

Antibiotic Resistance Genes from Whole Genome Sequences

Michael Feldgarden

NCBI/NLM/NIH

Michael.Feldgarden@nih.gov



U.S. National Library of Medicine
National Center for Biotechnology Information

Pathogens

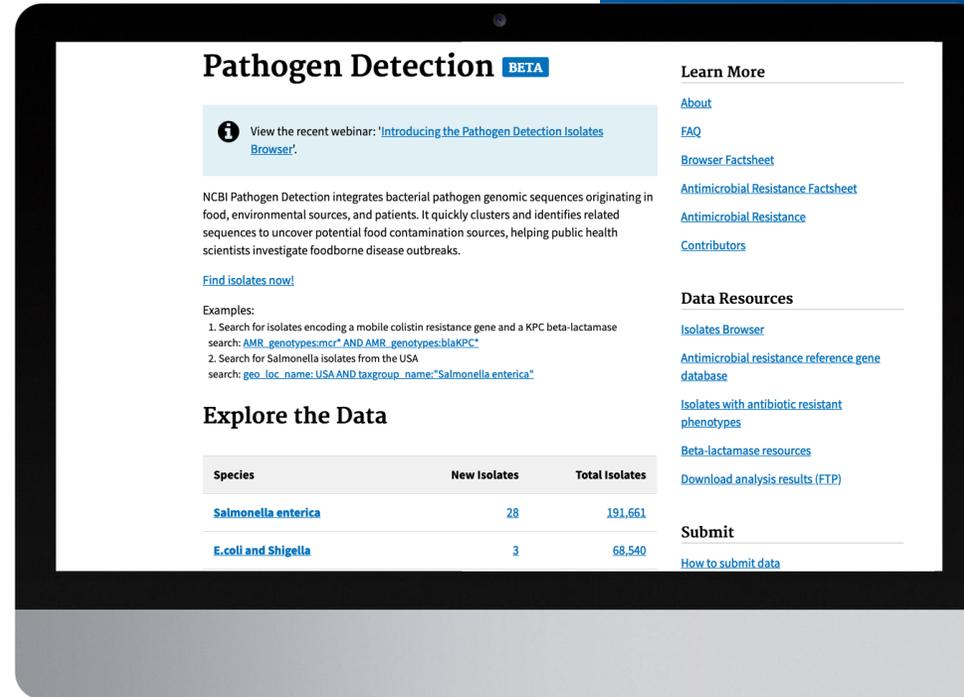
Enables public health scientists to quickly identify foodborne pathogens and antimicrobial resistance, accelerating the U.S. response time to outbreaks and food recalls.

