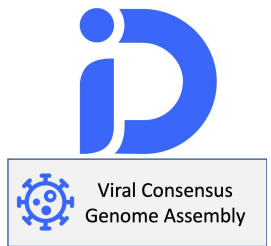


Jul 26, 2023

CZ ID Workflow for Assembling Viral Consensus Genomes

DOI

<https://dx.doi.org/10.17504/protocols.io.bp2l69ojklqe/v1>



Karyna Rosario Cora¹, Elizabeth Fahsbender¹, CZ ID Team¹

¹Chan Zuckerberg Initiative (CZI)



Karyna Rosario Cora

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://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>

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: June 05, 2023

Last Modified: July 26, 2023

Protocol Integer ID: 82896

Keywords: CZ ID, virus consensus genome, viral consensus genome assembly, viral consensus genome data through cz id, viral consensus genome pipeline, viral consensus genome data, assembling viral consensus, consensus genome, viral consensus, sequencing file, reference genome sequence, more about cz id, cz id, genome, whole genome sequence, metagenomic assay, bulk for any virus, target enrichment, optional primer bed file, virus, uploading illumina, pipeline, reference sequence

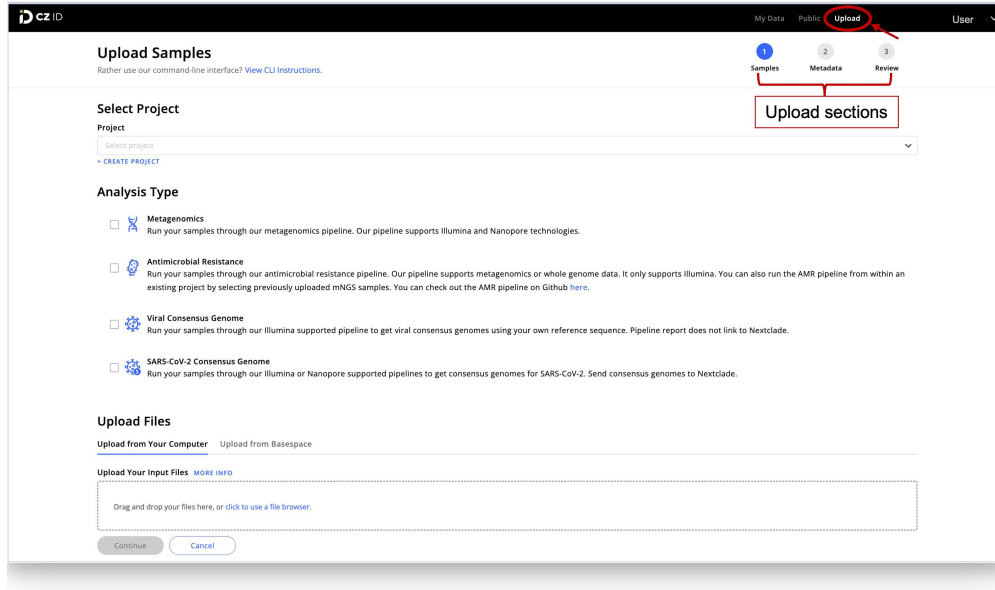
Abstract

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.

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.



The screenshot shows the 'Upload Samples' interface with three steps: 1. Samples, 2. Metadata, and 3. Review. The 'Select Project' section has a 'Project' dropdown menu. A red box highlights the 'Select project' text, and another red box highlights the 'Select an existing project from the dropdown menu' text. A red circle highlights the 'CREATE PROJECT' button. A red arrow points from the 'CREATE PROJECT' button to the 'New Project' modal. The 'New Project' modal has a 'Project Name' field with 'Test' entered. The 'Project Sharing' section has two options: 'Public Project' (selected) and 'Private Project'. The 'Project Description' field has 'Demo' entered. A red circle highlights the 'Create Project' button at the bottom of the modal. A red box highlights the 'Cancel' button.

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.

Analysis Type

☐ **Metagenomics**
Run your samples through our metagenomics pipeline. Our pipeline supports Illumina and Nanopore technologies.

☐ **Antimicrobial Resistance**
Run your samples through our antimicrobial resistance pipeline. Our pipeline supports metagenomics or whole genome data. It only supports Illumina. You can also run the AMR pipeline from within an existing project by selecting previously uploaded mNGS samples. You can check out the AMR pipeline on Github [here](#).

☒ **Viral Consensus Genome**
Run your samples through our Illumina supported pipeline to get viral consensus genomes using your own reference sequence. Pipeline report does not link to Nextclade.

