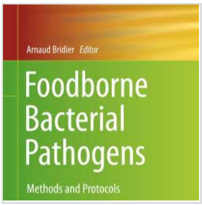

2

Feb 25, 2022

Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback V.2

Ruth Timme[1], Maria Sanchez[1], Marc Allard[1]

[1]US Food and Drug Administration

1

dx.doi.org/10.17504/protocols.io.bz7dp9i6

GenomeTrakr Springer Nature Books

Ruth Timme
US Food and Drug Administration

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

This protocol outlines the all the steps necessary to become a GenomeTrakr data contributor. GenomeTrakr is an international genomic reference database of mostly food and environmental isolates from foodborne pathogens. The data and analyses are housed at the National Center for Biotechnology Information (NCBI), which is a database freely available to anyone in the world. The Pathogen Detection browser at NCBI computes daily cluster results adding the newly submitted data to the existing phylogenetic clusters of closely related genomes. Contributors to this database can see how their new isolates are related to the real-time foodborne pathogen surveillance program established in the USA and a few other countries, and at the same time adding valuable new data to the reference database.

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

DOI

dx.doi.org/10.17504/protocols.io.bz7dp9i6

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

Citation: Ruth Timme, Maria Sanchez, Marc Allard Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback https://dx.doi.org/10.17504/protocols.io.bz7dp9i6

Ruth Timme, Maria Sanchez, Marc Allard 2022. Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.bz7dp9i6
Ruth Timme

collection

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

GenomeTrakr, WGS, Surveillance, Foodborne pathogens, Trackback

collection ,

Nov 18, 2021

Feb 25, 2022

55237

Introduction

1.1 Inception of GenomeTrakr Within the FDA Mission

In 2012 FDA began a pilot project called GenomeTrakr to build a public genomic reference database of historical food and environmental isolates of *Salmonella*. The goal of this project was to improve the accuracy and response time for identifying the causes of foodborne outbreaks, to identify harborage in facilities, and to aid in establishing preventative controls [1]. In this pilot WGS data were collected by a distributed set of public health laboratories, transferred to the FDA for quality screening, then uploaded under an umbrella BioProject at NCBI’s SRA database (Fig. 1). The result has been a continuously growing database of genomic sequence information and accompanying metadata (e.g., geographic location, source, and date) from food, environmental, and clinical isolates. Over 1000 *Salmonella* genomes were collected after the first year, around 10,000 by the second year, and now, after 5 years and multiple contributors, including other US agencies and Public Health England, the maturing *Salmonella* database is approaching 160,000 genomes [2]. After the initial success of sequencing *Salmonella*, the effort expanded to *Listeria monocytogenes* [3] in 2013, and soon thereafter pathogenic *Escherichia coli*/ *Shigella*, *Campylobacter jejuni*, *Vibrio parahaemolyticus*, and *Cronobacter*. The Pathogen Detection portal at

NCBI is now the central repository for foodborne pathogen genomes used for real-time surveillance in the US.

