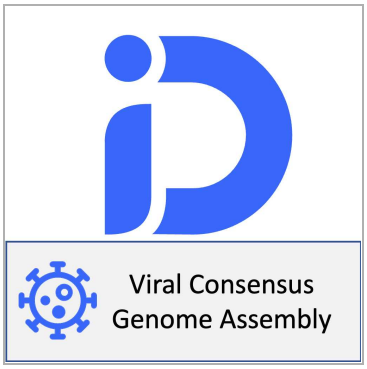


JUL 26, 2023

CZ ID Workflow for Assembling Viral Consensus Genomes

CZ ID
Karyna Rosario Cora[1], Elizabeth Fahsbender[1], Team[1]

[1]Chan Zuckerberg Initiative (CZI)

Karyna Rosario Cora

ABSTRACT

CZ ID's Viral Consensus Genome pipeline is designed to quickly assemble consensus genomes in bulk for any virus. Users can get started on their analysis by simply uploading Illumina sequencing files, providing a reference genome sequence, and including an optional primer BED file. The pipeline can be used with data obtained through primer spiking for target enrichment, PCR, whole genome sequence, or metagenomic assays. After uploading data, consensus genomes are automatically assembled against the user-provided reference sequence. This protocol describes how to upload, view, and download viral consensus genome data through CZ ID.

Click here to learn more about CZ ID's Viral Consensus Genome pipeline.

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.bp2l69ojklqe/v1

Protocol Citation: Karyna Rosario Cora, Elizabeth Fahsbender, CZ ID Team 2023. CZ ID Workflow for Assembling Viral Consensus Genomes. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bp2l69ojklqe/v1