90,000

Isolates analyzed
in the last year

>390,000

Records



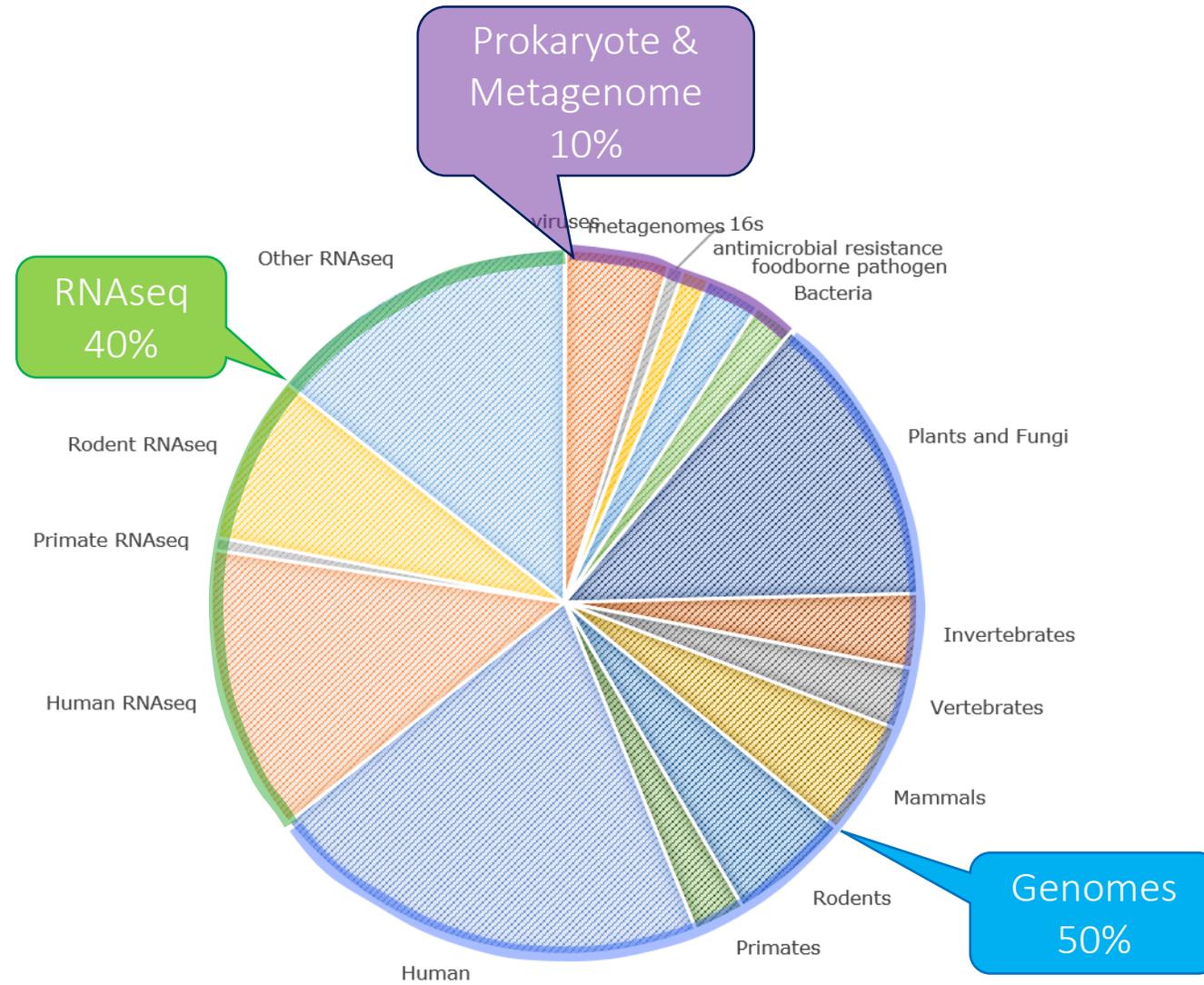
FDA has used the Pathogen Detection analysis results as part of 370 regulatory and compliance actions.

Achieved five-fold reduction of analysis time for Salmonella submissions.

Released National Database of Antimicrobial Resistant Organisms.

Developed and released AMRFinder tool to identify AMR genes in new isolates.

