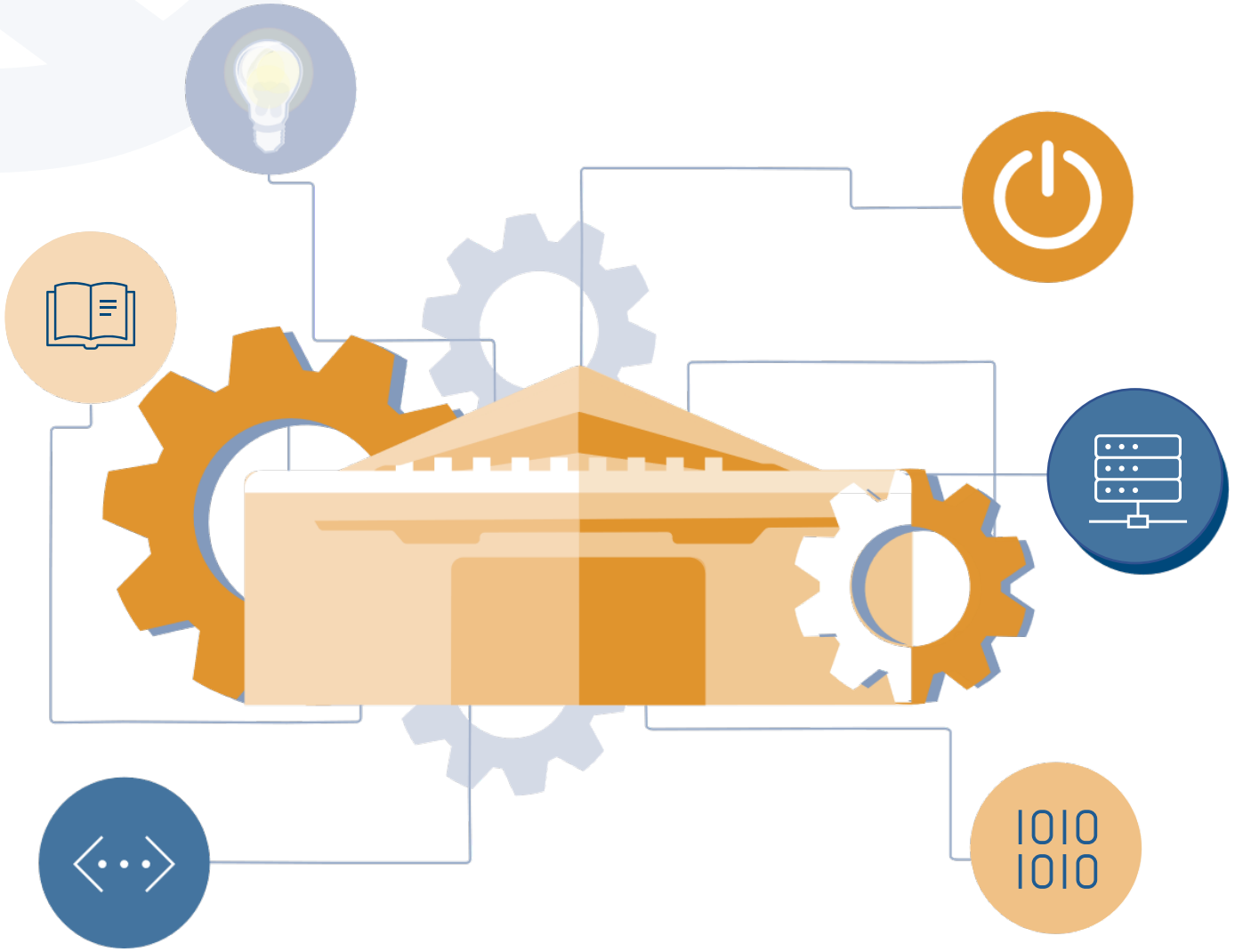
# Current State of Infrastructure and Interoperability of Pathogen Genomics Data in the U.S.




**Steve Sherry, Ph.D.**

**Acting Director, National Library of Medicine, NIH**

# Public sequence data supports pathogen surveillance and outbreak detection

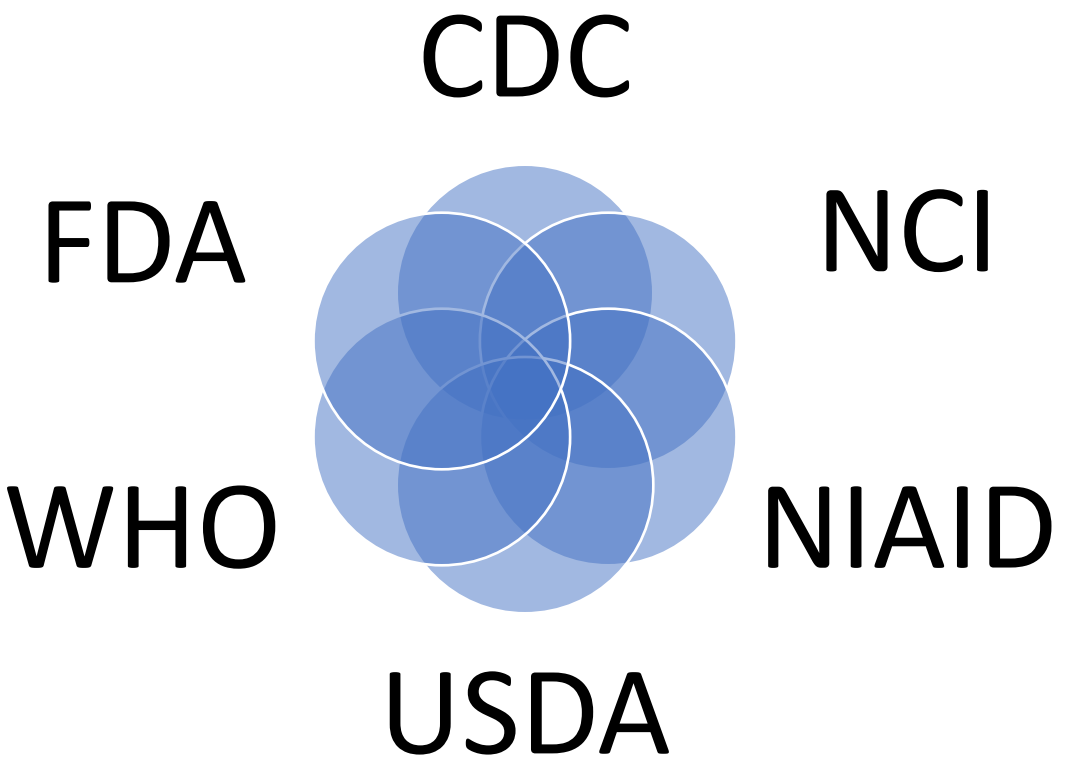
- Rapidly identify sources of outbreak to target interventions and remove threats
  - Foodborne illness
  - Contaminated medical devices
- Identify common sources of infection and target initiatives aimed at reduction
- Detecting emerging pathogen threats
  - Identify novel types or species that are beginning to cause disease
  - Identify antimicrobial resistance

# Public sequence data supports pathogen surveillance and outbreak detection

- Can accelerate research across multiple disciplines such as developing interventions or vaccines
- Monitoring ongoing pathogen evolution
  - Evolving resistance to vaccines
  - Evolving genetic resistance to therapeutics such as antibacterial or antiviral agents
  - Understanding pathogen evolution
    - what are the genetic components of virulence and pathogenicity in the pathogen that cause disease
    - What are the host or environmental factors that contribute to the spread of disease

