Running Theiagen's Snippy_Variants_PHB Workflow in Terra	
Document TG-SNP-01, Version 1	
Date:	Workflow Version:
01/21/2025	PHB v2.3.0

1. PURPOSE/SCOPE

This procedure describes the process of running Theiagen's Snippy_Variants_PHB pipeline using the Terra platform. Acceptable data types include single-end (SE) or paired-end (PE) reads in FASTQ format, or assembled sequences in FASTA format. A reference genome is provided as a workflow input for alignment to identify single-nucleotide polymorphisms (SNPs), multi-nucleotide polymorphisms (MNPs), and insertions/deletions (indels). An annotated GenBank file may be used as a reference genome together with user-specified genes of interest (as workflow inputs) to output mutations identified with the associated gene or annotation names.

2. REQUIRED RESOURCES

- Computer
- Internet connection: at least 10 and 5Mbps for download and upload speed, respectively
- Internet browser
  - Google Chrome, Firefox, or Edge
- Google account
- Terra account, linked to Google account
- Illumina SE or PE raw sequencing read files uploaded to Terra workspace
- Theiagen's Snippy_Variants_PHB workflow 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

*Prior to running the Snippy_Variants_PHB workflow, the workflow must be imported into the Terra workspace and a reference genome file must be available to use for analysis. For details see **Appendix 10.1** and **10.2**, respectively.*

4.1 Configure and Run the Snippy_Variants_PHB Workflow

1. Open Terra, navigate to the *workflows* tab, and select the *Snippy_Variants_PHB* workflow (Fig 1).
   a. *To import the workflow for the first time, refer to Appendix 10.1.*

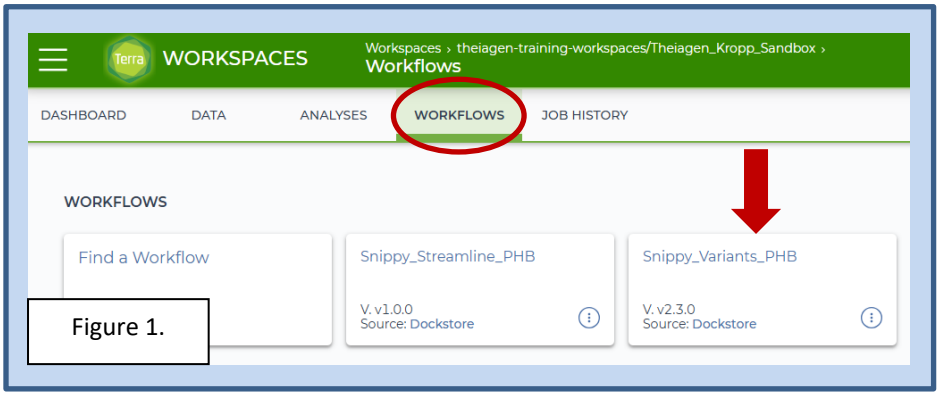

Figure 1.

2. In the version dropdown field, *select the workflow version* that was internally validated, or the latest version of the workflow (Fig 2).
3. *Uncheck call caching* (Fig 2).
4. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 2, b).
5. Choose the relevant *data table* under the select data table dropdown (Fig 2, c).
6. Click *select data* (Fig 2, d).