Protocol status: Working
We use this protocol and it's working

Created: Jun 05, 2023

Last Modified: Jul 26, 2023

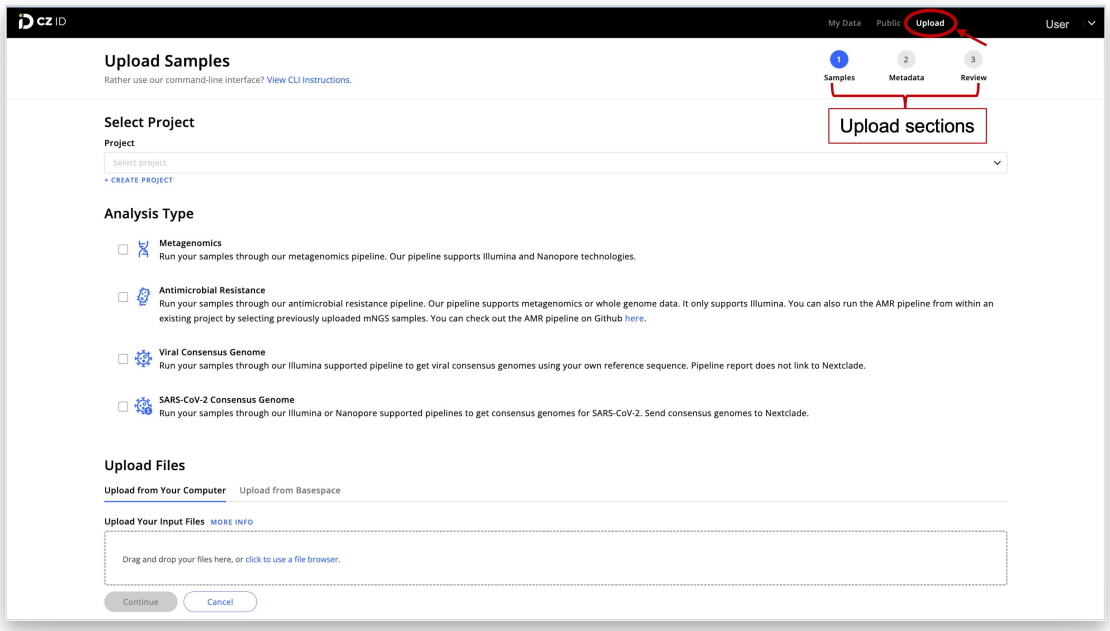
PROTOCOL integer ID: 82896

Keywords: CZ ID, virus consensus genome, viral consensus genome assembly

Upload Data

1 Log in to your CZ ID account.

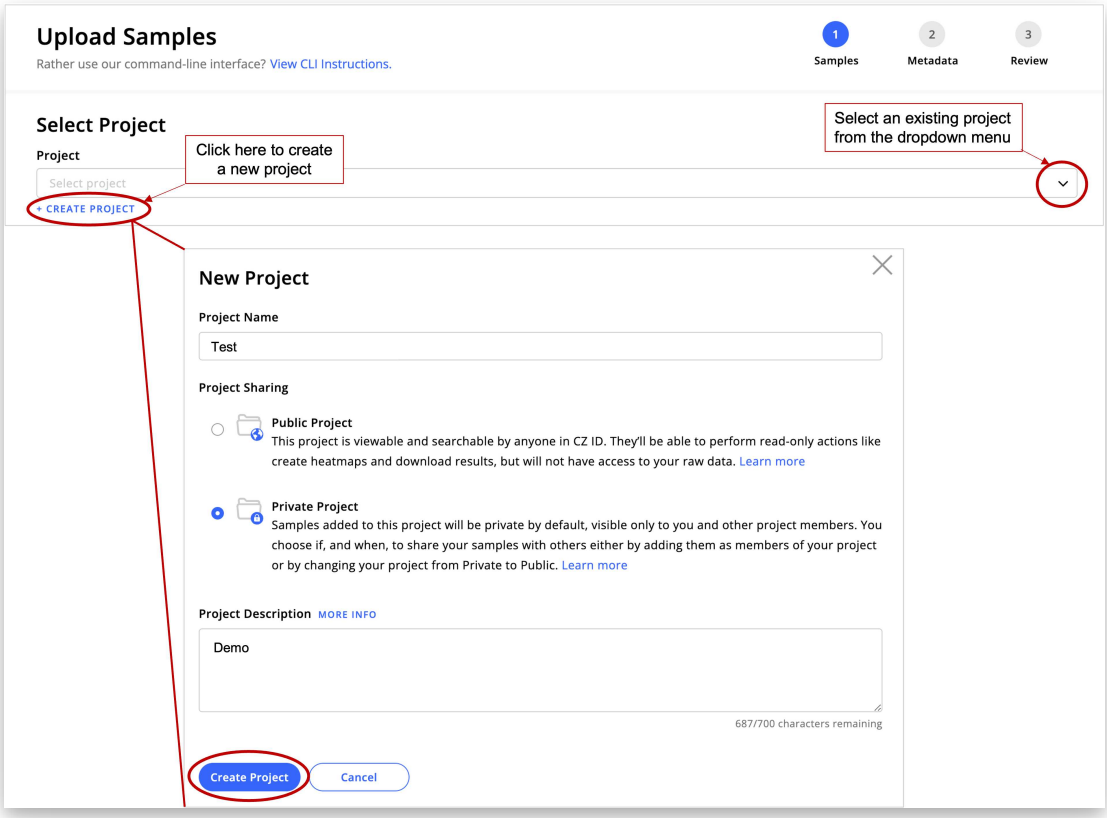
2 Navigate to the Upload page from the Discovery page by clicking the "Upload" link next to your username. Note that the upload process is divided into three general sections, including Samples, Metadata, and Review pages.



3 Select a project, analysis type, and sequencing files through the Upload Samples page.

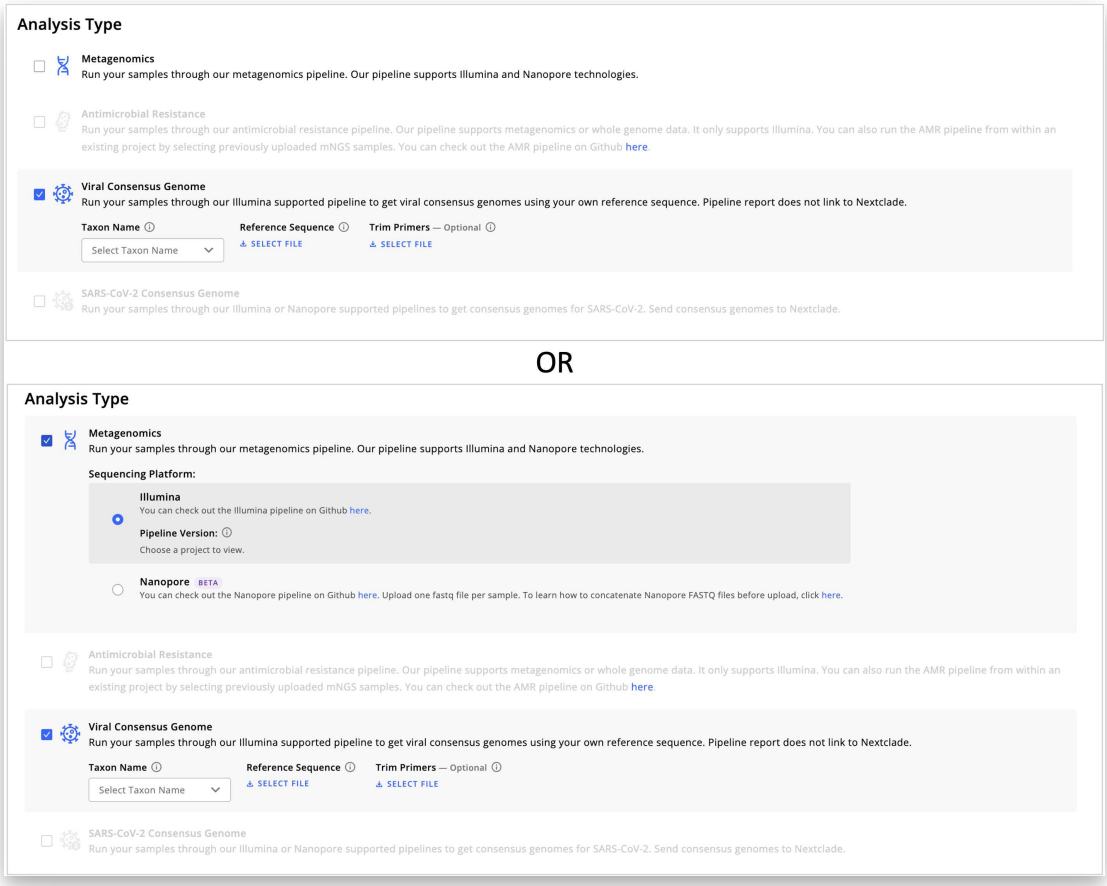
3.1 Selecting or Creating a Project

Samples uploaded to CZ ID are organized into projects. You can upload samples to an existing project or create a new one. When creating a new project, provide a project name, select the privacy status of the project, and provide a brief project description.



