Chapter 17

Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback

Ruth E. Timme, Maria Sanchez Leon, and Marc W. Allard

Abstract

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.

Key words GenomeTrakr, WGS, Surveillance, Genomics, Foodborne pathogens, Trackback

1 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

The original version of this chapter was revised. The correction to this chapter is available at https://doi.org/10.1007/978-1-4939-9000-9_22

Arnaud Bridier (ed.), *Foodborne Bacterial Pathogens: Methods and Protocols*, Methods in Molecular Biology, vol. 1918, https://doi.org/10.1007/978-1-4939-9000-9_17, © The Author(s) 2019

Fig. 1 GenomeTrakr data flow overview

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 food-borne pathogen genomes used for real-time surveillance in the US.

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 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-ES, 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.

2 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/bio-sample/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.

3 Internal QA/QC Pipeline

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). At the minimum submitters should establish the following quality control checks (Table 1).

3.1 Sequence Quality

1. The *average Q score* for R1 and R2 should be at Q30 or above, unless there is higher coverage (Table 1). The Q score averages are calculated as average Q score per filename (R1 and R2) as the mean Q of the file's Q histogram where the Q histogram is the count of each Q score appearing in that file.
2. The *estimated average coverage* should be above the taxon-specific determined thresholds listed in Table 1. Estimated coverage is calculated as the total number of bases in the reads divided by the assumed genome size. The following genome sizes are used for the most common foodborne pathogens.
 - *Salmonella enterica* = 4,700,000
 - *Listeria monocytogenes* = 3,000,000
 - *Escherichia coli* = 5,000,000
 - *Campylobacter* sp. = 1,600,000
 - *Vibrio parahaemolyticus* = 5,100,000
3. Pass = all metrics above the thresholds; Fail = any metric falling below the threshold.

Fig. 2 The GenomeTrakr quality control pipeline

Table 1
Sequence quality control checks and established thresholds

Quality metric	Listeria	Campylobacter	Escherichia	Salmonella
Average Read Quality (Q-score)	>28	>28	>28	>28
Expected Coverage	>20×	>20×	>40×	>30×

3.2 Sequence Integrity

1. *Kraken* [10]. We run every submission through Kraken against the mini kraken database. If the top hit does not match the submitted organism name the isolate will fail the sequence integrity check. Second, third, and fourth hits are manually evaluated if other downstream metrics point to contamination.
2. *SeqSero* [11]. We use SeqSero to determine serotype for *Salmonella* isolates. This serves as a confirmatory isolate ID for *Salmonella* (flagging intraspecies mix-ups). It also serves as a double check if sample mix-up is flagged with Kraken.
3. *ECTyper* [12]. We use ECTyper to determine serotype for *Escherichia coli* isolates. This serves as a confirmatory isolate ID for *E. coli* (flagging intraspecies mix-ups). It also serves as a double check if sample mix-up is flagged with Kraken.

4 NCBI Submission

1. NCBI's Pathogen Detection portal has general submission instructions that will supplement the GenomeTrakr steps included in this protocol. <https://www.ncbi.nlm.nih.gov/pathogens/submit-data/>.
2. Create a user account at NCBI: <https://www.ncbi.nlm.nih.gov/account/>
3. Navigate to NCBI's Submission Portal: <https://submit.ncbi.nlm.nih.gov>

5 BioProject

1. Create a New Data BioProject by Clicking on "BioProject" in the Submission Portal, <https://submit.ncbi.nlm.nih.gov>.
2. Submitter tab: populate with submitter info.
3. Project type tab:
 - (a) Project data type = Genome sequencing and assembly.
 - (b) Sample scope = multi-isolate.
 - (c) Click "Autogenerate locus tag prefix".
4. Target tab: Populate ONLY the Organism name here, usually *Genus species*, or just Genus if your laboratory does not determine species, e.g., *Campylobacter*.
5. General info tab: click "Release immediately following processing".