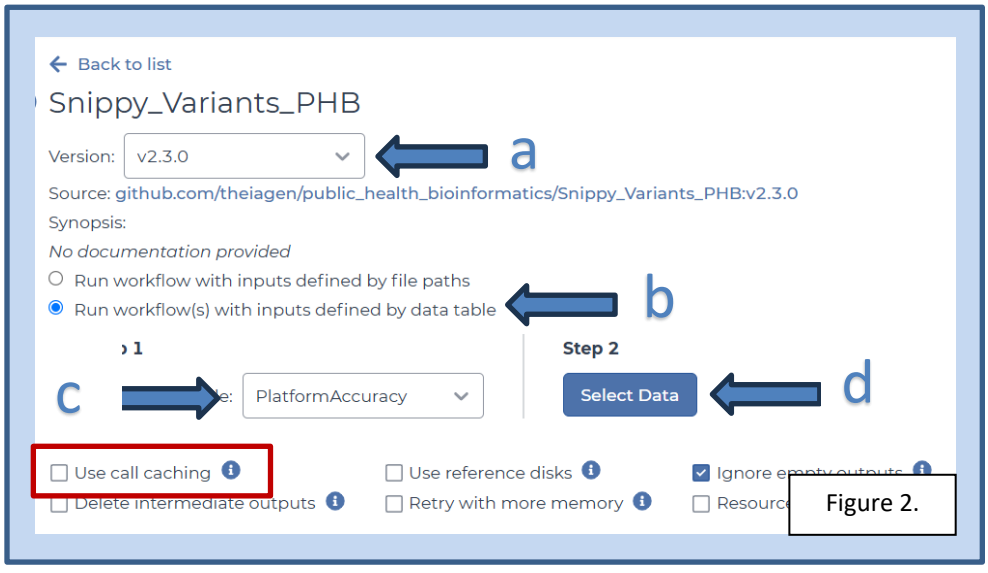

Figure 2.

7. In the pop-up window, *select each sample checkbox* to include in the analysis (Fig 3).
   a. *Click the down arrow and select all to process all specimens.*
   b. *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.*
   c. Scroll to the bottom and click *ok*.

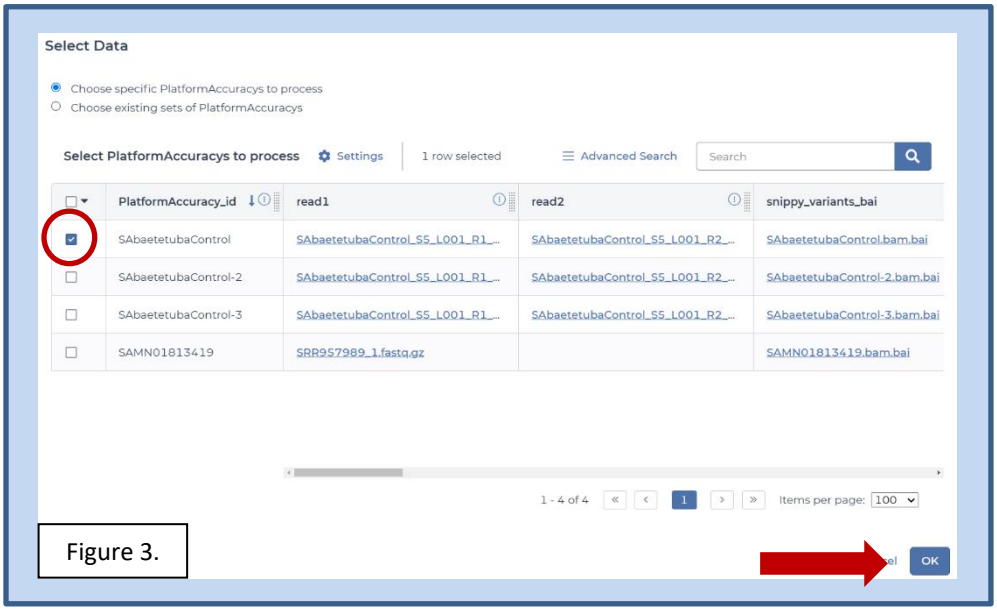

Figure 3.

8. In the first value field for the reference_genome_file variable, *click the folder icon* (Fig 4).

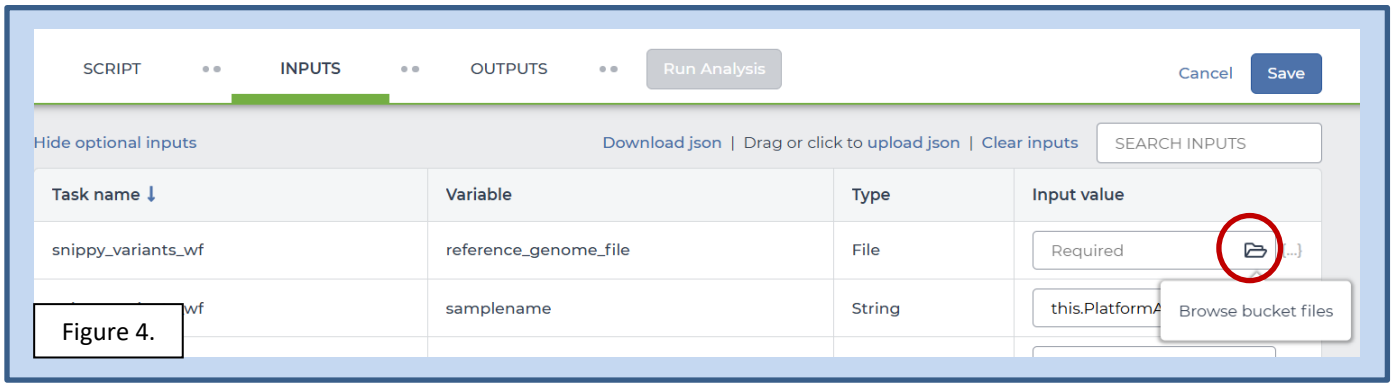

Figure 4.

