#14 Dashboard to Monitor Foodborne Pathogen Sequences Available via NCBI's Pathogen Detection Portal

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

Introductlion

FDA's GenomeTrakr (GT) program is a network of laboratories sequencing bacterial pathogens isolated from food and environmental samples. Data Management of whole genome sequencing (WGS) and associated metadata primarily occurs at the National Center for Biotechnology Information (NCBI). The NCBI Pathogen Detection (PD), which is essential for surveillance and outbreak investigation, analyzes data, builds SNP trees to identify relatedness among DNA sequences from bacterial isolates.

As the volume of WGS and associated metadata grows exponentially, dashboards are essential tools for exploratory analysis, providing a visual interface that allows users to interact with and analyze data in real-time. Dashboards are valuable resources for efficient data aggregation, rapid identification of trends and dynamic data exploration.

Abstract

WGS data hosted within NCBI are accessible through different databases, including NCBI Sequence Read Archive (SRA), BioSample, GenBank and PD. We transform raw metadata into an interactive and shareable dashboard that helps in uncovering trends, patterns, and offers insights of WGS metadata from foodborne pathogens.

The dashboard entitled -Characterization of Foodborne Sequences Available via NCBI Pathogen Detection Portal- enables the visualization of large amounts of metadata from NCBI PD for rapid comparison of foodborne pathogen submissions across years, source types, genomic relatedness, and geographical distribution.

The -Characterization of Foodborne Sequences Available via NCBI Pathogen Detection Portal- dashboard provides rapid access to large metadata sets and alleviates users' reliance on a time-consuming process of manually downloading metadata from multiple databases at NCBI. Real time availability of data in a dashboard format allows data exploration, tracking data releases and identification of underperforming areas, enabling timely intervention.

Materials and Methods

- We use Tableau to build a user-centric dashboard interface, applying data visualization best practices. Each visualization highlights five different components: pathogen diversity, pathogen diversity data object model (DOM) compliant, epi type, genomic relatedness and geographical distribution. Data was obtained from NCBI Sequencing Run Archive (SRA) and NCBI PD through access to Amazon Web Services (AWS) and ftp site, respectively.
- To visualize both the contextual data (SRA, BioSample) and cluster results (PD), Tableau integrates multiple data connectors, allowing us to feed the dashboard with NCBI data hosted on both the NCBI PD FTP site and Athena within Amazon Web Services (AWS).