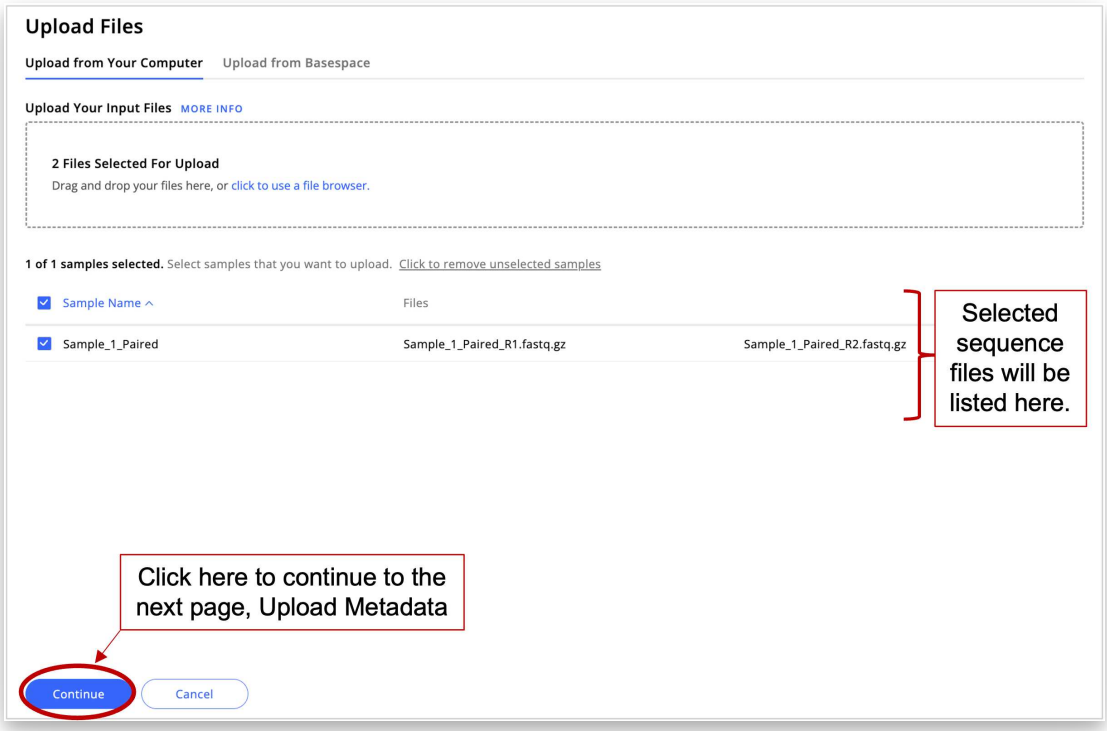
3.2 Selecting Analysis Type

For the analysis type you can select "Viral Consensus Genome" alone or choose to run metagenomic analysis at the same time. For Viral Consensus Genome, you will be prompted to provide a taxon name and a reference sequence file. You can opt to add a primer BED file to trim primers from reads during consensus genome assembly.



3.3 Uploading Sequence Files

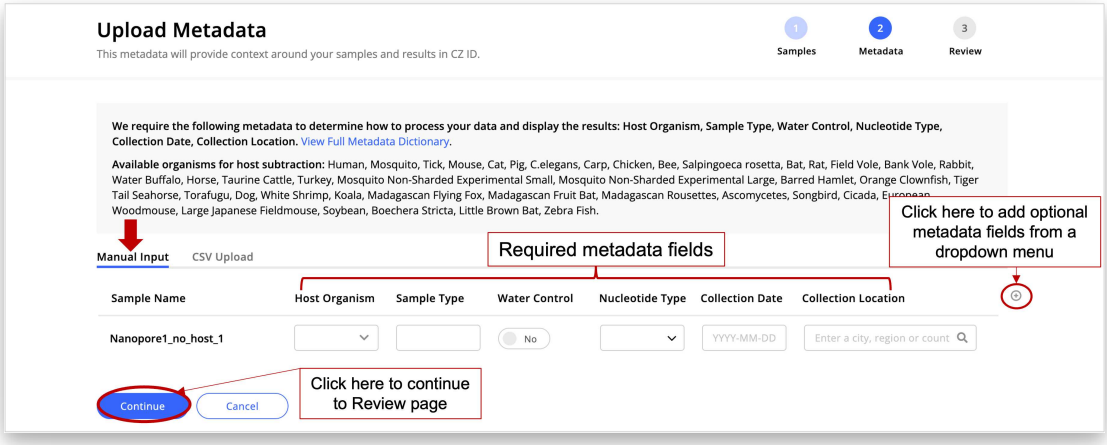
This is the final step within the Upload Samples page. Upload FASTQ (".fastq" or ".fq") or compressed FASTQ (".fastq.gz" or ".fq.gz") files directly from your computer (default) or retrieve sequencing files from BaseSpace. See Selecting Sequence Files for details. After selecting files, click the Continue button at the bottom of the screen to continue to the next page (Add Metadata).