6. Project Title: e.g., “GenomeTrakr Project: NY State Dept. of Health, Wadsworth Center”.
 - (a) Public Description: e.g., “Whole genome sequencing of cultured pathogens as part of a surveillance project for the rapid detection of outbreaks of foodborne illnesses”.
 - (b) Relevance = medical.
 - (c) Is your project part of a larger initiative which is already registered at NCBI? Click yes if you have an umbrella project established and include the accession here. This will properly link your data project with your overall umbrella project.
7. External links: Include a link to your laboratory’s website here.
8. BioSample tab: leave blank.
9. Publications tab: include publications if relevant.
10. Overview tab: check if everything looks correct and edit if necessary, then click “submit.”
11. BioProject accession will be available on the “my submissions” page of the Submission portal and usually starts with “PRJNAxxxxxx.”
12. For questions about establishing umbrella projects, linking data projects, or any other bioproject issue contact genomeprj@ncbinlm.nih.gov.

6 BioSample

Submit metadata to BioSample database:

1. Populate the combined pathogen package template with metadata for each isolate you intend to submit. Ensure that text is included for ALL mandatory fields. Include the word “missing” if data are not available for a given field (Table 2).
2. Click on “BioSample” from the home screen of NCBI’s submission portal, <https://submit.ncbi.nlm.nih.gov>.
3. Click on the “New submission” button at the top of the screen.
4. “Submitter” tab: populate submitter information.
5. “General Info” tab: click “release immediately following processing” and the “Batch/Multiple BioSamples.”
6. “Sample Type” tab: click Pathogen affecting public health, combined pathogen submission.
7. “Attributes” tab: click upload a file use Excel, then navigate to file for upload.

Table 2
The minimum set of metadata fields for food or environmental isolates

<i>Required fields</i>	<i>Description</i>
*sample_name	Sample Name is a name that you choose for the sample (or isolate in our case). It can have any format, but we suggest that you make it concise, unique and consistent within your laboratory, and as informative as possible. Every Sample Name from a single Submitter must be unique.
*bioproject_accession	The accession number of the BioProject(s) to which the BioSample belongs (PRJNAxxxxxx)
*attribute_package	Specify the pathogen type. Allowed values are “Pathogen.cl” (for clinical or host-associated pathogen) or “Pathogen.env” (for environmental, food or other pathogen). The value provided in this field drives validation of other fields.
*host	For Pathogen.cl: “Homo sapiens” if clinical isolate. C
*host_disease	For Pathogen.cl: Name of relevant disease, e.g., Salmonella gastroenteritis. Controlled vocabulary, http://bioportal.bioontology.org/ontologies/1009 or http://www.ncbi.nlm.nih.gov/mesh . Provide “missing” if unknown and leave blank for food/environmental isolates.
*organism	The most descriptive organism name for this sample (to the serovar, if possible).
*strain	This is the authoritative ID used in GenomeTrakr and usually the same as the sample_name listed above. It can have any format, but we suggest that you make it concise, unique and consistent within your laboratory, and as informative as possible. Strain names must be unique within the NCBI database.
*collected_by	Name of persons or institute who collected the sample
*collection_date	Date of sampling in ISO 8601 standard: “YYYY-mm-dd”, “YYYY-mm” or “YYYY” (e.g., 1990-10-30, 1990-10, or 1990)
*geo_loc_name	Geographical origin of the sample: http://www.insdc.org/documents/country-qualifier-vocabulary . Use a colon to separate the country or ocean from more detailed information about the location, e.g., “Canada: Vancouver”.
*isolation_source	Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived. For food isolates please provide generic descriptions of the food.
*lat_lon	This is a required field for NCBI, but not for GenomeTrakr. Include “missing” here if lat/long details are not provided.

8. “Overview” tab: Check over your submission for errors, then submit.
9. BioSample accessions will be available on the “my submissions” page of the Submission portal. Accessions will start with SAMNxxxxxxx.”

7 Sequence Submission

7.1 *Submit Raw Data to SRA Database*

1. Populate SRA's batch metadata table, downloaded from above.
 - (a) In the second Excel sheet, populate with the BioProject and BioSample accessions obtained from above.
 - (b) Populate remaining sheet according to instructions listed on first Excel sheet.
 - (c) Save spreadsheet under second sheet (SRA_data) as a TSV (tab-delimited file) for upload in the "SRA metadata" tab within the submission portal.
2. Click on "SRA" from the home screen of NCBI's submission portal.
3. "Submitter" tab: populate with personal info.
4. "General info" tab: Click yes for existing Bioproject and fill in with accession from above.
5. Click yes for existing BioSamples.
6. Click "release immediately following processing."
7. "SRA metadata" tab: Click "Upload a file using Excel or text format (tab-delimited)" and choose file saved above.
8. "Files" tab: check and resolve any validation errors.
9. "Overview" tab: check over entire submission, then click submit.

