Leveraging Public Information about Pathogens for Disease Outbreak Investigations

Tonia Korves
Matthew Peterson
Wenling Chang

March 2013

Department No.: G62, E52X
Project No.: 51MSR601-BA

Bedford, MA

MITRE

Abstract

With recent technological advances in DNA sequencing and access to information via internet databases, the amount of information about pathogen strains in the public domain is growing rapidly. This information can be leveraged to help identify the origins of pathogens that cause disease outbreaks. This report describes potential uses for public data in outbreak investigations, key data types, the formats and locations of pathogen data in public sources, and tools MITRE is designing for assembling and integrating information during disease outbreak investigations.

Acknowledgements

The authors thank Lynette Hirschman for advice and helpful suggestions on this work and document.

Introduction

Pathogens are capable of causing disease outbreaks with devastating consequences, including loss of life, societal disruption, and large economic costs. High consequence disease outbreaks can be naturally-occurring such as the devastating 1918 global influenza epidemic, accidentally-triggered such as the 2010-2012 Haitian cholera epidemic, and deliberately-caused such as the anthrax attacks in the fall of 2001. From national security, law enforcement, and public health perspectives, it is important to determine where an outbreak pathogen came from, and whether the outbreak was natural, accidental, or intentional. This information is critical for preventing repeat outbreaks and attacks, and apprehending any perpetrators.

One way to discover the source of a pathogen is to identify other pathogen strains with shared biological properties, such as genetic and phenotypic features. If a pathogen strain that is virtually identical to an outbreak pathogen strain is found, then institutions that possess this strain may be the source of outbreak. If closely related strains are found and come from a particular geographical region, then these point to a possible geographic origin of the outbreak strain.

Today there is growing amount of information about pathogen strains based on advances in DNA sequencing and access to information via internet databases and bioinformatics tools. While some of this information is stored in privately held databases within government agencies, public health organizations, private strain collections, and individual labs, there is also a growing amount of data about pathogen strains available in the public domain. This includes multiple types of data present in a variety of formats and sources. If this public data can be readily assembled and integrated, it could be leveraged to quickly identify the origins of outbreak-causing pathogens.

Here, we summarize the relevant information that resides in public sources, focusing on DNA data and information about pathogen strains. We discuss potential uses of public strain data in investigations, outline key data types for identifying strain origins, describe the locations and formats of strain data in public sources, and discuss tools MITRE is designing for integrating information from multiple sources.

Uses for Public Strain Information in Disease Outbreak Investigations

Public information about pathogens can be used by investigators to discover and evaluate potential sources of a pathogen. Specifically, this information can help address the following:

Is the pathogen a known laboratory strain? Determining whether an outbreak pathogen is a known laboratory strain is important for differentiating natural outbreaks from deliberate outbreaks and laboratory accidents. Well-known laboratory strains were used in two

significant bioattacks within the US; a laboratory strain of *Bacillus anthracis* used in anthrax attacks in 2001 [1], and a orderable laboratory strain of *Salmonella enterica* Typhimurium was used in the 1984 salad bar attacks [2]. In addition, accidental infections with lab strains have been identified in the past [3]. To assess whether an outbreak pathogen is a known laboratory strain, it is necessary to have information about orderable and published strains; much of this information is in the public domain.

What geographical location and type of environment might the pathogen have come from? Identifying the geographical region that a pathogen came from is important for determining whether an outbreak pathogen emerged locally or was introduced from somewhere else. For example, the *Vibrio cholerae* strain that caused the recent Haitian cholera outbreak was found to be more similar to strains from Nepal than to local strains, leading to the conclusion that it was likely introduced by a United Nations worker [4, 5]. Insight into the type of environment an outbreak strain came from, such as animal hosts versus humans, could also help identify the source. Some information about the locations and environments that pathogen strains came from, their biological features, and associations between these can be assembled from public databases and published papers.

What persons and institutions have related strains? Identifying laboratories that have pathogen strains related to an outbreak strain is important for obtaining strains for further investigation. For example, to evaluate candidate origins in the Haitian cholera outbreak, investigators had to obtain *V. cholerae* strains from Nepal, other countries and local sources. In the case of an accidental or deliberate laboratory release, it is important to identify persons and institutions that have used or have access to a strain. To find laboratories with particular strains, information is needed about strain name synonyms, places where the strain can be ordered or purchased, and institutions and people that have used that strain, such as from published papers.

Answering each of these questions requires putting together different types of data and data from multiple sources. This makes the assembly and integration of data from different sources critical for investigations.