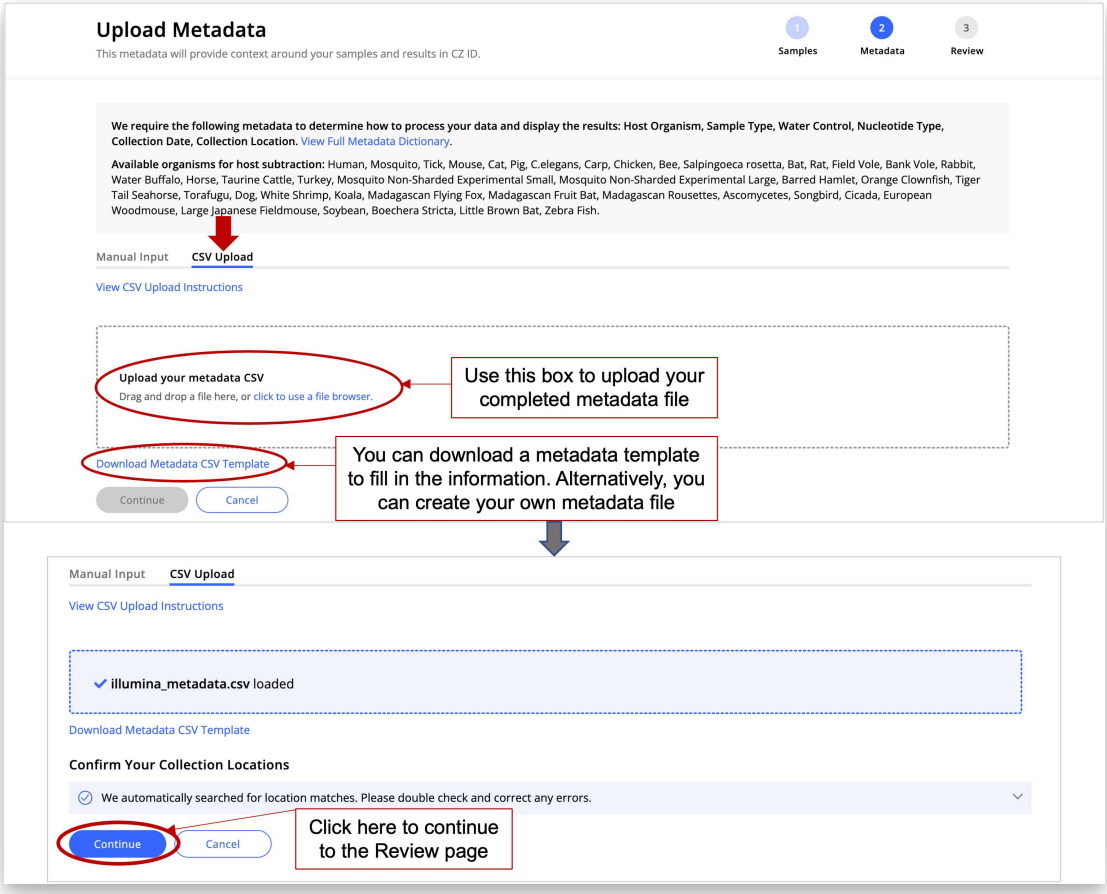
4 Add metadata through the Upload Metadata page.

4.1 Adding Metadata

You can enter metadata manually or by uploading a metafile file. Note that there are six required metadata fields, including: Host Organism, Sample Type, Water Control, Nucleotide Type, Collection Date, and Collection Location. See Adding Metadata for details. After adding metadata, continue to the next page (Review).



For manual metadata entry (Manual Input tab), enter the information in the provided table. By default, only required fields will be shown. However, you can add metadata fields by clicking the "plus" sign to the right of the table.

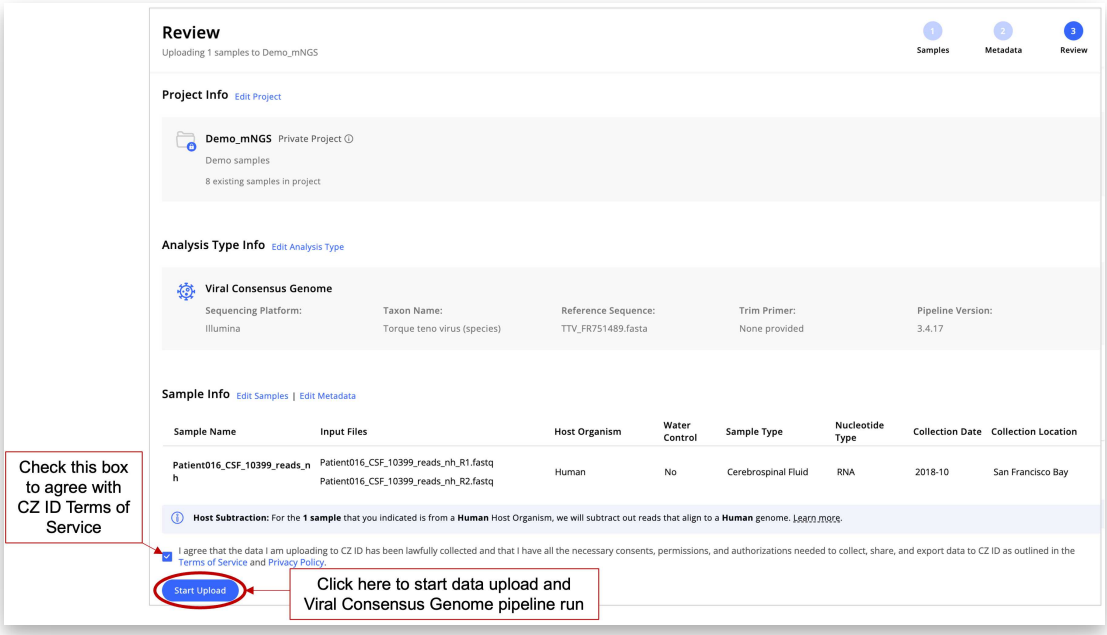


Add metadata by uploading a comma-delimited metadata file through the CSV Upload tab.

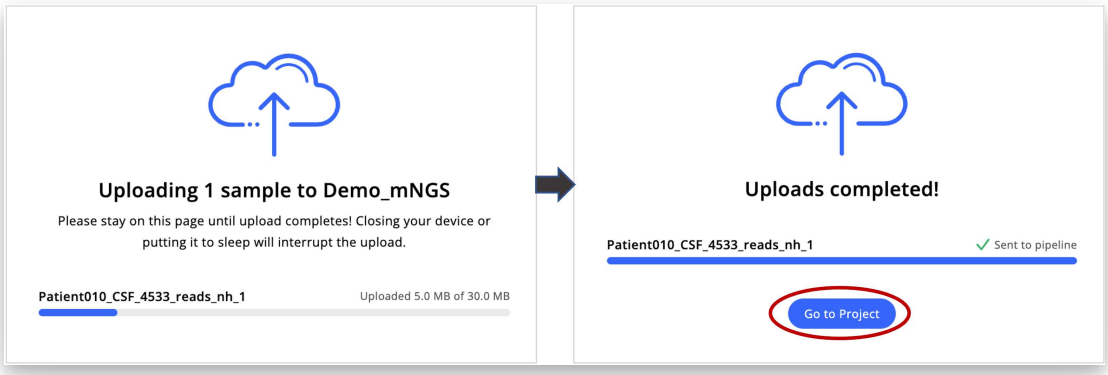
5 Review data and start upload through the Review page.

