PHA4GE Wastewater Genomic Surveillance Contextual Data Specification - Worked Examples

Created February 2024

Background

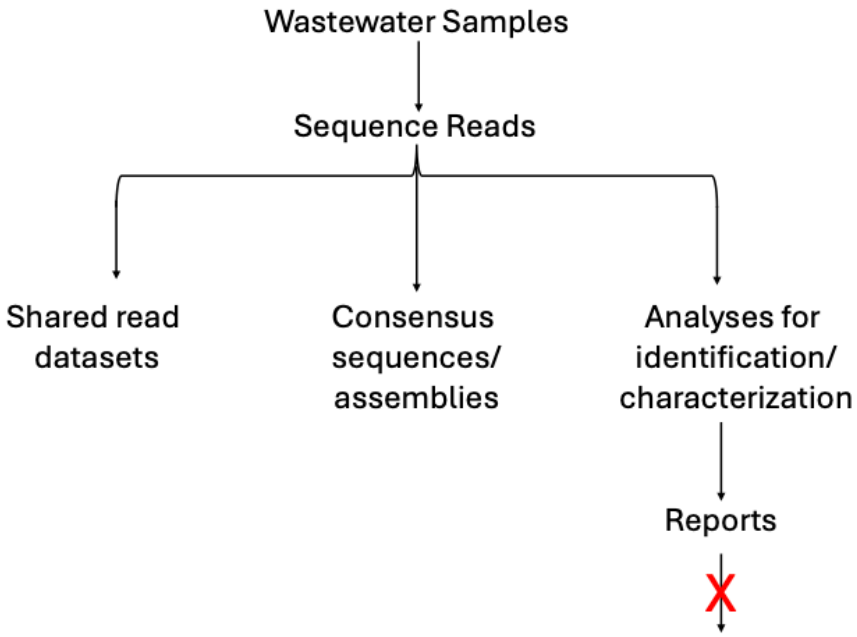
Wastewater-based genomics can be used to answer a wide variety of public health questions and may involve different types of samples, pathogen targets, methodologies, and partners associated with different contextual data collection and sequence data generation, as well as harmonization, integration, analysis and storage processes. Data can also be generated and shared publicly, within trusted networks, or retained within organizations for internal use only. Wastewater genomics is amenable to different sequencing assay types (amplicon, single isolate, metagenomics), and different datasets may be appropriate for different analyses depending on the scope of sample collection, characteristics of samples, quality and breadth of sequence data, etc. Parts of contextual data records may be contributed by different partners and so, at any time, may not be complete - and it is important to remember that not all partners are responsible for all aspects of its collection (e.g. sample collectors in the field are not responsible for downstream processing and sequencing attributes). Datasets can also be characterized according to different pre-analytical and post-analytical uses (Figure 1). Methods and tools can and should be captured as part of contextual data for improving reproducibility and interpretation of different datasets and analysis, e.g.

- raw reads that were quality trimmed with human host data removed - agnostic to pathogen or downstream analysis (processing and dehosting software should be captured)
- reads aligned to a reference taxonomic database to identify the presence of particular pathogens (taxonomic database and mapping software information should be captured),
- amplicon sequencing with determination of relative lineage abundances from mixed samples (amplicon scheme and analysis software information should be captured),
- assembled sequences characterized for attributing antimicrobial resistance determinants (assembly software and AMR detection software and databases should be captured).

As a result, wastewater contextual data records can vary - and can be as simple or as complex as is necessary according to those involved in producing those records.

Figure 1: Specification Scope. The Wastewater Contextual Data specification includes sample metadata and methods for generating sequence reads from samples. Raw and processed read datasets can be

stored and/or shared for different applications and studies (e.g. submitted to public repositories). Reads can also be used to generate consensus sequences (e.g. PrimalSeq-based sequencing) and assemblies (e.g. from isolates cultured from wastewater). Reads can also be mapped to reference databases (taxonomic, gene, marker, etc) in analyses producing reports describing different kinds of genomics-based identification/detection/characterization results. The specification provides fields for prioritized methods capture (e.g. software, databases), certain types of results (e.g. SARS-CoV-2 lineage calls), and report filenames. Results requiring more extensive organization should be structured using additional specifications (e.g. summarizing long lists of identified AMR genotypes).