Types of Strain Information for Disease Outbreak Investigations

Two general types of information are needed: biological information to evaluate similarity between samples, and contextual information, also known as metadata, to identify where strains came from, and the institutions and persons that possess strains.

The most basic biological information is the taxonomic identity of a pathogen, including species, subspecies, pathovar, and/or serovar. This information is critical for identifying sets of strains in information sources to investigate further. Beyond this, several kinds of biological properties have been used in disease outbreak investigations (Table 1). These include DNA properties, genetically-determined phenotypes, biochemical properties, strain variation within a sample, and the presence of other species in a sample. Each of these is well-suited for certain microbial forensic uses (Table 1). For searching information sources

for related strains, genetic and genetically determined properties are particularly useful because they can be measured consistently across growth conditions, provide varying levels of resolution, and often can be captured in standard formats in databases. Genetic data takes various forms, including whole genome sequence, Multi Locus Sequence Typing (MLST), Multi-Locus Variable number tandem repeat Analysis (MLVA), Pulse Field Gel Electrophoresis (PFGE), optical mapping, gene sequences, and virulence gene presence [6].

Table 1. Biological properties of pathogen samples and their uses in microbial forensics

General Biological Property	Specific Examples	State of Technologies	Uses for Microbial Forensics	Examples of Use in Outbreaks
DNA sequence properties of a strain	Whole genome sequence; multi-locus sequence type; electrophoresis gel patterns	Established and emerging	Identifying related strains and evaluating matches to particular strains. Identifying genetic engineering.	[4, 7-14]
Genetically-determined phenotypic properties	Serotype; toxin production; antibiotic resistance	Established	Identifying related strains for further investigation	[2, 15]
Biochemical properties	Trace chemicals, biochemical composition, protein expression; DNA methylation patterns	Established and emerging	Differentiating natural origins from laboratory production, and identifying methods of lab production	[1]
Strain variation in a sample	Presence of multiple morphotypes or DNA variants in a sample	Established and emerging	Evaluating matches to potential source samples	[12]
Trace organisms	Presence of other organisms in an outbreak sample, which could be assessed via metagenomic DNA sequencing	Emerging	Evaluating matches to potential source samples and possibly identifying source environments	attempted use: [1]

Metadata types about pathogen strains that are useful for disease outbreak investigations are outlined in Table 2 and fall into a few categories: information about the original collection of a strain (particularly important for discovering natural sources of strains), the people and institutions that have a strain, the biological data that is available for a strain, strain identifiers (critical for identifying particular strains across multiple information sources and institutions), and the provenance of these data (e.g. database entry identifiers and cross reference identifiers, critical for integrating information from different sources.)

Table 2. Metadata types for disease outbreak investigations

Metadata Category	Metadata Types
Sample Identification	Strain Name
	Alternate Strain Names
	Organism Name
	Species/subspecies
	NCBI Taxon ID
Biological Data Types Available	e.g. Whole genome sequence (WGS), Multi-locus Sequence Type (MLST), Serotype, Virulence Gene Sequences
	DNA Sequencing Status
	DNA Sequencing Platform, Depth, and Assembly Method
Original Collection of the Strain	Collection Location -Latitude and Longitude, Country, City
	Date of Collection
	Environment Type, e.b. Host/Human-Associated
	Host Disease Symptoms
	Host Information (gender, age, travel history, health status)
	Person who originally collected
	Institution which collected and Institution Location
	Publication on collection
	Sample storage info
	Isolation and growth conditions
Proximate Source of the Strain	Lab, person, vendor, or culture collection sample was obtained from
(applicable if not a novel collection)	Parent Strain Name
	Source Institution Name and Location
	Source Persons/POC
	Any prior source history information (including prior strain names)
Information Source Metadata	Information Source Name
	Data Source Entry Identifier
	Project Name or Paper title
	Entry Date
	Organizations that submitted the information and their locations
	Names of submitters and contributors, and Contact info
	Cross-referenced identifiers, e.g. NCBI Accession, PubMed Ids

Public Information Sources with Pathogen Strain Data

There are now many information sources about pathogen strains, focused on different types of data and employing diverse formats. An overview of major pathogen information sources, and the types, formats, and amount of information they contain is given in Figure 1. Brief descriptions of some important publicly available resources, their data formats, and their potential uses in outbreak investigations are given in Table 3.