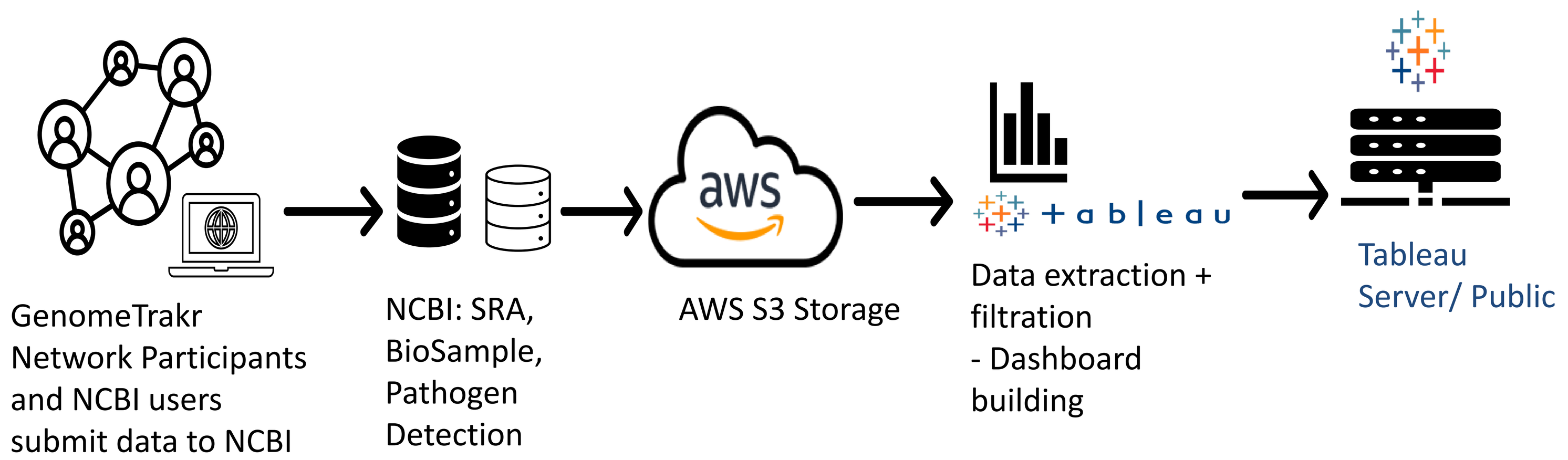
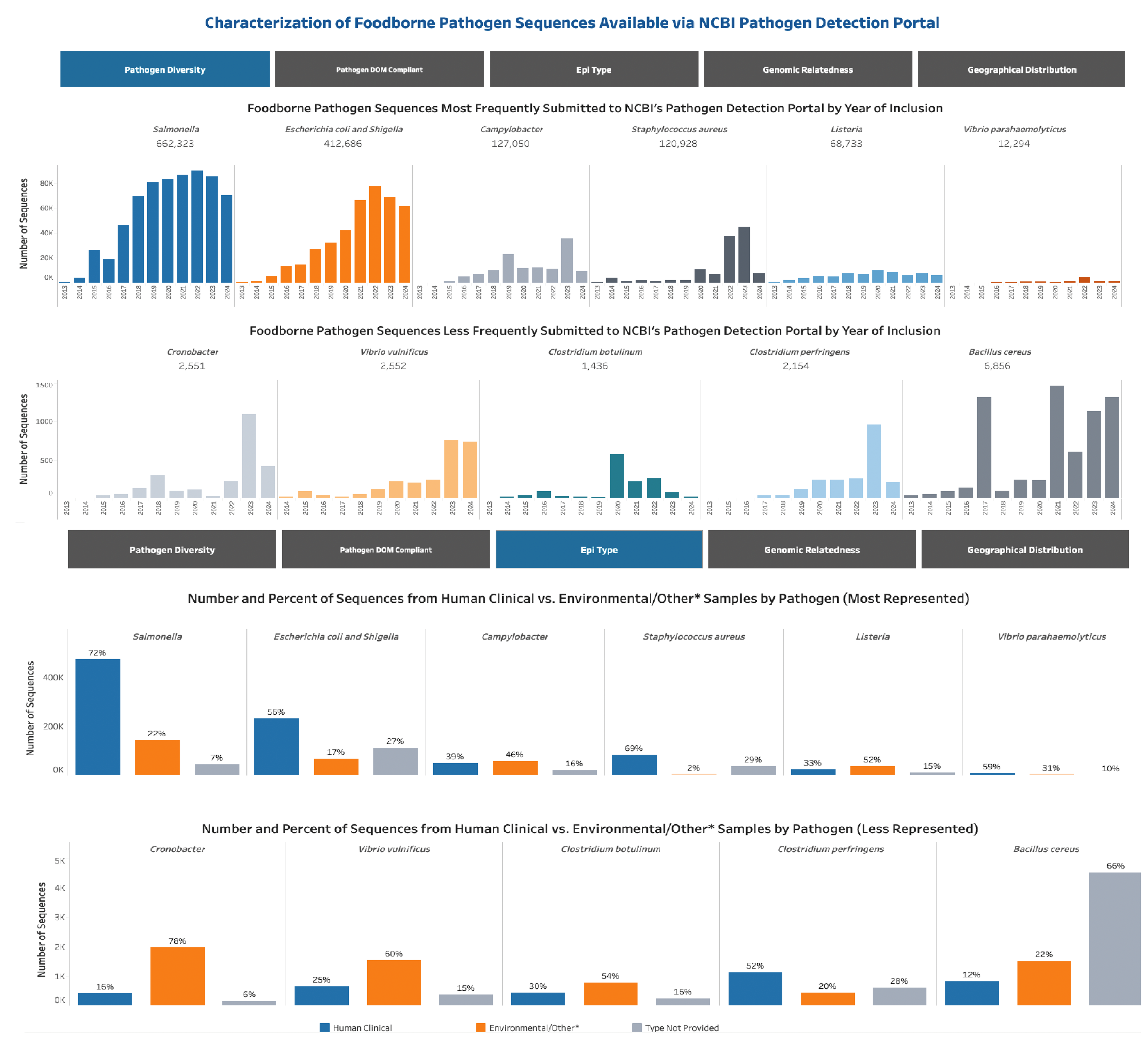


Figure 1. Data Sources for dashboard development

Results

We built a real-time dashboard that use bar charts, geographical distribution and tables to illustrate trends and changes over time for 1,653,237 WGS metadata records submitted to NCBI PD. Includes five different tabs: Pathogen Diversity, Pathogen DOM Compliant Epi Type, Genomic Relatedness and Geographical Distribution.