5.1 Reviewing Data

After adding metadata, you will be directed to the Review page where you can view samples and metadata ready to be uploaded. Review the project, sample, and analysis information. If you see an issue, you can edit your projects and your samples before uploading (note "Edit" links by each review section in the image below).



To begin uploading data to the Viral Consensus Pipeline, click “Start Upload” after accepting the CZ ID Privacy Policy and Terms of Service. Do not close the web page while samples are uploading to CZ ID servers. The upload will be canceled and you will have to re-start your upload. You will see an "Uploads completed" confirmation when your samples have been uploaded successfully. Once you see the confirmation, close your window or return to the Project page of interest to view the pipeline run status.

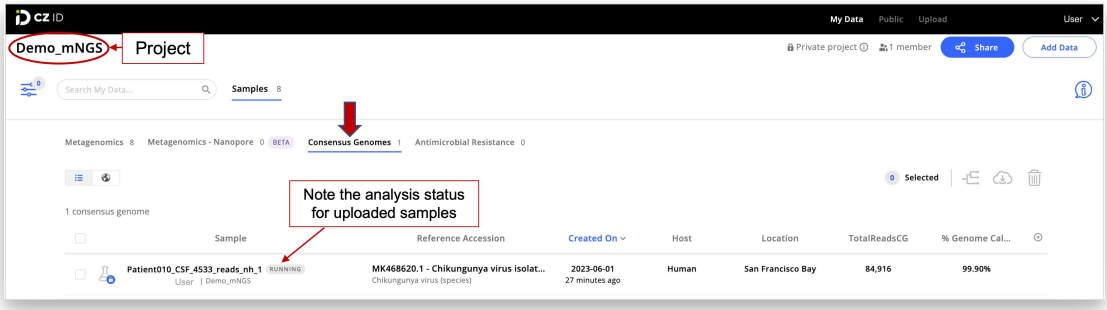


6 Check the status of the consensus genome.

6.1 Checking Genome Status

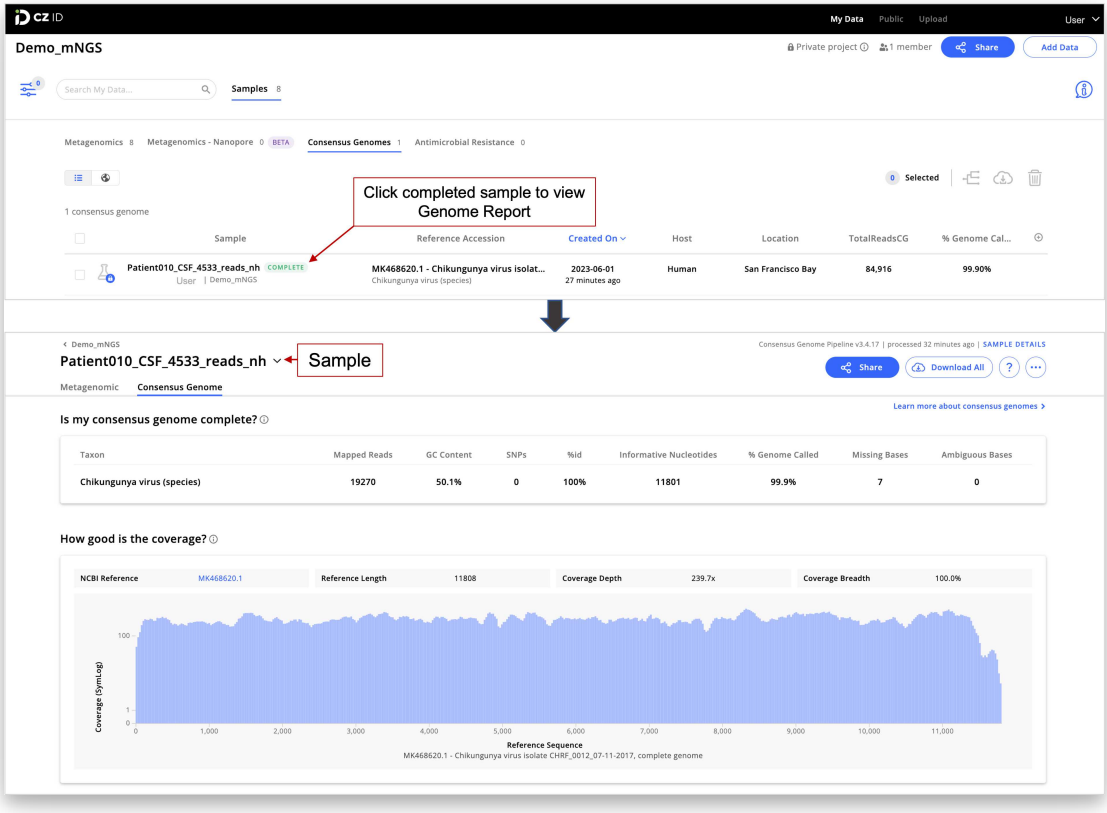
To view the status of your consensus genome, go to the Consensus Genome tab for the

Project page of interest.