Taxon Name ⓘ Reference Sequence ⓘ Trim Primers — Optional ⓘ
 [SELECT FILE](#) [SELECT FILE](#)

☐ **SARS-CoV-2 Consensus Genome**
Run your samples through our Illumina or Nanopore supported pipelines to get consensus genomes for SARS-CoV-2. Send consensus genomes to Nextclade.

OR

Analysis Type

☒ **Metagenomics**
Run your samples through our metagenomics pipeline. Our pipeline supports Illumina and Nanopore technologies.

Sequencing Platform:

☒ **Illumina**
You can check out the Illumina pipeline on Github [here](#).
Pipeline Version: ⓘ
Choose a project to view.

☐ **Nanopore** BETA
You can check out the Nanopore pipeline on Github [here](#). Upload one fastq file per sample. To learn how to concatenate Nanopore FASTQ files before upload, click [here](#).

☐ **Antimicrobial Resistance**
Run your samples through our antimicrobial resistance pipeline. Our pipeline supports metagenomics or whole genome data. It only supports Illumina. You can also run the AMR pipeline from within an existing project by selecting previously uploaded mNGS samples. You can check out the AMR pipeline on Github [here](#).

☒ **Viral Consensus Genome**
Run your samples through our Illumina supported pipeline to get viral consensus genomes using your own reference sequence. Pipeline report does not link to Nextclade.

Taxon Name ⓘ Reference Sequence ⓘ Trim Primers — Optional ⓘ
 [SELECT FILE](#) [SELECT FILE](#)

☐ **SARS-CoV-2 Consensus Genome**
Run your samples through our Illumina or Nanopore supported pipelines to get consensus genomes for SARS-CoV-2. Send consensus genomes to Nextclade.

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).

Upload Files

Upload from Your Computer Upload from Basespace

Upload Your Input Files [MORE INFO](#)

2 Files Selected For Upload
Drag and drop your files here, or [click to use a file browser](#).

1 of 1 samples selected. Select samples that you want to upload. [Click to remove unselected samples](#)

Sample Name	Files	Sample_1_Paired_R2.fastq.gz
Sample_1_Paired	Sample_1_Paired_R1.fastq.gz	Sample_1_Paired_R2.fastq.gz

Continue Cancel

Click here to continue to the next page, Upload Metadata

Selected sequence files will be listed here.

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**).

Upload Metadata

This metadata will provide context around your samples and results in CZ ID.

Samples Metadata Review

We require the following metadata to determine how to process your data and display the results: Host Organism, Sample Type, Water Control, Nucleotide Type, Collection Date, Collection Location. [View Full Metadata Dictionary](#).

Available organisms for host subtraction: Human, Mosquito, Tick, Mouse, Cat, Pig, C.elegans, Carp, Chicken, Bee, Salpingoeca rosetta, Bat, Rat, Field Vole, Bank Vole, Rabbit, Water Buffalo, Horse, Taurine Cattle, Turkey, Mosquito Non-Sharded Experimental Small, Mosquito Non-Sharded Experimental Large, Barred Hamlet, Orange Clownfish, Tiger Tail Seahorse, Torafugu, Dog, White Shrimp, Koala, Madagascan Flying Fox, Madagascan Fruit Bat, Madagascan Roussettes, Ascomycetes, Songbird, Cicada, Eukaryote, Woodmouse, Large Japanese Fieldmouse, Soybean, Boechera Stricta, Little Brown Bat, Zebra Fish.

Manual Input CSV Upload

Required metadata fields

Sample Name	Host Organism	Sample Type	Water Control	Nucleotide Type	Collection Date	Collection Location
Nanopore1_no_host_1			No		YYYY-MM-DD	Enter a city, region or count

Continue Cancel

Click here to continue to Review page

Click here to add optional metadata fields from a dropdown menu

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.

Upload Metadata
This metadata will provide context around your samples and results in CZ ID.

1 Samples 2 Metadata 3 Review

We require the following metadata to determine how to process your data and display the results: Host Organism, Sample Type, Water Control, Nucleotide Type, Collection Date, Collection Location. [View Full Metadata Dictionary.](#)

Available organisms for host subtraction: Human, Mosquito, Tick, Mouse, Cat, Pig, C.elegans, Carp, Chicken, Bee, Salpingoeca rosetta, Bat, Rat, Field Vole, Bank Vole, Rabbit, Water Buffalo, Horse, Taurine Cattle, Turkey, Mosquito Non-Sharded Experimental Small, Mosquito Non-Sharded Experimental Large, Barred Hamlet, Orange Clownfish, Tiger Tail Seahorse, Torafugu, Dog, White Shrimp, Koala, Madagascan Flying Fox, Madagascan Fruit Bat, Madagascan Rousettes, Ascomycetes, Songbird, Cicada, European Woodmouse, Large Japanese Fieldmouse, Soybean, Boechera Stricta, Little Brown Bat, Zebra Fish.