Some databases are designed principally for DNA sequence data. The primary repositories of sequence data in the United States are National Center for Biotechnology Information's (NCBI) Nucleotide database (GenBank; http://www.ncbi.nlm.nih.gov/genbank/), which houses information about gene sequences and complete and partially assembled genome sequences, and NCBI Sequence Reads Archive (SRA; http://www.ncbi.nlm.nih.gov/sra), which includes raw or minimally processed data from sequencing experiments. Information about the availability of sequences in these resources for particular species is now presented in NCBI Genome (http://www.ncbi.nlm.nih.gov/genome). Analogous sequence data repositories are the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) and the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/); together NCBI, ENA, and DDBJ form the International Nucleotide Sequence Data Collaboration (INSDC) and exchange sequence information daily. Major sequencing centers also post sequences on their own websites, both in addition to and in lieu of deposition to NCBI/EBI/DDBJ. Two leading sequencing centers for bacterial pathogen species that have sequence databases are Wellcome Sanger Trust Institute (in the UK; http://www.sanger.ac.uk/) and The Broad Institute (in the USA; http://www.broadinstitute.org/).

Other databases are designed for metadata about sequencing projects and strain collection. For pathogens, these include NCBI's recently updated BioProject database (http://www.ncbi.nlm.nih.gov/bioproject), NCBI's BioSample database which was created in 2011 (http://www.ncbi.nlm.nih.gov/biosample) [16], and Genomes Online (GOLD; http://www.genomesonline.org/cgi-bin/GOLD/index.cgi) [17]. Currently, the percent of data fields populated with original collection locations, dates, and environment types is low in these databases. When this information is present in BioProject and BioSample, it is found typically in free text in rather than in a structured format. However, efforts are being made to expand the capture of this information by introducing a few mandatory metadata fields for BioSample submission and to standardize data through the use of a controlled vocabulary for metadata types, called Minimum Information about any (x) Sequence (MIxS), which has been developed and promoted by the Genomic Standards Consortium (GSC) [18].

There are also pathogen-specific databases that focus on biological properties of genomes, standardization, and inclusion of metadata. PathoSystems Resource Integration Center (PATRIC; http://www.patricbrc.org/portal/portal/patric/Home) focuses on genome information for selected bacterial pathogens, with additional annotation, assembly, standardization, and comparison tools [19, 20]. The Pathogen-Annotated Tracking Resource Network (PATRN) system (www.patrn.net) focuses on foodborne bacterial pathogens and provides data manually curated from published papers in a structured,

searchable form [21]. This data includes laboratory results such as antibiotic resistance properties and various genetic marker assays and metadata. Some resources are focused on particular pathogens, for example, a database designed to support research on Tuberculosis hosted by the Broad Institute (http://tbdb.org/).

Databases that distinguish strains based on genetic markers are useful in identifying related strains for further study. This includes Multi Locus Sequence Typing (MLST) databases, a list of which can be found here: http://pubmlst.org/databases.shtml. MLST databases frequently contain a greater number of strains and more collection metadata than databases associated with genome sequence data. For some species, there are multiple MLST databases, which use different standards and genes, and consequently have different strain groupings. PulseNet uses Pulse Field Gel Electrophoresis for strain typing, and contains information about a very large number of strains, but is not publicly available. Other databases focus on pathogen virulence factors and antibiotic resistance genes. These include the Virulence Database at Lawrence Livermore National Lab (MvirDB; http://predictioncenter.llnl.gov/), Antibiotic Resistance Genes Database (ARDB; http://ardb.cbcb.umd.edu/), the Virulence Factor Database (VFDB; http://www.mgc.ac.cn/VFs/), and part of PATRIC.

Other resources are principally designed for conveying information about strains that can be ordered or purchased. These include catalogs of strains from culture collections, such as the American Type Culture Collection (ATCC; http://www.atcc.org/) and BEI Resources (http://www.beiresources.org/), and vendors that sell strains for diagnostic and research purposes. StrainInfo (http://www.straininfo.net/) is a database that aggregates information about strain names and strain synonyms from various global culture collections.

Finally, published papers are a critical source of information for biological, collection, person, and institution data, and contain much information in free text and tables that is not captured in structured databases. Abstracts for a good portion of biomedical literature can be found in MEDLINE (http://www.nlm.nih.gov/bsd/pmresources.html) or PubMed (http://www.ncbi.nlm.nih.gov/pubmed); full text articles for more recent publications can be found in PubMed Central (PMC; http://www.ncbi.nlm.nih.gov/pmc/), and journal websites serve as portals to this information.

