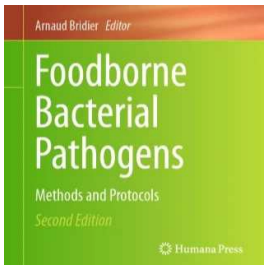
Sep 16, 2024 Version 4

A Comprehensive Guide to Quality Assessment and Data Submission for Genomic Surveillance of Enteric Pathogens V.4



DOI
dx.doi.org/10.17504/protocols.io.eq2lyprkplx9/v4

Ruth Timme[1], Marc Allard[1], Errol Strain[2], Tina Lusk Pfefer[3], Candace Hope Bias[3], Maria Sanchez[1]

[1]US Food and Drug Administration; [2]FDA; [3]US FDA

GenomeTrakr
Tech. support email: **genomeTrakr@fda.hhs.gov**

Ruth Timme
US Food and Drug Administration

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.eq2lyprkplx9/v4

External link: https://link.springer.com/protocol/10.1007/978-1-4939-9000-9_17

Collection Citation: Ruth Timme, Marc Allard, Errol Strain, Tina Lusk Pfefer, Candace Hope Bias, Maria Sanchez 2024. A Comprehensive Guide to Quality Assessment and Data Submission for Genomic Surveillance of Enteric Pathogens. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.eq2lyprkplx9/v4**Version created by **Ruth Timme**

Manuscript citation:
Timme, R.E. et al. (2025). A Comprehensive Guide to Quality Assessment and Data Submission for Genomic Surveillance of Enteric Pathogens. In: Bridier, A. (eds) Foodborne Bacterial Pathogens. Methods in Molecular Biology, vol 2852. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-4100-2_14

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

Created: February 26, 2024

Last Modified: September 16, 2024

Collection Integer ID: 95773

Keywords: GenomeTrakr, Surveillance, Foodborne pathogens, Open data

Disclaimer

Abstract

This document outlines the steps necessary to assemble and submit the standard data package required for contributing to the global genomic surveillance of enteric pathogens. Although targeted to GenomeTrakr laboratories and collaborators, these protocols are broadly applicable for enteric pathogens collected for different purposes. There are five protocols included in this chapter: (1) quality control (QC) assessment for the genome sequence data, (2) validation for the contextual data, (3) data submission for the standard pathogen package or Pathogen Data Object Model (DOM) to the public repository, (4) viewing and querying data at NCBI, and (5) data curation for maintaining relevance of public data. The data are available through one of the International Nucleotide Sequence Database Consortium (INSDC) members, with the National Center for Biotechnology Information (NCBI) being the primary focus of this document. NCBI Pathogen Detection is a custom dashboard at NCBI that provides easy access to pathogen data plus results for a standard suite of automated cluster and genotyping analyses important for informing public health and regulatory decision-making.

Although originally published as a Chapter in Methods and Protocols, Foodborne Bacterial Pathogens, the protocol has since been adapted and split into five separate protocols all of which are contained in this collection.

Guidelines

Direct link to method: https://link.springer.com/protocol/10.1007/978-1-0716-4100-2_14#citeas

1 Introduction

The landscape of food safety and public health is continually reshaped by the challenges posed by foodborne bacterial pathogens. At the US Food and Drug Administration, the GenomeTrakr Program coordinates a network of laboratories that perform genomic surveillance of foodborne pathogens collected from environmental sources, including food, food facilities, farms, water, etc. Genomic data collected from this network are publicly available at the National Institutes of Health, National Center for Biotechnology Information (NCBI) where they are analyzed alongside pathogens collected from humans and animals submitted from other US collaborating sequencing networks, like PulseNet [**1**,**2**,**3**], and other countries [**4**,**5**,**6**]. GenomeTrakr not only contributes data but also collaborates nationally and internationally to set and agree on a variety of data structure and quality control standards, along with providing methods and protocols for submitting data that adhere to those standards [**7**,**8**,**9**,**10**]. These standards and submission protocols ensure the integration of global enteric pathogen surveillance data within NCBI and the International Nucleotide Sequence Database Collaboration (INSDC) (Fig. **1**).

