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 decision-making 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 foodborne illness. 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 feed 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 identified 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.

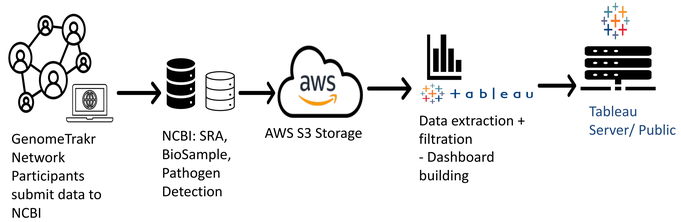

Figure 1. Data Sources for dashboard development

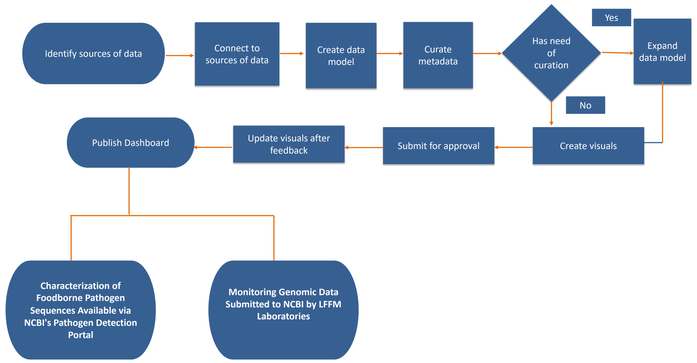

Figure 2. Foodborne pathogen dashboard development workflow

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.

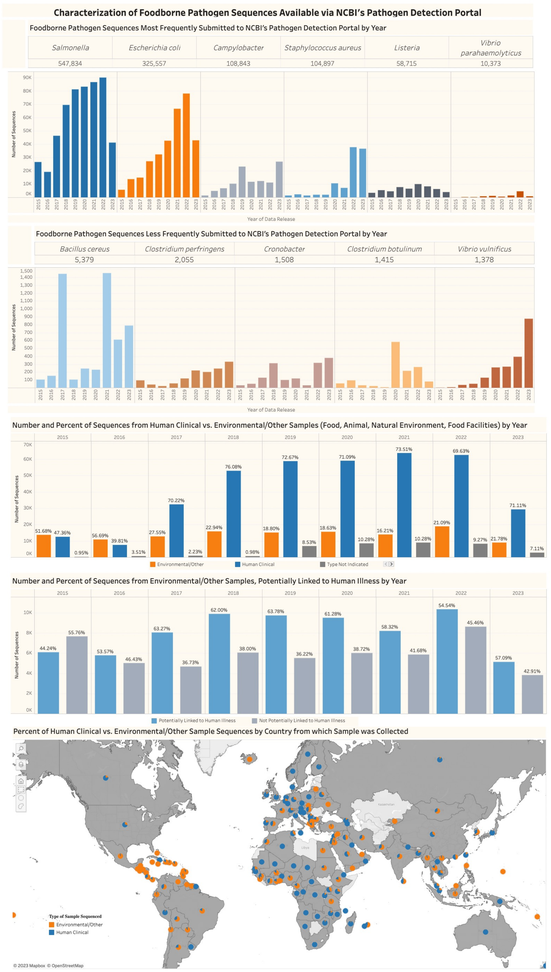

Figure 3. Characterization of Foodborne Pathogen Sequences Available via NCBI's Pathogen Detection Portal

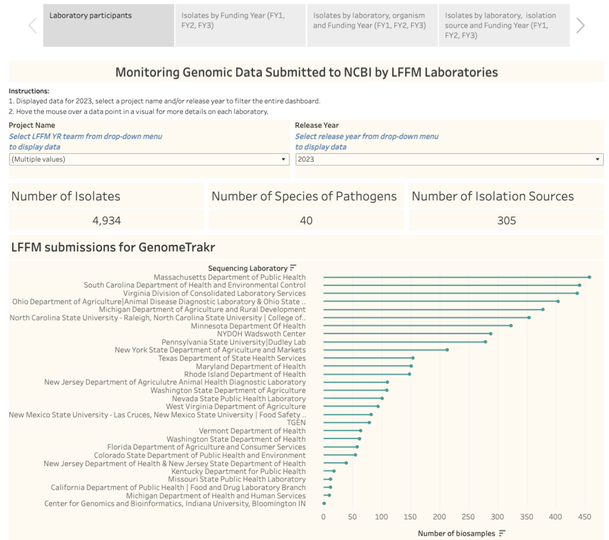

Figure 4. Monitoring Genomic Data Submitted to NCBI by LFFM Laboratories

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.