NCBI Pathogen Detection Pipeline

Background and live demo

William Klimke
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health
IFSH
6-7 May 2019
Chicago, IL
2013
FDA/CDC Real Time *Listeria* Project

FDA & CDC could leverage existing systems & work flows…

Could NCBI play a role?
Shared Network pathways and data streams for outbreak detection and investigations

Sampling of clinical/human, food, animal, and environmental bacterial isolates

**GenomeTrakr**
- USDA
- FSIS labs
- State agriculture & food regulatory labs

**PulseNet**
- FSIS & FDA labs
- food, environmental
- State health labs (clinical)

**Raw Genomic Sequence data**

**NCBI Submission Portal**
- Minimal metadata, everything publically accessible
- More extensive metadata, analyzed seq data (tools to translate) data shared among PulseNet labs only

**CDC analysis (all raw sequence goes into NCBI)**

**PulseNet National Database**
Raw data: hundreds of millions of base pairs in short reads

Intermediate data: assembly of a few million bases

Relevant data: a phylogenetic tree clustering isolates, SNP distances, annotated important (antimicrobial resistance) genes

Turn large data into useful data: within 24 hours
Analysis goals

1. Are these isolates clonally related?

2. What is the anti-microbial resistance gene repertoire of this isolate

Both results presented in easy to use web-interface
Pathogen Detection Overview

WGS reads from SRA
Assemble, quality filter
Cluster using wgMLST
Align, density filter
Assemblies from GenBank

SNP phylogenetic analysis (Maximum Compatibility - Cherry, 2017 - DOI: 10.1186/s12859-017-1520-4)
SKESA: strategic k-mer extension for scrupulous assemblies.

• Very fast de novo assembler
  • On our servers usually < 10 minutes
  • More and shorter contigs
  • Break contigs at ambiguities
SKESA assembler

• **High confidence** in the sequence that does assemble

• Slightly less genome representation

Souvorov et al., 2018: DOI: 10.1186/s13059-018-1540-z
SKESA accuracy vs. Spades and Megahit (quast)
56 SRA runs for 48 reference assemblies
Pathogen Detection Overview

WGS reads from SRA

Assemble, quality filter

Assemblies from GenBank

Rapid Reports/Cluster using wgMLST

Align, density filter

SNP phylogenetic analysis

< 1 hours

< 24 hours
Measurement of NCBI Pathogen Detection Pipeline Turnaround Time - 7 Day Window
Measurement of NCBI Pathogen Detection Rapid Reports Turnaround Time - 7 Day Window
NCBI’s Role in Combatting Antibiotic Resistance

1. Build AMR reference database (reference proteins and hidden Markov models (HMMs) and protein family hierarchy

2. Build AMRFinder tool to identify AMR proteins using reference database

3. Use AMRFinder to identify AMR proteins in all pathogen isolates integrated into NCBI Pathogen Browser

4. Capture antibiotic susceptibility test data (AST)

5. Integrate AST into NCBI Pathogen Browser
NCBI’s Role in Combatting Antibiotic Resistance

National Database of Antibiotic Resistant Organisms (NDARO)

Welcome to the NCBI National Database of Antibiotic Resistant Organisms (NDARO), a collaborative, cross-agency, centralized hub for researchers to access AMR data to facilitate real-time surveillance of pathogenic organisms.

AMRFinder Uses a Curated Database, HMMs and BLAST to Identify AMR genes

Proteins

HMMs and BLAST

Report on resistance genes in isolate

AMR database

4,810 resistance proteins
> 610 HMMs
34 drug classes resisted
~50% beta-lactamases
AMRFinder Has a Hierarchical Structure

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Resistance to carbapenems and other beta-lactam antibiotics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPC-2</td>
<td>Likely resistance to carbapenems and other beta-lactam antibiotics.</td>
</tr>
<tr>
<td>KPC family</td>
<td>Class A beta-lactamase of unknown specificity.</td>
</tr>
<tr>
<td>class A beta-lactamase</td>
<td>Prevents false-positive identification as a beta-lactamase. Not reported.</td>
</tr>
<tr>
<td>not beta-lactamase</td>
<td></td>
</tr>
</tbody>
</table>
Automating AMR Gene Annotation At Scale

**Pathogen Detection Pipeline (SRA):**
- GenomeTrakr
- PulseNet
- PHE
- FDA/CDC Antimicrobial Resistant Isolate Bank
- State laboratories
- Clinical laboratories

**Genbank assemblies:**
- General submissions

**AMRFinder: Combined BLAST/HMM AMR gene discovery**

- **Curated AMR gene sources:**
  - Domain experts
  - Large scale databases
    - FDA Center for Veterinary Medicine
    - ResFinder
    - The C.A.R.D. (~monthly exchanges)
- **Manual extraction from literature**

- **Curated HMM sources:**
  - ResFams
    - Select
    - Set cutoffs
  - New NCBI-built HMMs
    - Group sequences
    - Align
    - Build HMM
    - Set cutoffs

**Validation**

**Alleles/genes**
- 4,810

**HMMs**
- > 610

**Isolate Browser**
Presence of known AMR genes can be visualized and downloaded

**Surveillance alerts**
Identified colistin resistant *E. coli* without traditional phenotyping (Vasquez et al., 2016)

**AMR gene discovery**
Identified novel plasmid-borne fosfomycin gene (Rehman et al., 2017)

**PGAP Annotations**
Standardized annotation for all researchers
AMRFinder Release

NCBI has developed AMRFinder, a tool that identifies AMR genes using either protein annotations or nucleotide sequence. AMRFinder is used in the Pathogen Detection pipeline, and these data are displayed in NCBI’s Isolate Browser. AMRFinder can identify acquired antimicrobial resistance genes in either protein datasets or nucleotide data, including genomic data. AMRFinder relies on NCBI’s curated AMR gene database and curated collection of Hidden Markov Models. For more information on how AMRFinder operates, please see the available README file.