# Pathogen Genomics Priorities: Surveillance vs Outbreak

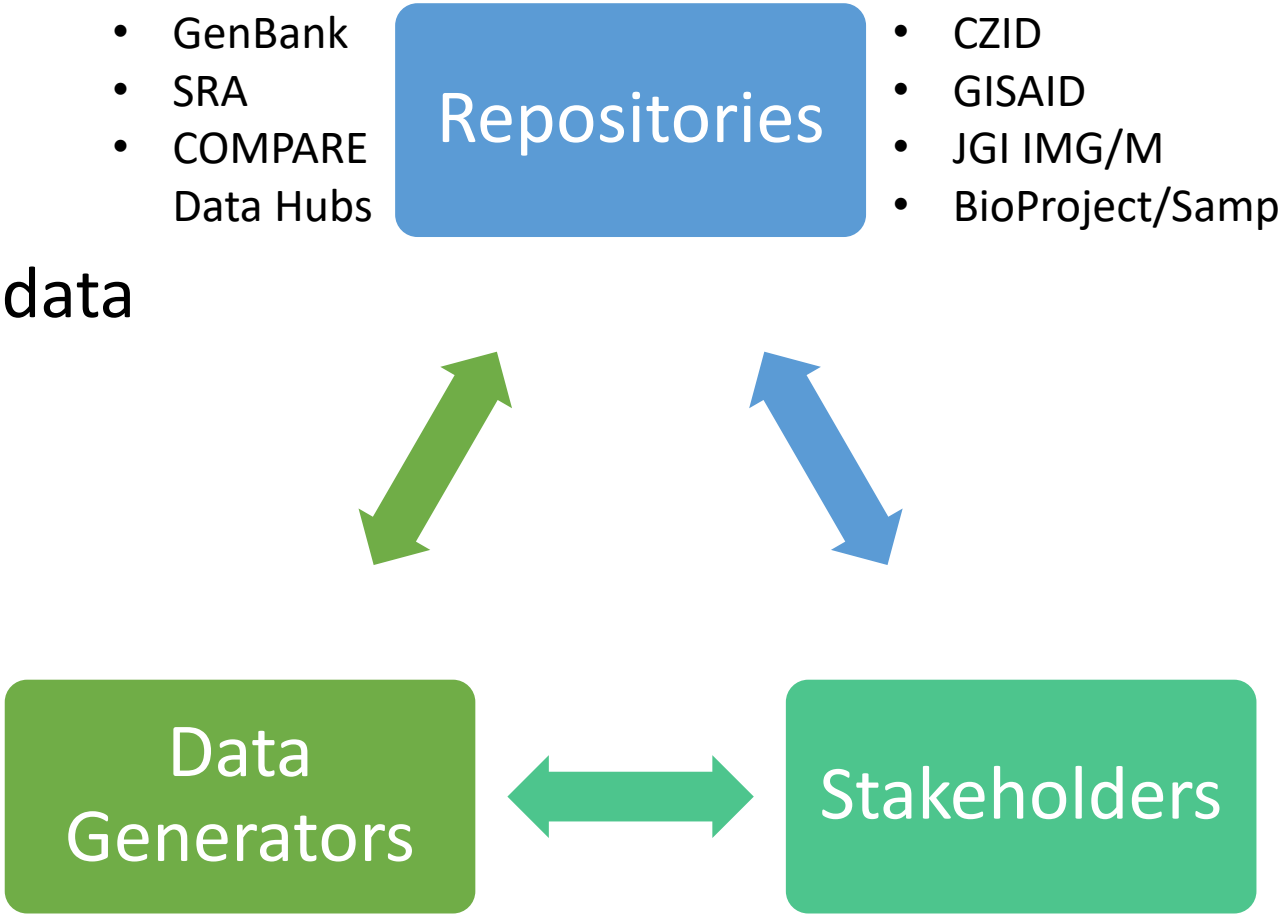
- Multiple groups maintain different lists of high-priority pathogens
  - Dynamic
  - Reflect different interests
- Prioritized pathogens can have different infrastructure and data management requirements:
  - chronic, ongoing relationship associated with monitoring
  - Shorter term, outbreak-related



# Developing a Pathogen Genomics Data Infrastructure

- Public sequence data
- Clinical data
- Research data

- GenBank
- SRA
- COMPARE Data Hubs




- CZID
- GISAID
- JGI IMG/M
- BioProject/Sample

- Public health
- Academia
- Government
- Non-profits
- Industry/pharma

# Pathogen Detection (PD) Development Timeline

**Funding support via NIAID from CARB**

**Establish minimal metadata standards for submission**

**Begin foodborne pilot project with *Listeria monocytogenes***

First release of Isolates Browser web interface for analysis results

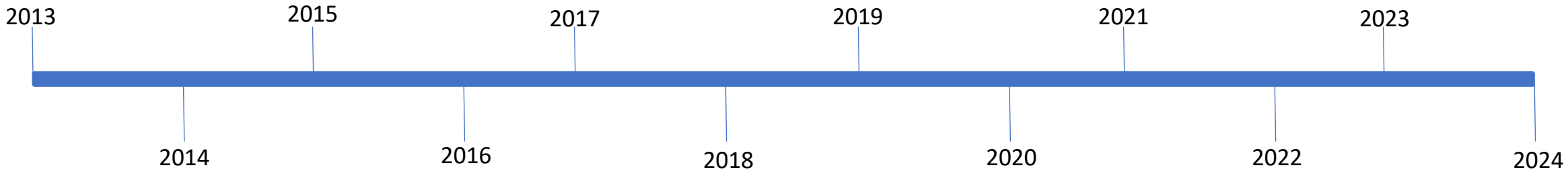
**Genomics for Food and Feed Safety (Gen-FS) Interagency Group Established**

Browser used for identification of international *L. monocytogenes* outbreak

**Development of ISO standard for WGS for food pathogens**

**Updated metadata standards for foodborne pathogens**

Work begins on backfilling 600K previously submitted samples to meet new standards




**Whitehouse initiative on Combating Antibiotic Resistant Bacteria (CARB)**

PD takes over beta-lactamase allele assignments from Lahey lab

First mobile colistin resistance (*mcr*) gene identified in China

By community request – PD team takes over *mcr* allele assignments

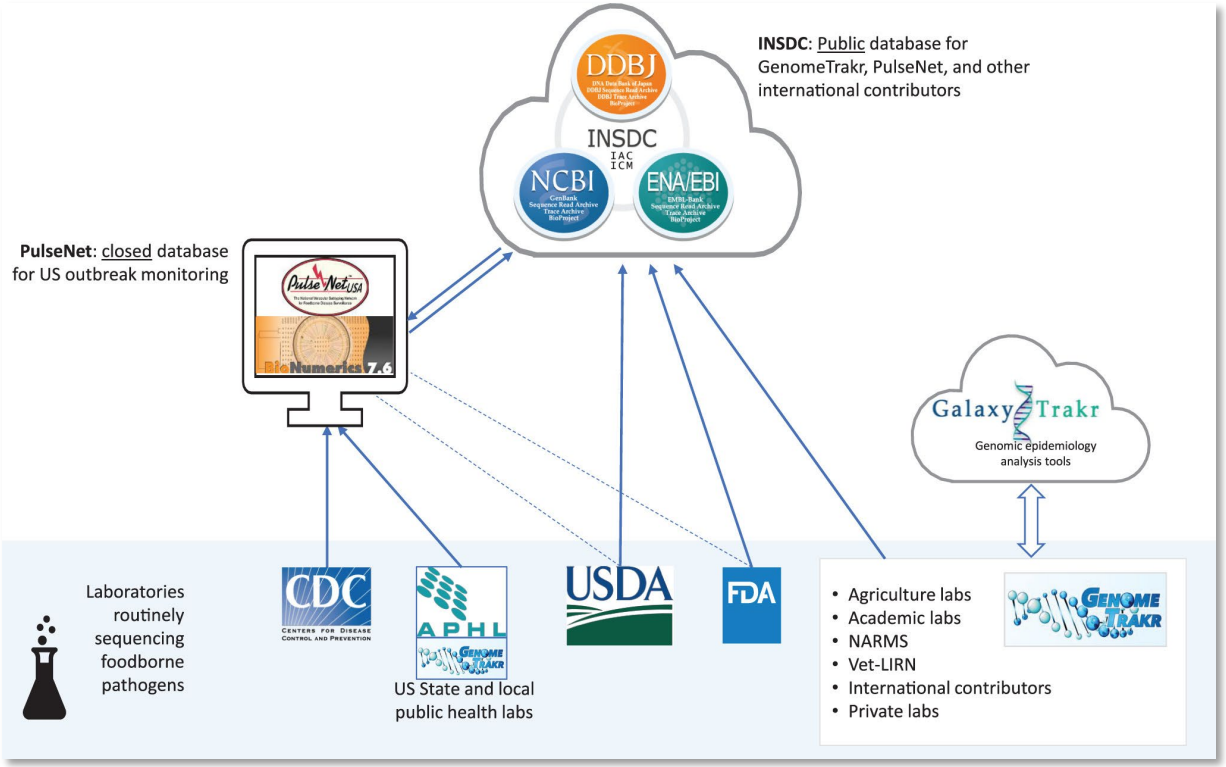
New *mcr*-9 family found frequently in *Salmonella isolates* in PD - FDA group showed it does not confer resistant – PD team downgrades the gene

Microbial resistance map released

# Interagency Collaboration for Genomics for Food and Feed Safety (Gen-FS)

- National Institutes of Health (NIH)
- Centers for Disease Control and Prevention (CDC)
- U.S. Food and Drug Administration (FDA)
- U.S. Department of Agriculture (USDA)
    - Food Safety and Inspection Service (FSIS)
    - Agricultural Research Service (ARS)
    - Animal and Plant Health Inspection Service (APHIS)

- Standardization of isolate metadata for submission
- Harmonized sequencing protocols across CDC PulseNet & FDA GenomeTrakr
- Harmonized quality assurance measures & accompanying quality control checks
- Ensures all WGS data generated by Gen-FS members in the Pathogen Detection database at NCBI meet the minimum quality standards

