One Health Enteric package:
Expanded and standardized metadata for enteric genomic epidemiology in the US

IFIC
THE 5TH HIGH-THROUGHPUT SEQUENCING (HTS) SYMPOSIUM

May 18-19, 2022
Ruth E. Timme
FDA, Center for Food Safety and Applied Nutrition
ruth.timme@fda.hhs.gov
Interagency Collaboration for Genomics for Food and Feed Safety (Gen-FS)

- National Institutes of Health (NIH)
- Centers for Disease Control and Prevention (CDC)
- 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).
Gen-FS Metadata workgroup contributors:

workgroup initiated in early 2020:

Ruth Timme
Maria Balkey
Kayla Pennerman
Christopher Grim
Michael Batz
Heather Tate
Karen Jinneman
Paul Morin
Amy Merrill
Gordon Martin
Yesha Shrestha
Sarah Peloquin
Olga Ceric

Cesar Morales
Jeffrey Levine
Mustafa Simmons
Jessica Hicks
Kimberly Cook
Jo Ann Van Kessel
Jim Bono
Beth Harris

Michael Feldgarden
Martin Shumway
John Anderson

Lavin Joseph
Steven Stroika

Jennifer Adams
**Goal**: For each pathogen, sample the entire One Health arena inhabited by that pathogen, (animal, environment, water, air, food processing plants, humans)
**Goal:** For each pathogen, sample the entire One Health arena inhabited by that pathogen, (animal, environment, water, air, food processing plants, humans)

Openly sharing data from this effort enables us to sample the One Health sphere.
US open genomic surveillance of enteric pathogens:
US open genomic surveillance of enteric pathogens:

Major US sequencing networks
Standard data enables global interoperability
# 2013 - current: INSDC Pathogen metadata package

**Core attributes:**
- Organism name
- Collected by
- Collection date
- Geographic location

**Human clinical package:**
- host
- host_disease
- host_age
- host_sex
- host_disease_state
- host_disease_stage
- host_subject_id
- host_tissue_sampled

**Environmental/animal/food package:**
- Isolation_source (free text)

- Impossible to do simple manual queries
- Not machine readable
- Data not in format to test broad-scale hypotheses, or to use in risk assessment or source attribution, etc.
New package requirements

• Meets FAIR principles (Findable, Accessible, Interoperable, Repeatable)
• Interoperable with existing Pathogen template, Food MIxS package and other related MIxS packages, and the IFSAC schema.
• Covers the One Health arena of enteric ecology (i.e. human, animals, food, built environment, farm, water, etc.)
• Applicable for academic research, industry, and regulatory applications
• Leverages existing ontologies (FoodOn, ENVO) and controlled vocabularies (minimize free text)
Examples of queries we’d like to enable:

• Return all qnrB19-containing *Salmonella* Typhimurium isolates collected from live pigs and pig farms
• Return all *Listeria monocytogenes* isolates collected from food contact surfaces in food manufacturing facilities
• Return all pathogen isolates collected from leafy green row crops in the United States
• Return all isolates collected from pre-harvest water sources from 2005-2010
• Return all *Salmonella* isolates collected from frozen breaded food products
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
One Health Enteric package scope

Food MIxS
231 terms

NCBI Pathogen
31 terms

One Health Enteric
90 terms

14 new terms
20 NCBI pathogen
47 MIxS terms
Examples of queries we’d like to enable:

• Return all *qnrB19*-containing *Salmonella* Typhimurium isolates collected from live pigs and pig farms.

• Return all *Listeria monocytogenes* isolates collected from food contact surfaces in food manufacturing facilities.

• Return all pathogen isolates collected from leafy green row crops in the United States.

• Return all isolates collected from pre-harvest water sources from 2005-2010.

• Return all *Salmonella* isolates collected from frozen, breaded food products.
Filtering at NCBI Pathogen Detection

- Salmonella enterica
- serovar Javiana
- collected from 2018 – present
- collected from “food”
### Filtering at NCBI Pathogen Detection

Filtering at NCBI Pathogen Detection

- **Salmonella enterica**
- sample from the US
- IFSAC category: fish

<table>
<thead>
<tr>
<th>Organism group</th>
<th>SNP cluster</th>
<th>Matched isolates</th>
<th>Matched clinical isolates</th>
<th>Matched environmental isolates</th>
<th>Total isolates</th>
<th>Minimal mid-diff</th>
<th>Minimal mid-name</th>
<th>Latest update</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salmonella enterica</td>
<td>FGSS0002947.5</td>
<td>10</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>2021-04-11</td>
<td></td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>FGSS0002405.16</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>127</td>
<td>0</td>
<td>2021-04-11</td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>FGSS0002716.4</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>789</td>
<td>2021-08-10</td>
<td></td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>FGSS0003678.1</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>210</td>
<td>2021-08-19</td>
<td></td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>FGSS0003953.183</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>28</td>
<td>2021-08-19</td>
<td></td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>FGSS0003764.2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>123</td>
<td>2021-08-19</td>
<td></td>
</tr>
</tbody>
</table>
Next steps

• **April 2022**: Package available for use at NCBI BioSample
• **2022-2025**: implementation across FDA, CDC, USDA, and NCBI-Pathogen Detection

- External samples from product testing collected (LFFM)
- Internal samples from facility inspections, food testing, and sampling assignments (FDA-ORA).
- Samples collected for research
SELECT the package that best describes your samples.