Fig 1: https://link.springer.com/protocol/10.1007/978-1-0716-4100-2_14/figures/1
Data flow for enteric pathogen genome surveillance, with the INSDC (NCBI for our purposes) as the public repository. This general workflow is used by GenomeTrakr laboratories, collaborators, and independent submitters

The protocols detailed in this chapter are designed to enable any laboratory—be it in public health, agriculture, academia, environmental, or industry—to contribute data toward global efforts in foodborne pathogen genomic surveillance. This contribution requires several steps, including (1) quality control assessment of genome sequence data, (2) validation of contextual data, (3) submission of standard genomic data packages to NCBI, (4) querying and tracking your submissions and analysis results within NCBI, and (5) data curation to ensure submissions maintain their public health relevance. These protocols adhere to and maintain the integrity and relevance of public health pathogen data within the global effort to reduce the burden of foodborne illness.

2 Materials

Materials described here cover the items that are needed inhand prior to initiating the protocols included in this chapter.

2.1 Sequence Data

1.Sequence data generated from a validated or verified whole genome sequence laboratory method, for example, GenomeTrakr methods [**11**], PulseNet methods [**12**], or other equivalent methods, performed on an isolated, pure culture-enteric pathogen, such as *Salmonella enterica*, pathogenic *Escherichia coli*, *Listeria monocytogenes*, *Campylobacter jejuni*, *Vibrio parahaemolyticus*, or *Cronobacter sakazakii*.
2.Raw fastq files with a 150 bp minimum read length generated from an Illumina platform (MiSeq, iSeq, and NextSeqs are all validated platforms).

2.2 Contextual Data (Metadata) Templates

1.Ensure that your laboratory can provide the minimum set of contextual data BEFORE any sequencing starts or submissions are prepared.

2.One Health Enteric BioSample template (download current version) from GitHub, https://github.com/CFSAN-Biostatistics/One_Health_Enteric_Package, or, GMVS, https://gmvs.fda.gov/Onehealthenteric.

3.NCBI's SRA metadata template: ftp://ftp-trace.ncbi.nlm.nih.gov/sra/metadata_table/SRA_metadata_acc.xlsx.

2.3 User Accounts to Access the Following Resources

1.NCBI Submission Portal: **https://submit.ncbi.nlm.nih.gov/subs/**.

2.GalaxyTrakr: https://galaxytrakr.org.

3 Methods

3.1 Quality Control Assessment and Characterization Screens

1.Sequence data should be thoroughly checked for quality control before submission to NCBI. GenomeTrakr screens for quality (sequence quality, coverage, etc.) and integrity (correct ID, no contamination, etc.) (Fig. **2**). GenomeTrakr has implemented open-source workflows in Galaxy [**13**] that laboratories can use to easily check data.

2.Sequence quality: Sequence quality is assessed at the read and at the assembly level. Analysis of genomic data for surveillance typically uses de novo assemblies as input to look for markers associated with foodborne outbreaks, toxins, and antimicrobial resistance (AMR).

3.*MicroRunQC* [**9**] De novo assemblies are created, and reads are mapped back to the resulting assembly, to obtain estimates for genome size, number of contigs, coverage, average read quality, and read insert size (Table **1**). Data that falls outside of the QC thresholds typically indicates a poor sequencing run but may also indicate problems with contamination or misidentified species. MicroRunQC also provides a seven-gene multilocus sequencing typing (MLST) prediction for common foodborne pathogens such as *Salmonella*, *Escherichia coli*, *Campylobacter*, *Listeria monocytogenes*, and others. Predictions of multiple alleles for individual MLST genes may indicate contamination or mixed cultures.
Protocols.io detailed protocol: Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow. Version 6. 2024. **dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v6**.

