	Analyzing Influenza Data in Terra using Theiagen's TheiaCoV Illumina PE and Augur Workflows		
	Document TG-FLU-PE, Version 1		
	Date:	Effective Date:	Workflow Version:
	8/4/2023	8/2023	PHB v1

1. PURPOSE/SCOPE

To standardize the process of analyzing influenza (Flu) next generation sequencing (NGS) data using Theiagen's TheiaCoV_Illumina_PE_PHB, Augur_Prep_PHB, and Augur_PHB workflows in Terra to determine typing, subtyping, and lineage designation. Acceptable data types include Illumina paired end (PE) raw read file format.

2. REQUIRED RESOURCES

- Computer
- Internet connection: at least 10 and 5Mbps for download and upload speeds, respectively
- Internet browser
  - Google Chrome, Firefox, or Edge
- Google account
- Terra account, linked to Google account
- Illumina paired end (PE) raw sequencing read files uploaded to Terra workspace
- Theiagen's TheiaCoV_Illumina_PE_PHB, Augur_Prep_PHB, and Augur_PHB workflows in Terra

3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-03	Uploading Local or SRA NGS Data & Creating a Results Metadata Table in Terra

4. PROCEDURE

4.1 CREATE A SAMPLE METADATA FILE (TSV FILE) FOR RAW READS

1. In Excel, *create a list* containing the following sample information:
   a. **For all analyses:**
      i. Column 1 header (Fig 1): *entity:kilifi_H3N2_id* where *kilifi_H3N2* is the name of the data table/group of samples to be analyzed
      ii. List all *sample IDs* in column 1

entity:kilifi_H3N2_id	accession	collection_date	continent	country	state	read1	read2
100734	A/Kilifi/131/201(	8/5/2010	Africa	Kenya	Kilifi	100734_R	100734_R
100954	A/Kilifi/132/201(	8/17/2010	Africa	Kenya	Kilifi	100954_R	100954_R
SRR11445941	SRR11445941	1/24/2017	Europe	Belgium	Brussels-C		
SRR13443360	SRR13443360	6/8/2018	Europe	Belgium	Brussels-C		
	76	3/11/2019	Europe	Spain	Catalonia		

Figure 1: Raw Read Metadata File.

b. For analysis from raw sequencing reads (Fig 1)

i. Use column headers: *read1* and *read2*

ii. List the *full cloud file paths* to read1 and read2 files

c. For analysis using SRA fetch (Fig 2):

i. Column header: *accession*, or similar

d. *Optional: remaining columns may be used to add metadata like run_id, additional lab results, sample collection information, demographic data, etc*

e. Do not include spaces in the headers

2. *Save as* a txt or tsv file
3. *Upload* to Terra workspace; see TG-TER-03 for details

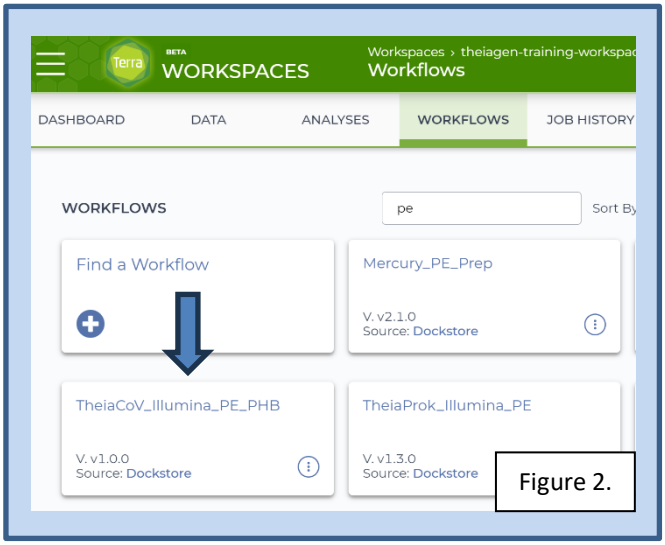

Figure 2.

4.2 RUNNING THE THEIACOV WORKFLOW

1. Open Terra and navigate to the *workflows* tab in the workspace containing flu data
2. Select the *TheiaCoV_Illumina_PE_PHB* workflow (Fig 2)
3. Choose the latest version of *version 1* in the version dropdown field, or the workflow version that was used during internal assay validation (Fig 3, a)
4. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 3, b)
5. Select the data table name under the select *root entity type* dropdown (Fig 3, c)
6. Click *select data* (Fig 3, d)

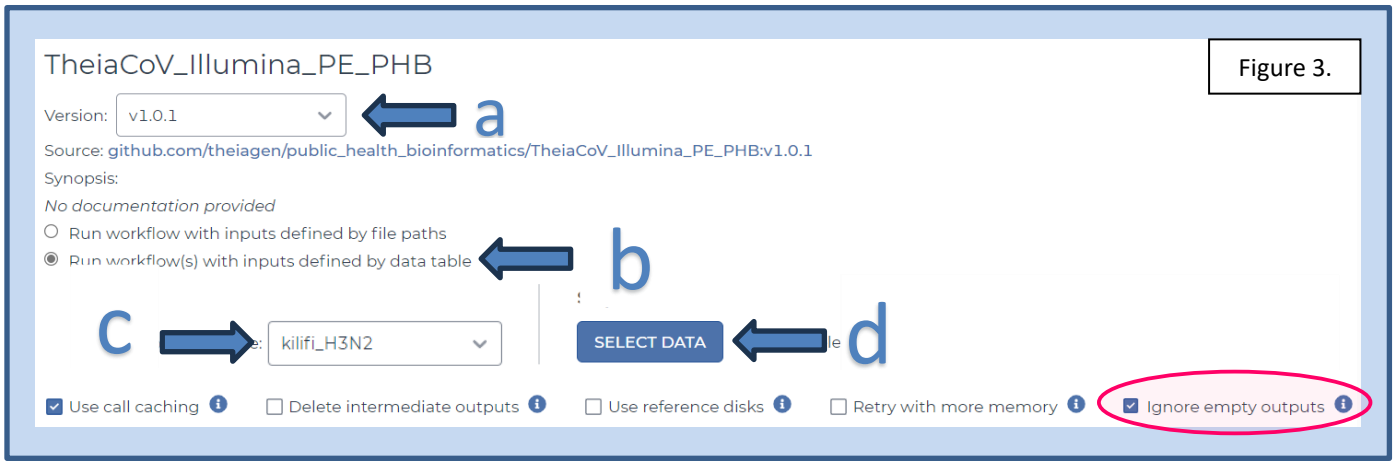

Figure 3.