Public SRA 5PB



The scale of data is growing

1. What tools will be needed to provide meaningful access to the growing amount of genomes.

2. Blast interfaces are becoming problematic with growing genome databases.

3. Metagenomic data: characterization and isolation of genes without reference to species.

Need for the STRIDES Initiative

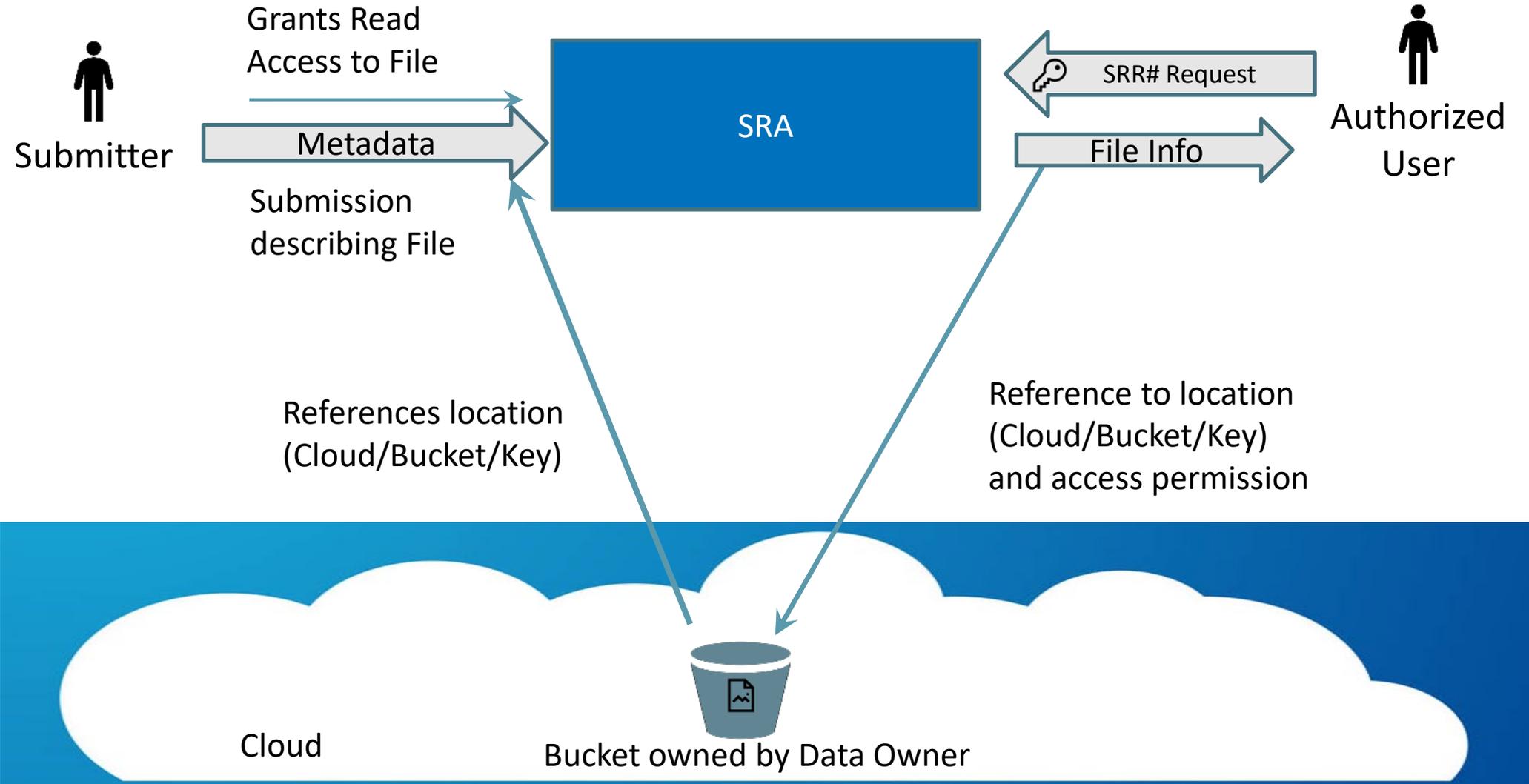
The NIH supports research projects that generate **tremendous amounts of biomedical data...**

...data have traditionally been stored and made available to the broader community via **public repositories** or at local institutions...

...model has become strained as the number of data-intensive projects, and the **amount of data generated in each project, continue to grow.**

<https://commonfund.nih.gov/strides>

SDDP Brokers Access to Cloud-Hosted Data



🔑 Billing/accounting

STRIDES Landing Page (Coming soon! - July, 2019)