Information Sources	Biological Data: Serotype	Biological Data: Assembled Whole Genome Sequence	Biological Data: Short Reads	Biological Data: MLST	Biological Data: Virulence Gene Profile	Metadata: Strain Name	Metadata: Strain Synonyms	Metadata: People and Institutions with Strain	Metadata: Original Collection Info	Number of Records per Pathogen Species: *Escherichia coli*	*Salmonella enterica*	*Bacillus cereus* and *B. anthracis*	*Burkholderia pseudomallei*
NCBI Nucleotide										383,628	199,150	28,198	36,820
Assembled Genomes/Contigs										580	294	135	31
NCBI SRA										3,701	2,423	309	200
NCBI BioProject										2,870	1,382	341	91
NCBI BioSample										4,086	3,458	206	218
GOLD										1,885	947	187	45
PATRIC										474	207	131	31
Broad Institute										139	0	59	0
Wellcome Trust Sanger Institute										1,050	2,815	0	454
PATRN										8,365	17,439	1,444	0
MLST Databases†										5,292	5,803	1,368	3,113
StrainInfo										6,848	1,015	428	44
ATCC Culture Collection										743	330	79	0
PubMed										314,154	37,754	11,135	1,575
Pub Med Central										144,719	10,597	6,214	1,203

Figure 1. Overview of information content of selected public information sources for pathogen strains. Dark blue—structured format; light blue—unstructured format (e.g. free text only or in tables with varying formats); gray-not present. Number of records based on searches on Feb 14, 2013 for PATRN and Jan 31-Feb 1, 2013 for all other sources. †Species record numbers obtained from: http://mlst.ucc.ie/mlst/dbs/Ecoli ; http://mlst.ucc.ie/mlst/dbs/Senterica; http://pubmlst.org/bcereus/; http://bpseudomallei.mlst.net/

Table 3. Public information sources with pathogen data and their potential uses for disease outbreak investigations

Information Source	Content	Download Formats	Potential Uses
Genetic Databases			
NCBI Nucleotide (GenBank)	Whole genome shotgun data, assembled genomes, individual genes, contributor information, sometimes with strain metadata as free text/non standardized form	GenBank data is available through FTP, as well as web services	Finding sequence matches, downloading sequences for comparative analyses. Can search with BLAST or based on metadata
NCBI Sequence Read Archive (SRA)	Raw DNA sequence data and links to other files for metadata. Often there are no strain names, and there are many pools of strains here. Also includes RNA-seq experiments and studies of mutants	Available via FTP or via web tools. For selected records, can use "Send to" to download an info file in Excel format.	Finding sequence matches (after some assembly) for genomes that have not been released in assembled formats
NCBI Genome	Provides a dendrogram of relationships between some strains with assembled genomes, a list of sequencing projects with BioProject cross refs, number of plasmids, plasmid sequence ids, and other basic information. Gives portal to BLAST only genomes of that species	Linked Nucleotide records can be downloaded via web interface or FTP	A way to quickly find the available and in progress genome sequences for a species in NCBI
MLST databases (separate ones for different species, run by various groups)	Multilocus Sequence Types (MLST), usually strain names and metadata; these databases contain many more strains than NCBI nucleotide records, and sometimes have well populated, structured metadata	Varies among databases. Some can be download as CSV	Finding related strains with same MLST type as a pathogen strain; finding the geographic distribution of previously collected, related strains; searching by metadata such as location.
Sequencing Center Databases			
Wellcome Trust Sanger Institute	Contains sequence data (frequently as raw reads) ahead of submission to public repositories. In some cases there is limited metadata. Site has data for *B. pseudomallei, Campylobacter jejuni*, *C. coli, E. coli*, *Salmonella*, *V. cholerae*, *Yersinia pestis*, and *Y. enterocolitica*	Sequence data downloadable via FTP. Also has BLAST function. Metadata appears to be put into EBI (European Bioinformatics Institute), with links to EBI sample pages	Downloading sequences for comparative analyses; possibly using linked strain metadata on Attributes pages in EBI (but many metadata fields are currently empty)

Information Source	Content	Download Formats	Potential Uses
The Broad Institute	Has sequence data, metadata, and gene annotations for strains used in Broad studies; includes draft assemblies for strains that are not assembled in NCBI, and also often has more metadata about these strains than in NCBI. Species of interest include: *Brucella*, *B. cereus*, *Listeria*, *Francisella*, 2011 *E. coli* outbreak and antibiotic resistant strains, *V. cholerae*	Data is organized by organism/project. Sequence data is downloadable for individual strains. Metadata format varies among organisms (eg, downloadable text, tables downloadable in Excel and as text, and tables embedded in webpage).	Downloading sequences for comparative analyses, including sequences and assemblies not available in NCBI; retrieving strain metadata not in NCBI
Sample and Project Databases			
NCBI BioProject	NCBI database for project metadata, including strain name, experimental methods, and submitter info	Can download in BioProject XML format via FTP. The XML files contain additional info compared to screen results, especially regarding submitter metadata.	Finding strains that people are working on, the kinds of data available or being collected for strains, submitter/researcher information, and sometimes metadata about the strain.
NCBI BioSample	A relatively new NCBI database for strain metadata	Can download BioSample in XML format via FTP	Finding strains that people are working on, including those for which there is no publication or released data yet; searching on and obtaining strain metadata (at least when it is better populated in the future).
Genomes OnLine (GOLD)	Lists metadata about complete and in progress sequencing projects, with links to where the sequence data can be found	Website does not provide a downloadable data file, but this may be available by request	Potentially, for finding strains based on metadata and pathovar information (but metadata content is currently rather sparse), and contact information for obtaining more data