7. In the pop-up window *select the checkbox* for each sample to be included in the analysis (Fig 4)
   a. *Click the checkbox dropdown and all to select all samples in the data table; if the checkbox at the top is checked, only the first 100 samples in the data table will be selected*
   b. *Additionally, a subset of samples may be chosen using the search bar to filter before selecting the checkbox at the top to only select samples matching the search criteria (Fig 4, highlight)*
   c. Scroll to the bottom and click *ok*
   d. *Optional: Check the box to* ignore empty outputs *(Fig 3)*

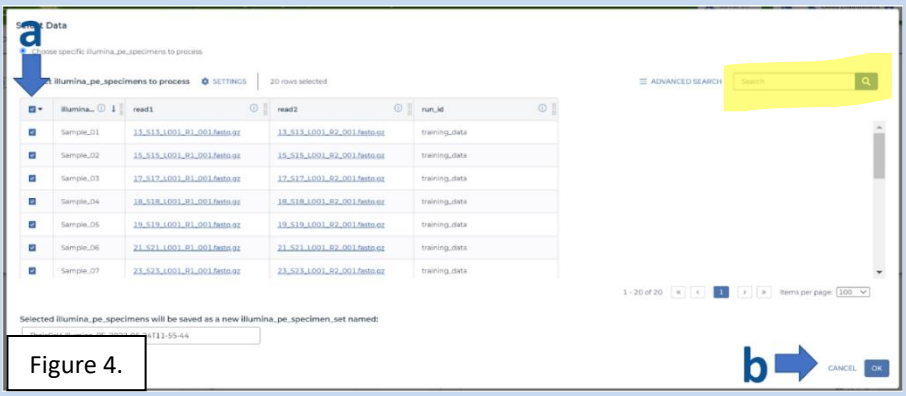

Figure 4.

8. Set the first three attributes in the table to *this.read1*, *this.read2* , and *this.kilifi_H3N2_id* (Fig 5), respectively
   a. Where *kilifi_H3N2* is the unique name of your data table in Terra
   b. In the *organism* attribute field enter *"flu"* in quotation marks (Fig 5)

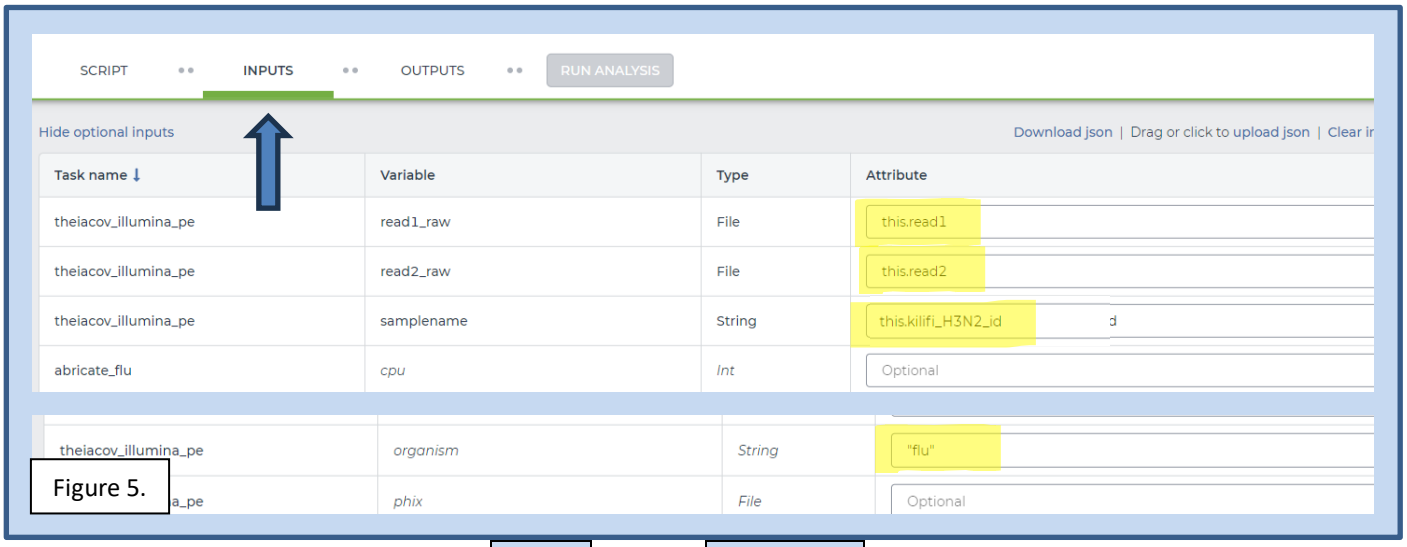

Figure 5.

9. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 6)
10. Click *save*
11. Launch the workflow by clicking *run analysis*; enter desired comments and click *launch*

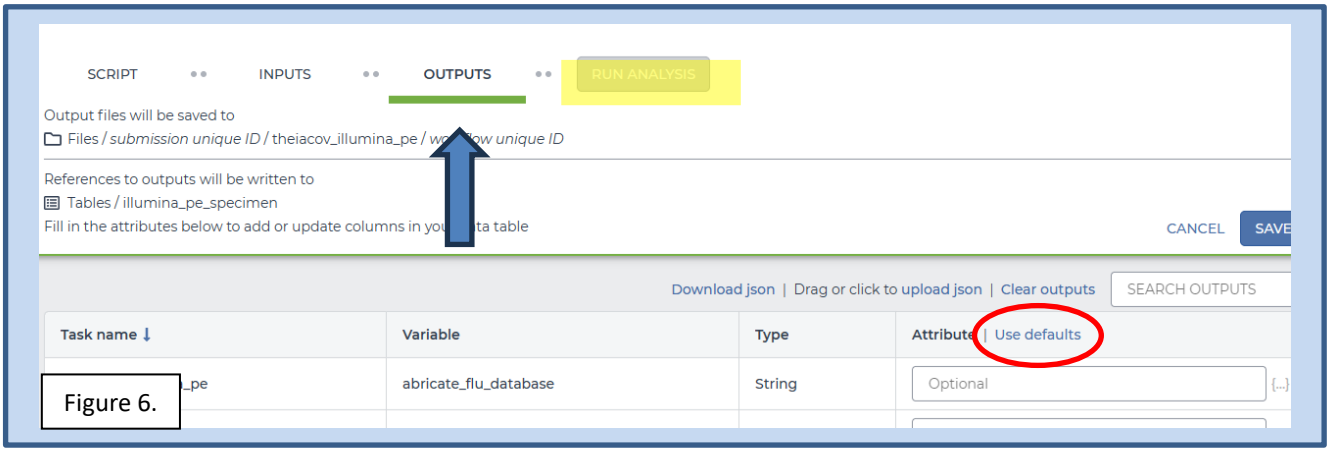

Figure 6.

4.3 RUNNING THE AUGUR PREP WORKFLOW FOR HA OR NA FLU PROTEINS

1. Confirm the TheiaCoV worklow has successfully completed by viewing the relevant job submission status in the *job history* workspace tab; green and red indicate successful and failed jobs, respectively, while blue represents unfinished jobs
2. Perform quality assessment per internal protocols and proceed with Augur Prep for samples passing quality control (QC) metrics
   a. *For samples that do not meet QC thresholds, resequence*
      i. *Samples not meeting thresholds may proceed to Augur Prep at the discretion of the lab*
3. Navigate to the *workflows* tab and select the *Augur_Prep_PHB* workflow
4. Choose the latest version of the workflow in the *version dropdown field*, or the workflow version that was used during internal assay validation
5. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 3, b)
6. Select the relevant data table name under the select *root entity type* dropdown (Fig 3, c)
7. Click *select data* (Fig 3, d)
8. In the pop-up window *select the checkbox* for each sample to be included in the analysis (Fig 7)
   a. Do not include any samples that are missing either *irma_ha_segment* or *irma_na_segment*
   b. *Click the down arrow and select all to process all samples*
   c. *Additionally, the search bar may be used to narrow down the sample list to only those matching search criteria (e.g. only "SRR" sample names or just "Type_A" samples, etc)*
   d. *Optional: name the output file something to differentiate it from other runs, e.g. HA_AugurPrep_YYYYMMDD.# or NA_AugurPrep_YYYYMMDD.# (Fig 6, c)*
   e. Scroll to the bottom and click *ok*