9. *Select the appropriate reference genome file* from the workspace (Fig 5).
   a. *To import reference genomes for the first time, see Appendix 10.2.*

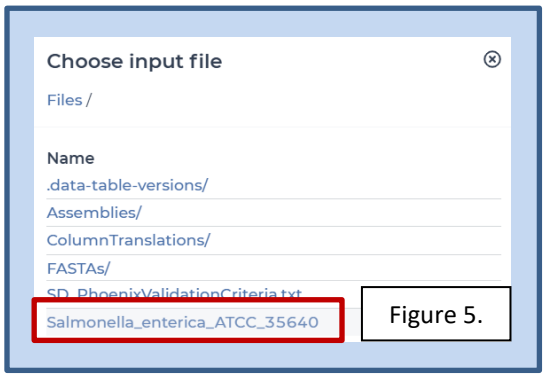

Figure 5.

10. Set the second input value for the `samplename` variable to *this.PlatformAccuracy_id*, where "PlatformAccuracy" is the name of the data table to be analyzed (Fig 6).
    a. *The `samplename` variable input should appear in the dropdown when clicking in the input value field. If it does not appear, check the data table name selected for analysis.*

Figure 6.

11. Scroll down through the input variable list and set the `read1` and `read2` variables to *this.read1* and *this.read2*, respectively where `read1` and `read2` are the data table column names containing read files (Fig 7).

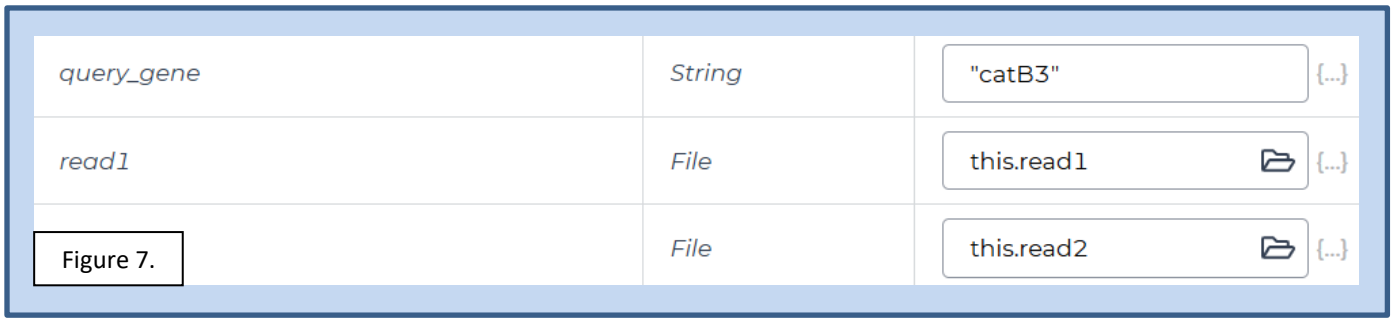

Figure 7.

12. Optional: In quotations, input a gene name or other annotation(s) of interest into the `query_gene` input variable; e.g. "catB3" (Fig 7).
    a. *The text must match the GenBank file exactly.*
13. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 8).
14. Click *save* (Fig 8).

