

Viral Consensus  
Genome Assembly

JUL 26, 2023

## OPEN ACCESS

**DOI:**

[dx.doi.org/10.17504/protocols.io.bp2l69ojklqe/v1](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)  
<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:** Jun 05, 2023

**Last Modified:** Jul 26, 2023

**PROTOCOL integer ID:**  
82896

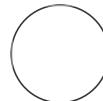
**Keywords:** CZ ID, virus  
consensus genome, viral  
consensus genome assembly

# CZ ID Workflow for Assembling Viral Consensus Genomes

CZ ID

Karyna Rosario Cora<sup>1</sup>, Elizabeth Fahsbender<sup>1</sup>, Team<sup>1</sup>

<sup>1</sup>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.

# Upload Data

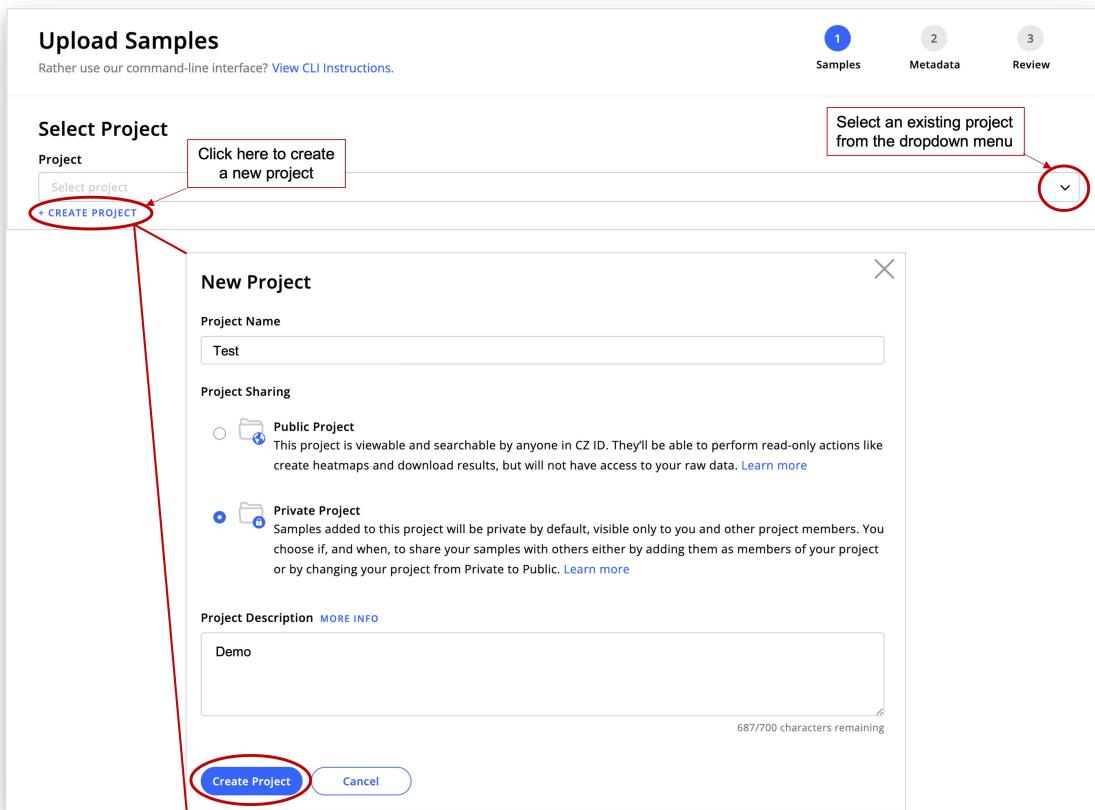
- 1 [Log in](#) to your CZ ID account.
- 2 Navigate to the Upload page from the [Discovery page](#) by clicking the "Upload" link next to your username. Note that the upload process is divided into three general sections, including Samples, Metadata, and Review pages.

The screenshot shows the 'Upload Samples' page on the CZ ID website. At the top, there's a navigation bar with 'My Data', 'Public', and 'Upload' (which is circled in red). Below the navigation, there are three numbered circles labeled 'Samples', 'Metadata', and 'Review'. A red box surrounds these three sections, with an arrow pointing to the 'Upload' section. The main form has sections for 'Select Project' (with 'Select project' and '+ CREATE PROJECT' buttons), 'Analysis Type' (with options for Metagenomics, Antimicrobial Resistance, Viral Consensus Genome, and SARS-CoV-2 Consensus Genome), and 'Upload Files' (with 'Upload from Your Computer' and 'Upload from Basespace' buttons, and a 'Drag and drop your files here...' input field). At the bottom are 'Continue' and 'Cancel' buttons.