<table>
<thead>
<tr>
<th>Packages</th>
<th>Packages for MAG submitters</th>
<th>Packages for metagenome submitters</th>
</tr>
</thead>
</table>

(Optional) Filter packages by organism name

Enter the full scientific name of your samples, e.g., *Escherichia coli*

To filter for relevant BioSample packages, enter the full scientific name of the organism of your samples.
- If your BioSamples are derived from a species not represented in NCBI's Taxonomy database, enter the genus-level name, e.g., *Escherichia*
- If your BioSamples are derived from more than one organism, enter the common species, genus, or family, e.g., *Enterobacteriaceae*
- If your BioSamples are metagenomic/environmental, or metagenome-assembled genomes (MAG), select the appropriate tab above

For more information about organism names, see [Organism information](#).

**NCBI packages**

- SARS-CoV-2: clinical or host-associated
  - Use for SARS-CoV-2 samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of SARS-CoV-2 cases.

- SARS-CoV-2: wastewater surveillance
  - Use for SARS-CoV-2 wastewater surveillance samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of SARS-CoV-2 cases.

- Pathogen
  - Use for pathogen samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of pathogens.

- Microbe
  - Use for bacteria or other unicellular microbes when it is not appropriate or advantageous to use *MAG, Pathogen, or Virus* packages.

- Model organism or animal
  - Use for multicellular samples or cell lines derived from common laboratory model organisms, e.g., mouse, rat, Drosophila, worm, fish, frog, or large mammals including zoo and farm animals.

- Metagenome or environmental
  - Use for metagenomic and environmental samples when it is not appropriate or advantageous to use *MAG* packages.

- Invertebrate
  - Use for any invertebrate sample.

- Human
  - WARNING: Only use for human samples or cell lines that have no privacy concerns. For all studies involving human subjects, it is the submitter's responsibility to ensure that the information supplied protects participant privacy in accordance with all

**GSC MixS packages for genomes, metagenomes, and marker sequences**

- MGS Cultured Bacterial/Archaeal
  - Use for cultured bacterial or archaean genomic sequences. Organism must have lineage *Bacteria* or *Archaea*.

- MGS Eukaryotic
  - Use for eukaryotic genomic sequences. Organism must have lineage *Eukarya*.

- MGS Viral
  - Use for virus genomic sequences. Organism must have lineage *Viruses*.

- MIMAG Metagenome-assembled Genome
  - Use for metagenome-assembled genome sequences produced using computational binning tools that group sequences into individual organism genome assemblies starting from metagenomic data sets. Organism cannot contain the term 'metagenome'. Use the MULVIG package for virus genomes.

- MIMARKS Specimen
  - Use for any type of marker gene sequences, e.g., 16S, 18S, 23S, 28S rRNA or COI obtained from cultured or voucher-identifiable specimens. Organism cannot contain the term 'metagenome'.

- MIMARKS Survey related
  - Use for any type of marker gene sequences, e.g., 16S, 18S, 23S, 28S rRNA or COI obtained directly from the environment, without culturing or identification of the organism. Organism must be a metagenome, where lineage starts with *unclassified sequences* and scientific name ends with 'metagenome'.

- MIMS Environmental/Metagenome
  - Use for environmental and metagenome sequences. Organism must be a metagenome, where lineage starts with *unclassified sequences* and scientific name ends with 'metagenome'.

- MISAG Single Amplified Genome
  - Use for single amplified genome sequences produced by

---

May 2022: OHE Package available within BioSample (soon!)
Custom packages with full picklists:

Version controlled source of custom packages for GenFS collab:
https://github.com/CFSAN-Biostatistics/One_Health_Enteric_Package

GenomeTrakr: entire package
LFFM product testing: food sub-package
Vet-LIRN: animal host and food sub-packages
FDA: food, facility, and env sub-packages

PulseNet clinical: human/animal host sub-package
Water sampling: farm/env sub-package

Guidance for populating GenomeTrakr metadata templates (BioSample and SRA) (Step 2.1)
Gen-FS Metadata workgroup contributors:
workgroup initiated in early 2020:

Ruth Timme
Maria Balkey
Kayla Pennerman
Christopher Grim
Michael Batz
Heather Tate
Karen Jinneman
Paul Morin
Amy Merrill
Gordon Martin
Yesha Shrestha
Sarah Peloquin
Olga Ceric

Cesar Morales
Jeffrey Levine
Mustafa Simmons
Jessica Hicks
Kimberly Cook
Jo Ann Van Kessel
Jim Bono
Beth Harris

Michael Feldgarden
Martin Shumway
John Anderson

Jennifer Adams

Lavin Joseph
Steven Stroika