### **Interactive influenza A virus workshop using the Influenza Research Database and ISU *FLU*ture: Tutorial 2018**

**Schedule**

1. 1:00pm – 1:30pm. Introduction to phylogenetics of influenza A virus in swine and ISU *FLU*ture
2. 1:30pm – 3:30pm. Interactive workshop covering:
   1. Interpreting an ISU-VDL diagnostic report
   2. Constructing a sequence file in FASTA format
   3. Annotating the sequences provided in diagnostic reports
   4. Determining the HA clade of diagnostic isolates
   5. Searching for similar sequences of a diagnostic IAV isolate
   6. Exploring the origin of a diagnostic IAV isolate
   7. Translating a sequence, and identifying genetic differences

**Interactive workshop**

We demonstrate the use of the ISU *FLU*ture and the Influenza Research Database (IRD) resources through a series of six use cases designed around sequences that might be encountered during a veterinary diagnostic investigation. Each use case is described as a series of steps performed using the ISU *FLU*ture or IRD website, with the text formatted using the following convention:

* “text in quotes” – text that needs to be entered into an input box on the website.
* **Bold text** – action buttons or options on the website that need to be clicked or selected.
* *Italic text* – questions, observations, comments, etc.

**I. Interpreting an ISU-VDL diagnostic report and constructing a sequence file in FASTA format**

Motivation

A typical IAV diagnostic workflow might include the capture of a specimen for an individual or group of symptomatic swine or from a swine surveillance program, the testing of the specimen using PCR for the generic IAV M gene and the subtype specific HA and NA genes, and, if positive, the sequencing of the HA genome segment as described in the diagnostic report.

Objective

Upon completion of this exercise, you will be able to interpret ISU-VDL diagnostic reports and extract sequence data for addition analyses.

Approach

The FASTA file format is a standard text file format recognized by many computational tools operating on nucleotide and protein sequence data.

The file consists of pairs of paragraph lines separated by carriage returns. The first paragraph line is annotated with a leading **>** symbol and is referred to as the description line (defline). The defline can contain any string of characters and delimiters and is often use to describe important metadata about the source of the sequence provided in the second paragraph line in the pair. The second line does not contain any leading annotation symbol and can be either a nucleotide or protein sequence. A single FASTA file can contain one or more defline + sequence pairs.

With the diagnostic report in hand, the first step is to generate a FASTA file for downstream use, as follows:

1. Open a standard text editing program (e.g. Notepad, TextEdit, BBedit).
2. Decide on which sequence metadata you would want to include in the defline (e.g. specimen identifier, subtype, year of isolation, host species, geographic location of sample collection, etc.) and what kind of delimiter to use (e.g. “|” ).
3. Construct the first paragraph defline by typing “>” followed by the metadata field values separated by the selected delimiter (“|”), e.g.:
   1. >XYZ12345|A/Swine/Nebraska/209/98|H3N2|1998|Swine|USA|Nebraska|Cluster\_I
4. Then cut and paste the sequence from the diagnostic report into the second paragraph line of the file.
5. Repeat for each of the provided diagnostic report sequences.
6. Save the file in simple txt format.
7. Replace the .txt in the file name with a .fasta extension.
8. Please create an additional clean FASTA file with sequence accession as the defline only.

**II. Annotate the sequences provided in ISU-VDL diagnostic reports**

Motivation

With the sequence in hand, a logical first step would be to *verify* the subtype determined by the PCR assays in the diagnostic report and to check if there are any potential sequencing artifacts.

Objectives

Upon completion of this exercise, you will be able to use the IRD sequence annotation tool to determine the influenza type, segment number, subtype (for HA, but the process also works for neuraminidase/NA) and translated amino acid sequence for your own IAV sequences.