- 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.

#### 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 Taxon Name](#) [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 Taxon Name](#) [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
<input checked="" type="checkbox"/> Sample_1_Paired	Sample_1_Paired_R1.fastq.gz    Sample_1_Paired_R2.fastq.gz

**Selected sequence files will be listed here.**

**Click here to continue to the next page, Upload Metadata**

**Continue**   **Cancel**

#### 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.

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 Rousetttes, Ascomycetes, Songbird, Cicada, European Woodmouse, Large Japanese Fieldmouse, Soybean, Boechera Stricta, Little Brown Bat, Zebra Fish.

**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

**Click here to add optional metadata fields from a dropdown menu**

**Continue**   **Cancel**

*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.*

The screenshot shows a two-step process for uploading metadata:

- Step 1: CSV Upload**
  - Header: "Upload Metadata". Subtext: "This metadata will provide context around your samples and results in CZ ID."
  - Section: "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 Rousetttes, Ascomycetes, Songbird, Cicada, European Woodmouse, Large Japanese Fieldmouse, Soybean, Boechera Stricta, Little Brown Bat, Zebra Fish."
  - Buttons: "Manual Input" (disabled), "CSV Upload" (highlighted with a red arrow pointing to it).
  - Link: "View CSV Upload Instructions".
  - Input area: "Upload your metadata CSV" (with a dashed box) and "Download Metadata CSV Template".
  - Buttons: "Continue" and "Cancel".
  - Text: "Use this box to upload your completed metadata file".
  - Text: "You can download a metadata template to fill in the information. Alternatively, you can create your own metadata file".
- Step 2: Review**
  - Header: "Manual Input" (disabled), "CSV Upload" (highlighted with a blue arrow pointing from the previous step).
  - Text: "✓ illumina\_metadata.csv loaded".
  - Link: "Download Metadata CSV Template".
  - Section: "Confirm Your Collection Locations".
    - Text: "We automatically searched for location matches. Please double check and correct any errors."
    - Checkboxes: "Continue" (highlighted with a red circle) and "Cancel".
    - Text: "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 [View](#)  
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](#)

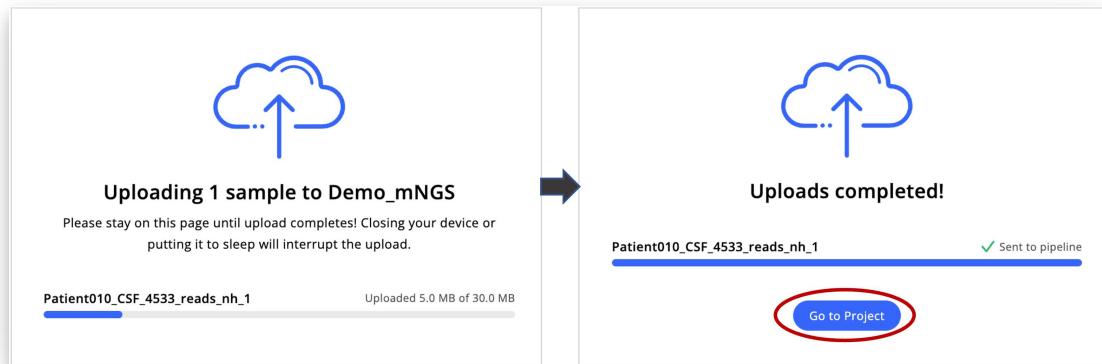
Sample Name	Input Files	Host Organism	Water Control	Sample Type	Nucleotide Type	Collection Date	Collection Location
Patient016_CSF_10399_reads_nh	Patient016_CSF_10399_reads_nh_R1.fastq Patient016_CSF_10399_reads_nh_R2.fastq	Human	No	Cerebrospinal Fluid	RNA	2018-10	San Francisco Bay

ⓘ 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.



## 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.

The screenshot shows the iDcz ID Project page for 'Demo\_mNGS'. The 'Consensus Genomes' tab is selected, showing one entry. A red arrow points to the sample name 'Patient010\_CSF\_4533\_reads\_nh\_1' which is labeled 'RUNNING'. A callout box with the text 'Note the analysis status for uploaded samples' is positioned over this sample row.