View Genome Report

7 Once the sample run is completed, click on the sample to view Consensus Genome Report page.

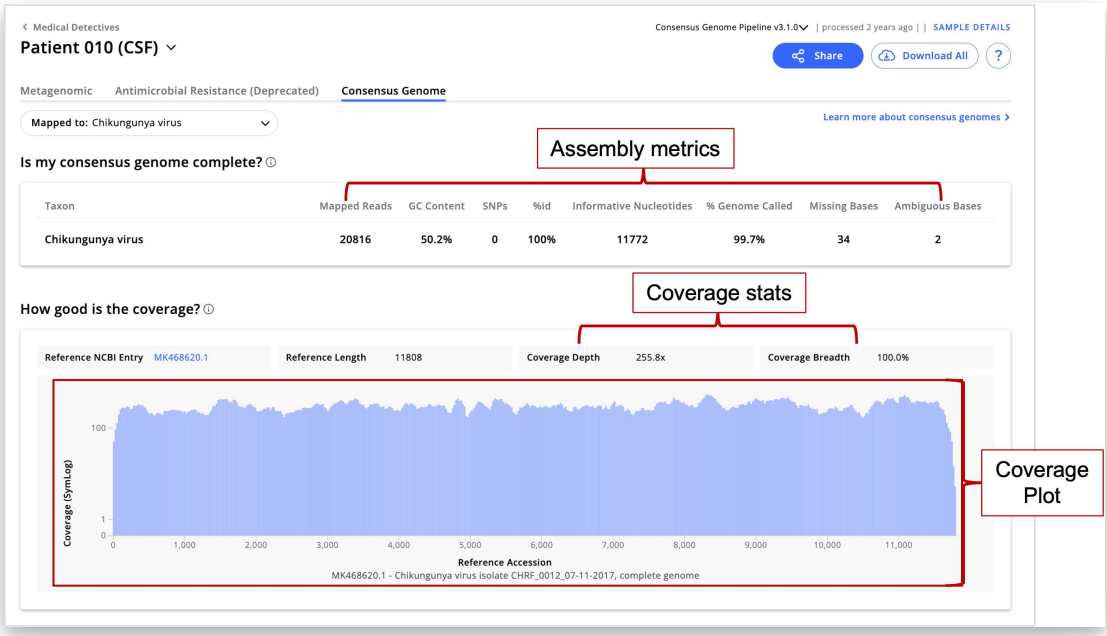


You will be directed to the Genome Report page after clicking on the sample name.

8 Review assembly metrics.

8.1 Assembly Metrics

You will be able to see various metrics on the Genome Report page. Use these metrics to asses the quality of the assembled genome.



Metrics include:

- **Coverage Plot** - Graph depicting the number of reads covering a given nucleotide of the reference sequence. The consensus genome must have >10 reads covering a specific genome site for a base to be called.
- **% Genome Called** - Refers to the percentage of the genome meeting thresholds for calling consensus bases.The closer this number is to 100%, the better.
- **SNPs** - Indicates the number of single nucleotide polymorphisms. SNPs represent single nucleotide variations between the reference accession and consensus genome.
- **Informative Bases** -Specifies the number of base calls (C, T, G, A) in the genome.
- **Ambiguous Bases** - If multiple sequencing reads support *more* than one nucleotide at a given site, those sites will be designated with an IUPAC ambiguity code. This metric specifies the number of non-C, T, G, A nucleotides in the consensus genome. The consensus genome pipeline only calls nucleotides that are detected at least at 75% frequency.
- **Mapped Reads** - Refers to the total number of reads that mapped to the reference genome.
- **GC content** - Percentage of G and C nucleotides in the consensus sequence. The GC content of the consensus sequence should be close to that of the reference sequence.

Download Data

9 Download virus consensus genome data, including consensus genome sequences (FASTA

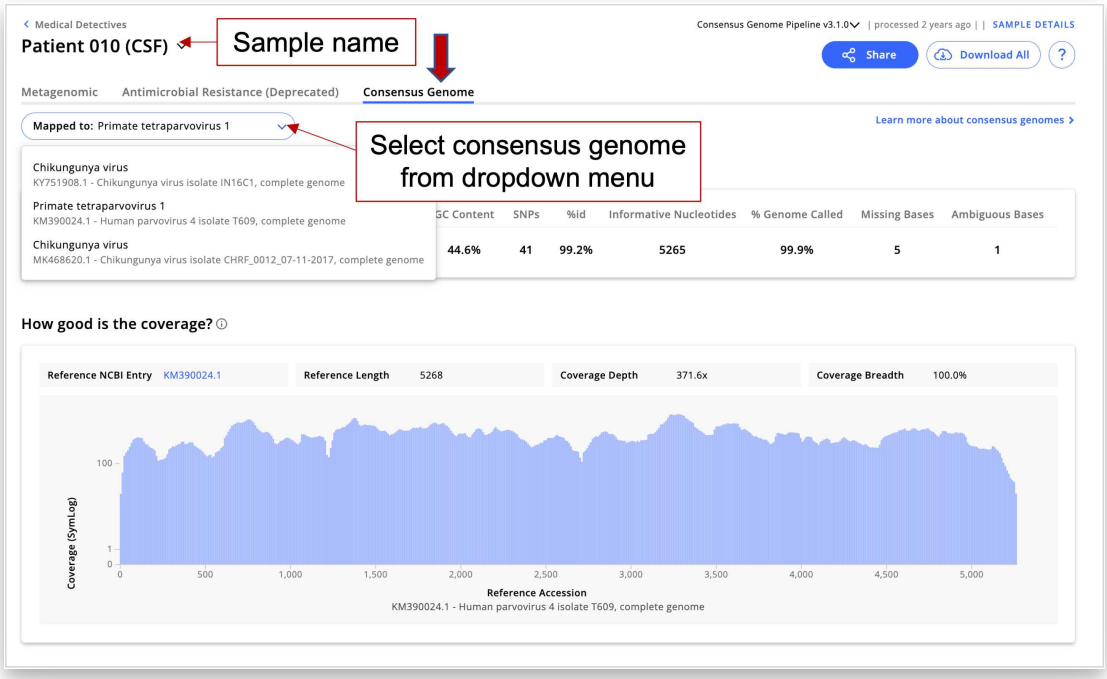
format) and intermediate files produced throughout the pipeline, through Genome Report and Project pages.