Approach

Input data: Your\_saved\_FASTA\_file.fasta

Tools: Use the sequence annotation tool on the IRD website (**https://www.fludb.org** > **Analyze & Visualize** > **Annotate Nucleotide Sequences)**

Questions

1. Write down the gene segment and subtype for each sequence.
   1. 2018059219
   2. 2018068786
   3. 2018067087
   4. 2018016010
   5. 2018073120
   6. 2017077766
   7. 2016028031
2. Did you notice any problems with any of the sequences?
   1. Is it possible to correct these data?
3. Download the Amino Acid FASTA for your annotated data and save the file (with AA in the title).

**III. Determine the HA clade of diagnostic isolates**

## Motivation

Determining the relationship between a diagnostic IAV isolate and other co-circulating and historical IAV strains is essential to track the transmission chain and determine whether your sequence represents a new influenza detection to your production system. Swine IAV of the H1 subtype have been divided into three major lineages arising from separate zoonotic spillover events, including 11 genetic clades arising through subsequent antigenic shift and drift. Swine of the H3 subtype have been divided into three major lineages arising from separate zoonotic spillover events, and one of the lineages has drifted to form 7 genetic clades.

## Objective

Upon the completion of this exercise, you will be able to use the ISU *FLU*ture HA identity tool to classify unknown sequences, determine where similar sequences are circulating, and likely neuraminidase pairing.

## Approach

Input data: Individual sequences in Your\_saved\_FASTA\_file.fasta OR Your\_saved\_Annotated\_AA\_FASTA\_file.fasta

Tools: The HA identity tool on *FLU*ture ([**http://influenza.cvm.iastate.edu**](http://influenza.cvm.iastate.edu) > **HA Identity Tool**) is powered by BLAST, which finds the most similar sequences in the ISU-VDL database to a query sequence. The tool limits the returned results to 100 per query, and all returned results are 96% similar to the query or higher. This is a rapid analytical approach, and output must be interpreted conservatively.

1. Navigate to the **HA Identity Tool**.
2. Paste in one sequence at a time into the text box.
3. Select either “nucleotide” or “amino acid” based on your sequence.
4. Click the search button.

## Questions

1. What HA clade is identified as the most similar to each of your sequences?
   1. 2018059219
   2. 2018068786
   3. 2018067087
   4. 2018016010
   5. 2017077766
   6. 2016028031
2. How many states has each virus been detected in?
   1. 2018059219
   2. 2018068786
   3. 2018067087
   4. 2018016010
   5. 2017077766
   6. 2016028031
3. What neuraminidase is each sequence most paired with?
   1. 2018059219
   2. 2018068786
   3. 2018067087
   4. 2018016010
   5. 2017077766
   6. 2016028031

**IV. Determine trends in the HA/NA clades of diagnostic isolates**

## Motivation

The ISU *FLU*ture database and interactive website provides a unique web resource that offers thirteen years of case summaries for IAV in swine. The primary utility of the ISU *FLU*ture database is near *real-time access* to the number of IAV detections and variation in HA and NA clades over time. Both publicly and privately funded case data are available within days after all appropriate diagnostic tests are complete. Using the data in *FLU*ture, veterinarians and vaccine developers can make informed decisions about vaccine design through matching vaccine components with currently circulating IAV strains.

## Objective

Upon the completion of this exercise, you will be able to use the ISU *FLU*ture website to identify the dominant HA/NA pairings of circulating IAV in swine

## Approach

The **Correlation, Time Series**, and **Heat Map** tools on *FLU*ture ([**http://influenza.cvm.iastate.edu**](http://influenza.cvm.iastate.edu)) allow dynamic exploration of data to address the questions posed below. In addition, these interfaces allow the user to control for the effects of variable sample size by using normalized data (“account by proportion”).