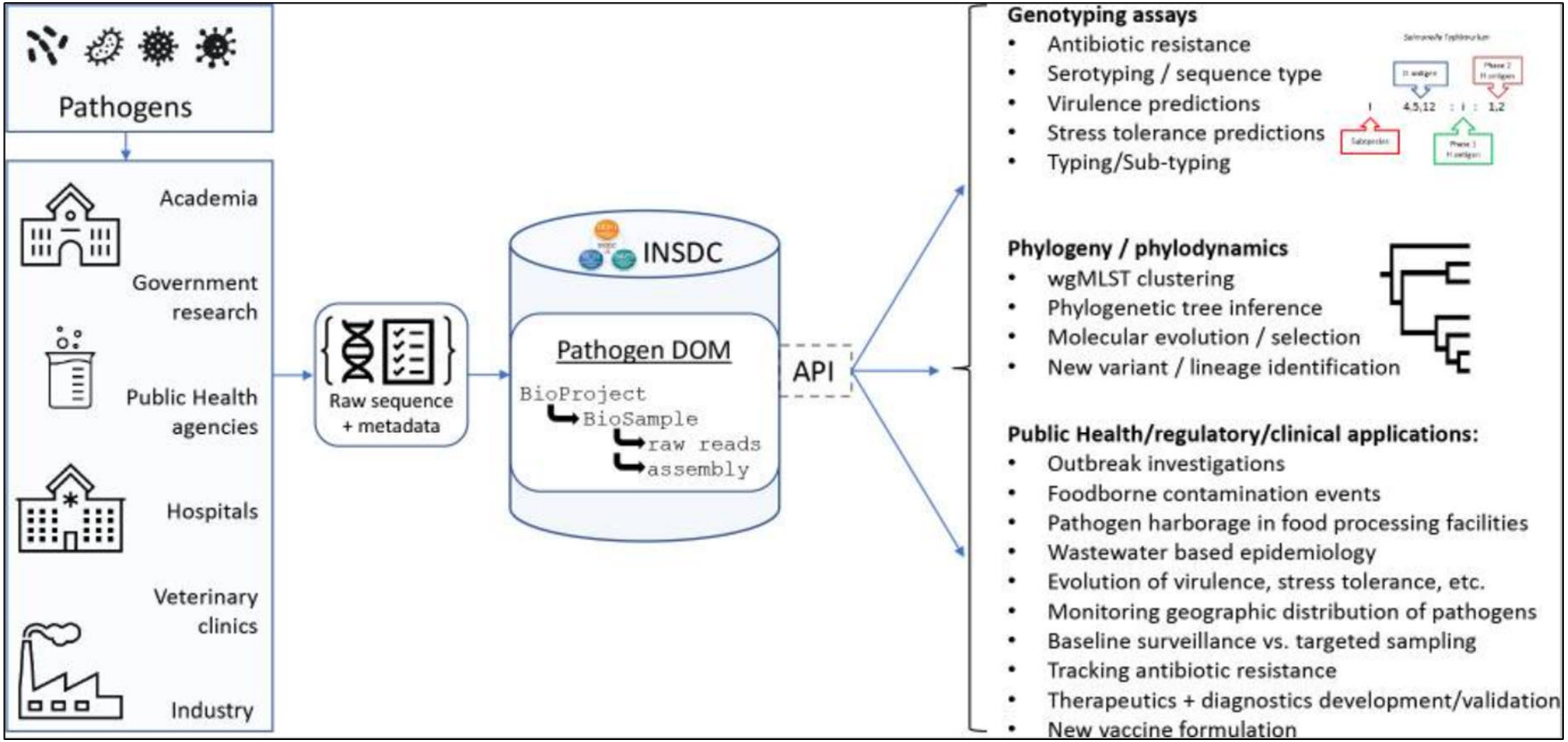


# PHA4GE

- PHA4GE- a collaborative community working to:
  - Advance the use of open data
  - Empower laboratories to analyze and govern their own data
  - Encourage standardized data structures and interchange formats

Public Health Alliance for Genomic Epidemiology

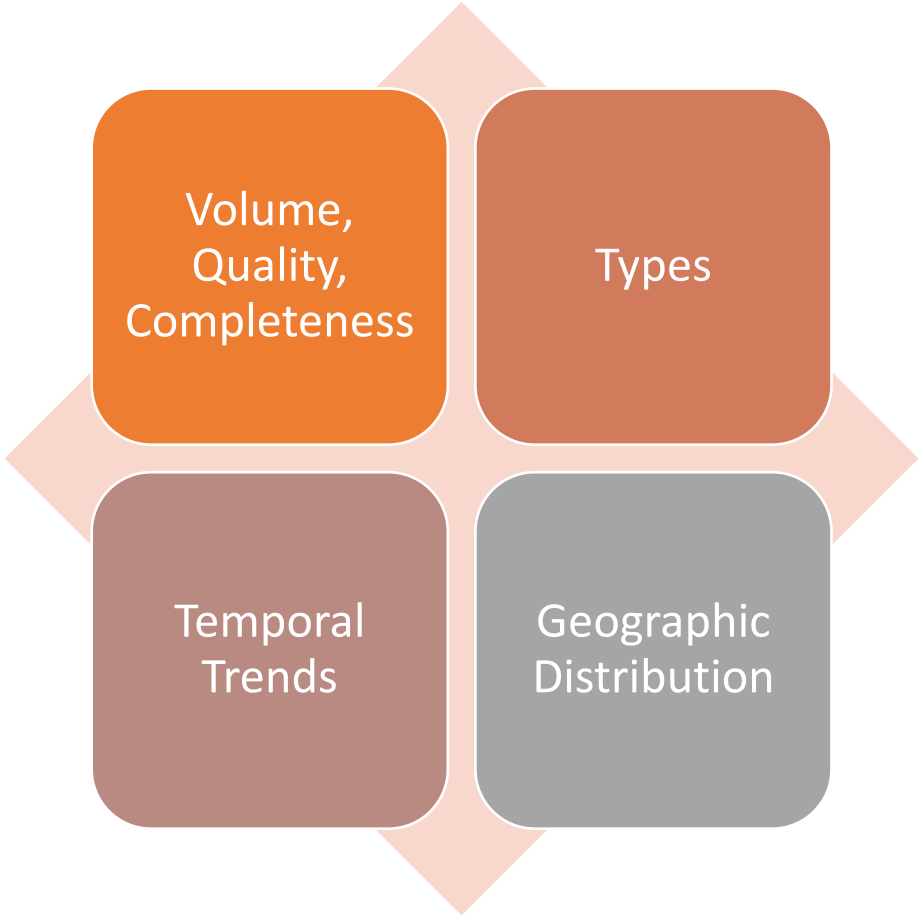
# The INSDC compliant Pathogen Data Object Model to structure genomic data for public health applications