## View Genome Report

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

The screenshot shows the iDcz ID Project page for 'Demo\_mNGS'. The 'Consensus Genomes' tab is selected, showing one entry. A red arrow points to the sample name 'Patient010\_CSF\_4533\_reads\_nh\_1' which is labeled 'COMPLETE'. A callout box with the text 'Click completed sample to view Genome Report' is positioned over this sample row. Below the table, a large downward arrow indicates the transition to the next step.

Consensus Genome Pipeline v3.4.17 | processed 32 minutes ago | SAMPLE DETAILS  
Share Download All ? ...

Learn more about consensus genomes >

Is my consensus genome complete? ○

Taxon	Mapped Reads	GC Content	SNPs	%id	Informative Nucleotides	% Genome Called	Missing Bases	Ambiguous Bases
Chikungunya virus (species)	19270	50.1%	0	100%	11801	99.9%	7	0

How good is the coverage? ○

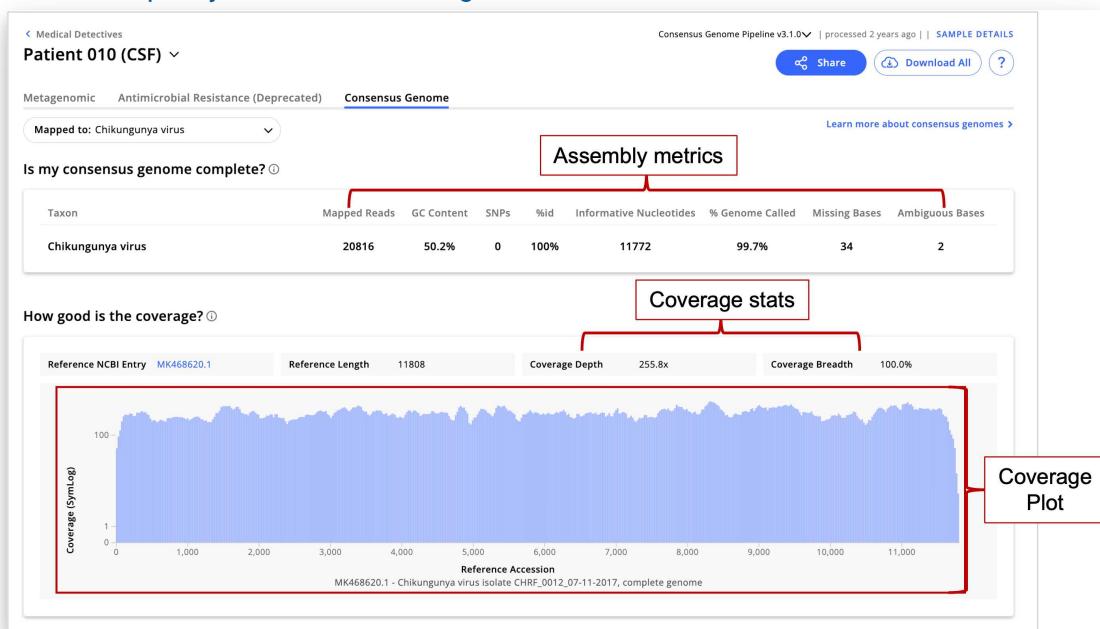
NCBI Reference: MK468620.1 Reference Length: 11808 Coverage Depth: 239.7x Coverage Breadth: 100.0%  
Reference Sequence: MK468620.1 - Chikungunya virus isolate CHRF\_0012\_07-11-2017, complete genome

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 [assess 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)

