Leveraging FDA's GenomeTrakr Data with Visualization Dashboards

Maria Balkey, Marc Allard, Tina Pfefer, Candace Hope Bias, Ruth E. Timme Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA

Abstract

FDA's GenomeTrakr (GT) is a public/private genomic epidemiology network for foodborne pathogen surveillance, specifically targeting pathogens isolated from food or environmental sources. The raw genome plus a small set of associated metadata are made publicly available at the National Center for **Biotechnology Information (NCBI).**

Within the NCBI, the Pathogen Detection (PD) database performs routine clustering and phylogenetic inference to identify relatedness among genomes from bacterial pathogens collected from different sources (food, animal, environmental or human).

These results play a crucial role in informing regulatory decisions made by FDA-CFSAN concerning instances of foodborne contamination.

However, managing laboratory participation, assessing source diversity, promoting adherence to metadata standards, and ensuring acceptance of GT data by NCBI pose challenges as the network's data contributions continue to grow.

To address these challenges, this project aimed to develop and implement GT dashboards that assess trends, identify gaps in metadata, and guide decisionmaking processes supporting foodborne pathogen surveillance.

Introduction

- FDA's GenomeTrakr (GT) program is a network of laboratories that sequence the genomes of bacterial pathogens isolated from various food and environmental samples.
- FDA funded Laboratory Flexible Funding Model (LFFM) cooperative agreement laboratories contribute to GenomeTrakr by sequencing and submitting genomes to NCBI as an effort to improve foodborne pathogen surveillance.
- The primary repository for the GT network genome sequences and associated metadata is the National Center for Biotechnology Information (NCBI).
- Data submitted to NCBI includes various metadata attributes such as: organism name, geographical location, collection date, isolate contributor and isolation source.
- NCBI PD performs real-time analyses that support the identification of sources of **6**odborne **li**ness. Cluster analysis unveils possible relatedness between genomes from clinical/human and environmental/other sources. NCBI PD integrates metadata, SNP analysis results and antibiotic susceptibility and resistance gene predictors.

Materials and Methods

- GT datasets hosted within NCBI are accessible through different databases, including NCBI Sequence Read Archive (SRA), BioSample, GenBank and NCBI PD.
- To visualize both the contextual data (SRA, BioSample) and cluster results (PD), we use Tableau, a platform that integrates multiple data connectors, allowing us to fed the GT dashboard with NCBI data hosted on both the NCBI PD FTP site and Athena within Amazon Web Services (AWS).
- Using a combination of metadata tags and established BioProjects, we can effectively visualize items of importance to the GT network, like source origin and diversity, gaps and inconsistencies in metadata, and potential linkages between different sample source types and human illness. We dentified the opportunity to leverage these data sources into dashboards with two goals:
- 1) to track data from broad contributors across the scientific community.
- 2) to internally track data submissions from GT laboratories which are awarded Laboratory Flexible Funding Model (LFFM) cooperative agreements.







Figure 2. Foodborne pathogen dashboard development workflow

Results and Discussion

We built two real-time dashboards: 1) Characterization of Foodborne Pathogen Sequences Available via NCBI's Pathogen Detection Portal and 2) Monitoring Genomic Data Submitted to NCBI by LFFM Laboratories. These two resources allow users to identify isolates submitted to NCBI, which are the isolation sources for the sequenced bacterial pathogens, and whether bacterial pathogens submitted to NCBI are implicated in human illness.



Figure 3. Characterization of Foodborne Pathogen Sequences Available via NCBI's Pathogen Detection Portal