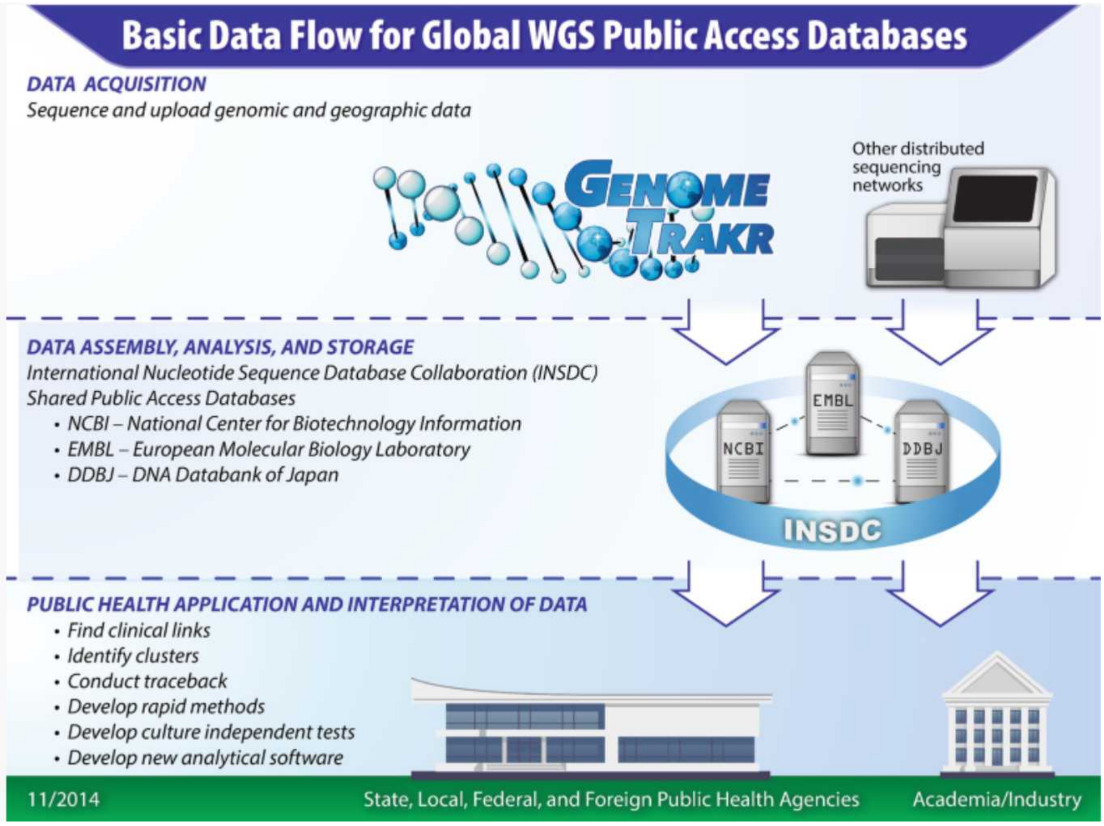
Basic Data Flow for Global WGS Public Access Databases
DATA ACQUISITION
Sequence and upload genomic and geographic data
GENOME TRAKR
Other distributed sequencing networks
DATA ASSEMBLY, ANALYSIS, AND STORAGE
International Nucleotide Sequence Database Collaboration (INSDC)
Shared Public Access Databases
• NCBI – National Center for Biotechnology Information
• EMBL – European Molecular Biology Laboratory
• DDBJ – DNA Databank of Japan
EMBL
NCBI
DDBJ
INSDC
PUBLIC HEALTH APPLICATION AND INTERPRETATION OF DATA
• Find clinical links
• Identify clusters
• Conduct traceback
• Develop rapid methods
• Develop culture independent tests
• Develop new analytical software
11/2014
State, Local, Federal, and Foreign Public Health Agencies
Academia/Industry

Fig. 1
GenomeTrakr data flow overview

1.2 Increased Role of Food/Environmental Isolate Contributions

Foodborne pathogen isolates collected by FDA field laboratories were a major contributor of the food and environmental isolates in the PulseNet PFGE database. These isolates largely come from FDA's regular sampling of imports, routine facility swabbing, and targeted high-risk food sampling assignments. Food and environmental isolates contributed from state public labs varies from state to state depending on sampling efforts and levels of collaboration with respective state agriculture lab(s). Some states contribute quite a few food and environmental isolates and others next to none. Because membership and contribution to the US PulseNet database is restricted to US public health labs that maintain certification, increasing the sources of food and environmental isolates from laboratories outside these members was not feasible.

Technology also played an important role in the importance of food and environmental isolates. Because the surveillance effort for PFGE is largely lead by epidemiological data, the food/environmental isolates played a secondary role to outbreak delineation (i.e., first define the scope of the outbreak, then use patient interviews to discover potential food sources, then target sampling for those suspect foods). This model works well for low resolution PFGE technology, but with a high-sensitivity technology such as WGS, these outbreak investigations can benefit when the underlying data plays a more forward role in the investigation. For example, a likely scenario under WGS is as follows: first a genomic signal is picked up with a clinical cluster matching a food/environmental isolate, then the full epidemiological investigation is launched in response, at the same time FDA launches additional inspections to understand the root cause of contamination along the farm to fork continuum. Because of the increased resolution of WGS, PulseNet is recognizing a greater number of smaller clusters. However, due to limited resources, those clusters that include a food or environmental isolate often get prioritized for follow-up over those that do not. This results in the food and environmental isolates potentially playing a more important role under a genomic surveillance network. The shift to storing the WGS data in an open, public database creates an opportunity to greatly increase the diversity of these isolates by targeting potential submitters outside the PulseNet community. FDA scientists recognized this advantage early on and worked to leverage the GenomeTrakr network to include non-PulseNet laboratories, such as state agriculture labs, academic labs, and international collaborators with the overall goal to more accurately capture the global population diversity within key foodborne pathogen species. This effort has resulted in a higher percentage of food and environmental isolates in the *Listeria* WGS database: as of April 2018 44% of PulseNet's PFGE database comprised food and environmental isolates compared to 69% at NCBI's Pathogen Detection database [2]. Ultimately, this will help to increase the probability of a food/environmental "match" for any new isolate being added to the database, supporting the FDA's mission to pinpoint the causes of foodborne outbreaks, to identify harborage in facilities, and to use WGS data to establish preventative controls.