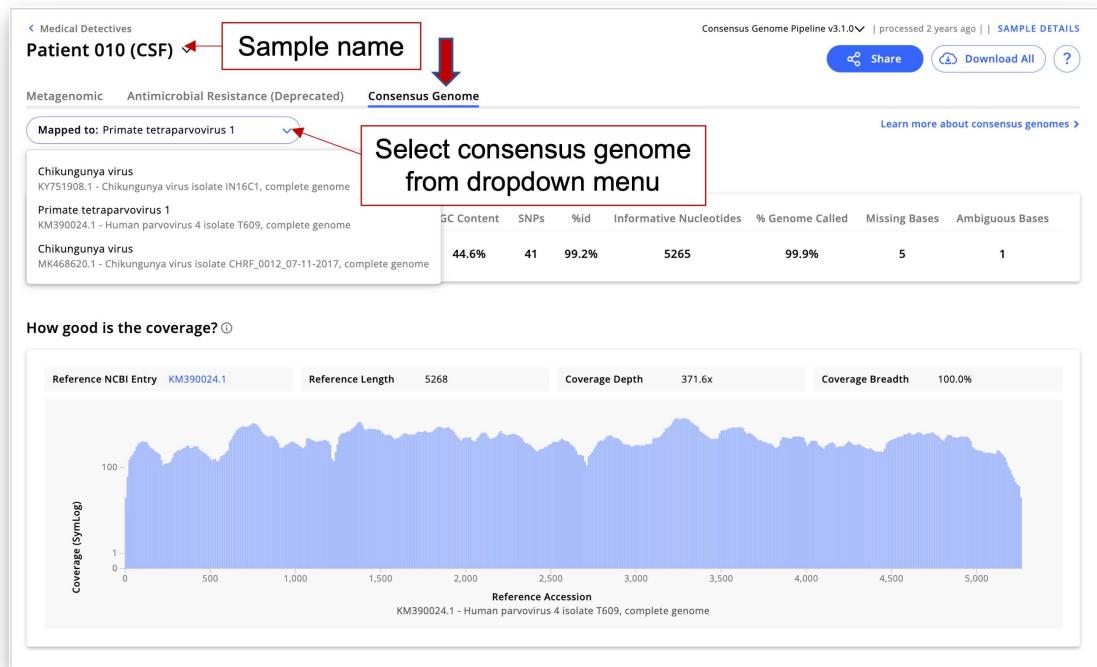
format) and intermediate files produced throughout the pipeline, through Genome Report and Project pages.

## 9.1 Downloading Data through Genome Report Page

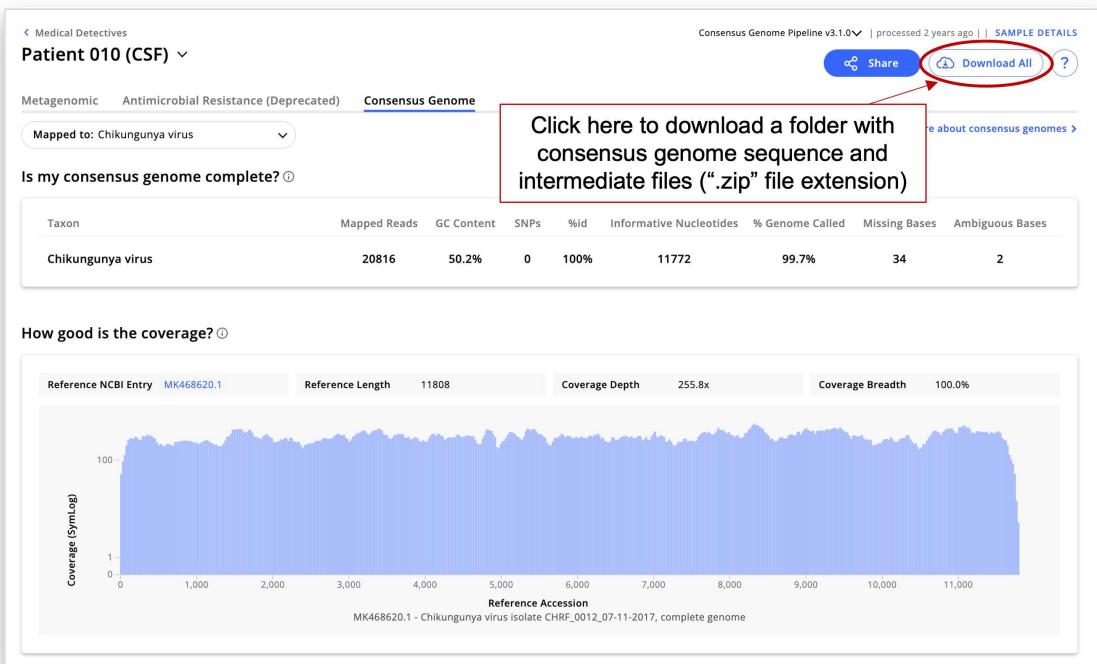
You can download data for a single consensus genome from the Genome Report page. Here you can download the consensus genome sequence and generated [intermediate files](#) in a single folder.