Information Source	Content	Download Formats	Potential Uses
Pathogen-Specific Databases			
PATRIC	Contains sequence, annotation, and metadata about important human pathogens. Includes information about virulence gene content, and a database with strain metadata. Contains a subset of the genomes in NCBI.	Structured strain metadata and virulence gene metadata, downloadable as text or Excel files. Downloadable sequence data	Comparison to well-characterized, sequenced strains; identifying gene and genomic island content of strains. Has a tool for finding genomes based on similarities. A source of collection dates and source countries in a structured format.
PATRN	Genetic and phenotypic properties of foodborne pathogen strains and metadata, curated from published papers. Extensive data on *Salmonella*, *E. coli*, and *Listeria*.	No direct downloading available; provides a private workspace and tools for working with these data and imported data.	Finding related strains based on a variety of biological properties, and analyzing relationships between strains based on these properties. Potentially much faster than a de novo search of published literature.
Culture Collection Databases			
StrainInfo	Contains strain names, synonyms, and relationships among isolates for strains available in culture collections. Also has links to other databases and publications, and links to sequences for these strains	There are three files downloadable in XML or CSV. Strain Passports, which contain strain synonyms, are not downloadable in bulk but can be downloaded individually in XML format	Finding strain name synonyms, strain isolate relationships, and stock centers where strains can be obtained from globally. The list of stock centers and synonyms is extensive, but not comprehensive. StrainInfo may also be useful for searching for sequence matches to culture collection strains.
American Type Culture Collection (ATCC)	Contains information about orderable strains in this collection, including contributor, and sometimes strain collection, history, and antigenic properties.	Not able to download directly.	Finding what strains can be ordered from this collection. Note, there are some strain identifiers here that are not included in StrainInfo.

Tools for Assembling Public Data for Disease Outbreak Investigations

To facilitate the use of this public data for disease outbreak investigations, MITRE is conducting internally funded research on methods for assembling, and integrating data relevant to a case. The research team is designing a set of tools for data assembly and integration for several key sources, with a focus on strain metadata. The basic approach is to import and parse relevant strain data into a single data platform designed for the capture of strain data needed for outbreak investigations. This platform is being built by creating modules in LabKey, a data management server for biological data [22]. LabKey is an extensible platform, allowing for additional functionality through the creation of modules, and the software can interface with previously deployed applications through the use of a variety of client APIs. LabKey also offers data via a user and role based permissions system, and visualization and collaboration tools. Because LabKey modules can be designed for various types of laboratory results, the use of this platform can also enable integration of lab results and data from public and non-public information sources. For additional analyses, LabKey data can be exported to other programs, such as Analyst's Notebook or Palantir. A feature of LabKey that makes it especially useful for outbreak analyses is the ability for a user to design their own data tables. These tables, known as "Lists" in LabKey, allow for the creation of an agile data model in support of an investigation by the end user.

The research team has developed methods for importing data into LabKey from several sources (Figure 2), including a program for automated import of MEDLINE data. To capture and integrate stain metadata, the team has built a database schema in PostgreSQL for the metadata types listed in Table 2. To automatically parse and import data from NCBI BioProject and BioSample to this database, a program has been written using Java, Python, AWK, and Shell scripts. Data from other sources, such as MLST databases, can be imported directly into LabKey as Lists and jointly queried with the metadata database. Tools are also being incorporated to map geographic collection locations of strains and the locations of laboratories that have published on them. To evaluate and further develop these tools, the team is working through a mock outbreak investigation of a *Salmonella enterica* strain.

