**Introductory molecular epidemiology exercise:**

**Demystifying genetic data and phylogenetic trees**

In this exercise, you will learn how to find, view and download published genetic sequence data via NCBI. By using some bioinformatics tools to manipulate a sample dataset, you will become familiar with basic concepts related to sequence alignment and interpretation of phylogenetic trees.

For this exercise, you will need access to the following online resources. The recommended browser for these is either Chrome or Safari. Internet Explorer will not work for all of them. Firefox may work, but hasn’t been tested on all.

* Archive of materials for the exercise: <https://github.com/cdcArmstrong/genomics_course> / Intro folder
* Pubmed: <https://www.ncbi.nlm.nih.gov/pubmed/>
* EMBL EBI Services: <https://www.ebi.ac.uk/services>
* Microreact: <https://microreact.org>
* Nextstrain: <https://nextstrain.org>

1. **Locate and download genetic sequence data from PubMed.**

The data in this example are from this article: Grubaugh ND et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 2017; 546:401-5

1. Go to PubMed (search “pubmed”).
2. Search for “grubaugh n[1au] AND zika”.
3. Select the record for this publication (“Genomic epidemiology reveals …”).
4. Look over the PubMed entry, including the headings on the right.
5. Note that article is publicly available—go to “PMC” entry and download.
6. **Link to, view and download nucleotide data**

Journals generally require authors to make sequence data for their articles public in NCBI/GenBank by the time of publication. NCBI has online tools for viewing and downloading sequence data.

1. Go to “Related information” on the right of the PubMed entry; two of the links are relevant to us:

* “SRA” (Sequence Read Archive—essentially, raw NGS data)
* “Nucleotide” (Genbank—final, processed sequence data)

1. Click on “Nucleotide” to go the GenBank (NCBI/Nucleotide) entries for this paper.

* Note that there are 38 entries in the Nucleotide listing.

1. Go to the first entry; review various aspects of the entry:
   * Various fields
   * Translation (amino acid sequence) of the one open reading frame
   * Nucleotide sequence
   * Click on “CDS” to view the coding sequence; note the start codon (ATG-AUG is the usual start codon in eukaryotes) and the stop codon (TAA).
   * Find links to BioSample and BioProject.
2. Go back to the Nucleotide listing and click on the second entry.
   * Scroll down to see that this sequence is a series of contigs aligned to a reference. Note that the gaps will appear as N’s in the download.
3. Go back to the Nucleotide listing and click the box to the left of the second entry (KY325483.1). Download and review genomic data.
   * In upper right, select “Send to” ->
     + “Complete record”
     + destination: “File”
     + format: “FASTA”
4. Open and review the FASTA file (using WordPad or similar text editor); note that gaps are prefilled with N’s, likely because the sequencing was low coverage and resulted in several contigs, which were then aligned against a reference.
5. **Align the sequence and create a phylogenetic tree**

The next steps employ online tools created by the European Molecular Biology Laboratory (EMBL). Google search for “EMBL EBI tools” or go to <https://www.ebi.ac.uk/services>.

1. Alignment
   * Select “DNA/RNA” (on right, “Browse by”).
   * Scroll down and open MAFFT aligner.
   * Browse your downloaded files to select “sequence.txt” for alignment.
   * Change output format to ClustalW
   * “Submit” (takes ~30 seconds)
   * Review alignment—note gaps at beginning and end (dashes) and in the middle (N’s—because the sequences we input had already been aligned against a reference)
2. Phylogenetic tree

* Click on the “Phylogenetic Tree” button.
* Look at both the “cladogram” and “real” renderings of the tree.
  + The cladogram shows the how the branches of the tree relate to each other without showing the actual branch lengths.
  + The “real” tree shows the branch lengths.
* Review the tree; note that it’s unrooted.
  + To save the phylogenetic tree file for use in part IV:
  + Click “Download phylogenetic tree and save as “tree.nwk” (in MacOS, be sure to Save As 🡪 Page Source [rather than as a “Web Archive”]).
  + Open “tree.nwk” and note format: parentheses, branch lengths.
  + Remove unnecessary characters with three simple find/replace actions: “lcl|”, “.1\_”, “.2\_” (replace all with null).