To enable researchers to incorporate these tools into their own analytical pipelines, NCBI has made the following resources publicly available:

1. A downloadable version of AMRFinder using CWL and Docker. See the README for installation instructions.
2. NCBI’s Bacterial Antimicrobial Resistance Reference Gene Database, which contains over 4,000 curated AMR protein sequences. There is also a tab-delimited table that lists accessions, gene names, and product names for all AMR genes in the reference database.
3. NCBI’s curated HMM collection of AMR protein sequences.

Note that this version of AMRFinder is an alpha version. If you have any questions about or experience problems running AMRFinder, please contact pd-help@ncbi.nlm.nih.gov

# AMR Reference Database

![Database Interface](https://www.ncbi.nlm.nih.gov/pathogens/isolates#/refgene/)

<table>
<thead>
<tr>
<th>#</th>
<th>allele</th>
<th>gene family</th>
<th>product name</th>
<th>type</th>
<th>subtype</th>
<th>class</th>
<th>subclass</th>
<th>refseq protei</th>
<th>refseq nucleic</th>
<th>genbank prot</th>
<th>genbank nuc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mcr-1</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.1</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_04958980</td>
<td>NG_050417.1</td>
<td>AKF16188.1</td>
<td>KP347127.1</td>
</tr>
<tr>
<td>2</td>
<td>mcr-1.10</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.10</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_09680744</td>
<td>NG_055583.1</td>
<td>ASK49940.1</td>
<td>MF179238.1</td>
</tr>
<tr>
<td>3</td>
<td>mcr-1.11</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.11</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_09690281</td>
<td>NG_056784.2</td>
<td>ATM29099.1</td>
<td>KY963650.2</td>
</tr>
<tr>
<td>4</td>
<td>mcr-1.12</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.12</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_10400696</td>
<td>NG_056412.1</td>
<td>BBD21811.1</td>
<td>LC337908.1</td>
</tr>
<tr>
<td>5</td>
<td>mcr-1.13</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.13</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_10954405</td>
<td>NG_057146.5</td>
<td>AVMB84739.1</td>
<td>MG384739.1</td>
</tr>
<tr>
<td>6</td>
<td>mcr-1.15</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.15</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_11878683</td>
<td>NG_061610.1</td>
<td>AXL06759.1</td>
<td>MG763687.1</td>
</tr>
<tr>
<td>7</td>
<td>mcr-1.2</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.2</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_00527407</td>
<td>NG_051170.1</td>
<td>CBY14852.1</td>
<td>LXQ0910002</td>
</tr>
<tr>
<td>8</td>
<td>mcr-1.3</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.3</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_07705486</td>
<td>NG_052661.1</td>
<td>ANJ15621.1</td>
<td>KU634208.1</td>
</tr>
<tr>
<td>9</td>
<td>mcr-1.4</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.4</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_07061106</td>
<td>NG_052664.1</td>
<td>APIM87143.1</td>
<td>KY041866.1</td>
</tr>
<tr>
<td>10</td>
<td>mcr-1.5</td>
<td>mcr-1</td>
<td>phosphoethanolamine−lipid A transferase MCR-1.5</td>
<td>AMR</td>
<td>AMR</td>
<td>COLISTIN</td>
<td>COLISTIN</td>
<td>WP_07651106</td>
<td>NG_052603.1</td>
<td>APIM84408.1</td>
<td>KY283125.1</td>
</tr>
</tbody>
</table>

Acknowledgements

Richa Agarwala
Azat Badretdin
Slava Brover
Joshua Cherry
Jinna Choi
Vyacheslav Chetvernin
Robert Cohen
Michael DiCuccio
Boris Fedorov
Mike Feldgarden
Lewis Geer
Dan Haft
Lianyi Han
Avi Kimchi
Michel Kimelman
William Klimke
Alex Kotliarov
Valerii Lashmanov
Aleksandr Morgulis
Eyal Moses
Chris O’Sullivan
Arjun Prasad
Edward Rice
Kirill Rotmistrovskyy
Alejandro A. Schaffer
Stephen Sherry
Sergey Shiryev
Martin Shumway
Oleg Shutov
Alexandre Souvorov
Tatiana Tatusova
Igor Tolstoy
Chunlin Xiao
Leonid Zaslavsky
Alexander Zasypkin
Lukas Wagner
Hlavina Wratko
Eugene Yaschenko
David Lipman
James Ostell
Kim Pruitt

CDC
FDA/CFSAN
GenFS
USDA-FSIS
PHE/FERA
NIHGRI
NIAID
WRAIR
Broad
Wadsworth/MDH
Vendors: PacBio, Illumina, Roche

pd-help@ncbi.nlm.nih.gov

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. http://www.ncbi.nlm.nih.gov
National Center for Biotechnology Information – National Library of Medicine – Bethesda MD 20892 USA
NCBI STRIDES Project

Promoting an ecosystem to advance genome research through cloud-centered compute environments
Figure 1. Growth of NCBI Data and Services, 1989-2017 Credit: NCBI
Public SRA 5PB

RNaseq 40%

Prokaryote & Metagenome 10%

Genomes 50%
SDDP Brokers Access to Cloud-Hosted Data

Submitter

Grants Read Access to File
Submission describing File

Metadata

SRA

Reference to location (Cloud/Bucket/Key) and access permission

Authorized User

File Info

SRR# Request

Bucket owned by Data Owner

Cloud

Billing/accounting