Manual Input **CSV Upload**

[View CSV Upload Instructions](#)

Upload your metadata CSV
Drag and drop a file here, or click to use a file browser.

Use this box to upload your completed metadata file

Download Metadata CSV Template

You can download a metadata template to fill in the information. Alternatively, you can create your own metadata file

Continue Cancel

Manual Input **CSV Upload**

[View CSV Upload Instructions](#)

✓ illumina_metadata.csv loaded

Download Metadata CSV Template

Confirm Your Collection Locations

☒ We automatically searched for location matches. Please double check and correct any errors.

Continue Cancel

Click here to continue to the Review page

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).

Review
Uploading 1 samples to Demo_mNGS

Project Info [Edit Project](#)

Demo_mNGS Private Project

Demo samples
8 existing samples in project

Analysis Type Info [Edit Analysis Type](#)

Viral Consensus Genome

Sequencing Platform: Illumina	Taxon Name: Torque teno virus (species)	Reference Sequence: TTV_FR751489.fasta	Trim Primer: None provided	Pipeline Version: 3.4.17
----------------------------------	--	---	-------------------------------	-----------------------------

Sample Info [Edit Samples](#) | [Edit Metadata](#)

Sample Name	Input Files	Host Organism	Water Control	Sample Type	Nucleotide Type	Collection Date	Collection Location
Patient016_CSF_10399_reads_nh	Patient016_CSF_10399_reads_nh_R1.fastq Patient016_CSF_10399_reads_nh_R2.fastq	Human	No	Cerebrospinal Fluid	RNA	2018-10	San Francisco Bay

Host Subtraction: For the 1 sample that you indicated is from a Human Host Organism, we will subtract out reads that align to a Human genome. [Learn more](#)

☒ I agree that the data I am uploading to CZ ID has been lawfully collected and that I have all the necessary consents, permissions, and authorizations needed to collect, share, and export data to CZ ID as outlined in the [Terms of Service](#) and [Privacy Policy](#).

[Start Upload](#)

Click here to start data upload and Viral Consensus Genome pipeline run

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.

Uploading 1 sample to Demo_mNGS

Please stay on this page until upload completes! Closing your device or putting it to sleep will interrupt the upload.

Patient010_CSF_4533_reads_nh_1

Uploaded 5.0 MB of 30.0 MB

Uploads completed!

Patient010_CSF_4533_reads_nh_1 ✓ Sent to pipeline

[Go to Project](#)

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.



Project

Consensus Genomes

Note the analysis status for uploaded samples

Patient010_CSF_4533_reads_nh.1 RUNNING

View Genome Report

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

Click completed sample to view Genome Report

Patient010_CSF_4533_reads_nh.1 COMPLETE

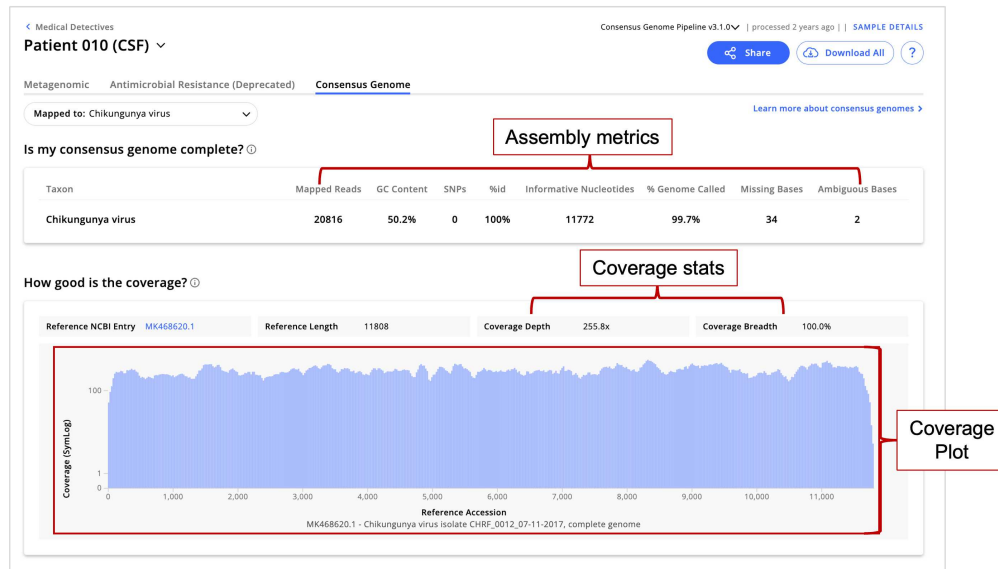
Sample

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.