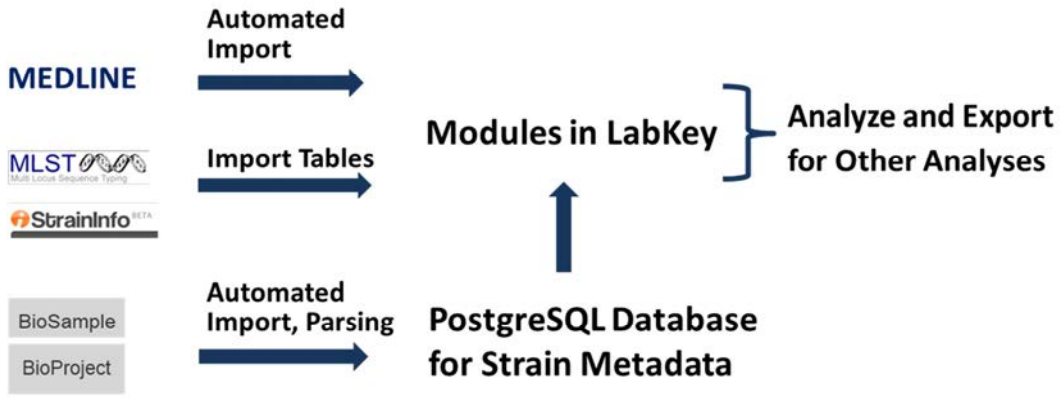

Figure 2. Overview of tools MITRE is designing for integration of pathogen strain data

Future Prospects and Challenges for Using Public Pathogen Data for Disease Outbreak Investigations

Given the anticipated growth in sequence data for pathogen strains due to sequencing technology advances and current efforts to enhance metadata capture, pathogen strain information is poised to become more plentiful and useful for disease outbreak investigations in the near future. Sequence data is expected to continue to grow exponentially. For example, Dr. W. Klimke of NCBI mentioned that 5 million new genome sequences are anticipated in the next five years (talk at Disease Outbreak Detection in the Genomics Era Meeting, Sept 2012). In addition, the "100K Genome Project" aims to sequence 100K additional foodborne pathogen genomes within the next five years, principally for bacteria pathogens (http://100kgenome.vetmed.ucdavis.edu/). It is anticipated that NCBI, with its European and Japanese partners, will continue to be the principal repository for sequence data. NCBI is also working on improving the submission of metadata, and there are international efforts to promote the capture and standardization of metadata for pathogen strains (Disease Outbreak Detection in the Genomics Era Meeting, Sept 2012). NCBI is also working on rapid assembly of submitted sequence reads into genomes while other researchers are developing methods for typing strains based on sequence read data [23, 24]; these efforts should make SRA records more readily useful for disease outbreak analyses.

Nevertheless, substantial challenges remain for capturing and integrating biological data for investigations. These challenges include: sequence data being present in different databases and in different formats, scientists not making sequence data or metadata public due to the effort involved in submitting data to NCBI and other public databases, metadata stored inconsistently in different metadata fields, the presence of important metadata in free text that is difficult to extract and integrate for structured analyses, and lack of standardization of strain names.

References

[1] National Research Council, "Review of the scientific approaches used during the FBI's investigation of the 2001 anthrax letters," 2011.

[2] T. J. Torok, R. V. Tauxe, R. P. Wise, J. R. Livengood, R. Sokolow, S. Mauvais, K. A. Birkness, M. R. Skeels, J. M. Horan, and L. R. Foster, "A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars," *JAMA,* vol. 278, pp. 389-395, Aug 1997.

[3] CDC, "Investigation update: Human *Salmonella* Typhimurium infections associated with exposure to clinical and teaching microbiology laboratories," 2012.

[4] C.-S. Chin, J. Sorenson, J. B. Harris, W. P. Robins, R. C. Charles, R. R. Jean-Charles, J. Bullard, D. R. Webster, A. Kasarskis, P. Peluso, E. E. Paxinos, Y. Yamaichi, S. B. Calderwood, J. J. Mekalanos, E. E. Schadt, and M. K. Waldor, "The origin of the Haitian cholera outbreak strain," *New Engl J Med,* vol. 364, pp. 33-42, 2011.

[5] R. S. Hendriksen, L. B. Price, J. M. Schupp, J. D. Gillece, R. S. Kaas, D. M. Engelthaler, V. Bortolaia, T. Pearson, A. E. Waters, B. Prasad Upadhyay, S. Devi Shrestha, S. Adhikari, G. Shakya, P. S. Keim, and F. M. Aarestrup, "Population genetics of *Vibrio cholerae*

from Nepal in 2010: Evidence on the origin of the Haitian outbreak," *mBio,* vol. 2, Jul/Aug 2011.