4.*Sequence composition and speciation*. Data that fails the MicroRunQC check is evaluated using tools like Kraken [**14**], sendsketch [**15**], and the Genome Taxonomy Database Toolkit (GTDB-Tk) [**16**] to investigate contamination, such as a mixture of two different *Campylobacter* strains and/or misidentified species, such as *L. innocua* mistaken for *L. monocytogenes*. Kraken provides read-level taxonomic estimates that can be used to identify contamination within and between species. Sendsketch provides an implementation of the min-hash algorithm, allowing for rapid comparisons of de novo assemblies to large prokaryote databases like NCBI RefSeq [**17**] to identify closely related genomes. Average nucleotide identity (ANI)-based tools, such as GTDB-Tk, have been used to define boundaries between bacterial species and can be used to classify de novo assemblies down to the genus and/or species level.

5.*Strain typing and subtyping:* Data that passes previously described QC checks is further analyzed using species-specific tools that identify markers associated with surface antigens and serotype classification. This analysis step can also identify intraspecies contamination or mixtures that may have been missed by other approaches. Tools used

for *Salmonella*, *E. coli*, and *L. monocytogenes* are listed below. For other organisms, we use speciation tools like sendsketch or GTDB-TK to confirm identity as described in the previous subheading.
1.*SeqSero2* [**18**]. SeqSero2 is a validated method at the FDA for *Salmonella* serotype determination [**19**]. This also serves as a confirmatory subtype for *Salmonella* (potentially flagging intraspecies mixups).
2.*SerotypeFinder* [**20**]. We recommend SerotypeFinder to determine serotype for *E. coli* isolates (not yet validated at the FDA). This serves as a confirmatory subtype for *E. coli*, potentially flagging intraspecies mixups.
3.*LisSero* [**21**]. We recommend LisSero to predict serogroups for *L. monocytogenes* (not yet validated at the FDA). LisSero provides predictions for the four main serotypes (1/2a, 1/2b, 1/2c, and 4b) associated with contaminated food and human illness.

6.Genomic surveillance includes identifying genes and point mutations of public health significance, including antimicrobial resistance, virulence, pathogenicity, and stress response. While the presence of a gene or mutation may not always correlate with phenotype, for example, some AMR genes may only be expressed under specific conditions, they are often tracked and reported for outbreaks and investigations related to food safety.
1.*AMRFinderPlus* [**22**]. De novo assemblies are screened using AMRFinderPlus to look for genes and point mutations from NCBI's Reference Gene Catalog (RGC). RGC includes genes/mutations associated with AMR and virulence, along with acid, biocide, metal, and stress resistance.

Fig 2: Quality control checkpoints for microbial pathogen WGS data**Full size image**

Table 1 Sequence quality control checks and established thresholds: Full size table

3.2 Metadata Validation (GMVS)

1.Two different templates capture the suite of sample and experimental contextual data needed for the pathogen genome submission (Fig. **3**) (*see* Subheading **2.2**): for sample-related information, we recommend the One Health Enteric (OHE) package [**23**], and for experimental information (information describing how the sequencing was performed), we recommend populating the default SRA metadata template. It is important to ensure that each piece of metadata gets submitted to the correct metadata attribute (or field) within the package, or the information will not get labeled properly in downstream applications and therefore will not be available to those interpreting the results.
Protocols.io detailed protocol: Guidance for populating and validating GenomeTrakr metadata templates (BioSample and SRA). Version 11. 2024. **dx.doi.org/10.17504/protocols.io.eq2ly3x1pgx9/v11**

2.*Populate the OHE package.* The full package covers the major sample categories relevant for surveillance of bacterial foodborne pathogens represented as subpackages that cover human/animal hosts, food products, food facilities, and other environmental sources. Although the full package contains 82 attributes, only a subset of these are required: 12 core attributes, plus a set of 2–3 attributes for each subpackage (Table **2**). We recommend users download the most appropriate subpackage for their sample, then populate the mandatory and recommended fields and any others they can to describe their sample.

3.*Validate your metadata* at the GenomeTrakr Metadata Validation System: https://gmvs.fda.gov.

4.*Populate the SRA metadata template.* Follow guidance in **step 3** of "Guidance for populating and validating GenomeTrakr metadata templates (BioSample and SRA).