For Scientists who want to get to and use data in the cloud

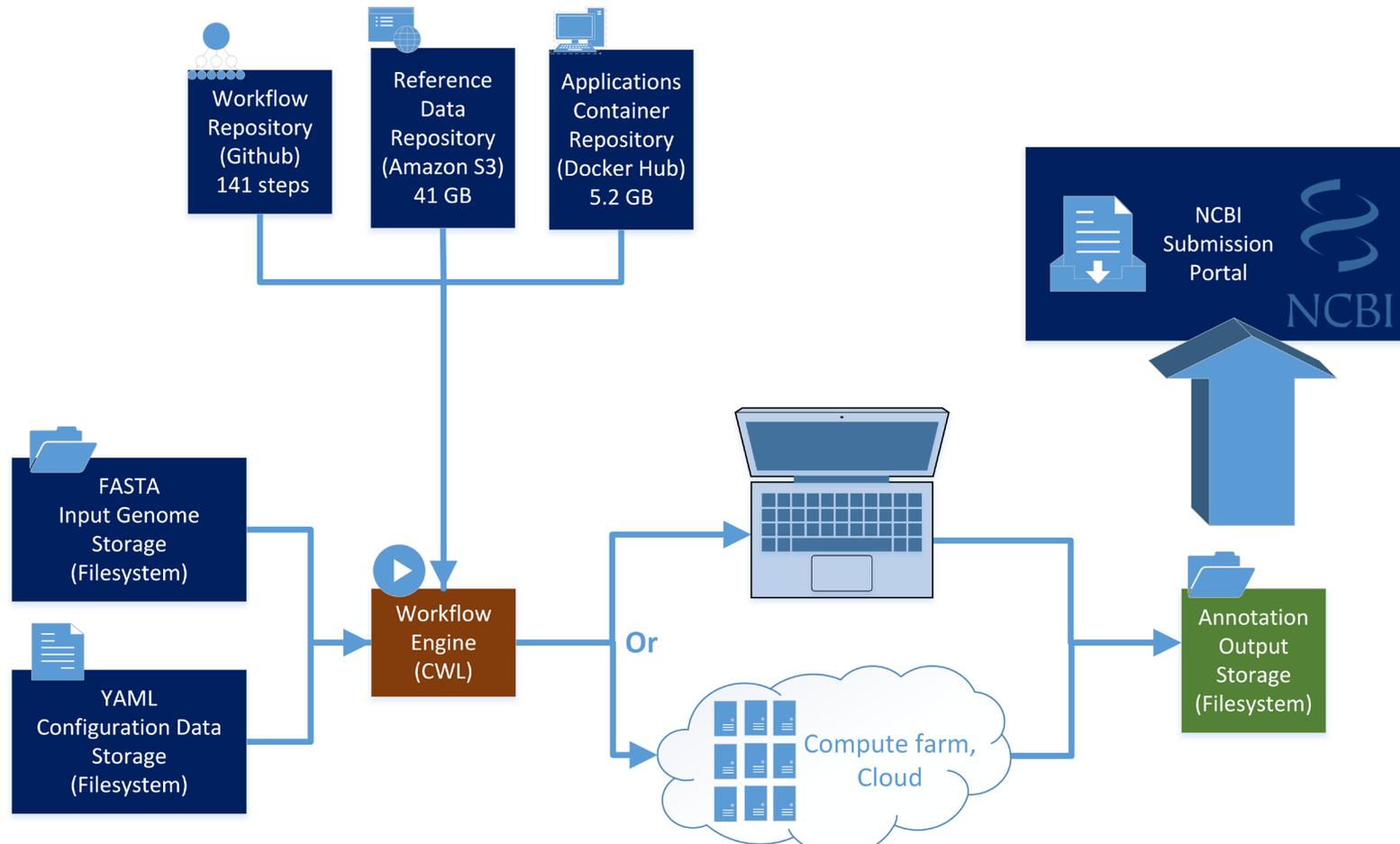
- Where to go for help/get started
- Intro to tools / tutorial
 - Intro to SRA Run Selector to SRA Toolkit
 - Tutorial - SRA Run Selector + Toolkit - STRIDES data
 - Blast Plus intro + tutorial
 - PGAPx
- Cost explanations
 - Cost ranges/estimates
 - Cost prediction tool(s).

- Details on identity management, hackathon schedules, technical details, etc.

Example of cloud-based computing tools: PGAP

The Prokaryotic Genome Annotation Pipeline is now available for download.
You can annotate your genomes on your own machine, a local cluster or the Cloud!

Poster P398
Fri 11am -12pm,4-5 pm
MBP10 - Tools for Genetics & Genomics



The package includes:

- ✓ A Docker image containing the applications
- ✓ The CWL workflow
- ✓ The reference data
- ✓ `cwltool`, the reference implementation for CWL