[6] B. A. Sabat AJ, Nashev D, Sá-Leão R, van Dijl JM, Laurent F, Grundmann H, Friedrich AW, on behalf of the ESCMID Study Group of Epidemiological Markers (ESGEM), "Overview of molecular typing methods for outbreak detection and epidemiological surveillance,"*Euro Surveill,* vol. 18, p. 20380, 2013.

[7] J. L. Gardy, J. C. Johnston, S. J. H. Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. M. Jones, F. S. L. Brinkman, R. C. Brunham, and P. Tang, "Whole-genome sequencing and social-network analysis of a tuberculosis outbreak," *New Engl J Med,* vol. 364, pp. 730-739, 2011.

[8] P. Laksanalamai, L. A. Joseph, B. J. Silk, L. S. Burall, C. L. Tarr, P. Gerner-Smidt, and A. R. Datta, "Genomic characterization of *Listeria monocytogenes* strains involved in a multistate listeriosis outbreak associated with cantaloupe in US," *PLoS ONE,* vol. 7, p. e42448, 2012.

[9] E. K. Lienau, E. Strain, C. Wang, J. Zheng, A. R. Ottesen, C. E. Keys, T. S. Hammack, S. M. Musser, E. W. Brown, M. W. Allard, G. Cao, J. Meng, and R. Stones, "Identification of a salmonellosis outbreak by means of molecular sequencing," *New Engl J Med,* vol. 364, pp. 981-982, 2011.

[10] A. Mellmann, D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S. F. McLaughlin, J. K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P. M. Brzoska, R. L. Moore, S. Guenther, J. M. Rothberg, and H. Karch, "Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology," *PLoS ONE,* vol. 6, p. e22751, 2011.

[11] D. A. Rasko, D. R. Webster, J. W. Sahl, A. Bashir, N. Boisen, F. Scheutz, E. E. Paxinos, R. Sebra, C.-S. Chin, D. Iliopoulos, A. Klammer, P. Peluso, L. Lee, A. O. Kislyuk, J. Bullard, A. Kasarskis, S. Wang, J. Eid, D. Rank, J. C. Redman, S. R. Steyert, J. Frimodt-MÃ¸ller, C. Struve, A. M. Petersen, K. A. Krogfelt, J. P. Nataro, E. E. Schadt, and M. K. Waldor, "Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany," *New Engl J Med,* vol. 365, pp. 709-717, 2011.

[12] D. A. Rasko, P. L. Worsham, T. G. Abshire, S. T. Stanley, J. D. Bannan, M. R. Wilson, R. J. Langham, R. S. Decker, L. Jiang, T. D. Read, A. M. Phillippy, S. L. Salzberg, M. Pop, M. N. Van Ert, L. J. Kenefic, P. S. Keim, C. M. Fraser-Liggett, and J. Ravel, "*Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation," *Proc Natl Acad Sci U S A,* vol. 108, pp. 5027-5032, Mar 2011.

[13] E. S. Snitkin, A. M. Zelazny, P. J. Thomas, F. Stock, N. C. S. Program, D. K. Henderson, T. N. Palmore, and J. A. Segre, "Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing," *Sci Transl Med,* vol. 4, p. 148ra116, Aug 2012.

[14] A. M. Wright, S. B. Beres, E. N. Consamus, S. W. Long, A. R. Flores, R. Barrios, G. S. Richter, S.-Y. Oh, G. Garufi, H. Maier, A. L. Drews, K. E. Stockbauer, P. Cernoch, O. Schneewind, R. J. Olsen, and J. M. Musser, "Rapidly progressive, fatal, inhalation anthrax-like infection in a human: Case report, pathogen genome sequencing, pathology, and coordinated response," *Arch Pathol Lab Med,* vol. 135, pp. 1447-1459, 2011.

[15] M. N. E. Scheutz F, Frimodt-Møller J, Boisen N, Morabito S, Tozzoli R, Nataro JP, Caprioli A. , "Characteristics of the enteroaggregative Shiga toxin/verotoxin-producing *Escherichia coli* O104:H4 strain causing the outbreak of haemolytic uraemic syndrome in Germany, May to June 2011," *Euro Surveill,* vol. 16, 2011.

[16] T. Barrett, K. Clark, R. Gevorgyan, V. Gorelenkov, E. Gribov, I. Karsch-Mizrachi, M. Kimelman, K. D. Pruitt, S. Resenchuk, T. Tatusova, E. Yaschenko, and J. Ostell, "BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata," *Nucleic Acids Res,* vol. 40, pp. D57-D63, Jan 2012.

[17] I. Pagani, K. Liolios, J. Jansson, I. M. A. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides, "The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata," *Nucleic Acids Res,* vol. 40, pp. D571-D579, 2011.