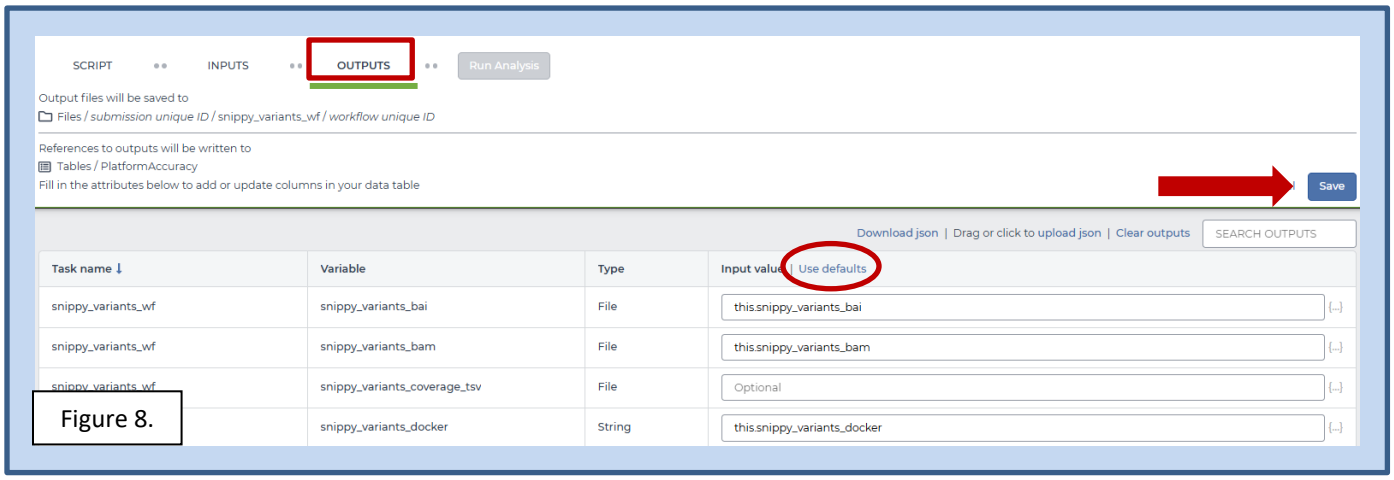

Figure 8.

15. Click *run analysis*, *enter comments* as needed, and select *launch* (Fig 9).

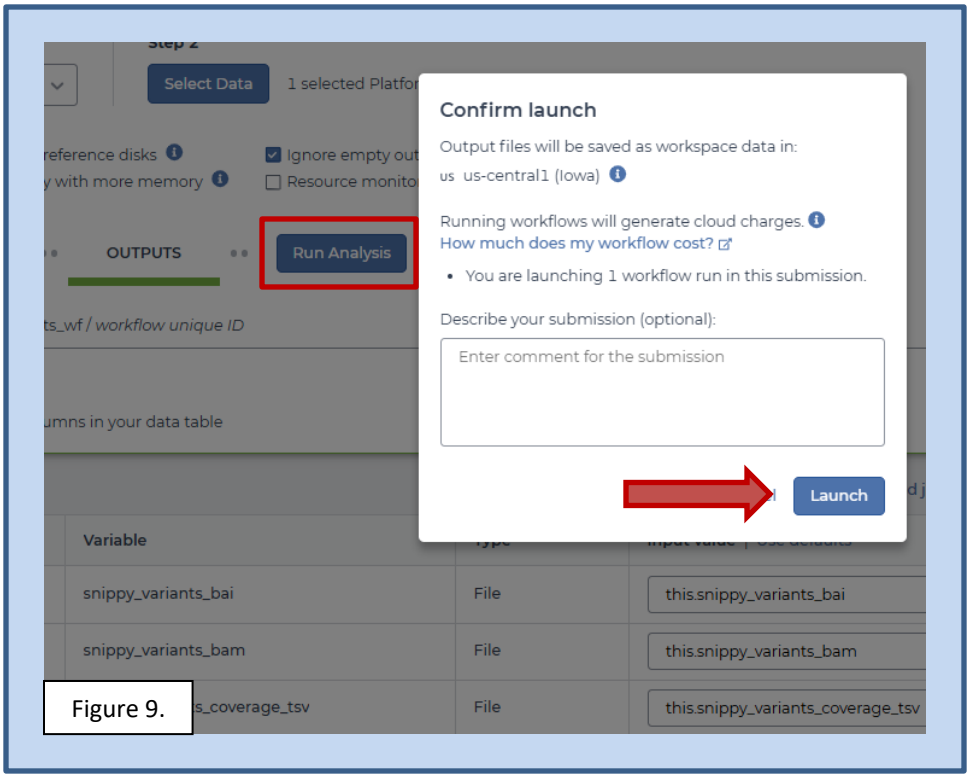

Figure 9.