The following worked examples represent a set of mock scenarios and suggested data structures derived from the PHA4GE Wastewater Contextual Data Specification that have been used to record examples of wastewater genomic surveillance contextual data. The intention of these worked examples is to demonstrate how the specification can be used in different ways to harmonize and future-proof a wide variety of data across different use cases - highlighting key templates, fields and terms. The worked examples provided below represent mock data based on real situations (the details provided DO NOT represent actual public health data, events or processes, and are meant solely for training purposes). Fields considered “required” in the specification are highlighted in yellow, while “recommended” and “optional” fields are highlighted in purple and white respectively. Recommended templates for each of the scenarios are also provided.

How to Use the PHA4GE Wastewater Contextual Data Specification: Scenarios and Recommended Data Structures

Scenario 1: Partial record from a sample collector

A technician collected a wastewater sample (sample ID CVW-6758900) at a municipal wastewater treatment plant in the UK on February 28 2023 as part of the UKHSA's Environmental Wastewater Monitoring programme. The targets of interest were environmental pollutants, however the sample was later used for pathogen surveillance. The sample is awaiting sequencing. The technician provides a partial contextual data record describing sample collection to the sequencing lab, who will provide further details at a later date. The record submitted to the sequencing lab is provided below. This record highlights the minimal "required" contextual data in the template (12 fields), and highlights what to do when not all the information can be supplied.

Template: Pathogen-Agnostic

specimen collector sample ID: CVW-6758900
sample collected by: UK Health Security Agency
geo_loc_name (country): United Kingdom [GAZ:00002637]
geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Not Provided [GENEPIO:0001668]
purpose of sampling: Wastewater chemical surveillance [GENEPIO:0100870]
sample collection date: 2023-02-28
environmental site: Wastewater treatment plant [ENVO:00002272]
environmental material: Wastewater [ENVO:00002001]
purpose of sequencing: Not Provided [GENEPIO:0001668]
sequenced by: Not Provided [GENEPIO:0001668]
sequenced by contact name: Not Provided [GENEPIO:0001668]
sequenced by contact email: Not Provided [GENEPIO:0001668]
sequencing instrument: Not Provided [GENEPIO:0001668]

Scenario 2: Sequencing cultured organisms isolated from wastewater

A wastewater sample (WD-1234-i9v) was collected as part of a research project examining the utility of wastewater surveillance for identifying and monitoring cholera in the community. The sample was collected on January 3 2022 from a school latrine (sampling site ID AAA123). Cholera isolates were cultured from the sample (microbiological method as per Mtonga et al, 2018; doi: 10.1371/journal.pntd.0007642), and one of the isolates was sequenced. The sample

was sequenced by the Zambia Ministry of Health (project lead: Cheelo Mtonga; mtongam@moh.zambia.org). The library was prepared using an Nextera XT DNA Library Preparation Kit and sequenced using an Illumina MiSeq. Raw reads were quality filtered and primer sequences trimmed using Trimmomatic. Paired-end reads were assembled *de novo* to construct a draft genome using the SPADES v.3.11.1 software. The quality of de novo assemblies was assessed using Quast (v.4.5) and the data passed the research project's quality control processes. The metadata, sequence reads and assembly were then uploaded to NCBI (BioProject PRJNA608678). The associated contextual data record is provided below. This record highlights identifier and attribution tracking, sampling strategy capture, structuring of sample metadata, high-level quality control annotations, as well as microbiological, assembly and sequencing methods tagging.