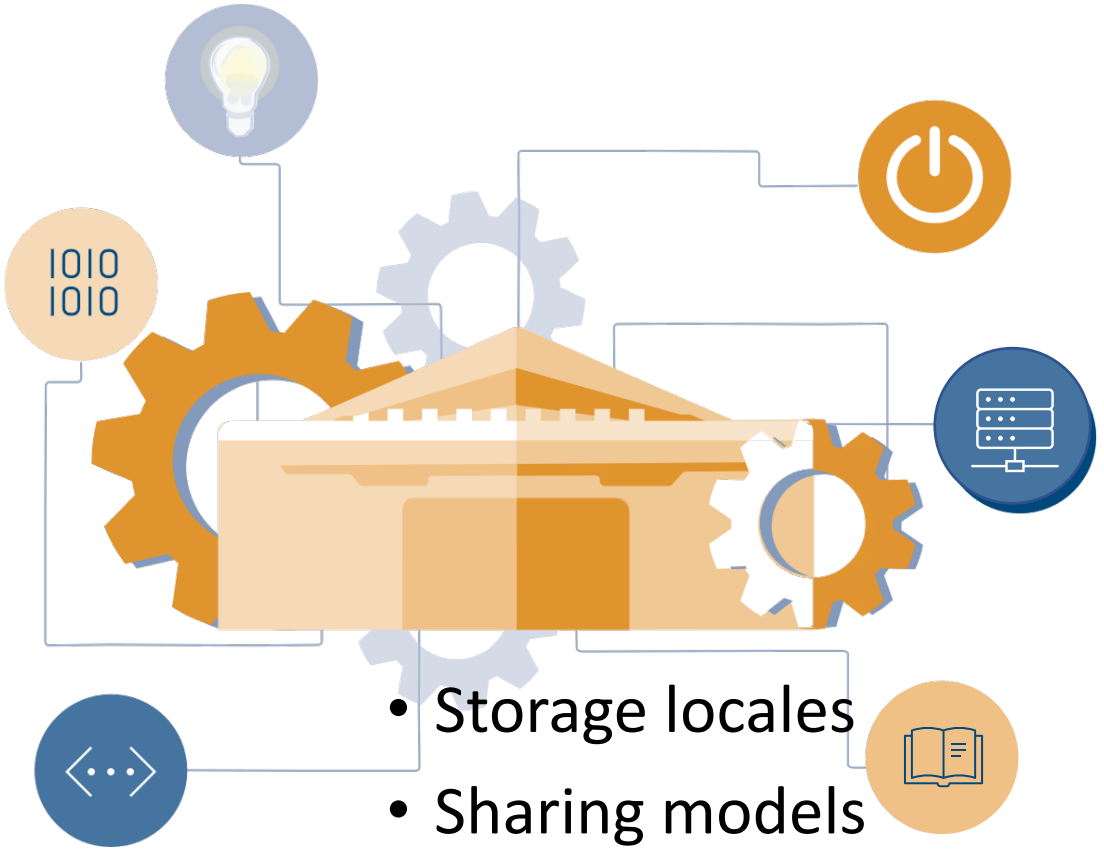




https://doi.org/10.1099/mgen.0.001145

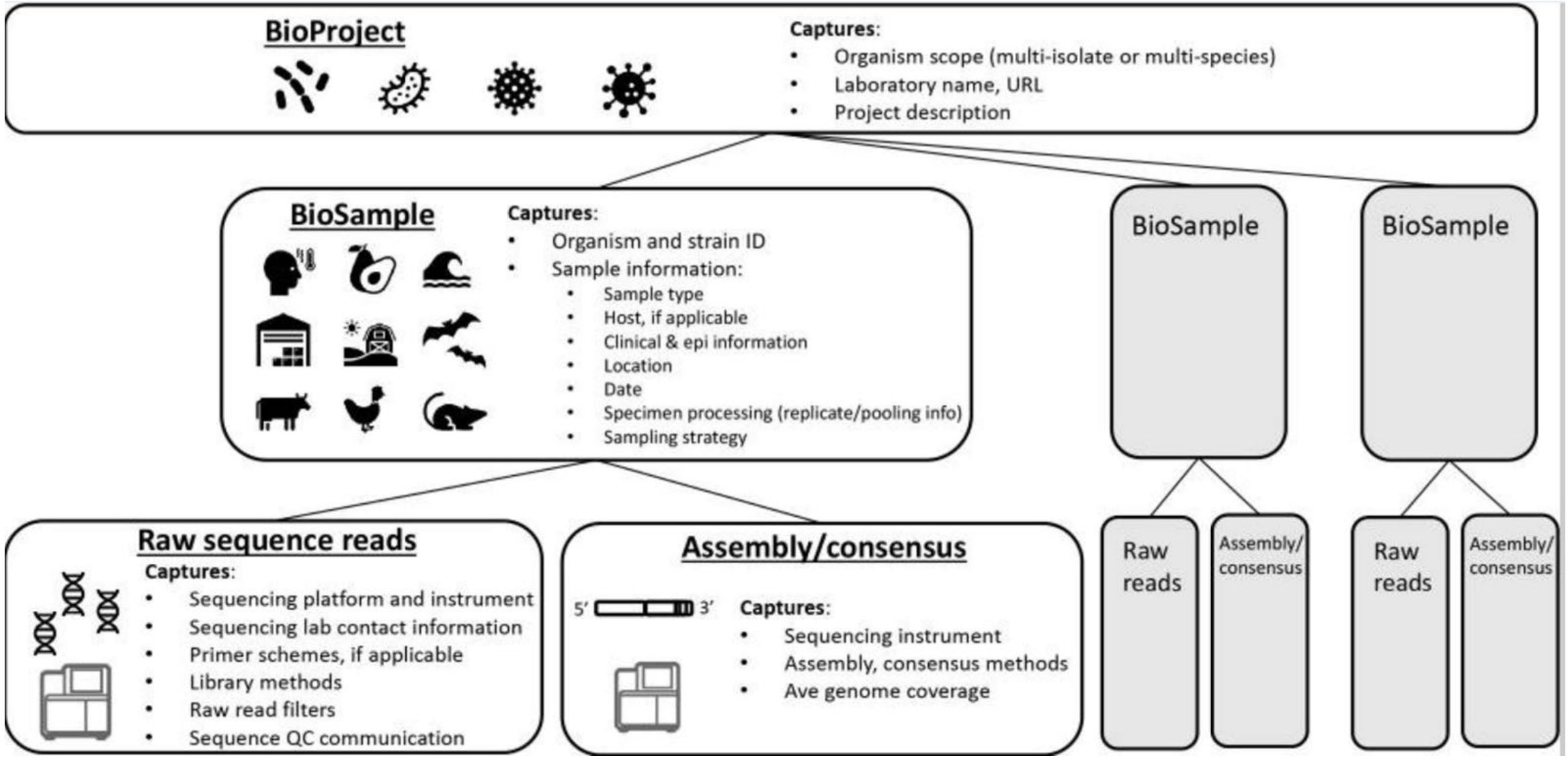
# Variables in PG Data Management and Reuse




- Storage locales
- Sharing models
- Exchange mechanisms

# The INSDC compliant Pathogen Data Object Model to structure genomic data for public health applications





https://doi.org/10.1099/mgen.0.001145

# One Health Enteric Package

BioSample package developed under Gen-FS metadata working group captures the full One Health sample space for enteric microbes

## *One Health Enteric package scope*

**CORE attributes**

- Isolate identifiers
- Collected by
- Date of collection
- Geographic location
- Sampling purpose
- Sampling device
- Project name
- IFSAC category
- Source type
- sequenced by

**Human/animal host**

- Host
- Host disease
- Host sex + age
- Host tissue sampled
- Animal environment
- Antimicrobials in food
- Animal housing system

**Food samples**

- Geographic origin
- Intended consumer
- Collection site description
- Food product type
- Label claims
- Food source
- Food processing types
- Food preservation process
- Food additives
- Food contact surface
- Food container wrapping
- Food container integrity

**Food facility**

- Facility type
- Building setting
- Food processed
- Facility location
- Monitoring zone
- Indoor sampling surface
- Surface material
- Surface material cond.
- Surface orientation
- Surface temperature
- Biocide used
- Animal intrusion

**Farm and Environment**

- ENVO triad
- Farm type
- Plant growth medium
- watering method
- Relative loc of sample
- Fertilizer administration
- Food cleaning process
- Sanitizer used
- Farm equip. used
- Water samples
- Extreme weather event
- Mechanical damage

12

*Developed by international group of experts*

*Consists of core attributes and optional attributes depending on source*

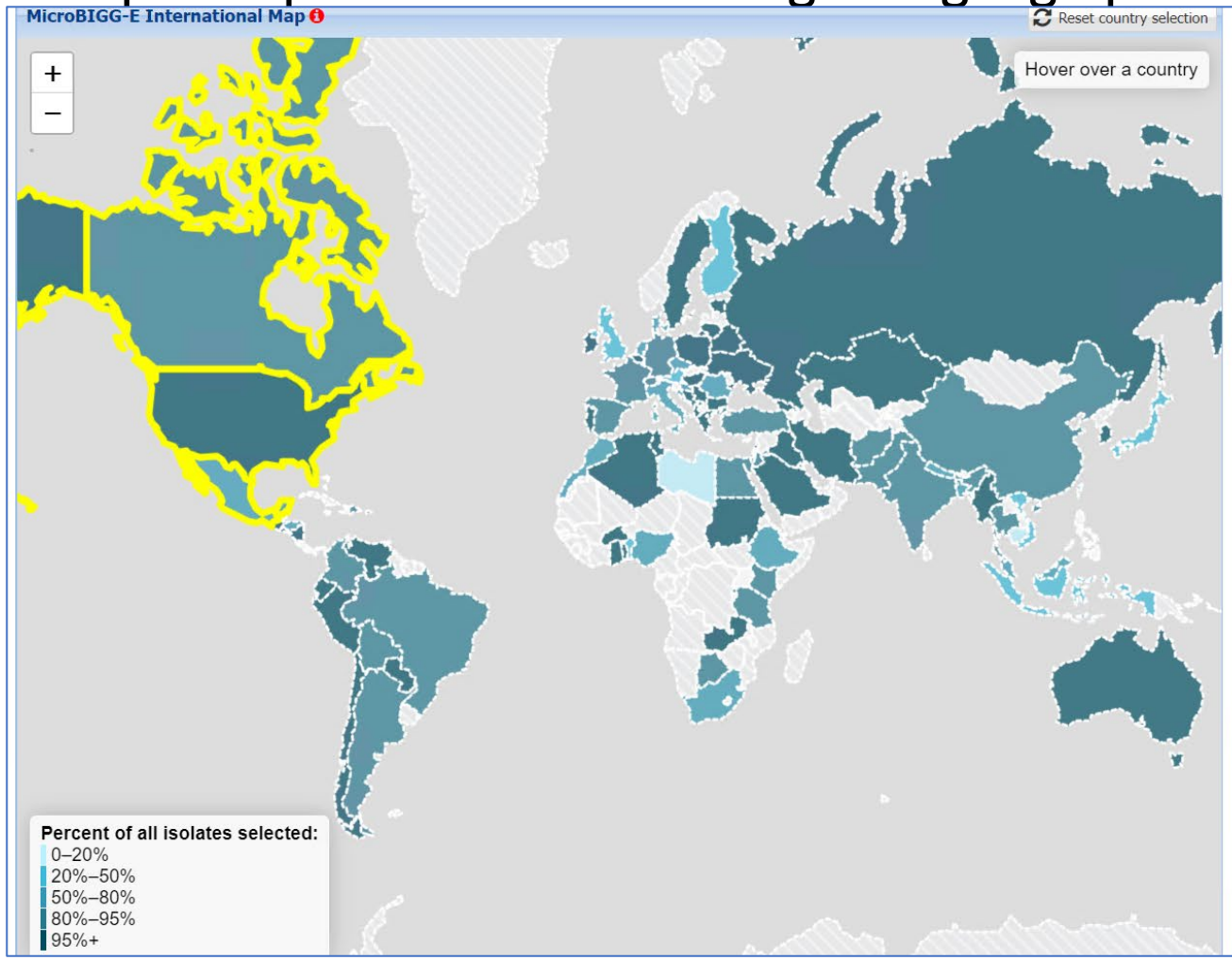
*Aim is to better provide outbreak and traceback investigations of bacterial pathogens*