5. QUALITY RECORDS

- Raw read files
- Workflow version and input parameters
- Reference sequence and metadata, as applicable
- Sample read, assembly, and result-specific QC metrics
- All workflow outputs relevant to results, including tool and database versions

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. LIMITATIONS

- "Mutations identified by the Snippy_Variants_PHB workflow are highly dependent on the choice of reference genome. Mutations cannot be identified in genomic regions that are present in your query sequence and not the reference."[1]
- "The outputs from samtools coverage (found in the snippy_variants_coverage_tsv file) may differ from the snippy_variants_percent_ref_coverage due to different calculation methods. samtools coverage computes genome-wide coverage metrics (e.g., the proportion of bases covered at depth ≥ 1), while snippy_variants_percent_ref_coverage uses a user-defined minimum coverage threshold (default is 10), calculating the proportion of the reference genome with a depth greater than or equal to this threshold."[1]

8. REFERENCES

[1]"Phylogenetic Construction – Snippy Variants." Theiagen Public Health Bioinformatics, Theiagen Genomics, 3 Dec. 2024, https://theiagen.github.io/public_health_bioinformatics/latest/workflows/phylogenetic_construction/snippy_variants/?h=snip. Accessed 23 Jan. 2025.

9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	1/2025

10. APPENDICES

10.1 Find and Import the Snippy_Variants_PHB Workflow

1. Navigate to the *workflows* tab of the workspace (Fig 10).
2. Users that already have the Snippy_Variants_PHB workflow in their workspace can *select the workflow* (Fig 10) and proceed to the Configure and Run the Snippy_Variants_PHB Workflow section of this SOP.

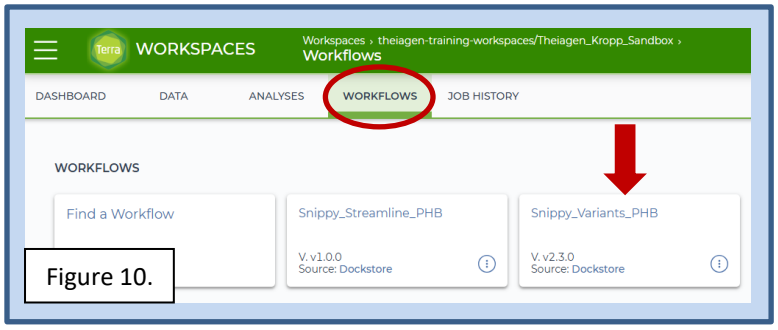

Figure 10.

   a. To import the workflow, click *find a workflow* (Fig 10).
   b. In the pop-up window, click *Dockstore.org* in the bottom left (Fig 11).
   c. Click *Organizations* in the banner at the top and search for *Theiagen* using the search box (Fig 12).

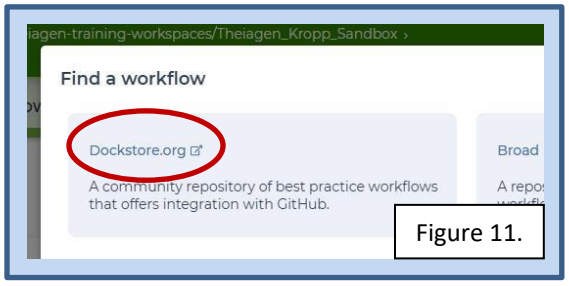

Figure 11.

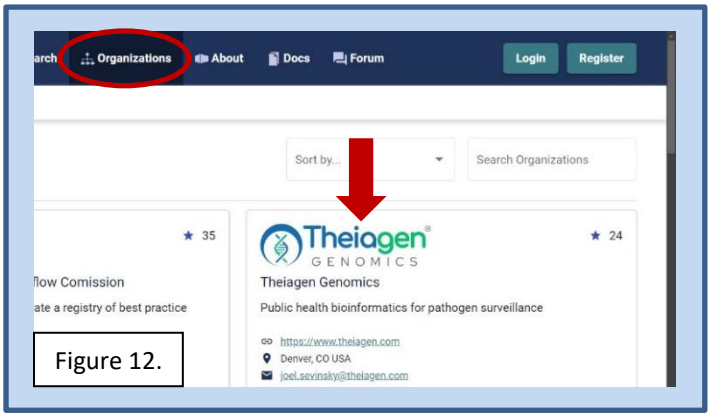

Figure 12.

d. Click the *Public Health Bioinformatics (PHB)* collection (Fig 13)and using *ctrl* + *f* on Windows search for *snippy* (Fig 14).