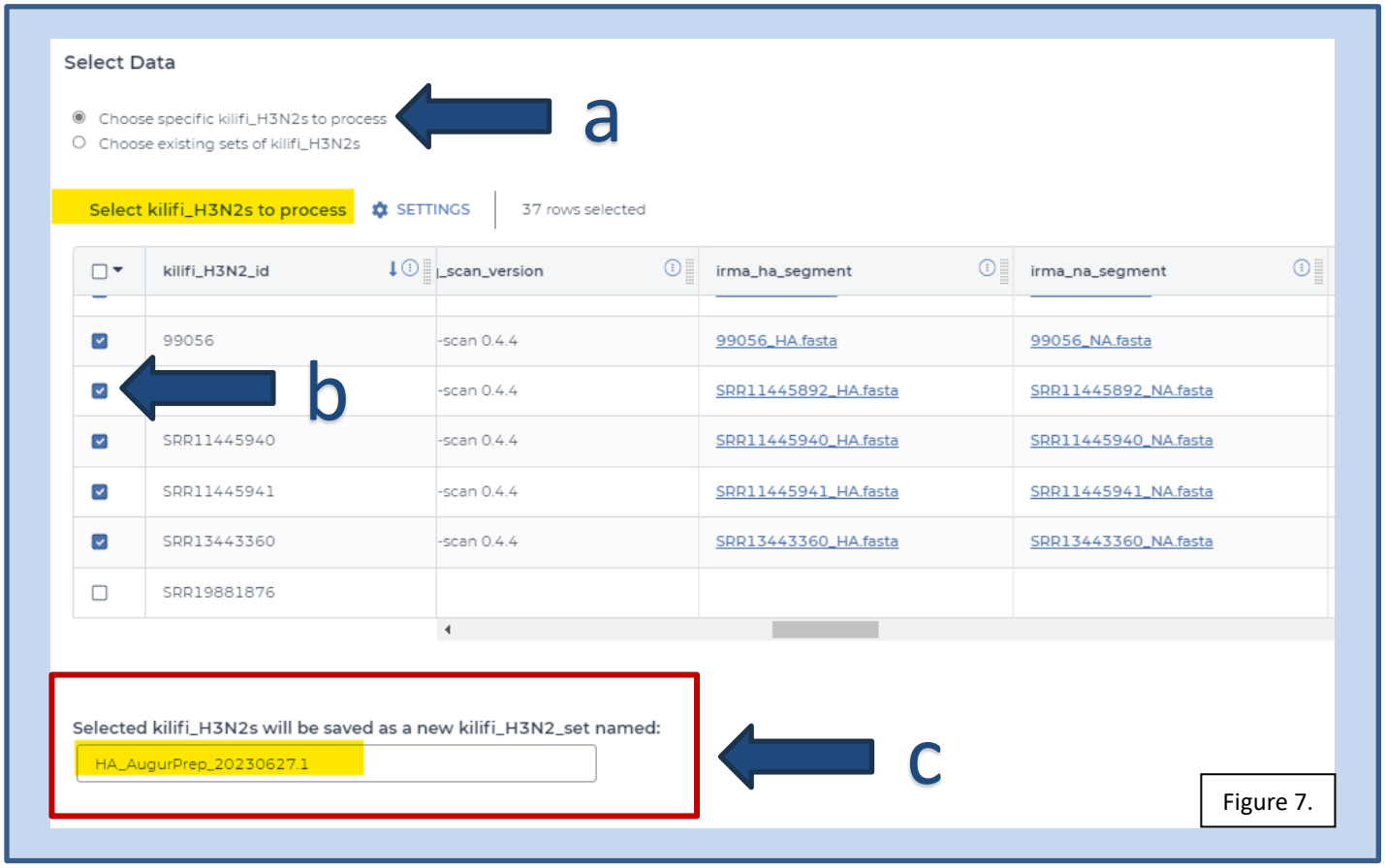

Figure 7.

9. Set the first five input attributes in the table to the following, respectively (Fig 8)
   a. *this.irma_ha_segment*
      i. When running Augur Prep on the NA segment, enter *this.irma_na_segment*
   b. *this.collection_date*
   c. *this.continent*
   d. *this.country*
   e. *this.state*
      i. Input text shown in grey indicates this is variable; text must your Terra data table column headers, excluding the prefix "this." (Fig 1)
10. In the *organism* attribute field enter *"flu"* in quotation marks (Fig 5)
11. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 6); click *save*
12. Launch the workflow by clicking *run analysis*; enter desired comments and click *launch*
13. Repeat section 4.3 to run Augur Prep for the NA protein with the following changes:
    a. When selecting data to run, *enter a different output file name* to distinguish the NA Augur run from the HA run, e.g. NA_AugurPrep_YYYYMMDD.# (Fig 7, c)

b. Specify *this.irma_na_segment* in the first input field rather than using this.irma_ha_segment as before

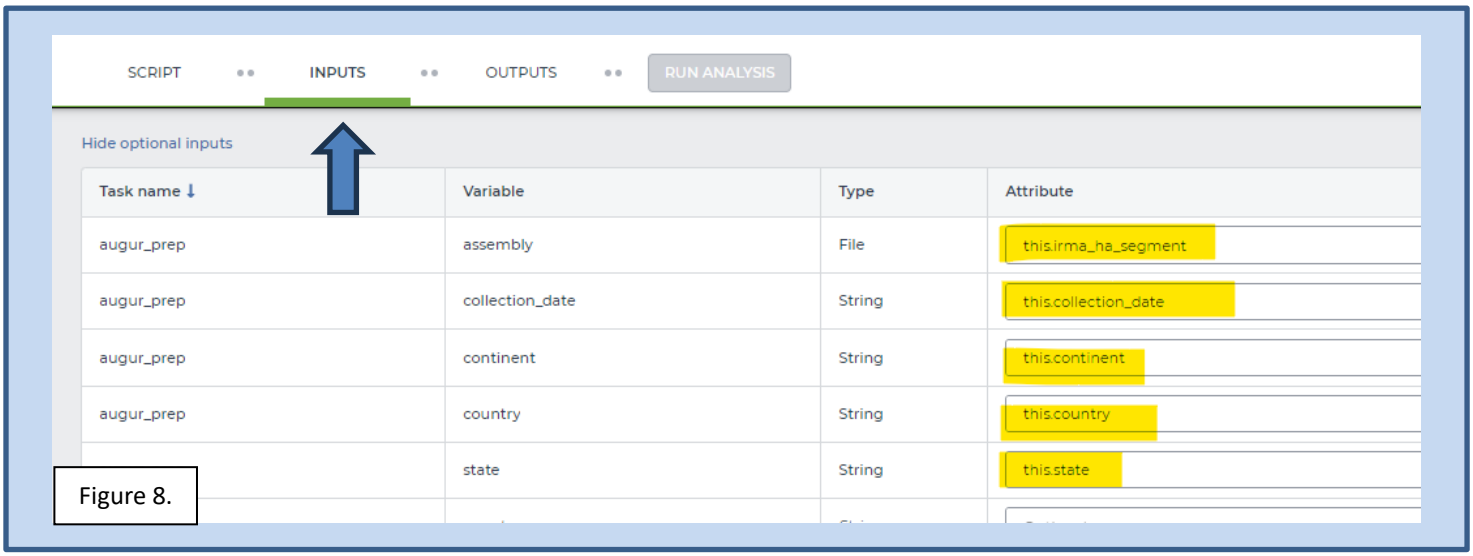

Figure 8.