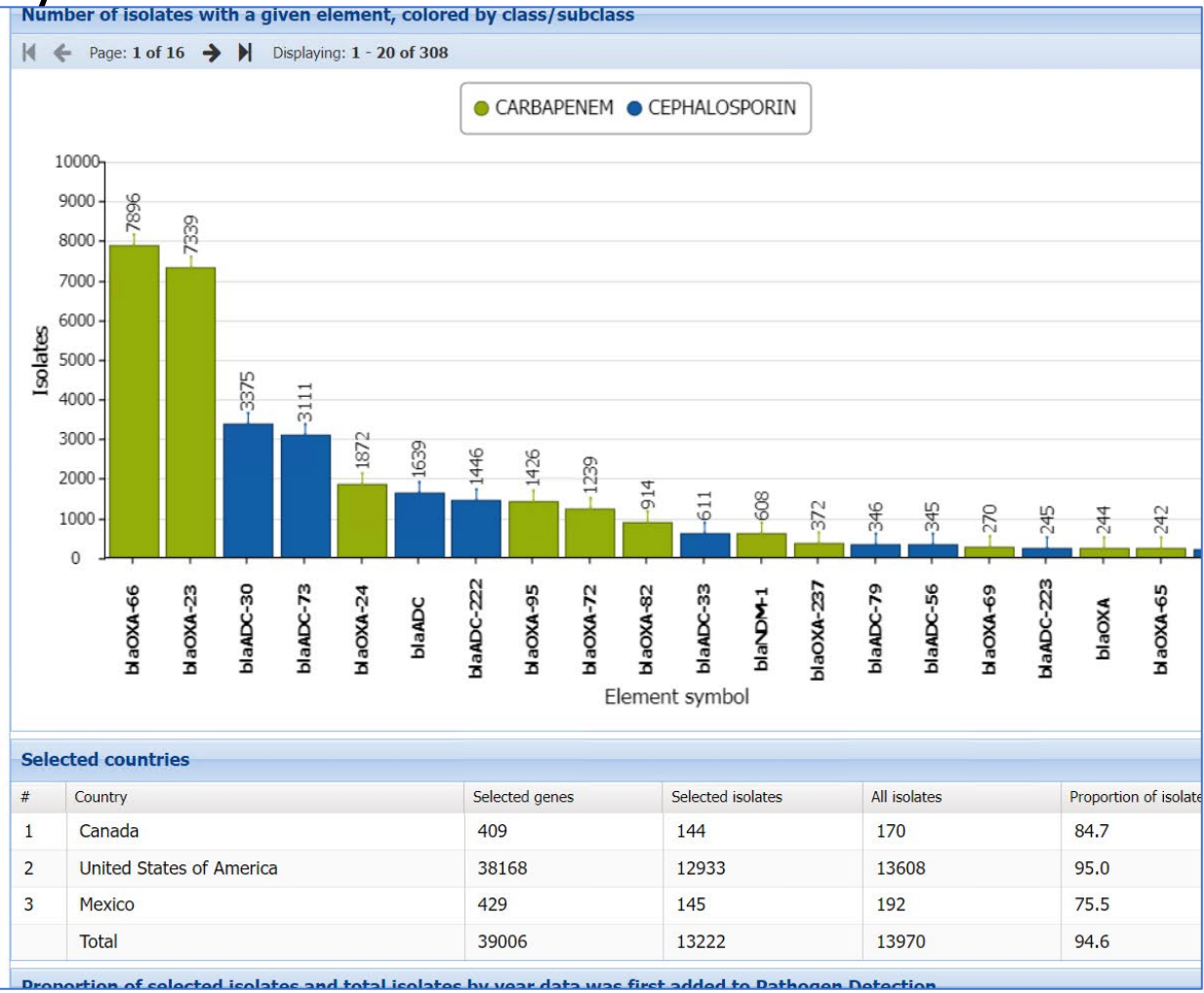
*It includes using ontologies for certain attributes to provide better structure for comparison than free text*

# Selected submissions to Pathogen Detection for past year

| Category | Organism | No. of Isolates | No. of Countries | % Missing Location |
|---|---|---|---|---|
| Foodborne | *Salmonella enterica* | 95, 433 | 32 | 18.95 |
| | *Escherichia coli* + *Shigella* spp. | 73, 319 | 94 | 15.23 |
| | *Campylobacter* spp. (*jejuni* and *coli*) | 15, 999 | 30 | 11.61 |
| | *Listeria monocytogenes* | 8, 293 | 12 | 9.73 |
| | *Cronobacter sakazakii* | 945 | 12 | 14.71 |
| Clinical | *Acinterobacter baumannii* | 14, 555 | 56 | 0.99 |
| | *Klebsiella pneumoniae* | 19, 224 | 77 | 5.09 |
| | *Staphylococcus aureus* | 16, 740 | 62 | 2.80 |
| | *Streptococcus pneumoniae* | 31, 037 | 41 | 51.93 |

# Graphic representation of AMR genes geographically



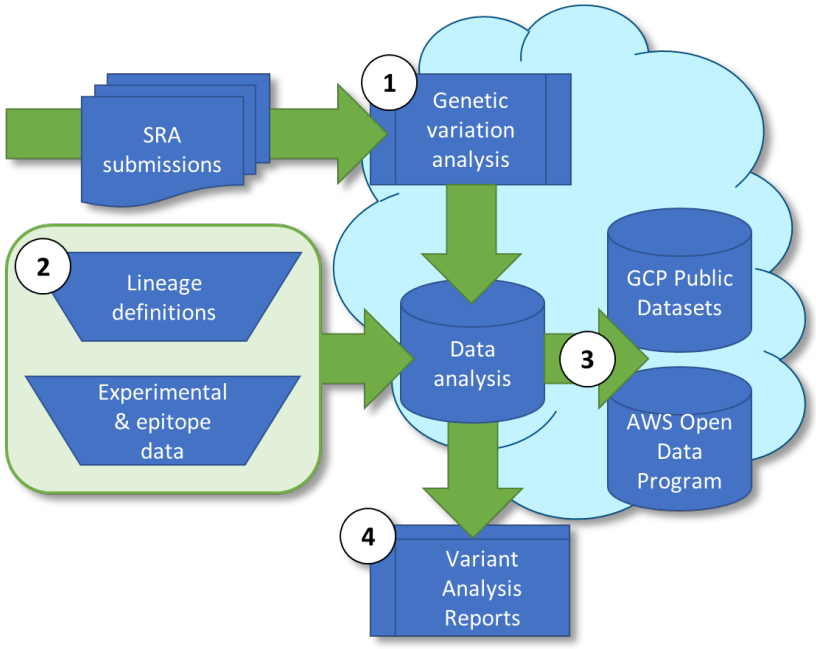




**Selected countries**

| # | Country | Selected genes | Selected isolates | All isolates | Proportion of isolate |
|---|---|---|---|---|---|
| 1 | Canada | 409 | 144 | 170 | 84.7 |
| 2 | United States of America | 38168 | 12933 | 13608 | 95.0 |
| 3 | Mexico | 429 | 145 | 192 | 75.5 |
| | Total | 39006 | 13222 | 13970 | 94.6 |

Carbapenem and cephalosporin resistance in *Acinetobacter baumannii* in North America from data submitted by end of 2023