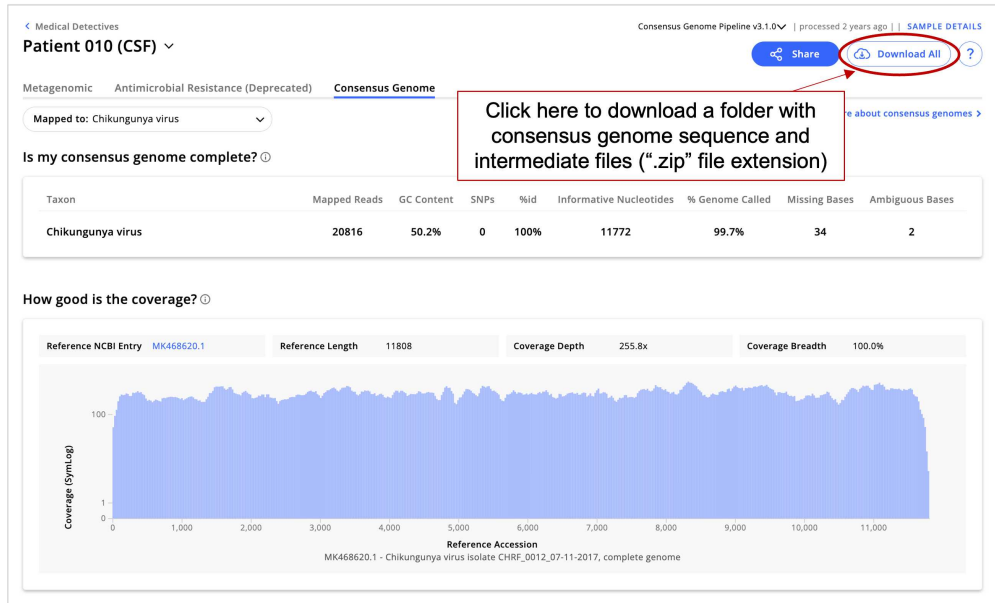
To download a folder with consensus genome data:

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.

The screenshot shows the 'Consensus Genome Pipeline v3.1.0' interface. At the top, the sample name 'Patient 010 (CSF)' is displayed. Below it, the 'Consensus Genome' tab is selected. A dropdown menu is open, showing a list of genomes to select from. A red box highlights the dropdown menu with the text 'Select consensus genome from dropdown menu'. Below the dropdown, a table displays genome statistics for three entries: Chikungunya virus, Primate tetraparvovirus 1, and Chikungunya virus. The table has columns for GC Content, SNPs, %id, Informative Nucleotides, % Genome Called, Missing Bases, and Ambiguous Bases. Below the table, a section titled 'How good is the coverage?' shows a coverage plot for the selected genome, KM390024.1. The plot shows coverage depth on a log scale (1 to 100) across the reference accession (0 to 5,000). The plot indicates a coverage depth of 371.6x and a coverage breadth of 100.0%.

	GC Content	SNPs	%id	Informative Nucleotides	% Genome Called	Missing Bases	Ambiguous Bases
Chikungunya virus KY751908.1 - Chikungunya virus isolate IN16C1, complete genome	44.6%	41	99.2%	5265	99.9%	5	1
Primate tetraparvovirus 1 KM390024.1 - Human parvovirus 4 isolate T609, complete genome							
Chikungunya virus MK468620.1 - Chikungunya virus isolate CHRF_0012_07-11-2017, complete genome							

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.