Monitoring Genomic Data Submitted to NCBI by LFFM Laboratories Project area end/or release year to flirer the entre dambara. Project Name Release Year Solid LFM YB team from drop-down menu Aumobic of Consolid team from drop-down menu Solid LFM YB team from drop-down menu Aumobic of Consolid team for the down of team for team for team for the down of team f	<	Laboratory participants	lsolate: FY2, FY	s by Funding Year (FY1, 3)	Isolates and Fun	by labor ding Yea	ratory, org r (FY1, FY	ganism ′2, FY3)	lsolat source FY2, F	:es by la .e and Fi FY3)	borator unding \	ry, isola Year (FY	tion 1,	>	
httrutions: L Displayed data for 2023, select a project name and/or release year to filter the entire dashbaord. Project Name Select LFMY VE trainin from drop-down menu to display data (Multiple values) A, 934 A0 Number of Isolates A, 934 A0 CEFFM submissions for GenomeTraits Sequencing Laboratory F Mease Vear Sequencing Laboratory S Mease Vear Mease Vear New Very State Department of Apriculture and Market New Very State Department of Apriculture and Market New Very State Department of Apriculture and Market New Very State Department of Apriculture Animal Health Disgoratic Laboratory Wear North Carolina State Department of Apriculture Animal Health Disgoratic Laboratory Wear North Carolina State Department of Apriculture Animal Health Disgoratic Laboratory Wear North Carolina State Department of Apriculture Animal Health Disgoratic Laboratory Measing State Department of Apriculture Animal Health Disgoratic Laboratory Mease Vear New Xearse State Duble Health Disgoratic Laboratory Measing State Department of Apriculture Animal Health Disgoratic Laboratory Mease Vear New Xearse Department of Apriculture Animal Health Disgoratic Laboratory Measing State Department of Apriculture Animal Health Disgoratic Laboratory Measing St		Monitorin	g Gen	omic Data Subm	nitted to	NCB	l by LF	FM La	bora	torie	s				
Project Name Solact LFFM YR tearm from drop-down menu to display data 2023 Number of Isolates 4,934 40 305 LFFM submissions for GenomeTrak Control family for the solation for the solati	nstructions: Displayed data for 2023, select a project name and/or release year to filter the entire dashboard. 2. Hove the mouse over a data point in a visual for more details on each laboratory.														
(Multiple values) 2023 • Number of Isolates Number of Species of Pathogens Number of Isolation Sources 4,934 40 305 Sequencial laborator F Massachusetts Department of Public Patho Sequencial laborator F Massachusetts Department of Public Patho North Carolina Department of Public Patho Number of Agriculture Animal Disese Diagnostic Laboratory & Ohio Department of Agriculture Animal Disese Diagnostic Laboratory & Ohio Department of Agriculture and Rural Development. Numeter of Agriculture and Rural Development. New York State Department of Fublic Patho New York State Department of Fublic Patho New York State Department of Fublic Patho New Mexico State University Public Patho New Mexico State Department of Agriculture and Consumer Services Consumo State Department of Fublic Patho New York State Department of Fublic Patho New Mexico State Department of Agriculture and Consumer Services	Project Na Select LFF to display	ame FM YR tearm from drop-down menu v data		Release Year Select release year from drop-down menu to display data											
Number of Isolates Number of Species of Pathogens Number of Isolation Sources 4,934 40 305 DEFM Submissions for GenomeTrats Seven gaborators of Mathogens Number of Number of Subter Subt	(Multiple	values)		•	2023									•	
4,934 40 305	Number of Isolates Num			Number of Spe	f Species of Pathogens				Number of Isolation Sources						
LFFM submissions for GenomeTrak:	4,934			40				305							
Sequencing Laboratory F Massachusetts Department of Public Health South Carolina Department of Consolidated Laboratory Services Ohio Department of Agriculture and Rural Development Michigan Department of Agriculture and Rural Development North Carolina State University [College of. Minnesota Department of Health New York State Department of State Health Services Maryland Department of Health New York State Department of State Health Services Maryland Department of Health New York State Department of State Health Services Maryland Department of Health New Jersey Department of Agriculture Animal Health Diagnostic Laboratory West Virginia Department of Health New Mexico State University Food Safety Wernont Department of Health Wernont Department of Health New Jersey Department of Agriculture and Consumer Services Colorado State Department of Public Health and Environment New Jersey Department of Agriculture and Consumer Services Colorado State Department of Health and Environment <td colsp<="" th=""><th colspan="14">LFFM submissions for GenomeTrakr</th></td>	<th colspan="14">LFFM submissions for GenomeTrakr</th>	LFFM submissions for GenomeTrakr													
Number of Diosamples =	Sequencing Laboratory - Massachusetts Department of Health South Carolina Department of Health and Environmental Control Urignia Division of Consolidated Laboratory & 0hio State . Miningan Department of Agriculture and Rural Development North Carolina State University College of. Miningan Department of Agriculture and Rural Development NVDOH Wadswoth Center Pennsylvania State University College of. Maryland Department of Health New York State Department of Health New Jersey Department of Agriculture and Department of Health New Jersey Department of Agriculture and Department of Agriculture New Mexico State University - Las Cruces, New Mexico State University Food Safety . Wermont Department of Agriculture and Consumer Services Colorado State Department of Public Health Laboratory West Virginia Department of Agriculture New Jersey Department of Agriculture and Consumer Services Colorado State Department of Public Health and Human Services Center for Genomics and Bioinformatics, Indiana University, Bloomington IN 0 50 100 150 200 250 300 350 400 45											450			

Figure 4. Monitoring Genomic Data Submitted to NCBI by LFFM Laboratories

Conclusion

- As the GT network continues to expand its data contributions, there is a growing need to address challenges related to tracking laboratory participation, describing and quantifying diversity of sample sources collected by our network, and improving contextual data in the database.
- To tackle these issues, this project developed a suite of GT dashboards that provide insights and facilitate decision-making processes supporting CFSAN's oversight of the GT program.
- These dashboards will provide rapid access to large metadata sets and relieve data users of the time-consuming process of manually querying metadata from multiple databases at NCBI and pulling it into Microsoft Excel or R statistical package.
- This streamlined approach saves valuable time and enhances the overall data exploration and analysis experience.