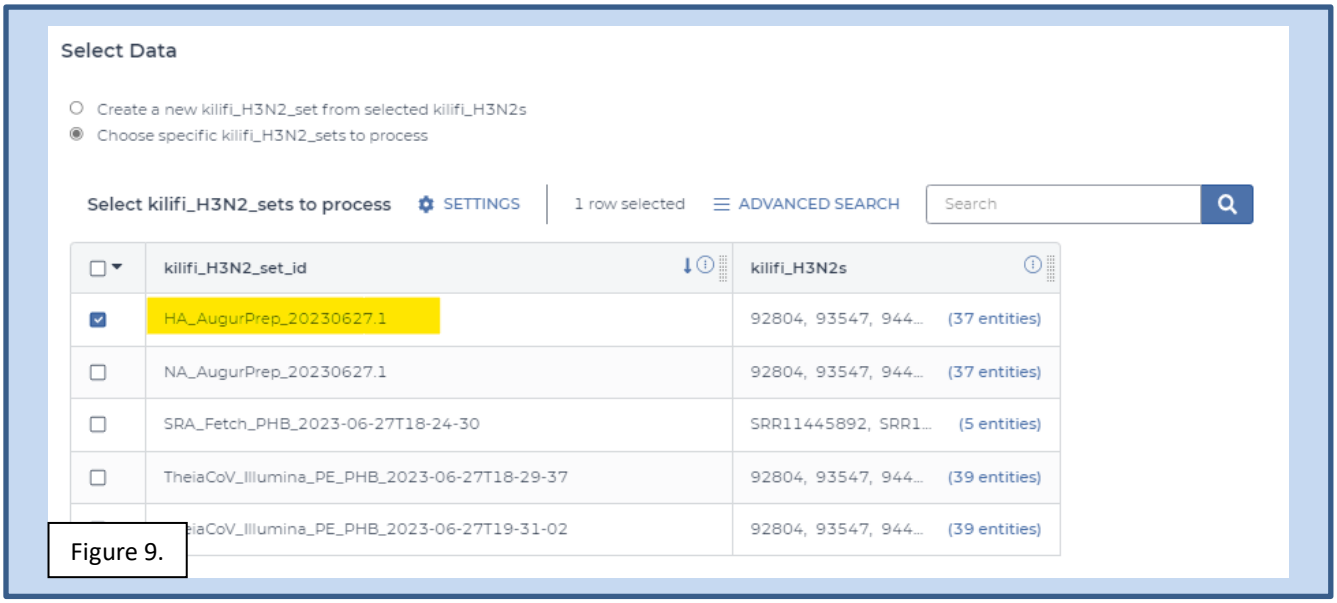

Figure 9.

4.4 RUNNING THE AUGUR [ANALYSIS] WORKFLOW

1. Open the *Augur_PHB* workflow in the workspace containing flu data
2. Choose latest version of the workflow in the *version dropdown field*, or the workflow version that was used during internal assay validation
3. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 3, b)
4. Select the relevant SET data table under the select *root entity type* dropdown (Fig 3, c)
   a. E.g. Kilifi_H3N2_set
5. Click *select data* and choose the output file name from the Augur Prep workflow previously ran (e.g. HA_AugurPrep_YYYYMMDD.#); if Augur Prep was run twice for the HA and NA protein segments, the Augur Analysis workflow must be run twice, as well (they cannot be run together)
6. Click on the *inputs* tab to specify settings
7. Set the first three attributes in the table to *this.kilifi_H3N2s.irma_ha_segment* , *"kilifi_H3N2_HA"* , and *this.kilifi_H3N2s.augur_metadata*, respectively (Fig 10)
   a. Where *kilifi_H3N2* is the unique name of your data table in Terra for all three attribute fields
8. Set the flu_segment, flu_subtype, and organism optional fields to *"HA"* , *"H3N2"* , and *"flu"* , respectively (Fig 10)
9. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 6); click *save*
10. Launch the workflow by clicking *run analysis*; enter desired comments and click *launch*

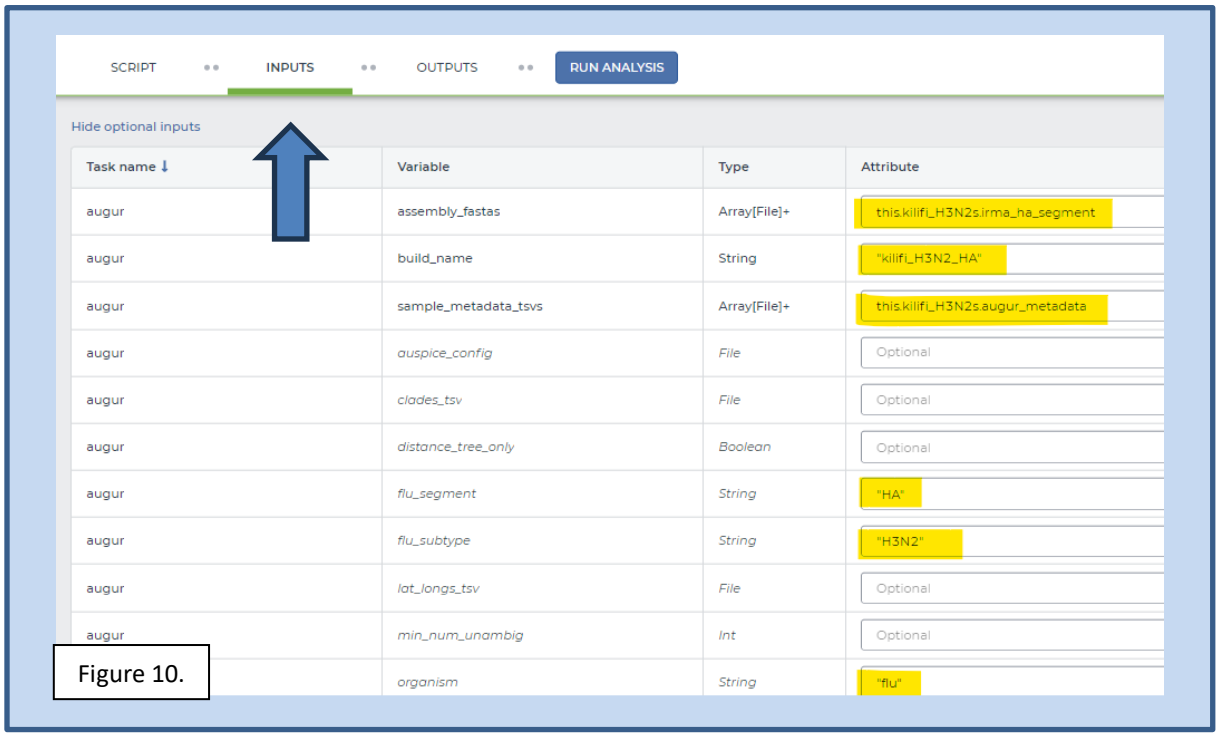

Figure 10.