"Table 2 Mandatory and conditionally mandatory fields within the One Health Enteric metadata packageFull size table

3.3 NCBI Submission

1.The NCBI submission protocol provides guidance on how to submis a standardized data package to NCBI, compliant with the Pathogen Data Object Model (Pathogen DOM) [**10**] (Fig. **3**). Ensure both sequence data and metadata pass quality control/validation checks prior to submission. If you have sequences to submit that belong to more than one BioProject, create a separate submission + metadata table for each of your BioProjects. NCBI will perform validation upon submission.

Protocols.io detailed protocol: NCBI submission protocol for microbial pathogen surveillance. Version 10. 2024. **dx.doi.org/10.17504/protocols.io.4r3l284pql1y/v10**

Fig. 3. Overview of the Pathogen Data Object Model (Pathogen DOM) detailing the scope of BioProjects, BioSamples, and Sequence Read Archive (SRA) submissions along with associated metadata standards for enteric pathogen surveillance. Abbreviations: Pkg Package **Full size image**

3.4 View Data at NCBI

1.The NCBI is a comprehensive and multifaceted genomic database that encompasses a wide array of individual databases. Each of these databases includes extensive collections and provides specialized portals for querying different types of data. To facilitate the tracking of laboratory submissions and the capture of automated clustering and genotyping results, we developed a straightforward protocol. This protocol outlines a few simple steps for constructing queries using a list of genome IDs.
Protocols.io detailed protocol: Querying for Bacterial Pathogen Genomic Data at NCBI. Version 1. 2024. **dx.doi.org/10.17504/protocols.io.36wgq3kbklk5/v1**

3.5 Curate Data at NCBI

1.The shift toward genomic surveillance and open pathogen tracking systems necessitates expanded duties for laboratory scientists and public health experts involved in collecting data for public health decision making. In addition to aquiring skills needed to collect and analyze these data, laboratories should also implement standard procedures for consistent, accurate data recording and updating, crucial for the utility of these public resources for pathogen surveillance. Appointing a dedicated person or team responsible for data management, maintaining documented curation protocols, and establishing a routine assessment for curation needs can streamline this proces. Ensuring timely updates and effective communication among different teams involved in sample handling and genomic data collection is essential for maintaining current and reliable public records.

Protocols.io detailed protocol: NCBI Bacterial Pathogen Data Curation Protocol: SOP for Editing GenomeTrakr Submissions. Version 5. 2024. https://doi.org/10.17504/protocols.io.36wgq5jb5gk5/v5

Disclaimer

Files

SEARCH

Protocol

NAME

Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow

VERSION 6

CREATED BY

Ruth Timme
US Food and Drug Administration

OPEN →

Protocol

NAME

Guidance for populating and validating GenomeTrakr metadata templates (BioSample and SRA)

VERSION 11

CREATED BY

Ruth Timme
US Food and Drug Administration

OPEN →

Protocol

NAME

NCBI submission protocol for microbial pathogen surveillance

VERSION 10

CREATED BY

Ruth Timme
US Food and Drug Administration

OPEN →

Protocol

NAME

Querying for Bacterial Pathogen Genomic Data at NCBI

VERSION 1

CREATED BY

Maria Balkey

US Food and Drug Administration

OPEN →

Protocol

NAME

NCBI Bacterial Pathogen Data Curation Protocol: SOP for Editing GenomeTrakr Submissions

VERSION 5

CREATED BY

Ruth Timme
US Food and Drug Administration

OPEN →

Protocol references

Timme, R.E. *et al.* (2025). A Comprehensive Guide to Quality Assessment and Data Submission for Genomic Surveillance of Enteric Pathogens. In: Bridier, A. (eds) Foodborne Bacterial Pathogens. Methods in Molecular Biology, vol 2852. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-4100-2_14

Timme R.E., Sanchez Leon M., Allard M.W. (2019) Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. In: Bridier A. (eds) Foodborne Bacterial Pathogens. Methods in Molecular Biology, vol 1918. Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9000-9_17