Template: Pathogen-Agnostic

specimen collector sample ID: WD-1234-i9v
BioProject accession: PRJNA608678
sampling site ID: AAA123
sample collected by: Zambia Ministry of Health
geo_loc_name (country): Zambia [GAZ:00001107]
geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Vibrio cholerae [NCBITaxon:666]
purpose of sampling: Research [GENEPIO:0100003]
sample collection date: 2022-01-03
environmental site: School [ENVO:03501130]
environmental material: Wastewater [ENVO:00002001]
wastewater system type: Latrine [ENVO:01000519]
microbiological method: doi: 10.1371/journal.pntd.0007642
purpose of sequencing: Research [GENEPIO:0100003]
sequencing assay type: Whole genome sequencing assay [OBI:0002117]
library preparation kit: Nextera XT DNA Library Preparation Kit
sequenced by: Zambia Ministry of Health
sequenced by contact name: Cheelo Mtonga
sequenced by contact email: mtongam@moh.zambia.org
sequencing instrument: Illumina MiSeq [OBI:0002003]
raw sequence data processing method: Trimmomatic
sequence assembly software name: SPAdes
sequence assembly software version: 3.11.1
quality control method name: Quast
quality control method version: 4.5
quality control determination: Sequence passed quality control [GENEPIO:0100563]
sequence submitted by: Cheelo Mtonga
sequence submitter contact email: mtongam@moh.zambia.org

Scenario 3: Identifying single pathogens using amplicon approaches

The Global Viral Disease Eradication Initiative uses wastewater samples from different parts of the world to identify and characterize pathogens causing priority diseases such as measles, polio and hepatitis. A wastewater sample (CAM-LAG-00123) was collected from a wastewater treatment lagoon in Cambodia on February 15 2023 as part of a baseline surveillance assessment. The sample was subdivided and the different subsamples were used as technical replicates during method development and optimization. A sequencing library was prepared for one of the replicates (subsample CAM-LAG-00123-3b) using a proprietary enrichment kit (GVDE enrichment kit) and a primer amplicon panel specific for 20 different diseases, including Poliovirus types 1, 2 and 3 (GVDE Viral Surveillance Panel). The library was sequenced using an Illumina NovaSeq 6000 (sequencing lab contact: Dr. Maya Lee; mlee@gvde.org). The reads were filtered, trimmed, dehosted, and a Poliovirus 1 genome was de novo assembled using a suite of tools available in the RAMPART (v1.2.0) platform. The associated contextual data record is provided below. This record highlights experimental replicate, amplicon scheme and dehosting software and assembly annotations.

Template: Pathogen-Agnostic

specimen collector sample ID: CAM-LAG-00123
specimen collector subsample ID: CAM-LAG-00123-3b
sample collected by: The Global Viral Disease Eradication Initiative
geo_loc_name (country): Cambodia [GAZ:00006888]
geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Poliovirus 1 [NCBITaxon:12080]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
scale of sampling: Community-level surveillance [GENEPIO:0100874]
sample collection date: 2023-02-15
environmental site: Waste stabilization pond (lagoon) [ENVO:03600076]
environmental material: Wastewater [ENVO:00002001]
wastewater system type: Wastewater stabilization pond [ENVO:03600076]
experimental specimen role type: Technical replicate [EFO:0002090]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Amplicon sequencing assay [OBI:0002767]
library preparation kit: GVDE enrichment kit
amplicon pcr primer scheme: GVDE Viral Surveillance Panel
sequenced by: The Global Viral Disease Eradication Initiative
sequenced by contact name: Maya Lee
sequenced by contact email: mlee@gvde.org
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
raw sequence data processing method: RAMPART 1.2.0
dehosting method: RAMPART 1.2.0
sequence assembly software name: RAMPART
sequence assembly software version: 1.2.0

Scenario 4: Identifying single pathogens using metagenomic approaches (alternative to Scenario 3)

The Global Viral Disease Eradication Initiative uses wastewater samples collected in different parts of the world to identify and characterize pathogens causing priority diseases such as measles, polio and hepatitis. A wastewater sample (CAM-LAG-00123) was collected from a wastewater treatment lagoon in Cambodia on February 15 2023 as part of a baseline surveillance assessment. The sample was subdivided and the different subsamples were used as technical replicates during method development and optimization. A sequencing library was prepared for one of the replicates (subsample CAM-LAG-00123-4a) using a MagMAX Wastewater Ultra Nucleic Acid Isolation Kit, with Virus Enrichment. The library was sequenced using an Illumina NovaSeq 6000 (sequencing lab contact: Dr. Maya Lee; mlee@gvde.org) and the reads were filtered, trimmed, dehosted using a suite of tools available in the RAMPART (v1.2.0) platform. Total filtered and dehosted reads were mapped to a custom reference taxonomic database (GVDEdb 3.4.5) and Poliovirus 1 was identified with 85x coverage across 90% of the Poliovirus 1 reference genome. The associated contextual data record is provided below. This record highlights alternative methods to identifying pathogens in complex samples by capturing “organism” information from taxonomic analysis.