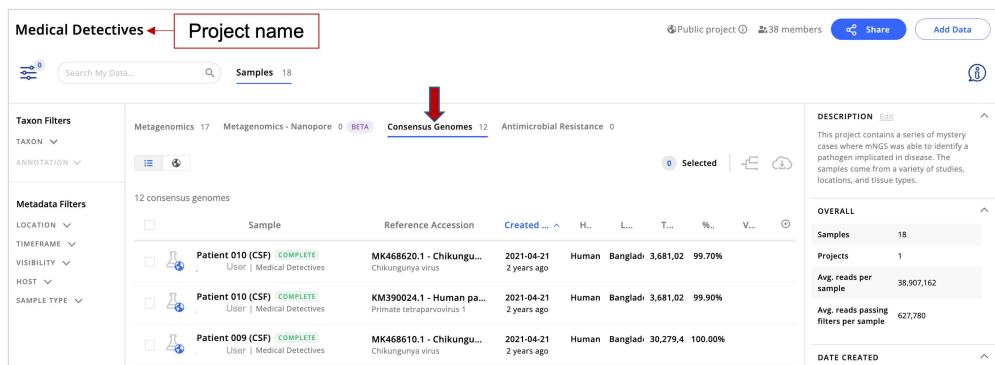


9.2 Downloading Data through Project Pages

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.

The screenshot shows the 'Consensus Genomes' section of the protocols.io interface. A table lists 12 consensus genomes. Two genomes are selected, indicated by blue checkmarks in the 'Selected' column. A red box highlights the 'Download' icon (a cloud with a download arrow) in the top right corner of the table. Another red box highlights the 'Select genomes' button in the left sidebar. A third red box highlights the 'Download icon will be active after selecting genomes' text.

Sample	Reference Accession	Created	Host	Location	Coverage
Patient 010 (CSF) - COMPLETE	MK468620.1 - Chikungunya virus	2021-04-21 2 years ago	Human	Bangladi	3,681,02 99.70%
Patient 010 (CSF) - COMPLETE	KM390024.1 - Human pa...	2021-04-21 2 years ago	Human	Bangladi	3,681,02 99.90%
Patient 009 (CSF) - COMPLETE	MK468610.1 - Chikungunya virus	2021-04-21 2 years ago	Human	Bangladi	30,279.4 100.00%

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

The first screenshot shows the 'Select a Download Type' dialog box with four options: 'Consensus Genome (consensus.fa)', 'Sample Metadata (sample_metadata.csv)', 'Consensus Genome Overview (.csv)', and 'Intermediate Output Files'. The 'Consensus Genome Overview (.csv)' option is selected. The second screenshot shows the same dialog box with the 'Consensus Genome Overview (.csv)' option selected and the 'Include sample metadata in this table' checkbox checked. A red circle highlights the 'Start Generating Download' button in the second screenshot.

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

The screenshot shows the CZ ID interface for the 'Medical Detectives' project. The top navigation bar includes 'My Data', 'Public', 'Upload', and a 'User' dropdown menu. The 'User' menu is open, showing options like 'Downloads', 'Help Center', 'Contact Us', 'Terms of Use', 'Privacy Notice', and 'Logout'. The 'Downloads' option is circled in red. Below the navigation bar, the project details for 'Medical Detectives' are shown, including a search bar, filters, and a table of samples. An arrow points from the 'Downloads' menu item to the 'Downloads' page below. The 'Downloads' page has a 'Back' link and a table of generated files. The table has columns for 'Download', 'Date', 'Count', and 'File Size'. Two files are listed: 'Consensus Genome Overview' and 'Reads (Non-host)'. The 'Download' column for the first file has a 'DOWNLOAD FILE' link circled in red. A red box labeled 'File generation status' points to the 'COMPLETE' status of the first file. Another red box labeled 'Click here to download files' points to the 'DOWNLOAD FILE' link for the first file.

Medical Detectives

Public project 38 members

Search My Data... Samples 18

Taxon Filters

TAXON Metagenomics 17 Metagenomics - Nanopore 0 BETA Consensus Genomes 12 Antimicrobial Resistance 0 BETA

ANNOTATION 12 consensus genomes

Metadata Filters

LOCATION TIMEFRAME VISIBILITY HOST SAMPLE TYPE

DESCRIPTION

This project contains a series of mystery cases where mNGS was able to identify a pathogen implicated in disease. The samples come from a variety of studies, locations, and tissue types.

OVERALL

Samples 18
Projects 1
Avg. reads per sample 38,907,162
Avg. reads passing filters per sample 627,780

Downloads

Back

Download	Date	Count	File Size
Consensus Genome Overview COMPLETE	2023-03-01 5 minutes ago	2 Consensus Genomes	590 Bytes
Reads (Non-host) COMPLETE	2023-02-24 5 days ago	1 Metagenomic	89.5 MB

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