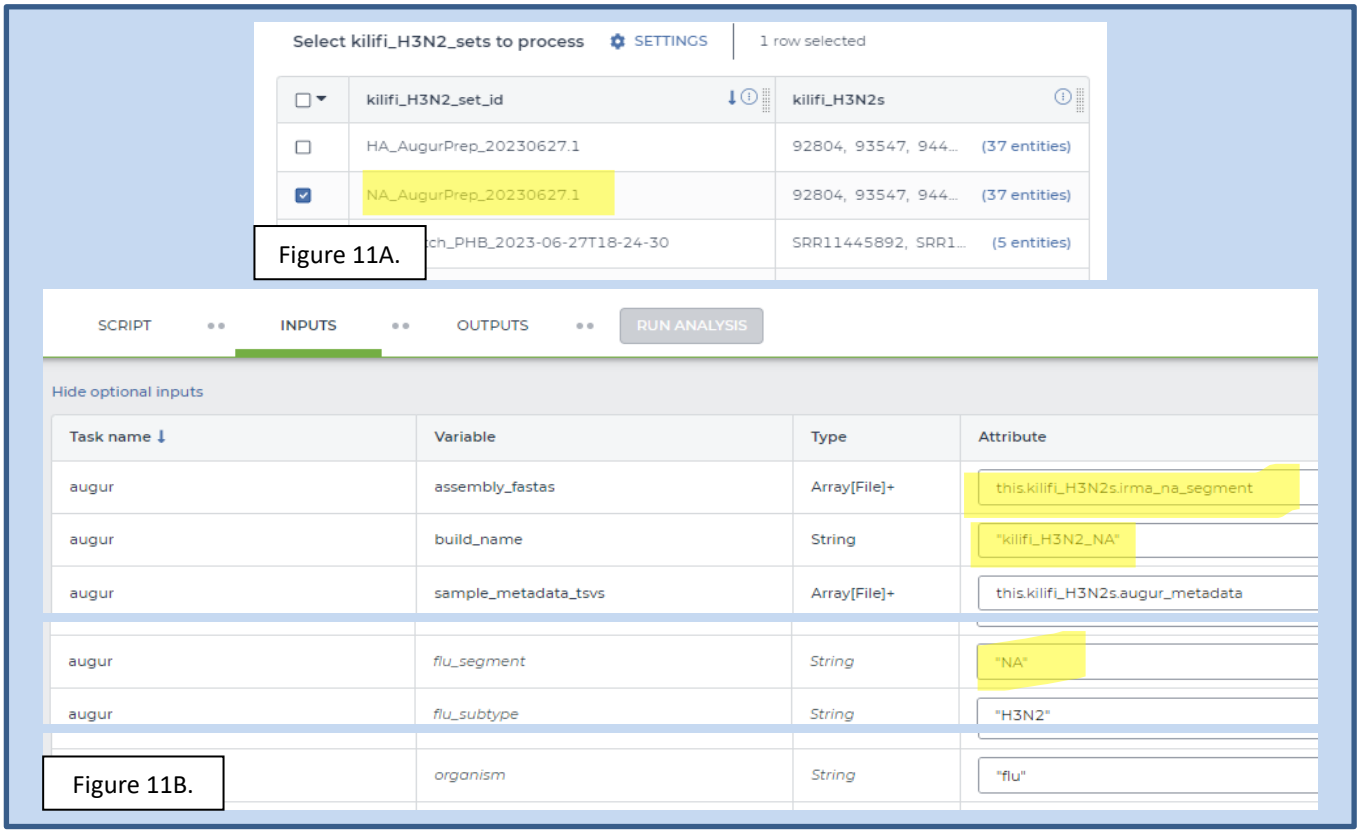

Figure 11A.

Figure 11B.

11. Repeat section 4.4 for the NA protein with the following changes:
    a. When selecting data to run, choose the *NA Augur Prep output file* (Fig 11A)
    b. Specify *this.kilifi_H3N2s.irma_na_segment*, *"kilkifi_H3N2_NA"* and *"NA"* in the first, second, and seventh input fields (Fig 11B)

4.5 DETERMINING FLU TYPING, SUBTYPING, AND LINEAGE

1. Navigate to the *data* tab of the Terra workspace containing flu data
2. *Open the data table* by clicking on the name of the data table in the left sidebar

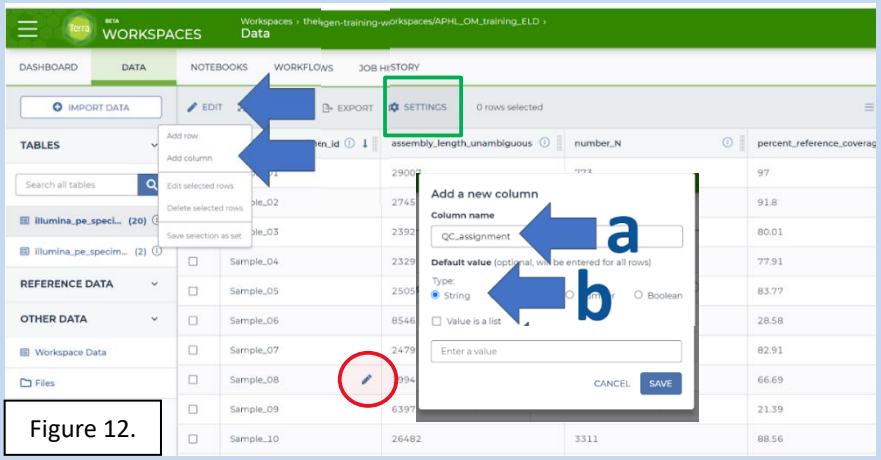

Figure 12.