Template: Pathogen-Agnostic

specimen collector sample ID: CAM-LAG-00123
specimen collector subsample ID: CAM-LAG-00123-4a
sample collected by: The Global Viral Disease Eradication Initiative
geo_loc_name (country): Cambodia [GAZ:00006888]
geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Poliovirus 1 [NCBITaxon:12080]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
scale of sampling: Community-level surveillance [GENEPIO:0100874]
sample collection date: 2023-02-15
environmental site: Waste stabilization pond (lagoon) [ENVO:03600076]
environmental material: Wastewater [ENVO:00002001]
wastewater system type: Wastewater stabilization pond [ENVO:03600076]
experimental specimen role type: Technical replicate [EFO:0002090]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Whole virome sequencing assay [OBI:0002768]
library preparation kit: MagMAX Wastewater Ultra Nucleic Acid Isolation Kit, with Virus Enrichment
sequenced by: The Global Viral Disease Eradication Initiative
sequenced by contact name: Maya Lee

sequenced by contact email: mlee@gvde.org
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
raw sequence data processing method: RAMPART 1.2.0
dehosting method: RAMPART 1.2.0
read mapping software name: Bowtie2
read mapping software version: 2.5.3
taxonomic reference database name: GVDEdb
taxonomic reference database version: 3.4.5

Scenario 5: Characterizing AMR in wastewater samples

A global sewage project collects samples from a wide variety of locations around the world in order to establish baselines for antimicrobial resistance (i.e. prevalence, distribution, types of resistance). The national public health laboratory of the UAE collected a wastewater sample from an airport sewer system in Dubai on June 6 2023 (sample UAErt3-478-0091). The sample was sequenced using an Oxford Nanopore GridION instrument, and the reads are quality filtered and trimmed using BBduk2 v1.5. To assign resistance genes to pathogen species, contigs were assembled using metaSPAdes (v3.13.0) and then characterized via mapping to the CARD resistance database (v3.2.9) using the Resistance Gene Identifier software (v6.0.3). A CARD report summarizing results is stored in the system for longitudinal comparisons (filename: 20230606_XYB-123_results.txt). The associated contextual data record is provided below. This record highlights alternative methods to identifying pathogens and AMR determinants in complex samples by capturing information for contig assembly, as well as taxonomic analysis, and AMR detection software and reference databases information.

Template: AMR

specimen collector sample ID: UAErt3-478-0091
sample collected by: The National Public Health Laboratory of the United Arab Emirates
geo_loc_name (country): United Arab Emirates [GAZ:00005282]
geo_loc_name (state/province/territory): Emirate of Dubai
organism: Not Applicable [GENEPIO:0001619]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
scale of sampling: Institution-level surveillance [GENEPIO:0100875]
sample collection date: 2023-06-06
environmental site: Airport [ENVO:03501122]
environmental material: Wastewater [ENVO:00002001]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Whole metagenome sequencing assay [OBI:0002623]
sequenced by: The National Public Health Laboratory of the United Arab Emirates
sequenced by contact name: Not Provided [GENEPIO:0001668]
sequenced by contact email: mlee@gvde.org
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
number of total reads: 21300465

minimum post-trimming read length: 150
number of contigs: 1500201
raw sequence data processing method: RAMPART 1.2.0
dehosting method: RAMPART 1.2.0
sequence assembly software name: Bowtie2
sequence assembly software version: 2.5.3
AMR analysis software name: Resistance Gene Identifier
AMR analysis software version: 6.0.3
AMR reference database name: Comprehensive Antibiotic Resistance Database (CARD)
AMR reference database version: 3.2.9
AMR analysis report filename: 20230606_XYB-123_results.txt

