.*
* Create a **reference.gb** file in the config folder following steps 7-9 in the tutorial.
  + A good way to find a reference sequence is to use the **RefSeq** search function in Genbank: <https://www.ncbi.nlm.nih.gov/refseq/>



* + Remember to download as a **GenBank** (.gb) file and edit using a text editor as directed in step 8 of the tutorial.



* + Delete the existing **reference.gb** file and replace with yours. Don’t forget to edit the new reference genome, by replacing **mat\_peptide** by CDS, and **product** by gene, making sure that the elements are vertically aligned as in the original file (see below). Also, at gene fields, replace long or generic gene names by standard acronyms.

A screenshot of a cell phone

Description automatically generated

* Edit the **auspice\_config.json** file in the config folder and the **snakemake** file with your project information as outlined in steps 10 and 11 of the tutorial.

**Step 3**: Run augur and auspice

* Log into your terminal or linux environment
* Navigate to your working directory with all of your updated files
* Activate nextstrain using:
  + **conda activate nextstrain**
* Run augur using:
  + **snakemake export**
* If you have error messages, they are most likely due to improper data formatting. Double check your data and use the error messages to direct you to your problem. Please attempt to solve on your own before contacting an instructor.
* Tip, if you make updates to your metadata and need to re-run augur via snakemake, run this to clear your old results files:
  + **snakemake clean**
* When augur finishes without errors, run auspice using:
  + **auspice view**
* Open a web browser and view the results using the link generated in the command line, likely:
  + [**http://localhost:4000**](http://localhost:4000)
* Use **Ctrl + C** to exit the auspice visualization process.

**Step 4**: Create a webpage using Nextstrain Community

* The easiest way to share our phylogenetic data is to send a URL. Thus, Nextstrain created a function to visualize your metadata and tree json files using GitHub repositories:
  + <https://nextstrain.org/docs/contributing/community-builds>
* First, if you don’t already have one, create a GitHub account:
  + <https://github.com/>
* Then, create a repository. Here is a tutorial:
  + <https://help.github.com/en/github/getting-started-with-github/create-a-repo>
* *\*\*Important\*\** – the repository must be named exactly the same as your **\_tree.json** and **\_meta.json** files in your auspice folder
  + E.g. if your files are **ZIKV-PR\_tree.json** and **ZIKV-PR\_meta.json**, then your repository must be called **ZIKV-PR**
* Create a folder in your GitHub repository named **auspice**
* Add your your **\_tree.json** and **\_meta.json** files in your auspice GitHub folder
  + E.g., see here: <https://github.com/grubaughlab/ZIKV-PR/tree/master/auspice>
* Access your data via **nextstrain.org/community**/**github-name**/**github-repo**.
  + E.g., see here: <https://nextstrain.org/community/grubaughlab/ZIKV-PR>
* Share your results with the world!

**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 tutorial data. Then answer the following questions.**

**E4-1**: Provide a URL link to your Nextstrain community page. (10 points; you’ll be scored on the completeness, aesthetics, and functionality of your website)

**E4-2:** Using screenshots of your Nextstrain page, walk me through your major findings. (Max 400 words & 4 screenshots; 7 points)

**E4-3**: Compared to your maximum likelihood tree from exercise 3, how much new information does your time-resolved phylogeny and phylogeographic analysis reveal? (Max 200 words; 3 points)