## Questions

1. From *2015 to* *2017*, approximately what percentage of IAV positive samples in the ISU-VDL database are from each of the subtypes? **(**[**http://influenza.cvm.iastate.edu**](http://influenza.cvm.iastate.edu) **> Correlation > Subtype per year** or **> Time Series)**
   1. H1N1 \_\_\_\_\_\_
   2. H1N2 \_\_\_\_\_\_
   3. H3N2 \_\_\_\_\_\_
2. In *2017*, what were the top 4 HA clades circulating? Is there a particular time of the year when H1 viruses are circulating? Is this the same or different for H3 viruses? ([**http://influenza.cvm.iastate.edu**](http://influenza.cvm.iastate.edu) **> Correlation > HA Clade Frequency Detection per year** or **> Time Series)**
   1. \_\_\_\_\_\_
   2. \_\_\_\_\_\_
   3. \_\_\_\_\_\_
   4. \_\_\_\_\_\_
   5. Are there any seasonal trends?
3. Based upon the data processed by the ISU-VDL, what *H3 and H1 HA* clades are found in Ohio? ([**http://influenza.cvm.iastate.edu**](http://influenza.cvm.iastate.edu) **> Correlation > H3/H1 Clade Frequency Detection per Pig’s Origin State)**
4. a) When were human-to-swine-2016 viruses first detected in the U.S. swine population? What proportion of the IAV diagnostic cases does this clade represent?

b) Of the six diagnostic cases you annotated and identified in Section II and III, which clade is the most frequently detected from 2010 to present?

([**http://influenza.cvm.iastate.edu**](http://influenza.cvm.iastate.edu) **> Time Series > H3/H1/HA Clade Frequency Detection of Detection over Time**)

* 1. Human-to-swine-2016 detection \_\_\_\_\_\_\_\_
  2. Human-to-swine-2016 proportion \_\_\_\_\_\_\_\_
  3. Most frequently detected HA clade from 2010 to present

1. What are the top 4 HA-NA pairings from *2015 to present*? Are there any HA clades that are no longer worth considering in a multivalent vaccine? ([**http://influenza.cvm.iastate.edu**](http://influenza.cvm.iastate.edu) **> Heat Map**)

**V. Search for similar sequences of a diagnostic IAV isolate**

## Motivation

# The BLAST algorithm is one of the original comparative genomics analysis tools that calculates a sequence similarity score after performing an alignment of two nucleotide or protein sequences. Thus, using BLAST as an approach to search a database of sequences to identify those database sequences most similar to the query sequence is an excellent strategy to identify sequences with close evolutionary relationships.

# The IRD implementation of BLAST allows the user to search through selected sequence databases compiled to anticipate specific user preferences (e.g. Nucleotides for segment 4 hemagglutinin of the H1 subtype), or to allow users to provide their own custom BLAST databases through their private workbench account (i.e., create a set of your own data to BLAST against). We will search for the closest sequences of isolate *2018016010* using BLAST and see if your query sequence is worth investigating further.

Objective

Upon the completion of this exercise, you will be able to use the Influenza Research Database (IRD) BLAST tool to search against IRD-compiled IAV sequence databases or your own custom datasets.

Approach

Input data: a) isolate *2018016010*; b) Your\_saved\_FASTA\_file.fasta; and c) a sequence search as a working set.

Tools: Use the IRD BLAST tool ([**https://www.fludb.org/**](https://www.fludb.org/) **> Workbench > View** the desired dataset **> Run analysis > Identify Similar Sequences (BLAST)**) or ([**https://www.fludb.org/**](https://www.fludb.org/) **> Analyze & Visualize > Identify Similar Sequences (BLAST)**)

## Questions

1. What hosts are infected by strains most similar to isolate *2018016010*? What does this information suggest?
2. Create an *Arkansas* swine IAV HA working set (i.e., the same process can be used for your production system/area of interest), upload Your\_saved\_Fasta\_file.fasta and BLAST isolate 2018067087 to determine if this is something new to *Arkansas*.

**VI. Explore the origins of a diagnostic IAV isolate**

Motivation

Phylogenetic analysis is widely used to trace the possible origin of a new virus and track the transmission chain. Sequence-associated metadata describing the host species, isolation date, geographic location, etc. of an IAV diagnostic sample can be extremely helpful in distinguishing between local virus transmission/circulation versus incursion from distant sources. In this exercise, we will explore the use of phylogenetic analysis and sequence metadata to identify a possible incursion event.