Scenario 6: SARS-CoV-2 surveillance (rich contextual data)

Untreated, fast moving, wastewater is continuously collected in a municipal sewer system starting on Nov 1 2023 for 72hrs. The sewer system, which collects rainwater as well as household and institutional waste, is part of a routine surveillance program for tracking community-level SARS-CoV-2 variants (sewer site ID WWSC2-ABC-b) in order to establish baseline norms. The sewer is located near a hospital and the hospital's effluent is piped into the sewer system. Five Moore swabs from the site of collection are pooled (sample ID BW-WW-12345). It rained the day before sample collection (5cm of rain). The wastewater catchment area serves approx 800 000 people in a suburban area (Mississauga, Ontario, Canada). The ambient air temperature at the time of collection was 15 degrees Celsius. The water was 8 degrees Celsius at the time of collection, and 3 degrees Celsius when it was received by the sequencing lab. The instantaneous flow rate is 3 cubic meter per second (m^3/s), with 8% total suspended solids. The sample was collected by the Region of Peel regional authority, and sequenced by the Public Health Ontario provincial health laboratory (contact: Johnny Bloggs; jbloggs@provlab.ca). A watershed shapefile delineating the geographical coordinates covered by the sewer system is available. The presence of SARS-CoV-2 was first detected using qPCR (N1 gene, Ct value of 22). The amplicon-based sample was sequenced on an Illumina MiSeq on Jan 18 2024 using the ARTIC V5 400bp primer scheme (artic-v5.3.2_400), and consensus sequences were generated using ViralRecon software v1.23 and lineage assignments were performed using pUShER (v1.2.6). The rich contextual data record for the sequence is provided below. This record is for the public health laboratory's use only, and many details were removed when sharing data according to organization-specific data sharing policies. The associated contextual data record is provided below. This record highlights how rich contextual data can be captured using the specification - including catchment details such as geographical coordinates and population ranges, activity upstream of sampling that may affect results, how to record longitudinal sampling events, capture of environmental conditions and measurements, associated laboratory testing results (Ct values), and lineage designations.

Template: SARS-CoV-2

specimen collector sample ID: BW-WW-12345

sampling site ID: WWSC2-ABC-b
sample collected by: Region of Peel Regional Authority
geo_loc_name (country): Canada [GAZ:00002560]
geo_loc_name (state/province/territory): Ontario [GAZ:00002563]
geo_loc_name (city): Mississauga
watershed shapefile availability: Watershed shapefile available
organism: Severe acute respiratory syndrome coronavirus 2 [NCBITaxon:2697049]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
presamping activity: Healthcare activity [NCIT:C16205]
presampling activity details: hospital effluent piped into sewer system
sample collection date: 2023-11-01
sample collection time duration value: 72
sample collection time duration unit: Hour [UO:0000032]
scale of sampling: Community-level surveillance [GENEPIO:0100874]
specimen processing: Pooling specimens [OBI:0600016]
specimen processing details: 5 Moore swabs pooled from same sewer
environmental site: Sewer [ENVO:01000924]
environmental material: Wastewater [ENVO:00002001]
environmental material properties: Untreated [GENEPIO:0101009]; Fast fluid flow rate [GENEPIO:0101006]
wastewater system type: Combined sewer system [ENVO:03501453]
collection device: Moore swab [GENEPIO:0100949]
collection method: Passive composite sampling [GENEPIO:0100955]
populated area type: Suburban [GSSO:011077]
water catchment area human population measurement value: 800 000
water catchment area human population bin: 100,000 - 1,000,000 people
presampling weather conditions: Rain [ENVO:01001564]
precipitation measurement value: 5
precipitation measurement unit: centimeter (cm) [UO:0000015]
ambient temperature measurement value: 15
ambient temperature measurement unit: degree Celsius (C) [UO:0000027]
instantaneous flow rate measurement: 3
cubic meter per second (m^3/s): cubic meter per second (m^3/s)
total suspended solids (TSS) measurement value: 8
total suspended solids (TSS) measurement unit: Percent (%) [UO:0000187]
sample temperature value (at collection): 8
sample temperature unit (at collection): degree Celsius (C) [UO:0000027]
sample temperature value (when received): 3
sample temperature unit (when received): degree Celsius (C) [UO:0000027]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Amplicon sequencing assay [OBI:0002767]
sequencing date: 2024-01-18
sequenced_by: Public Health Ontario
sequenced_by_contact_name: Johnny Bloggs