7.2 *Submit Assembled Data to Genomes Database*

1. Navigate to <https://submit.ncbi.nlm.nih.gov/subs/genome/>.
2. Download batch upload template, titled "Genome Info File Template": <https://submit.ncbi.nlm.nih.gov/templates/>.
3. Provide content for the following fields: biosample_accession, assembly_method, assembly_method_version, genome_coverage, sequencing_technology, and filename. Save as an excel or tab-delimited text file.
4. Click "New Submission" and follow prompts. Note: only one BioProject can be specified per submission.
5. Upload Genome Info template in the "Genome Info" tab.
6. Check submission and submit.

8 View and Curate Data at NCBI

8.1 *Check Your Data Within Each Database at NCBI*

1. *BioProject*: Are your submissions properly linked to the BioProject accession you provided? Search for your project using the PRJNA accession: <https://www.ncbi.nlm.nih.gov/bioproject>

2. *BioSample*: Does the metadata look correct in the BioSample records? Search using the SAMN accessions: <https://www.ncbi.nlm.nih.gov/biosample>
3. *SRA*: Are your raw data submissions available in the sequence read archive? Search using the SRR run accessions: <https://www.ncbi.nlm.nih.gov/sra>
4. *Pathogen Detection*: For every BioProject linked to the Pathogen Detection pipeline, NCBI will automatically pick up your submissions and run them through their internal QC pipeline before getting clustered. The QC plus the clustering pipeline add 1-3 days to the processing time after submission.
 - (a) Search for your data using the Strain name submitted in the BioSample record: <https://www.ncbi.nlm.nih.gov/pathogens>.
 - (b) Isolates that fail NCBI's QC thresholds will be listed in the exceptions file posted at NCBI's FTP site:
ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/<pathogenName>/PDG0000000XX.XXXX/Exceptions/PDG0000000XX.XXXX.reference_target.exceptions.tsv
 For example: <ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Salmonella/PDG000000002.1216/Exceptions/PDG000000002.1216.exceptions.tsv>
 - (c) If your isolate appears on this list please retract the data from SRA by sending an email to sra@ncbi.nlm.nih.gov (see 8.2). Re-isolation or re-sequencing might be required, depending on the quality issue flagged. When new data are generated follow the original SRA submission instructions listed previously.

8.2 Curation of NCBI Data

1. Curation is done entirely through email. Maintaining current and updated data is an extremely important part of running a valid surveillance database and NCBI expects multiple requests/day or week to update and retract data. Do not hesitate to send these emails!
2. *BioSample*: email corrections, updates, and retractions to bio-samplehelp@ncbi.nlm.nih.gov. Emails should have the request in the email body, e.g., "please retract the following BioSamples, or please update the attached biosamples." Updates are performed by attaching a tab-delimited text file with the BioSample accessions in the first column, with subsequent columns containing updated information. Ensure the exact same header names are used here as were included in the original BioSample submission, e.g. strain, organism, collected_by, isolation_source, collection_date, geo_loc_name, etc. You will receive a confirmation email that the updates were performed.

3. SRA: email retractions or questions to sra@ncbi.nlm.nih.gov. Email should include a list of SRR accessions to retract and reason for retraction (sample mix-up, quality of data, etc.)
4. Pathogen Detection browser: email questions to pd-help@ncbi.nlm.nih.gov.
5. BioProject: email updates, creation of Umbrella projects, questions, and retractions to genomeprj@ncbi.nlm.nih.gov.

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>
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>
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>
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
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>
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>
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>
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: ultra-fast metagenomic sequence classification using exact alignments. *Genome Biol* 15:R46. <https://doi.org/10.1186/gb-2014-15-3-r46>
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>
12. Laing C ECTyper (and easy typer). In: GitHub. https://github.com/phac-nml/ecoli_serotyping. Accessed 25 Jun 2018

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.