3. View *settings* above the data table (Fig 12), select *none* (Fig 13)
4. Select the following columns: *abricate_flu_subtype*, *abricate_flu_type*, *irma_subtype*, *irma_subtype*, *nextclade_clade*, and nextclade_lineage
   a. *Optional: save this column group for future use by clicking the* save this column selection *field,* naming it *(e.g. FluTyping), and clicking* save
   b. *IRMA is used to produce a consensus and variants assembly; abricate is used to confirm IRMA typing results*
5. Click *done*
6. Determine the type, subtype, and lineage for each sample by viewing the corresponding columns
7. Identify the Pangolin lineage for each sample
   a. In the data table, find the column titled *pango_lineage*; nomenclature will be similar to the following: B.1.167
8. Follow lab-specific resulting and reporting procedures, as applicable

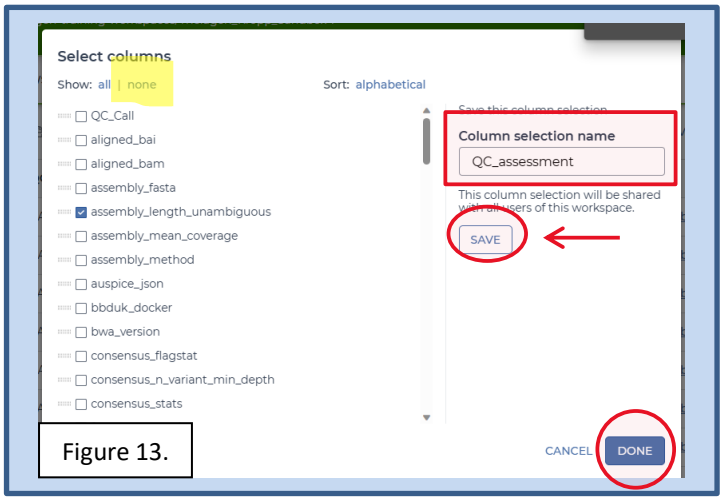

Figure 13.

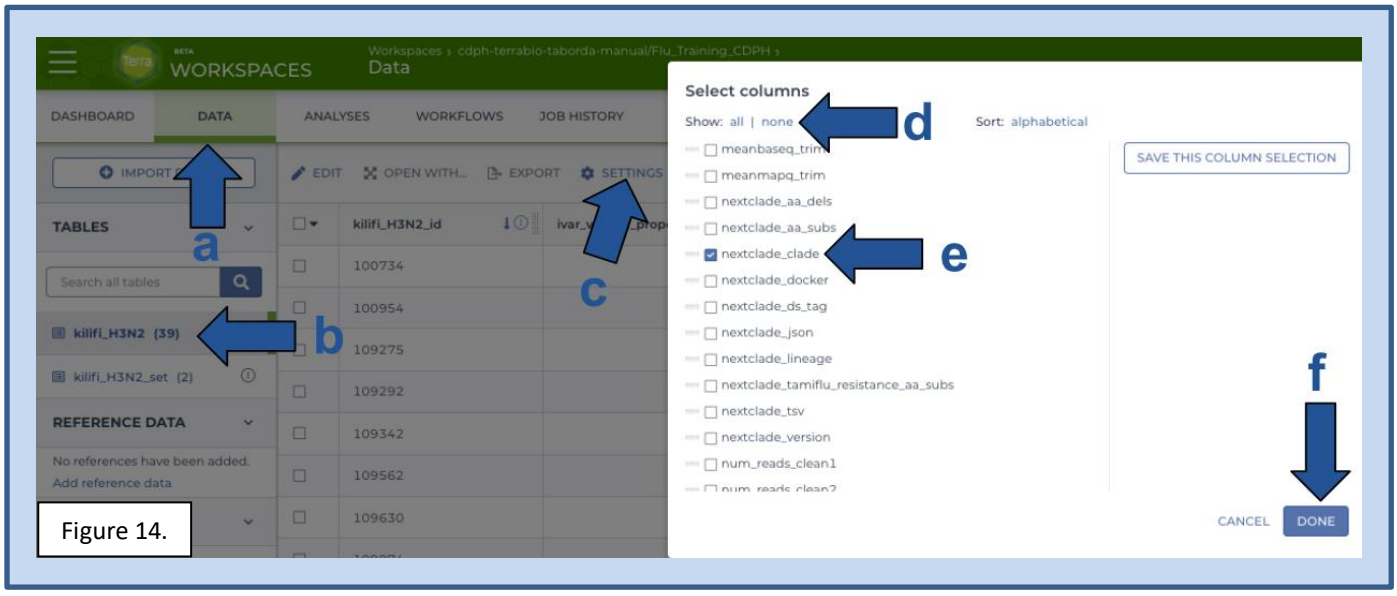

Figure 14.