[18] P. Yilmaz, R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I. Karsch-Mizrachi, A. Johnston, G. Cochrane, R. Vaughan, C. Hunter, J. Park, N. Morrison, P. Rocca-Serra, P. Sterk, M. Arumugam, M. Bailey, L. Baumgartner, B. W. Birren, M. J. Blaser, V. Bonazzi, T. Booth, P. Bork, F. D. Bushman, P. L. Buttigieg, P. S. G. Chain, E. Charlson, E. K. Costello, H. Huot-Creasy, P. Dawyndt, T. DeSantis, N. Fierer, J. A. Fuhrman, R. E. Gallery, D. Gevers, R. A. Gibbs, I. S. Gil, A. Gonzalez, J. I. Gordon, R. Guralnick, W. Hankeln, S. Highlander, P. Hugenholtz, J. Jansson, A. L. Kau, S. T. Kelley, J. Kennedy, D. Knights, O. Koren, J. Kuczynski, N. Kyrpides, R. Larsen, C. L. Lauber, T. Legg, R. E. Ley, C. A. Lozupone, W. Ludwig, D. Lyons, E. Maguire, B. A. Methe, F. Meyer, B. Muegge, S. Nakielny, K. E. Nelson, D. Nemergut, J. D. Neufeld, L. K. Newbold, A. E. Oliver, N. R. Pace, G. Palanisamy, J. Peplies, J. Petrosino, L. Proctor, E. Pruesse, C. Quast, J. Raes, S. Ratnasingham, J. Ravel, D. A. Relman, S. Assunta-Sansone, P. D. Schloss, L. Schriml, R. Sinha, M. I. Smith, E. Sodergren, A. Spor, J. Stombaugh, J. M. Tiedje, D. V. Ward, G. M. Weinstock, D. Wendel, O. White, A. Whiteley, A. Wilke, J. R. Wortman, T. Yatsunenko, and F. O. Glockner, "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications," *Nat Biotech,* vol. 29, pp. 415-420, 2011.

[19] T. Driscoll, J. L. Gabbard, C. Mao, O. Dalay, M. Shukla, C. C. Freifeld, A. G. Hoen, J. S. Brownstein, and B. W. Sobral, "Integration and visualization of host-pathogen data related to infectious diseases," *Bioinformatics,* vol. 27, pp. 2279-2287, 2012.

[20] J. J. Gillespie, A. R. Wattam, S. A. Cammer, J. L. Gabbard, M. P. Shukla, O. Dalay, T. Driscoll, D. Hix, S. P. Mane, C. Mao, E. K. Nordberg, M. Scott, J. R. Schulman, E. E. Snyder, D. E. Sullivan, C. Wang, A. Warren, K. P. Williams, T. Xue, H. Seung Yoo, C. Zhang, Y. Zhang, R. Will, R. W. Kenyon, and B. W. Sobral, "PATRIC: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species," *Infect Immun,* vol. 79, pp. 4286-4298, Nov 2011.

[21] G. Gopinath, K. Hari, R. Jain, M. K. Mammel, M. H. Kothary, A. A. Franco, C. J. Grim, K. G. Jarvis, V. Sathyamoorthy, L. Hu, A. R. Datta, I. R. Patel, S. A. Jackson, J. Gangiredla, M. L. Kotewicz, J. E. LeClerc, M. Wekell, B. A. McCardell, M. D. Solomotis, and B. D. Tall, "The Pathogen-annotated Tracking Resource Network (PATRN) system: A web-based resource to aid food safety, regulatory science, and investigations of foodborne pathogens and disease," *Food Microbiol,* 2013.

[22] E. Nelson, B. Piehler, J. Eckels, A. Rauch, M. Bellew, P. Hussey, S. Ramsay, C. Nathe, K. Lum, K. Krouse, D. Stearns, B. Connolly, T. Skillman, and M. Igra, "LabKey Server: An

open source platform for scientific data integration, analysis and collaboration," *BMC Bioinformatics,* vol. 12, p. 71, 2011.

[23] M. Inouye, T. Conway, J. Zobel, and K. Holt, "Short read sequence typing (SRST): multi-locus sequence types from short reads," *BMC Genomics,* vol. 13, p. 338, 2012.

[24] M. V. Larsen, S. Cosentino, S. Rasmussen, C. Friis, H. Hasman, R. L. Marvig, L. Jelsbak, T. Sicheritz-PontÃ©n, D. W. Ussery, F. M. Aarestrup, and O. Lund, "Multilocus sequence typing of total-genome-sequenced bacteria," *J Clin Microbiol,* vol. 50, pp. 1355-1361, Apr 2012.