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.

Medical Detectives Project name

Consensus Genomes

Sample	Reference Accession	Created ...	H...	L...	T...	%...	V...
Patient 010 (CSF) - Chikungunya virus	MK468620.1 - Chikungunya virus	2021-04-21 2 years ago	Human	Bangladeshi	3,681,02	99.70%	
Patient 010 (CSF) - Human parvovirus 1	KM390024.1 - Human parvovirus 1	2021-04-21 2 years ago	Human	Bangladeshi	3,681,02	99.90%	
Patient 009 (CSF) - Chikungunya virus	MK468610.1 - Chikungunya virus	2021-04-21 2 years ago	Human	Bangladeshi	30,279,4	100.00%	

**DESCRIPTION** Edit

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

DATE CREATED

2. Select genomes to download and click Download icon.

Taxon Filters  
Metagenomics 17 Metagenomics - Nanopore 0 BETA Consensus Genomes 12 Antimicrobial Resistance 0

Annotation  
Metadata Filters  
Location  
Timeline  
Visibility  
Host  
Sample Type

12 consensus genomes

Sample	Reference Accession	Created ...	H...	L...	T...	%...	V...
Patient 010 (CSF) COMPLETE User   Medical Detectives	MK468620.1 - Chikungu...	2021-04-21 2 years ago	Human	Bangl...	3,681,02	99.70%	
Patient 010 (CSF) COMPLETE User   Medical Detectives	KM390024.1 - Human pa...	2021-04-21 2 years ago	Human	Bangl...	3,681,02	99.90%	
Patient 009 (CSF) COMPLETE User   Medical Detectives	MK468610.1 - Chikungu...	2021-04-21 2 years ago	Human	Bangl...	30,279,4	100.00%	

DESCRIPTION Edit  
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

DATE CREATED

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

Select a Download Type  
2 consensus genomes selected

- Consensus Genome (consensus.fa)  
Download multiple consensus genomes as separate or a single file.
- Sample Metadata (sample\_metadata.csv)  
User-uploaded metadata, including sample collection location, collection date, sample type
- Consensus Genome Overview (.csv)  
Consensus Genome QC metrics (e.g. % genome coverage, mapped read #, SNP #) and other summary statistics
- Intermediate Output Files  
Intermediate output files including BAM files, coverage plots, QUAST report and more. [Learn More](#)

Start Generating Download  
Downloads for larger files can take multiple hours to generate.

Select a Download Type  
2 consensus genomes selected

- Consensus Genome (consensus.fa)  
Download multiple consensus genomes as separate or a single file.
- Sample Metadata (sample\_metadata.csv)  
User-uploaded metadata, including sample collection location, collection date, sample type
- Consensus Genome Overview (.csv)  
Consensus Genome QC metrics (e.g. % genome coverage, mapped read #, SNP #) and other summary statistics  
 Include sample metadata in this table
- Intermediate Output Files  
Intermediate output files including BAM files, coverage plots, QUAST report and more. [Learn More](#)

Start Generating Download  
Downloads for larger files can take multiple hours to generate.

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

My Data Public Upload User  
Public project 38 members Share Help Center Contact Us Terms of Use Privacy Notice Logout

Medical Detectives

Search My Data... Samples 18

Taxon Filters  
Metagenomics 17 Metagenomics - Nanopore 0 BETA Consensus Genomes 12 Antimicrobial Resistance 0 BETA

Annotation  
Metadata Filters  
Location  
Timeline  
Visibility  
Host  
Sample Type

12 consensus genomes

Sample	Reference Accession	Created ...	H...	L...	T...	%...	V...
Patient 010 (CSF) COMPLETE User   Medical Detectives	MK468620.1 - Chikungu...	2021-04-21 2 years ago	Human	Bangl...	3,681,02	99.70%	
Patient 010 (CSF) COMPLETE User   Medical Detectives	KM390024.1 - Human pa...	2021-04-21 2 years ago	Human	Bangl...	3,681,02	99.90%	

DESCRIPTION Edit  
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

File generation status	Date	Count	File Size	Action
COMPLETE	2023-03-01 5 minutes ago	2 Consensus Genomes	590 Bytes	<a href="#">DOWNLOAD FILE</a>
COMPLETE	2023-02-24 5 days ago	1 Metagenomic	89.5 MB	<a href="#">DOWNLOAD FILE</a>

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