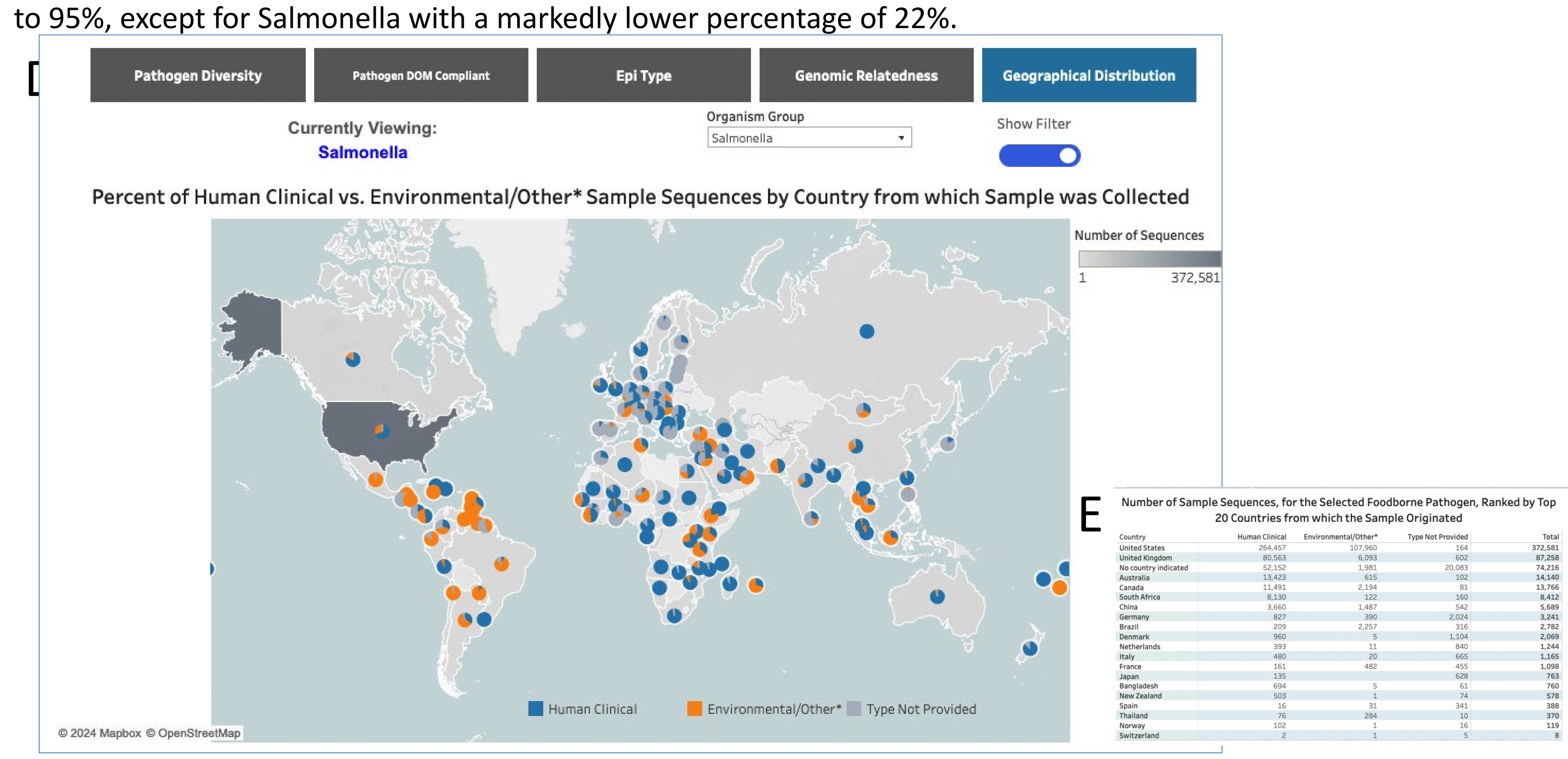


A. Pathogen Diversity displays the number of sequences published through NCBI's Pathogen Detection portal each year for 11 different foodborne pathogens. The Pathogen DOM (data object model) Compliant tab illustrates a similar distribution obtained by filtering for only those sequences that include raw data.

B. Epi Type presents cumulative inclusions for the same 11 pathogens, differentiated by NCBI sample category (Human Clinical and Environmental/Other). This illustrates the relationship between human clinical prevalence and overall representation in the dataset.

C: Genomic Relatedness: compares the percent of environmental/other sequences that fall within a given genomic distance of the nearest human clinical sample across the 11 most frequent pathogens.

For each pathogen, four grouped bars represent the different genomic relatedness in SNP distance for each pathogen. The 50 SNPs' category dominates the genomic relatedness across pathogens, with percentages up to 95%, except for Salmonella with a markedly lower percentage of 22%.



D. Interactive comparison of percent of environmental/other sequences across multiple countries, featuring filters that allow users to view specific pathogens. Additionally, a table is included, listing the top 20 countries by number of submitted sequences, providing a comprehensive view of demographic data from which the sample originated.

Conclusions

- Dashboards are an essential resource to communicate complex data in a straightforward and understandable way. There is a growing need to develop clear and concise visual representation of data. To tackle this need, this project developed a dashboard to identify trends, patterns, and outliers that might be missed in the metadata associated to whole genome sequencing data in the NCBI PD.
- This exploratory analysis of WGS data submitted to NCBI PD shows that nearly 100,000 records were submitted without raw sequencing data. Within organism groups, sequences with unassigned epi types ranged from 7 to 66% of all sequences due to submissions using incorrect BioSample metadata packages. More than 50 SNPs includes above 30% of data for organisms other than *Salmonella*.
- We expect these dashboard statistics will prompt NCBI submitters to address gaps in previously submitted metadata to allow for better performance of NCBI PD algorithms and more powerful analyses based on cleaner, more accurate data.