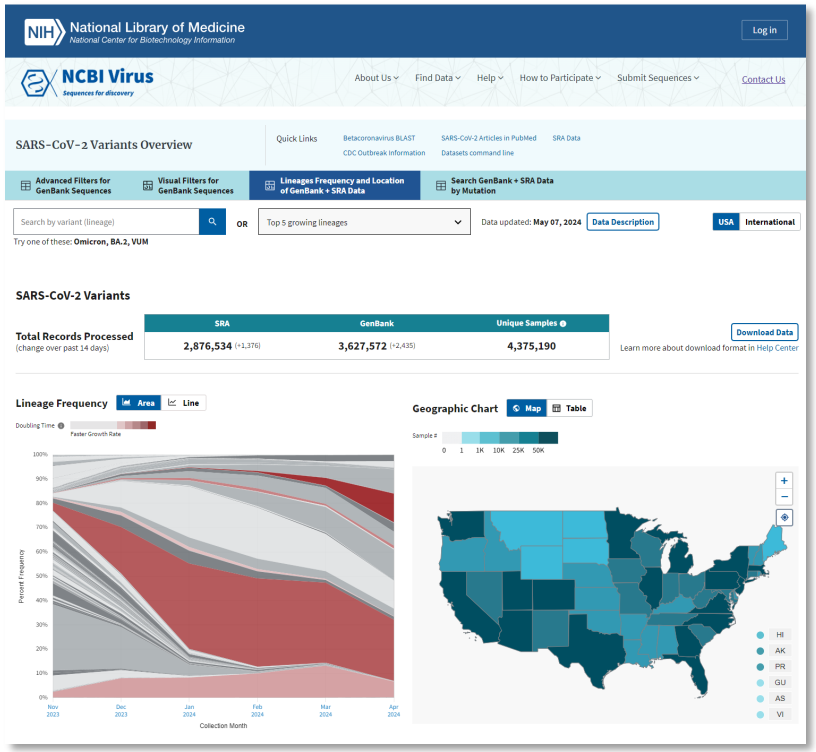
# Support of viral pathogen genomics

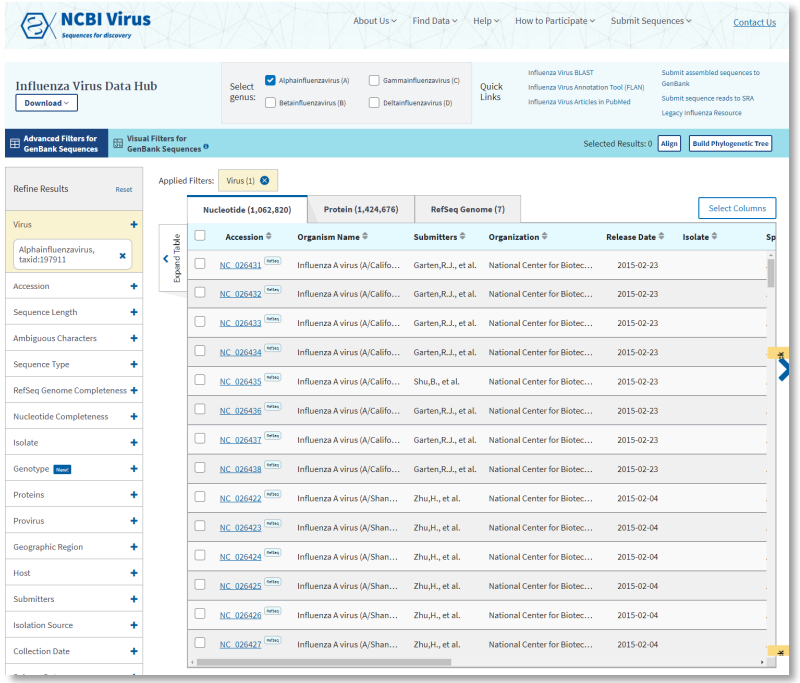



SARS-CoV-2 variation analysis pipeline

*Connor, R, et al. Viruses. 2024 Mar 11;16(3):430. doi: 10.3390/v16030430.*

SARS-CoV-2 Variants Overview



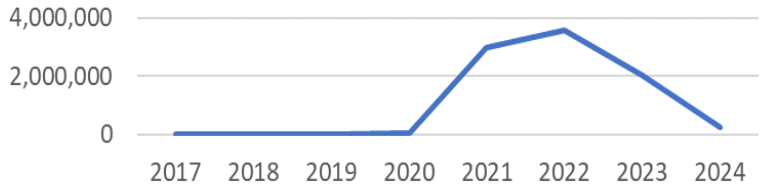

Influenza Virus Data Hub

# Selected viral genomic submissions to INSDC in the past year

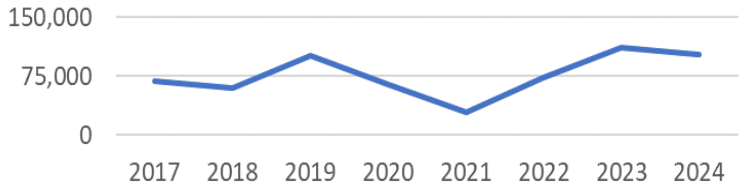
| Virus | **GenBank Records** | No. of Countries | % Missing Location | **SRA Records** | No. of Countries | % Missing Location |
|---|---|---|---|---|---|---|
| SARS-CoV-2 | 1,080,335 | 58 | 0.0% | 428,591 | 59 | 0.3% |
| Influenza A virus | 144,282 | 57 | 2.0% | 6,370 | 20 | 4.4% |
| Influenza B virus | 27,448 | 16 | 0.6% | 368 | 6 | 0.0% |
| Dengue virus | 5,416 | 72 | 0.0% | 1,610 | 40 | 0.2% |
| African swine fever virus | 709 | 35 | 0.4% | 63 | 10 | 20.6% |
| Measles virus | 1,104 | 25 | 0.0% | 54 | 3 | 5.6% |
| Ebolaviruses | 207 | 2 | 0.0% | 14 | 2 | 0.0% |
| Marburg virus | 1 | 1 | 0.0% | 8 | 2 | 0.0% |
| Rift Valley fever virus | 101 | 8 | 1.0% | 68 | 3 | 0.0% |
| Zika virus | 141 | 11 | 6.0% | 674 | 4 | 15.4% |
| HIV-1 | 51,391 | 49 | 12.2% | 1,590 | 11 | 9.1% |

# Variability in Submissions of Selected Viral Organisms

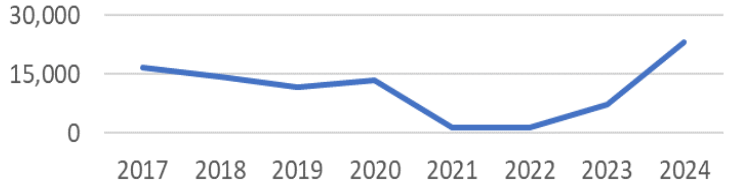
SARS-CoV-2



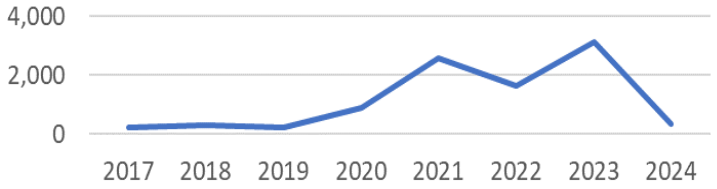

Influenza A virus



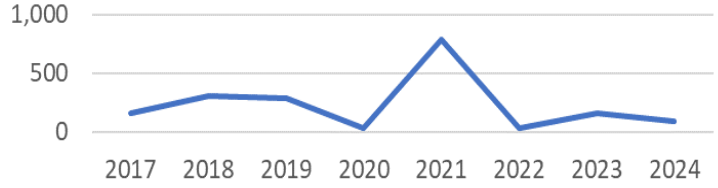

Influenza B virus



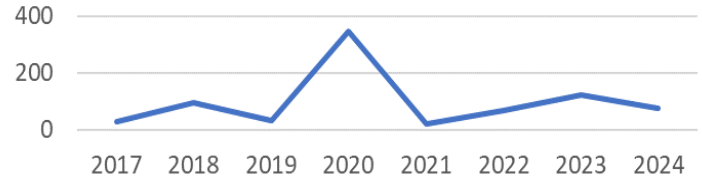

African swine fever virus



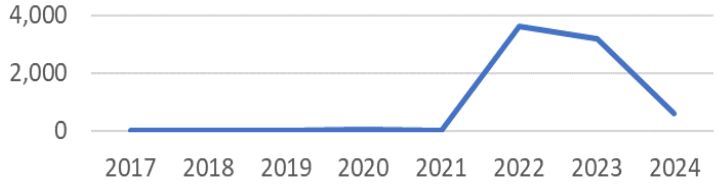

Ebolaviruses



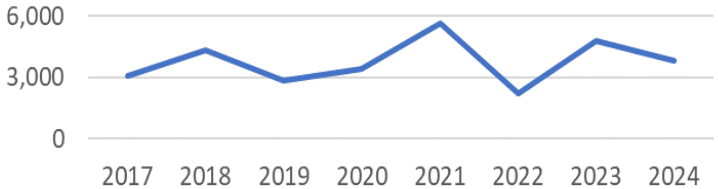

Rift Valley fever virus




MPOX virus




Dengue virus



# Data Interoperability: status and challenges

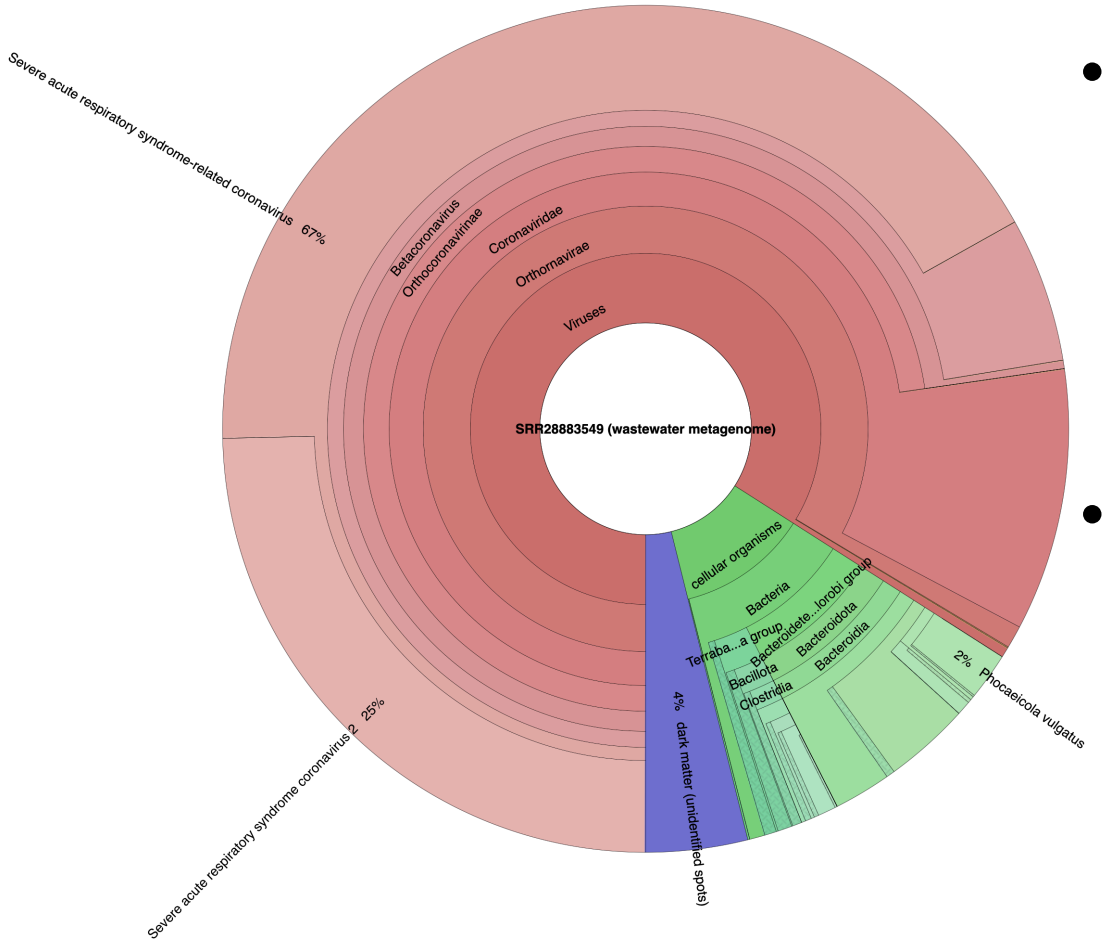
Public access repositories received ~50% of global data generated during COVID-19 pandemic

