**Introductory molecular epidemiology exercise:**

**Demystifying genetic data and phylogenetic trees**

* During this exercise, the participant 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>
  + 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>
* Find the Zika article in PubMed (Grubaugh ND et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 2017; 546:401-5)
  + Go to PubMed (search “pubmed”)
  + Search for “grubaugh n[1au] AND zika”
  + Select the record for this publication (“Genomic epidemiology reveals …”)
  + Look over PubMed entry, including the headings on the right
  + Note that article is publicly available—go to “Nature” entry and download
* Link to and review nucleotide data
  + 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)
  + Click on “Nucleotide” to go the GenBank (NCBI/Nucleotide) entries for this paper
  + Note that there are 38 entries in the Nucleotide listing
  + Go to first entry; review various aspects of the entry:
    - Various fields
    - Translation of the one ORF
    - Sequence
    - Click on “CDS” to view the coding sequence; note the start codon (ATG-AUG is the usual start codon in eukaryotes) and stop codon (TAA)
    - Show links to BioSample and BioProject
  + Return to Nucleotide listing and go to second entry; note that this is a series of contigs that have been aligned to a reference; note that the gaps will appear as N’s in the download
  + Return to the Nucleotide listing
* Download and review genomic data
  + In upper right, select “Send to” ->
    - “Complete record”;
    - destination: “File”;
    - format: “Fasta”
  + Open and review fasta file; 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.
* Alignment
  + Google search for “EMBL EBI tools”
  + Filter on “DNA/RNA” (on right, “Browse by”)
  + Open MAFFT aligner
  + Select “sequence.txt” file 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)
* Phylogenetic tree
  + Click on “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 tree
  + Note that it’s unrooted
  + 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”; note format—parentheses, branch lengths
  + There is a clean version of this in the archive (“tree.clean.nwk”), or this can be cleaned with three simple find/replace: “lcl|”, “.1\_”, “.2\_” (replace all with null)
* Import into MicroReact
  + For this, the participant will need two files:
    - The tree file just created (tree.clean.nwk)
    - The data file (from online archive: “grubaugh\_data.csv”), which is a CSV file with some relevant data from the GenBank records. The accession number has been labeled “Id”.
  + 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”
* Manipulating the tree in Microreact
  + Note several aspects of tree:
    - It’s unrooted (we’ll get back to what that means later)
    - There’s a distance scale at the lower left
    - How to look at distance between isolates (by following horizontal branch lengths
    - Several sequences (at left are identical)
    - There’s also a timeline on the bottom; the figure would include a map if we also had lat/long data
  + Manipulate the tree
    - Reroot the tree in the middle
    - Click the “Show controls” icon in upper right
    - Show 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 show the impact
  + Exploring the tree
    - Unrooted
      * In general, 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
      * With unrooted trees, it’s useful to have an outgroup to be able to root the tree
    - Color labels by host (default)
      * Point out that mosquitos are interspersed with humans
      * Note two main clades (definition of clade: “any monophyletic group”)
        + Smaller clade (5 isolates), distantly related
        + Larger clade, more closely related, with mosquito pools interspersed
        + One suggests repeated introductions
        + The other suggests local transmission
      * Country
        + 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
        + The paper (Grubaugh et al.) suggests 4 introductions into the US.
  + Exploring other trees on Microreact
    - Go to Kat Holt’s *K pneumoniae* ST307 microreact, <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
      * Walk students through a metadata block
  + Exploring trees on Nextstrain
    - 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—this is a segmented virus
      * look at the H1N1pdm tree—why does it only go back to 2013?
    - TB
      * Note the long evolutionary timeline
      * Estimated evolutionary rate (which is 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)
    - 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
    - MERS-CoV (<http://nextrain.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?