9.1 Downloading Data through Genome Report Page

You can download data for a single consensus genome from the Genome Report page. Here you can download the consensus genome sequence and generated intermediate files in a single folder.

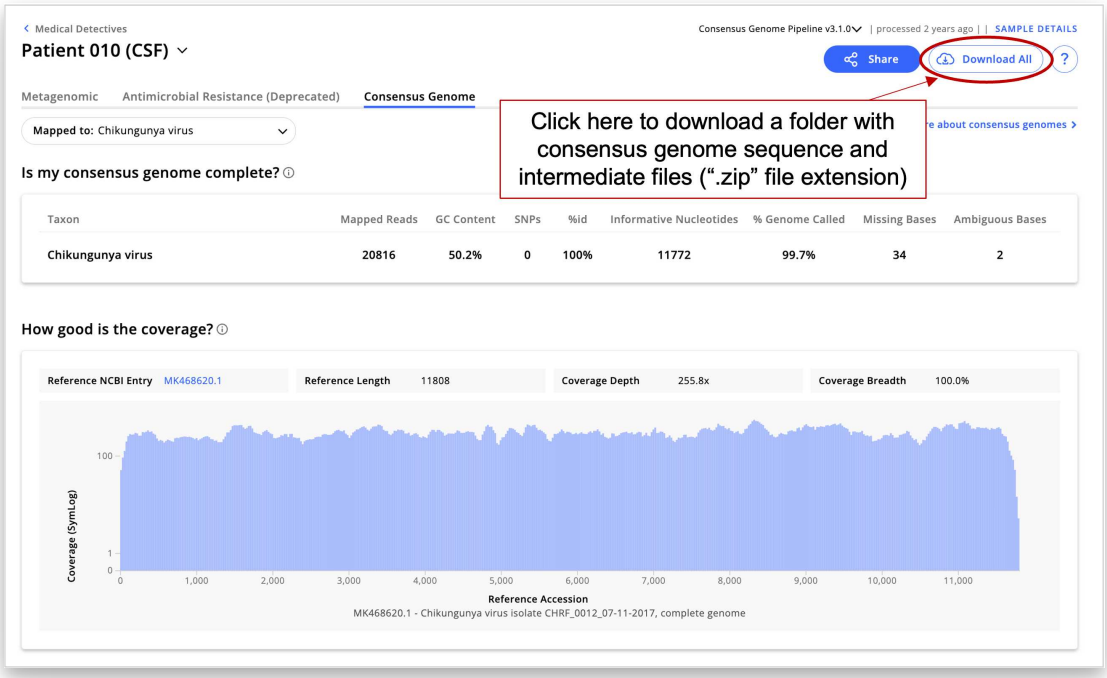
To download a folder with consensus genome data:

1. Navigate to the Consensus Genome tab for the sample and, if multiple genomes have been assembled, select the genome of interest from the dropdown menu.



2. To download all the data associated with the selected consensus genome, click the "Download

All" button on the right-hand side of the page.

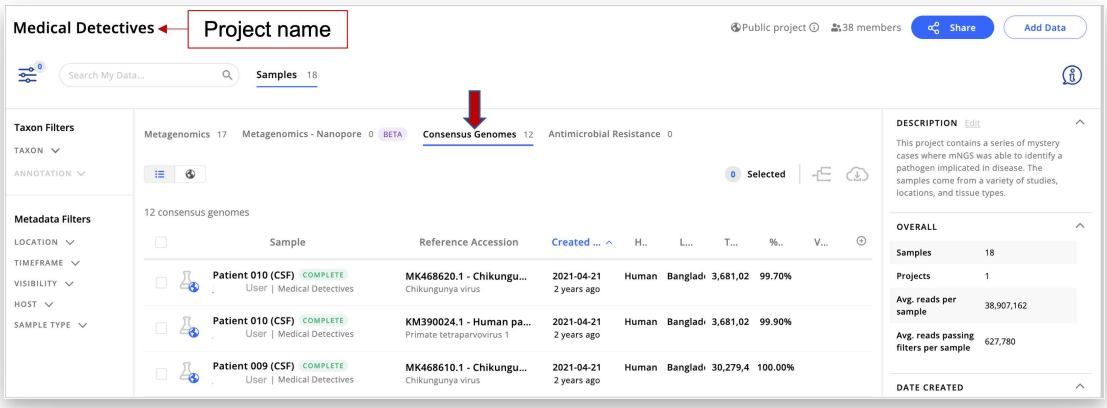


9.2 Downloading Data through Project Pages

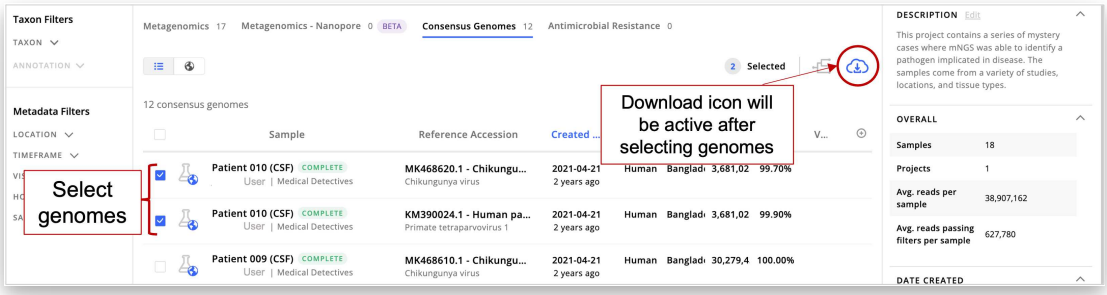
You can download data for a single or multiple consensus genomes (bulk download) from the Consensus Genome tab for a project of interest. From this tab you can download the consensus genome sequence, assembly metrics, sample metadata, and intermediate files.