sequenced_by_contact_email: jbloggs@provlab.ca
sequencing instrument: Illumina MiSeq [OBI:0002003]
amplicon pcr primer scheme: artic-v5.3.2_400 [GENEPIO:0100856]
amplicon size: 400
consensus sequence software name: ViralRecon
consensus sequence software version: 1.23
diagnostic measurement method: Quantitative real time polymerase chain reaction (qPCR) [OBI:0000893]
gene name: N gene (orf9) [GENEPIO:0100153]
diagnostic target presence: Diagnostic target present [GENEPIO:0100987]
diagnostic measurement value: 22
diagnostic measurement unit: Cycle threshold (Ct) [GENEPIO:0100657]
lineage/clade name: JN.1
lineage/clade analysis software name: pUShER
lineage/clade analysis software version: 1.2.6

Scenario 7: SARS-CoV-2 lineage determination using Freya Analysis

A wastewater sample (sample ID BW-WW-12345) was collected in a municipal sewer system on Nov 1 2023 as part of a routine surveillance program for tracking community-level SARS-CoV-2 variants (sewer site ID WWSC2-ABC-b) in order to establish baseline norms. The sample was collected by the Region of Peel regional authority, and sequenced by the Public Health Ontario provincial health laboratory (contact: Johnny Bloggs; jbloggs@provlab.ca). The presence of SARS-CoV-2 was first detected using qPCR (N1 gene, Ct value of 22). The amplicon-based sample was sequenced on an Illumina NovaSeq 6000 on Jan 18 2024 using the ARTIC V5 400bp primer scheme (artic-v5.3.2_400). Lineage assignments were performed using Freya 1.5.0 (full analysis available in file aggregated-WWSC2-ABC-b_1234.tsv). The associated contextual data record is provided below. This record is for the public health laboratory's use only, and many details were removed when sharing data according to organization-specific data sharing policies. This record highlights how lineage identifications using Freyja analyses can be captured using the specification - including results details such as lineage/clade names (multiple lineages can be summarized and separated by a semicolon) and analysis filenames, as well as methods details such as lineage analysis software name and version number.

Template: SARS-CoV-2

specimen collector sample ID: BW-WW-12345
sampling site ID: WWSC2-ABC-b
sample collected by: Region of Peel Regional Authority
geo_loc_name (country): Canada [GAZ:00002560]
geo_loc_name (state/province/territory): Ontario [GAZ:00002563]

organism: Severe acute respiratory syndrome coronavirus 2 [NCBITaxon:2697049]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
sample collection date: 2023-11-01
scale of sampling: Community-level surveillance [GENEPIO:0100874]
environmental site: Sewer [ENVO:01000924]
environmental material: Wastewater [ENVO:00002001]
populated area type: Suburban [GSSO:011077]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Amplicon sequencing assay [OBI:0002767]
sequencing date: 2024-01-18
sequenced_by: Public Health Ontario
sequenced_by_contact_name: Johnny Bloggs
sequenced_by_contact_email: jbloggs@provlab.ca
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
amplicon pcr primer scheme: artic-v5.3.2_400 [GENEPIO:0100856]
amplicon size: 400
diagnostic measurement method: Quantitative real time polymerase chain reaction (qPCR) [OBI:0000893]
gene name: N gene (orf9) [GENEPIO:0100153]
diagnostic target presence: Diagnostic target present [GENEPIO:0100987]
diagnostic measurement value: 22
diagnostic measurement unit: Cycle threshold (Ct) [GENEPIO:0100657]
lineage/clade name: B.1.617.2; B.1.2; AY.6; Q.3; EG.5; JN.1
lineage/clade analysis filename: aggregated-WWSC2-ABC-b_1234.tsv
lineage/clade analysis software name: Freyja
lineage/clade analysis software version: 1.5.0
breadth of coverage: 91
depth of coverage: 100