Objectives

Upon completion of this exercise, you will be able to build a phylogenetic tree on a set of sequences to infer their evolutionary relationships.

Approach  
A typical phylogenetic analysis workflow involves the following steps:

1. Upload private sequences to your personal workbench space;
2. Search for sequences in IRD and save desired sequences as a working set in your private workbench space;
3. Combine multiple datasets within your private workbench;
4. Build a phylogenetic tree on a set of sequences to infer their evolutionary relationships.

Input data:

1. FASTA file from <https://anderson-lab.com/teaching/>
2. IRD Metadata file from <https://anderson-lab.com/teaching/>
3. HA sequences in the delta2 clade from swine, USA, date range from 2017 to Present with duplicates removed.

Download link for data: https://www.icloud.com/iclouddrive/0gQhO0Fmia4CiTipzdg3rZqZg#mckean%5F2018.zip

Tools:

1. Upload the sequence file (mckean\_2018.fasta) to your workbench using the **Workbench** > **Upload Data/Metadata** option. Select **Upload Sequence Associated Metadata** > **Option 2** (IRD\_metadata\_template\_ISU\_workshop\_2018.xlsx)
2. Use the Swine H1 Clade search interface (**Search Data** > **Search Sequences** > **Swine H1 Clade**) to search for HA-H1N2 sequences in the delta2 clade from swine, USA, date range 2017 to Present, and Remove Duplicates using **Advanced Options**. Then save the retrieved sequences to Workbench using **Add to Working Set** > **Create a new working set with the selected items**.
3. Within your personal workbench, combine the saved working set with the AASV workshop dataset (Select datasets > **More Actions** > **Combine**).
4. View your dataset (**Display Settings > items per page 200**) and select *2018067087* (uncheck the other uploaded sequences) and all other data.
5. Align your data (**Run Analysis > Align Sequences (MSA)**)
6. Build a phylogeny tree on the combined dataset (**Run Analysis** > **Generate Phylogenetic Tree > Build Tree**).
7. Use the IRD tree viewer to visualize the tree and infer the origin of isolate *2018067087*. (If you cannot find *2018067087* use the Y+ to expand the tree, and the search interface to highlight the tree.

Questions

1. How many delta2 viruses are in the IRD? How many in 2017? How may are in Iowa?
2. Find the closest isolates to *2018067087* what states are they from?
3. Can you infer the possible origin of the diagnostic isolate?

**VII. Identify variation in amino acids at important positions.**

Motivation

Six amino acid positions (145, 155, 156, 158, 159 and 189, referred to as the antigenic motif or “key sites”; H3 numbering) in the globular head of hemagglutinin (HA1 domain) play an important role in defining the antigenic phenotype of swine Clade IV (C-IV) H3N2 IAV. Mutation at one of these positions may cause significant antigenic drift; multiple mutations in these 6 sites almost always causes significant antigenic change.

Objectives

Upon completion of this exercise, you will be able to translate nucleotide sequence data to amino acids, trim the signal peptide, and identify variation between two H3 sequences.

Approach  
A typical approach involves:

1. Upload private sequences to your personal workbench space;
2. Align the sequences and investigate sequence variation;

Input data: Your\_saved\_Annotated\_AA\_FASTA\_file.fasta from section II.

Tools:

1. Upload the sequence file (Your\_saved\_Annotated\_AA\_FASTA\_file.fasta) to your workbench using the **Workbench** > **Upload Data/Metadata** option. DO NOT USE EITHER BUTTON TO INCLUDE METADATA, if you click a button, refresh the Upload Data/Metadata page.
2. View the data and select 2017077766 and 2016028031. Then **Run Analysis > Align Sequences (MSA).**
3. Trim the signal peptide.

Questions

1. What is the antigenic motif of each of the sequences?
   1. 2016028031
   2. 2017077766
2. See [open access] Bolton et al. (in press), Influenza and Other Respiratory Viruses. <https://onlinelibrary.wiley.com/doi/full/10.1111/irv.12610>
   1. Do you think your two viruses are antigenically similar or different?