4.6 VISUALIZING THE AUGUR TREE IN AUSPICE

1. Navigate to the *workspace data tab* and select the *data table* containing flu data
2. Click *settings* and *none*; then select only the nextclade_clade column (Fig 14)

3. *Select the checkbox* for each sample containing a value in the nextclade_clade column, including "unassigned" values (Fig 15, g)
4. Click *export* and *download as TSV* to download the metadata file (Fig 15, h)
5. *Save the file* as "[DATA_TABLE_NAME]auspice_metadata_YYMMDD.tsv"
6. In the flu Terra workspace, *open the "set" data table* with flu data; nomenclature is the same as the original data table plus the suffix "_set" (e.g. flu_H3N2_set)
7. Under the column titled auspice_input_json, *click to download* the file corresponding to the sample set of interest
   a. A new window will open in the browser; *right click* and *save* this file
8. In a new browser window *open https://auspice.us/*; *drag and drop* the auspice input json file onto the webpage
9. *Drag and drop* the metadata file onto the webpage
   a. The sample names in the auspice output json and the metadata file must match; if they don't, open the metadata file and manually edit the names
10. In auspice, click the *color by* dropdown and choose *nextclade_clade*
    a. *Unassigned clades may be older clades that are not currently assigned a clade by nextclade; these are displayed in black in the auspice timetree*
11. *View the timetree and map* with geographical coordinates of the analyzed strains (Fig 17)
12. Repeat section 4.6 with the NA auspice tree and corresponding metadata

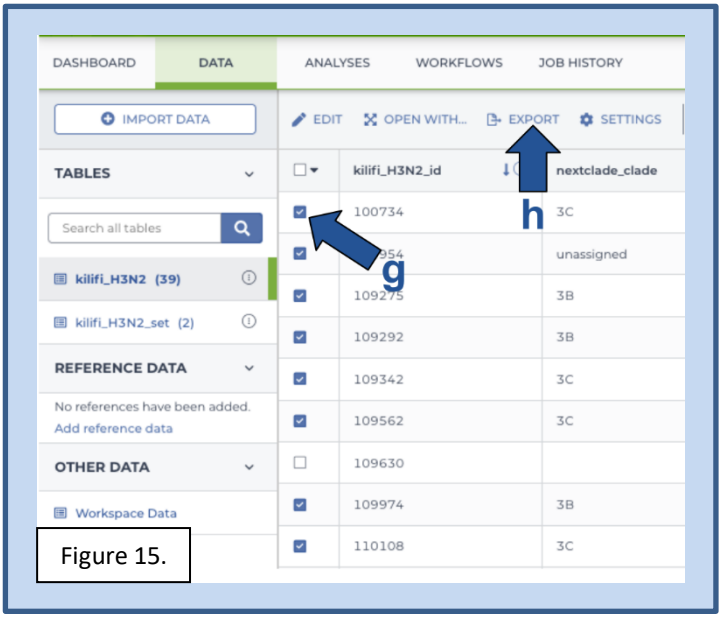

Figure 15.

Figure 16.

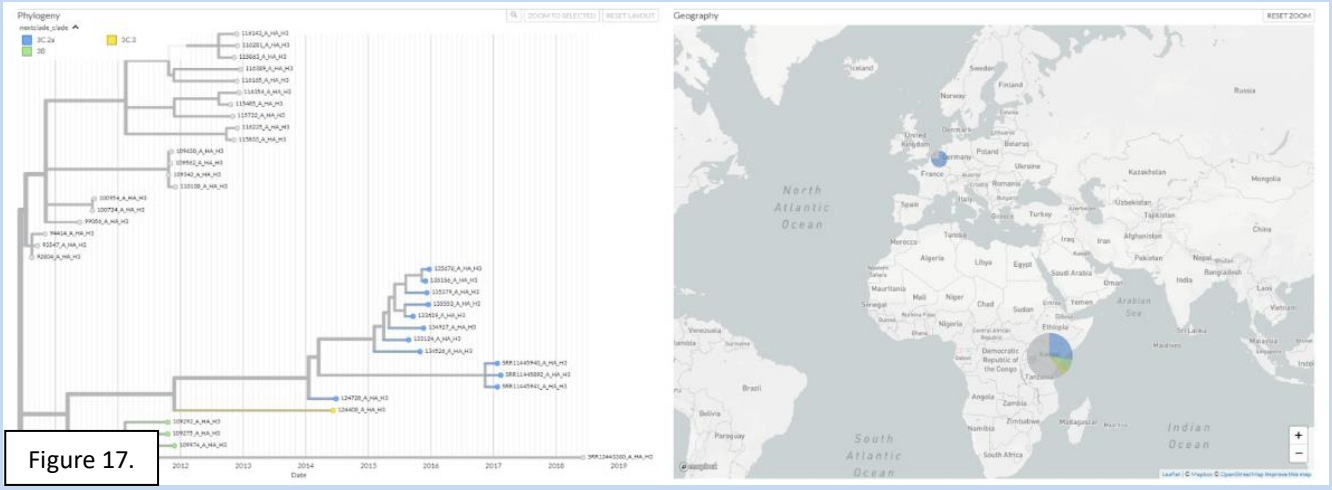

Figure 17.

5. QUALITY RECORDS

- Raw read files
- Sample read and assembly QC metrics
- Nextclade_clade and pango_lineage determinations
- Auspice timetree and map

6. TROUBLESHOOTING

- Consult with internal staff familiar with this procedure or contact support@theiagen.com for troubleshooting inquiries
- For document edit requests, contact support@theiagen.com

7. INTERFERENCES

N/A

8. REFERENCES

None

9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	7/2023

10. APPENDICES

10.1 USING EXCEL TO ALIGN TSV SAMPLE NAMES TO AUSPICE.JSON FILE NAMES

1. Open the metadata tsv file titled "[DATA_TABLE_NAME]auspice_metadata_YYMMDD.tsv" in excel
2. Select column B, right click, and insert a column between columns A and B
3. In cell B2, enter the following formula relevant to the protein segment being analyzed
   a. For HA protein segments: *=A2&"_A_HA_H3"*
   b. For NA protein segments: *=A2&"_A_NA_N2"*
4. With cell B2 selected, click and drag the green plus at the bottom right of the green box down to copy the formula to the remaining samples
5. Select column C, right click, and insert a new column between columns B and C
6. Select and copy column B contents
7. Select column C and right click to paste values into column C
8. Copy and paste cell A1 into C1
9. Delete columns A and B leaving only the pasted sample name values and nextclade_clade
10. Save the file and return to auspice; drag and drop or upload the new metadata file into auspice to add corresponding sample metadata