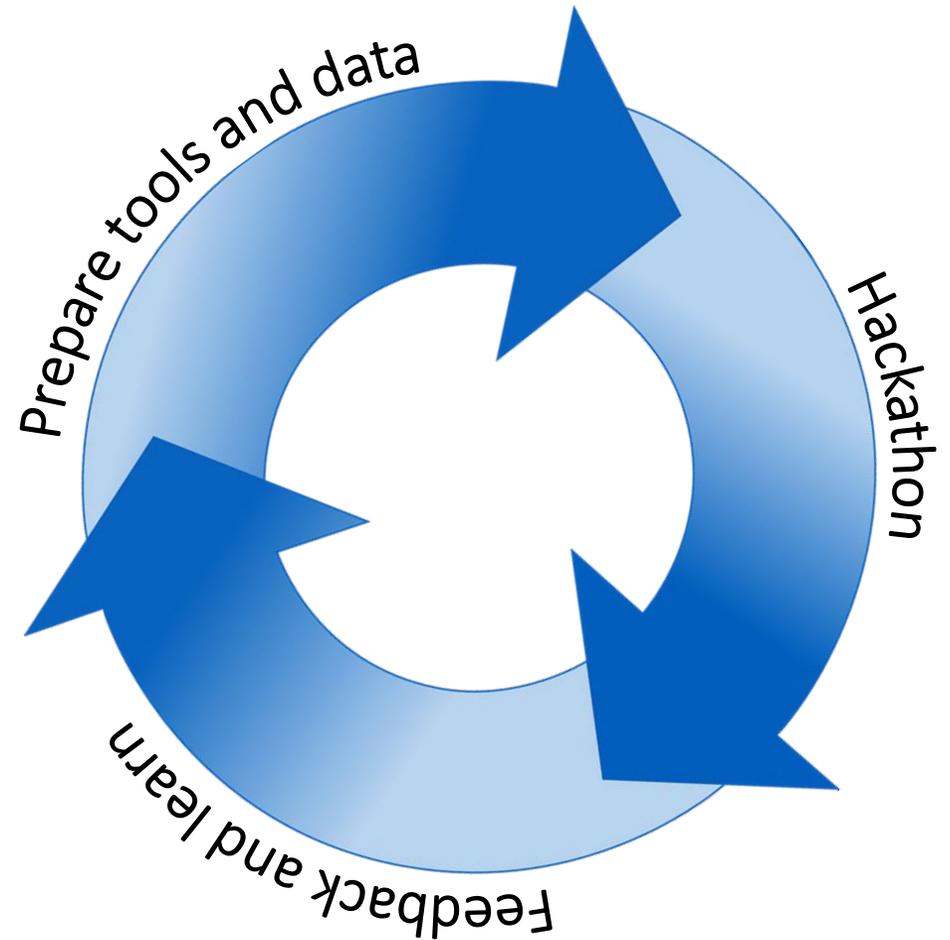
The results are:

- ✓ Compliant with GenBank submission requirements
- ✓ Conforming to PGAP executed at NCBI

Prokaryotic Genome Annotation Pipeline: Saturday: 12:30-12:45 PM

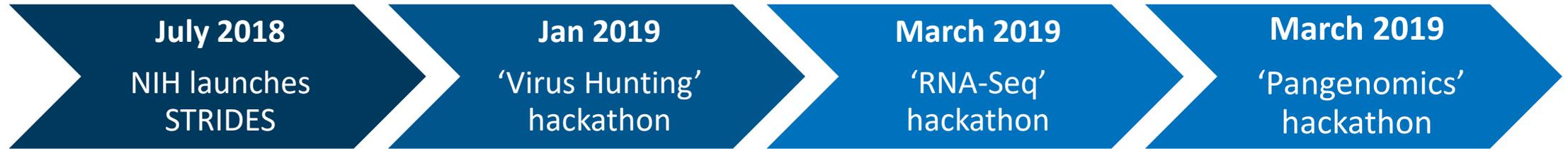
The NCBI hackathon approach

- Each hackathon framed by scientific questions
- Tools & data necessary to answer questions delivered to cloud
- Lessons learned from hackathon are integrated into data and tool improvements
- Stable workflows then exposed in educational workshops



Contact us at info@ncbi.nlm.nih.gov

STRIDES Hackathon progress



- Viruses and Microbes
 - Representational contigs created from SRA data to aid in analysis
 - Viral content in metagenomic SRA datasets identified
 - Facilitate identification of samples containing organism(s) of interest
 - Support biological analysis across large numbers of SRA samples
- Human Genetics
 - Tools and references to build “normalized” RNA-seq data representations and support cross sample analysis

Contact us at info@ncbi.nlm.nih.gov

NCBI is interested in getting feedback on what else is needed

- Browsable phylogenetic trees for tens to hundreds of thousands of genomes?
- Presence of biosystems and systems biology?
- PGAPx annotation testers