- Standards
  - Variable
  - Minimal required information (collection burden vs. utility)
- Data Exchange
  - Limited by restrictive repository egress policies
- Lack of oversight/regulation
  - No performance goals for data interoperability

# Pathogen Genomics Infrastructure: Status and Challenges

- Distribution
  - Interoperability
  - Latency
  - Scalability
  - Repository access policies
  - Public communications timelines
- Sustainability: current trends
- Potential opportunities for AI

# Wastewater and environmental biosurveillance




- All sequences submitted to SRA analyzed by the Sequence Taxonomic Analysis Tool (STAT)
  - Identifies the organismal content of each sequence sample – species, genera, families
  - Results published on web, AWS, and GCP and can be used to monitor viruses and organisms
- Public Wastewater Surveillance Consortium
  - Formed by NCBI to foster interagency collaboration
  - Wastewater and environmental data, metadata, and analysis standards and best practices

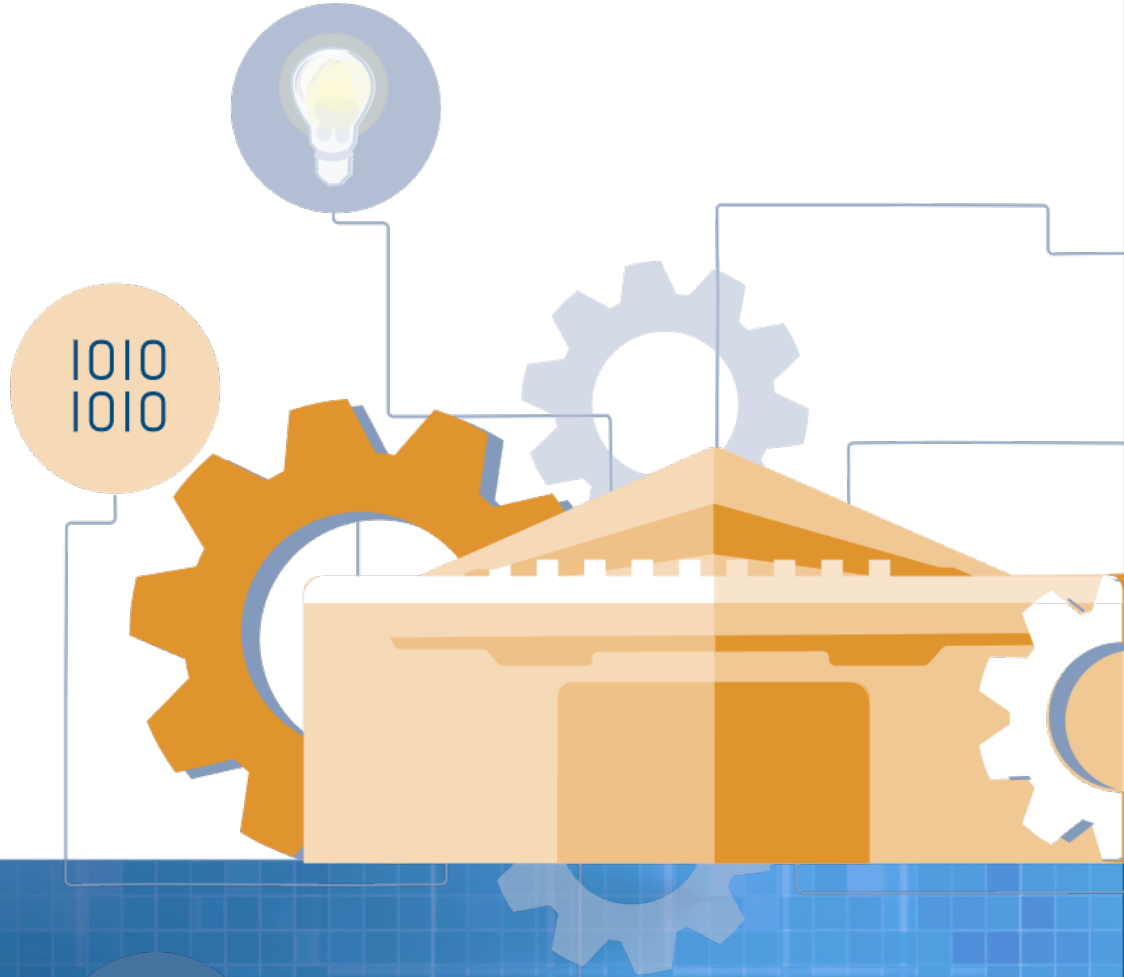
*Katz KS, et al. Genome Biol. 2021 Sep 20;22(1):270. doi: 10.1186/s13059-021-02490-0.*

# NCBI pathogen genomics lessons learned

- SRA and GenBank are critical to sharing SARS-CoV-2 sequence data and descriptive metadata with the scientific and public health communities
- SRA data critical to data validation and scientific insights
- Standardized data reduction key to making large scale data useful for downstream analyses
- Community tools like the NCBI developed “human sequence scrubber” necessary to support pathogen data
- Partnership-driven analysis best practices reduce burden on individual researchers or groups

# Questions?

# Extra Slides



# International Nucleotide Sequence Database Collaboration (INSDC)


- GenBank and SRA at National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health in Bethesda, Maryland, USA
- DNA Data Bank of Japan (DDBJ) at the National Institute of Genetics in Mishima, Japan
- European Nucleotide Archive (ENA) at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK

NCBI

DDBJ
DNA Data Bank of Japan

ENA
European Nucleotide Archive



Arita M, Karsch-Mizrachi I, Cochrane G. **The international nucleotide sequence database collaboration**. *Nucleic Acids Research. 2021 Jan;49(D1):D121-D124*. DOI: 10.1093/nar/gkaa967.

# Sequence Data Generation

| PURPOSE | BIOLOGICAL SAMPLE |
| --- | --- |
| Research | Isolated |
| Public Health | Environmental |
| Surveillance | |

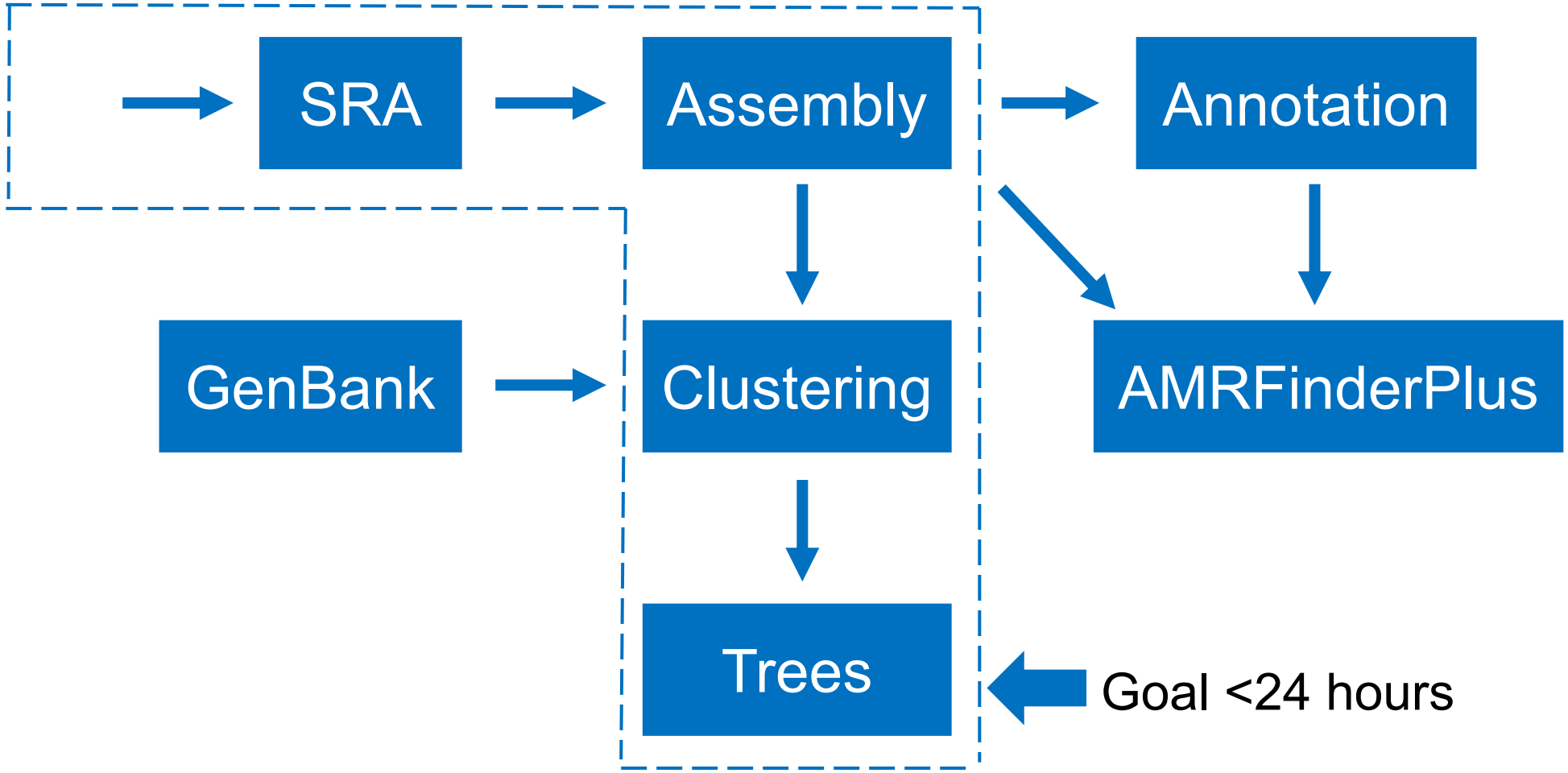
# Combating Antibiotic Resistant Bacteria (CARB)