1.3 GenomeTrakr Data Flow and Open Source Analysis Pipelines

Sequence data are generated at one of more than 40 GenomeTrakr laboratories, then transferred immediately to our data center at FDA-CFSAN. Newly generated sequence data are processed through our internal quality control pipeline where metrics are accessed for data quality (sequence quality, and sequence coverage) and integrity (correct species and serovar assignment). Data that passes predetermined thresholds are submitted to the short-read archive (SRA) at NCBI where they are processed through NCBI's Pathogen Detection analysis pipeline. Within a couple days the sequences will then appear in the Pathogen Detection browser where results of nightly cluster analyses are available for searching and browsing. On average GenomeTrakr submits over 1000 isolates per month to the Pathogen Detection pipeline.

FDA monitors the public Pathogen Detection site daily looking for mission-relevant clustering results, such as a close match between a food isolate and an isolate collected from a clinical patient or an environmental swab isolate match to an isolate collected from the same location in a previous year. Upon seeing results like these one of the FDA data scientists will download the sequence data associated with a particular cluster, then rerun the SNP analysis using CFSAN's open source SNP pipeline [4]. Depending on the nature of the cluster, appropriate stakeholders will be

Citation: Ruth Timme, Maria Sanchez, Marc Allard Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback
https://dx.doi.org/10.17504/protocols.io.bz7dp9i6

contacted for follow-up. For example, a cluster showing clonal isolates collected from the same facility over multiple years might be sent to the FDA's Office of Compliance where the data will be added to ongoing facility investigations. Similarly, new food + clinical matches might be forwarded to the FDA's Coordinated Outbreak and Response (CORE) team or perhaps a state lab might be contacted if the cluster appears to be contained within state boundaries. A regulatory response by the FDA will include all the evidence gathered across a full investigation, which might include site visits, epidemiological evidence, as well as supporting data from WGS cluster analysis. This three-legged stool of evidence from epidemiology, site investigations, and WGS provide support for FDA regulatory decisions.

1.4 WGS Data Collection and Analysis—Validation and Harmonization

Genomics for Food Safety (Gen-FS) is a working group in the USA [5], with representatives from the CDC, FDA, USDA and NCBI. The Gen-FS working groups carefully harmonize quality management systems across GenomeTrakr and PulseNet, including Quality Assurance (QA) measures and accompanying quality control (QC) checks, to ensure all WGS data in the Pathogen Detection database meet the Gen-FS minimum quality standards. In addition, all downstream analyses, including cluster analysis presented through the Pathogen Detection website, outbreak analyses from PulseNet, and identifying the source of contamination events from GenomeTrakr, are harmonized such that the results are accurate and comparable across the different analysis pipelines. Benchmark datasets derived from empirical data [6] as well as simulated data [7] are used in this harmonization effort. Gen-FS also runs an annual multilab proficiency test (PT) across the PulseNet/GenomeTrakr lab network [8] which measures proficiency for each laboratory and also serves as a multilab validation exercise by accessing the accuracy and reproducibility in the WGS data collection across the whole network.

The Global Microbial Identifier (GMI) [9] is an international organization dedicated harmonizing the multiple in-country efforts of genomic pathogen surveillance. GMI is working toward a global system to aggregate, share, mine and translate genomic data for microorganisms in real time. Multiple GMI working groups have agreed on minimum metadata standards, proposed quality control standards on the sequence data, and produced benchmark datasets for validating analysis pipelines. Additional efforts for global harmonization include developing guidance documents on the value of WGS technologies, and the value of sharing these data, both with the Food and Agricultural Organization (FAO) and World Health Organization (WHO) (see http://www.fao.org/food/food-safety-quality/a-z-index/wgs/en/).

The GenomeTrakr database is a free, open-access, database for consumption and contribution. Contributors do not have to be affiliated with the FDA, Gen-FS, or GMI to submit data and use NCBI's Pathogen Detection portal to view clustering results. This chapter outlines all the steps necessary to independently submit data to the GenomeTrakr database at NCBI, including WGS quality standards, NCBI data submission, and finally how to view and curate your data and cluster results at NCBI.