To download consensus genome files of interest:

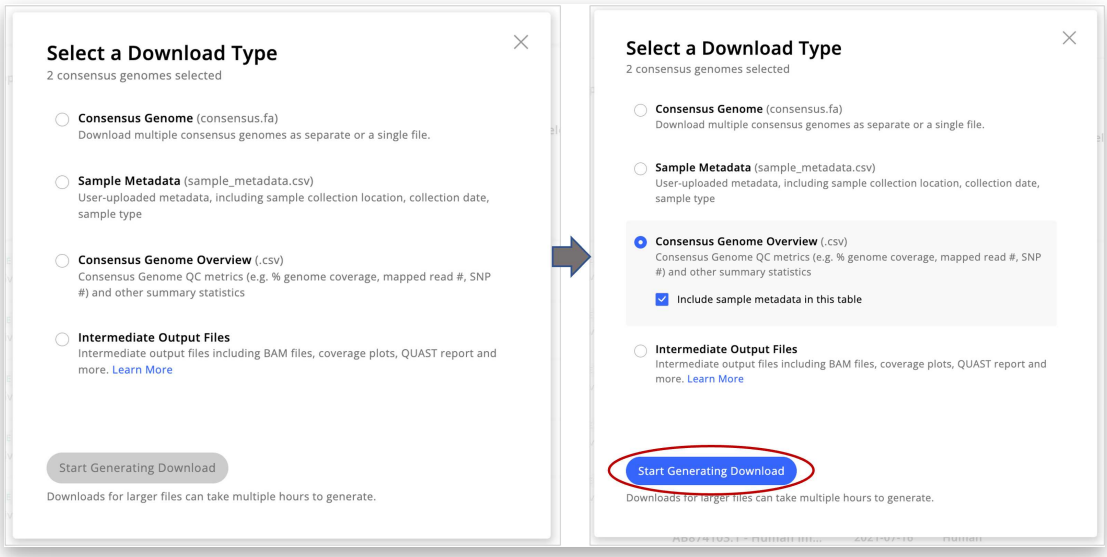
1. Navigate to the Consensus Genomes tab for the Project page of interest.



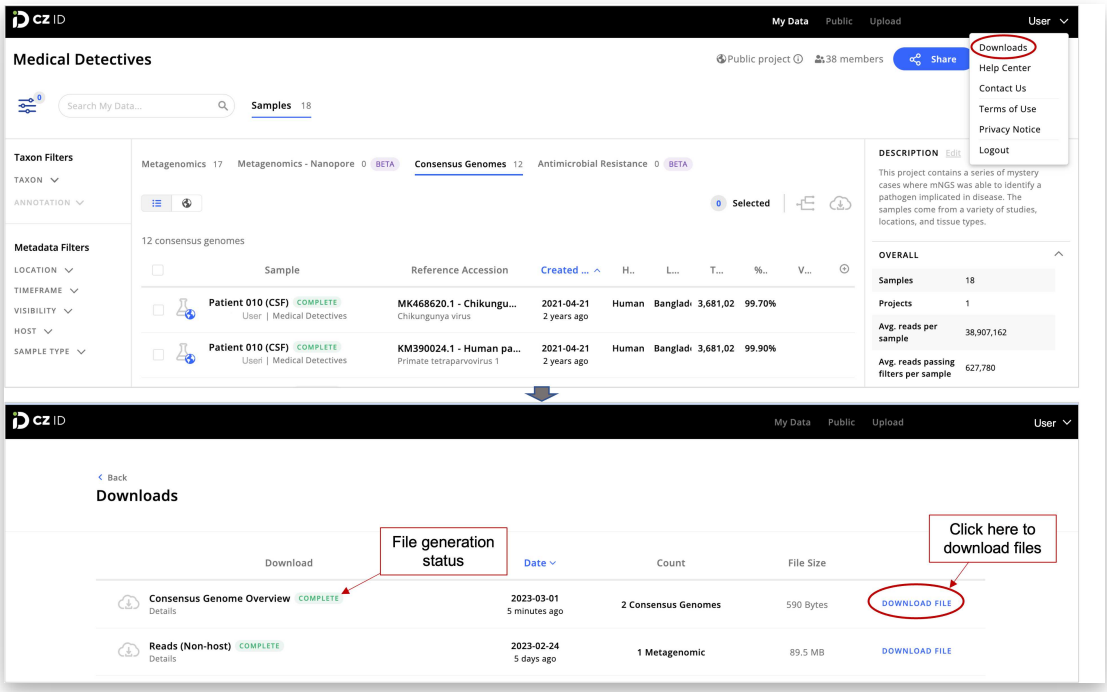
2. Select genomes to download and click Download icon.



3. Select the download type of interest from the dialog box.



4. To view file status and download files, navigate to the Downloads page through the username dropdown menu.



Use the dropdown menu by your username on the right-hand side of the page to go to the Downloads page.