1. **Work with the phylogenetic tree in Microreact**

Microreact is open source software for visualizing and sharing genomic epidemiology data. This example combines the sequence data file from section **III** with epidemiologic data.

1. For this section, two files are needed:
   * The tree file, as created in **III**, or a clean version downloaded from the archive (“tree.clean.nwk”).
   * The data file from the online archive (“grubaugh\_data.csv”), which is a CSV file containing some relevant data from the GenBank records. The accession number is labeled “Id”.
2. Open microreact.org (if first time at site, accept cookie policy).
   * Scroll down and click on the “Upload your project” tile
   * Drag the two files (“tree.clean.nwk” and “grubaugh\_data.csv”) to the web browser to create the “microreact”.
3. Note several aspects of the tree in Microreact.
   * It is unrooted.
   * There is a distance scale at the lower left.
   * Look at the distance between isolates by following horizontal branch lengths.
   * Several sequences at left are identical.
   * There is a timeline on the bottom; the figure would also include a map if we had lat/long data.
4. Manipulate the tree.
   * Reroot the tree in the middle.
   * Click the “Show controls” icon in upper right.
   * Show the tree in different forms (and explain why it’s the same).
   * Circular
   * Radial (arguably most appropriate for an unrooted tree)
   * Diagonal (cladogram)
   * Heirarchical (vertical – horizontal tree rotated 90° clockwise)
   * Return to the “Rectangular” (horizontal)
   * Show how to manipulate size
   * Show leaf labels (select id).
   * Rotate a branch and observe the impact.
5. Explore the tree.
   * In general, there are two ways to make a tree:
   * Distance-based (neighbor-joining, for example); unrooted, but often can be rooted by including an “outgroup”
   * Sequence based (ML, or Bayesian); rooted
   * Color labels by host (default).
   * Note that mosquitos are interspersed with humans.
   * Note that there are two main clades (definition of clade: “any monophyletic group”):

* Smaller clade (5 isolates), distantly related: suggests repeated introductions
* Larger clade, more closely related, with mosquito pools interspersed: suggests local transmission
  + Change to “Colour by country”; “Label by country”
  + Shows recent travel history, if any (or “USA”).
  + The one labeled (“USA or Cuba”) is from someone who had returned 3 weeks earlier from Cuba.
  + Note timeline: imports were from early on—presumably before there was transmission within the US.
  + Grubaugh et al. suggest 4 introductions into the US.

1. **Exploring other trees on Microreact**

Other projects are available on Microreact.

1. Go to Kat Holt’s *K pneumoniae* ST307 project, <https://microreact.org/project/ryiY_FlfQ>
   * Turn off map (globe in upper right) and metadata table (lower pane)
   * Add metadata blocks
   * Turn “Metadata headers” on (button in upper left with pull-down menu)
   * Turn “Align leaf labels” on (this setting is found under “Nodes & Labels”)
   * Select a few metadata variables: Sample, CatB4, TetA, OXA-9 KPC-2, KPC3
   * Close the metadata block menu
2. **Exploring trees on Nextstrain**

Nextstrain is another open source platform for visualizing and sharing genomic epidemiology data. Several examples are available:

1. Seasonal Flu (H3N2)
   * Note how quickly it diverges.
   * Note how new clades are constantly emerging and replacing older clades.
   * Note that this is based on HA genes (influenza is a segmented virus).
   * Look at the H1N1pdm tree: why does it only go back to 2013?
2. TB
   * Note the long evolutionary timeline
   * Estimated evolutionary rate (displayed in the “Clock” tree option):
   * For flu: 3.3e-3 (3300 per million per year)
   * For TB: 1.3e-8 (0.013 per million per year; note: the rate estimated by other means is 10-fold this ~1e-7, or one mutation per genome every 2 years)
3. Zika
   * Note location of US strains (can be done by mousing-over the US entry in the legend)—most of these are from the report we used earlier.
   * Look at central American strains: the tree suggests they were already circulating before the first strains were discovered in Brazil.
4. MERS-CoV (<https://nextstrain.org/mers>)
   * The question here is whether MERS is being sustainably transmitted in humans, or whether human cases result from exposure to camels.
   * What conclusions do you draw from the tree?