To address the growing threat of antibiotic resistance, the U.S. Government released the National Strategy for CARB in September 2014

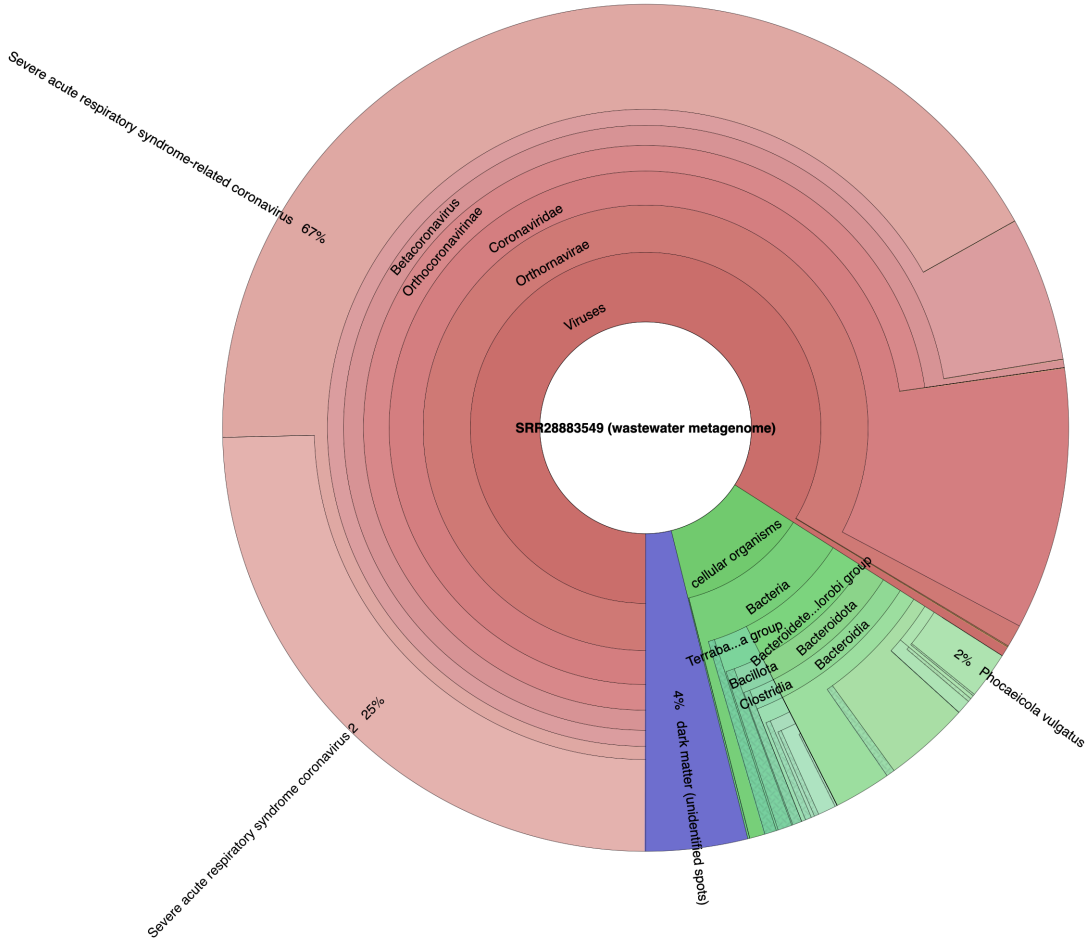
NLM receives support via NIAID for this effort

- Developed the National Database of Antibiotic-Resistant Organisms (NDARO) - a centralized hub for researchers to access AR data to facilitate real-time surveillance of pathogenic organisms.
    - developed and maintains a curated reference database of AR, virulence, and stress response genes
    - developed AMRFinderPlus to identify these genes in bacterial genomes
    - provides web interfaces on the presence of these genes in more than 1.9 million isolates along with associated metadata

# Pathogen Detection (PD) Pipeline



# Wastewater and environmental biosurveillance

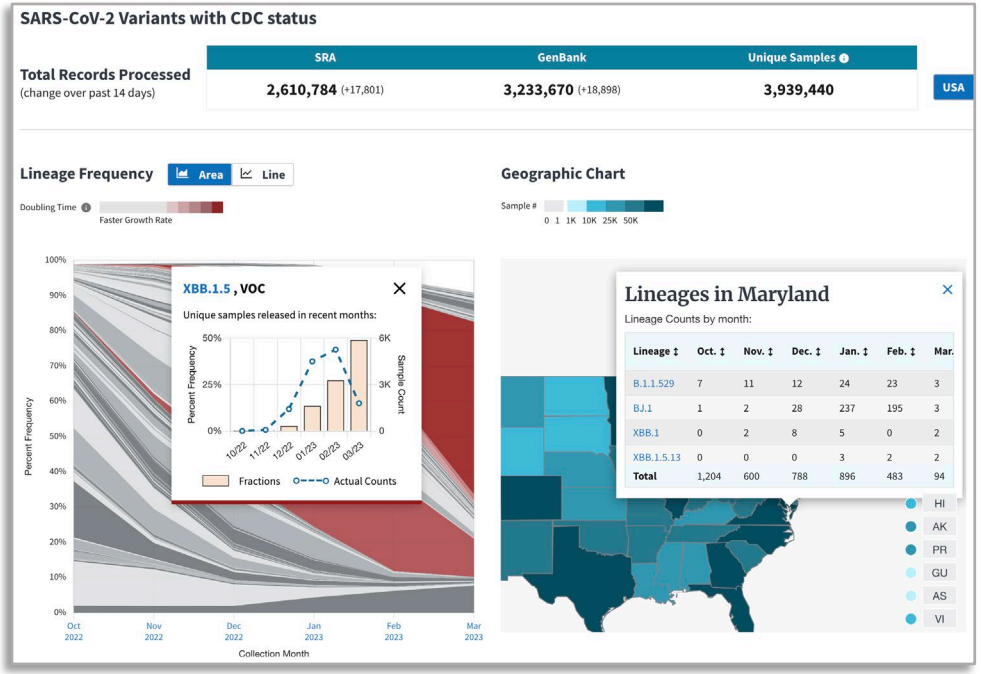



- All sequences submitted to SRA analyzed by the Sequence Taxonomic Analysis Tool (STAT)
  - Identifies the organismal content of each sequence sample – species, genera, families, etc.
  - Results published on web, AWS, and GCP and can be used to monitor viruses and organisms
- Public Wastewater Surveillance Consortium
  - Formed by NCBI to foster interagency collaboration
  - Wastewater and environmental data, metadata, and analysis standards and best practices

*Katz KS, et al. Genome Biol. 2021 Sep 20;22(1):270. doi: 10.1186/s13059-021-02490-0.*

# Lessons learned from NCBI pathogen genomics support

- Partnership-driven analysis best practices reduce burden on individual researchers or groups
- SRA and GenBank were critical to sharing SARS-CoV-2 sequence data and descriptive metadata with the scientific and public health communities
- SRA data critical to data quality and validation
  - Mutations, linkage, and recombination
- Standardized data reduction key to making the data useful for downstream analyses
  - 500 TB compressed into in easy-to-use tables
  - 1.4 M pathogen samples compressed into actionable data in short time frame
- Tools like the NCBI developed “human sequence scrubber” necessary to support pathogen data



# Selected viral genomic submissions to INSDC in the past year

| | GenBank Records | | | SRA Records | | |
|---|---|---|---|---|---|---|
| **Virus** | No. of Records | No. of Countries | % Missing Location | No. of Records | No. of Countries | % Missing Location |
| SARS-CoV-2 | 1,080,335 | 58 | 0.0% | 428,591 | 59 | 0.3% |
| Influenza A virus | 144,282 | 57 | 2.0% | 6,370 | 20 | 4.4% |
| Influenza B virus | 27,448 | 16 | 0.6% | 368 | 6 | 0.0% |
| Dengue virus | 5,416 | 72 | 0.0% | 1,610 | 40 | 0.2% |
| African swine fever virus | 709 | 35 | 0.4% | 63 | 10 | 20.6% |
| Measles virus | 1,104 | 25 | 0.0% | 54 | 3 | 5.6% |
| Ebolaviruses | 207 | 2 | 0.0% | 14 | 2 | 0.0% |
| Marburg virus | 1 | 1 | 0.0% | 8 | 2 | 0.0% |
| Rift Valley fever virus | 101 | 8 | 1.0% | 68 | 3 | 0.0% |
| Zika virus | 141 | 11 | 6.0% | 674 | 4 | 15.4% |
| HIV-1 | 51,391 | 49 | 12.2% | 1,590 | 11 | 9.1% |