References

1. Allard MW, Strain E, Melka D et al (2016) Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol 54:1975–1983. https://doi.org/10.1128/JCM.00081-16
CrossRef PubMed PubMedCentral Google Scholar

2. NCBI Pathogen Detection Homepage.https://www.ncbi.nlm.nih.gov/pathogens. Accessed 16 Jun 2018

3. Jackson BR, Tarr C, Strain E et al (2016) Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis 63:380–386. https://doi.org/10.1093/cid/ciw242
CrossRef PubMed PubMedCentral Google Scholar

4. Davis S, Pettengill JB, Luo Y et al (2015) CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Comput Sci 1:e20. https://doi.org/10.7717/peerj-cs.20
CrossRef Google Scholar

5. CDC (2015) Annual report to the secretary, Department of health and human services. Improving governmental coordination and integration, interagency collaboration on genomics and food safety (Gen-FS), Section 1, pp. 22–23
Google Scholar

6. Timme RE, Rand H, Shumway M et al (2017) Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. PeerJ 5:e3893. https://doi.org/10.7717/peerj.3893
CrossRef PubMed PubMedCentral Google Scholar

7. McTavish EJ, Pettengill J, Davis S et al (2017) TreeToReads–a pipeline for simulating raw reads from phylogenies. BMC Bioinformatics 18:178. https://doi.org/10.1186/s12859-017-1592-1
CrossRef PubMed PubMedCentral Google Scholar

8. Timme RE, Rand H, Sanchez Leon M et al (2018) GenomeTrakr proficiency testing for foodborne pathogen surveillance: an exercise from 2015. Microb Genom 57:289. https://doi.org/10.1099/mgen.0.000185
CrossRef Google Scholar

9. Global Microbial Identifier (2011) The global microbial identifier homepage.http://www.globalmicrobialidentifier.org. Accessed 25 Jun 2018

10. Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.

Citation: Ruth Timme, Maria Sanchez, Marc Allard Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback https://dx.doi.org/10.17504/protocols.io.bz7dp9i6

https://doi.org/10.1186/gb-2014-15-3-r46
CrossRef PubMed PubMedCentral Google Scholar

11. Zhang S, Yin Y, Jones MB et al (2015) Salmonella serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol 53:1685–1692.
https://doi.org/10.1128/JCM.00323-15
CrossRef PubMed PubMedCentral Google Scholar

12. Laing C ECTyper (and easy typer). In: GitHub.https://github.com/phac-nml/ecoli_serotyping. Accessed 25 Jun 2018

Materials

Materials described here cover the formats of sequence data, accompanying metadata, and quality control thresholds being utilized by GenomeTrakr for the submission of raw sequence data.

1. *Project creation:* Establish an umbrella BioProject(s) that will hold one or multiple data BioProjects (e.g., one for each pathogen species). Email the umbrella accession to pd-help@ncbi.nlm.nih.gov and ask to have it linked to the Pathogen Detection pipeline.
2. *Metadata:* Download the combined pathogen package template from NCBI: https://www.ncbi.nlm.nih.gov/biosample/docs/templates/packages/Pathogen.combined.1.0.xlsx
3. *Sequence files.* Files with the following formats are accepted.

-Raw fastq files generated from an Illumina platform instrument (MiSeq, NextSeq, HiSeq, etc.). Download SRA's batch metadata table: ftp://ftp-trace.ncbi.nlm.nih.gov/sra/metadata_table/SRA_metadata_acc.xlsx.

-FASTA formatted complete genomes or draft assemblies with contigs below 200 bp removed.

DISCLAIMER:

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

This protocol outlines the all the steps necessary to become a GenomeTrakr data contributor. GenomeTrakr is an international genomic reference database of mostly food and environmental isolates from foodborne pathogens. The data and analyses are housed at the National Center for Biotechnology Information (NCBI), which is a database freely available to anyone in the world. The Pathogen Detection browser at NCBI computes daily cluster results adding the newly submitted data to the existing phylogenetic clusters of closely related genomes. Contributors to this database can see how their new isolates are related to the real-time foodborne pathogen

Citation: Ruth Timme, Maria Sanchez, Marc Allard Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback https://dx.doi.org/10.17504/protocols.io.bz7dp9i6

surveillance program established in the USA and a few other countries, and at the same time adding valuable new data to the reference database.

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

FILES

Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow
Version 3
by Ruth Timme, US Food and Drug Administration

Guidance for populating GenomeTrakr metadata templates (BioSample and SRA)
Version 7
by Ruth Timme, US Food and Drug Administration

NCBI submission protocol for microbial pathogen surveillance
Version 5
by Ruth Timme, US Food and Drug Administration

NCBI data curation protocol - SOP for editing GenomeTrakr submissions
Version 2
by Ruth Timme, US Food and Drug Administration

Citation: Ruth Timme, Maria Sanchez, Marc Allard Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback https://dx.doi.org/10.17504/protocols.io.bz7dp9i6