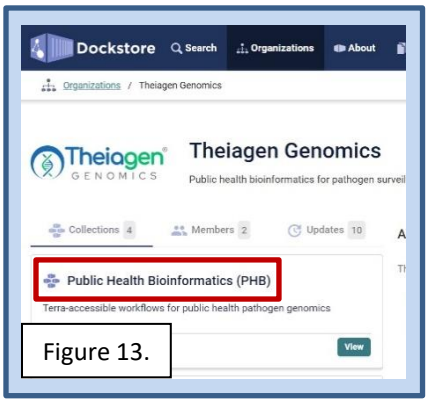

Figure 13.

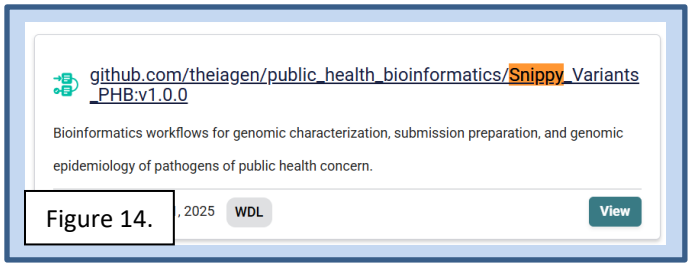

Figure 14.

e. Click on the *Terra icon* (Fig 15) to import the workflow into a Terra workspace.

f. *Select the workspace* in the destination workspace dropdown field and click *Import* (Fig 16).

Figure 15.

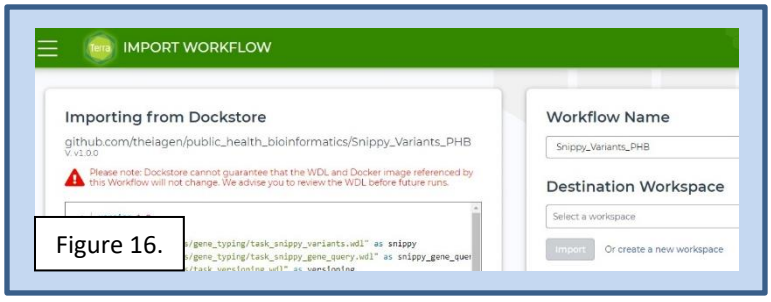

Figure 16.

10.2 IMPORT THE REFERENCE GENOME FASTA FILE

1. In the Terra workspace data tab, click on the *folder* icon on the right-hand side (Fig 17).

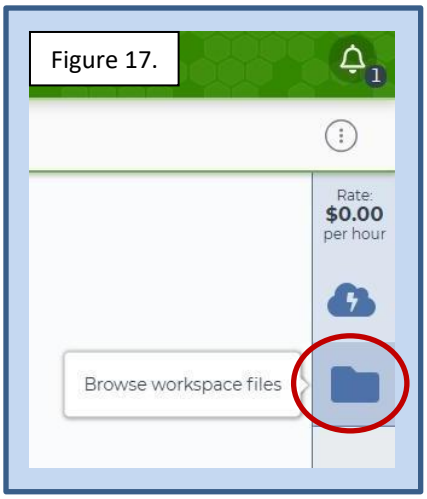

Figure 17.

2. Click *upload* and select the corresponding reference genome file in fasta (e.g. .fa, .fasta) or full GenBank (.gbk) file format (Fig 18).
3. Choose the relevant file and click *open* to import the file.

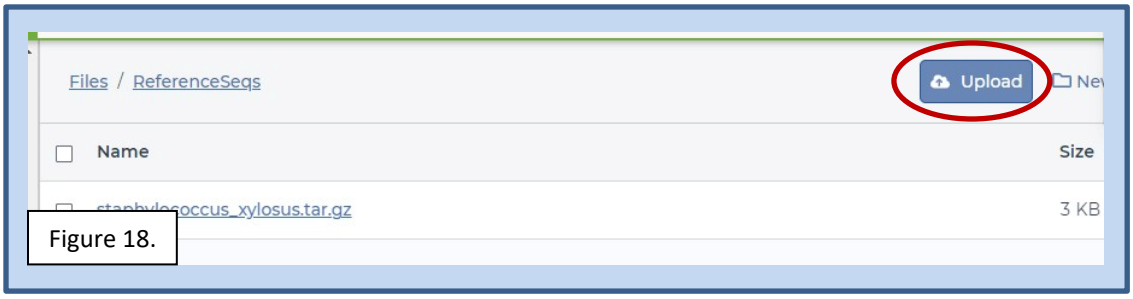

Figure 18.