Pathogens

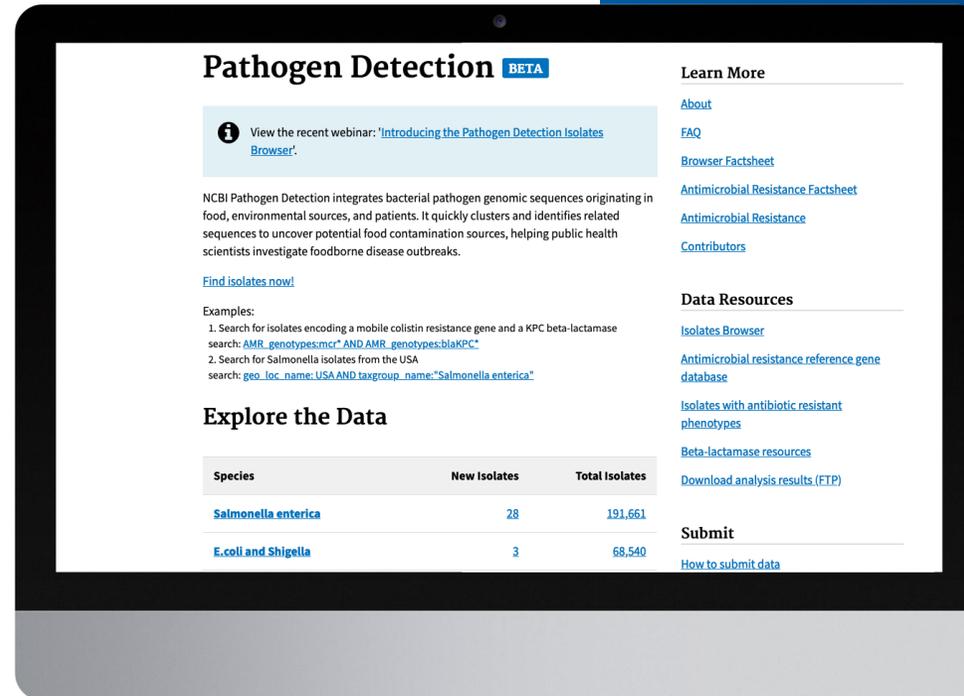
Enables public health scientists to quickly identify foodborne pathogens and antimicrobial resistance, accelerating the U.S. response time to outbreaks and food recalls.

90,000

Isolates analyzed
in the last year

363,164

Records



FDA has used the Pathogen Detection analysis results as part of 370 regulatory and compliance actions.

Achieved five-fold reduction of analysis time for Salmonella submissions.

Released National Database of Antimicrobial Resistant Organisms.

Developed and released AMRFinder tool to identify AMR genes in new isolates.

NDARO

[Health](#) > [Pathogen Detection](#) > National Database of Antibiotic Resistant Organisms (NDARO)

Search page ^ v

National Database of Antibiotic Resistant Organisms (NDARO)

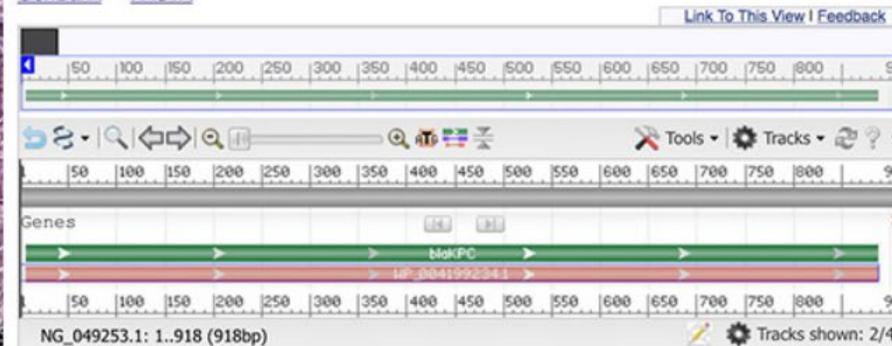
Welcome to the NCBI National Database of Antibiotic Resistant Organisms (NDARO), a collaborative, cross-agency, centralized hub for researchers to access AMR data to facilitate real-time surveillance of pathogenic organisms.



Klebsiella pneumoniae blaKPC gene for carbapenem-hydrolyzing class A beta-lactamase KPC-2, complete CDS

NCBI Reference Sequence: NG_049253.1

[GenBank](#) [FASTA](#)



From left to right: Multi-drug resistant *Salmonella enterica*, kpc2 carbapenem resistance gene

National Action Plan for Combating Antibiotic-Resistant Bacteria (CARB)

- NCBI is partnered with several outside agencies, including the FDA, CDC, USDA, WHO, PHE, and others to take the following steps:
 - To increase standardization, NCBI has developed and maintains a **curated database of AMR genes**.
 - To make AMR-related data more widely available, NCBI is **collecting genetic and antibiotic susceptibility data**.
 - To make more effective use of bacterial genomic data, NCBI has **developed AMRFinder to identify AMR genes in bacterial genomes**.
 - To assist researchers and public health officials, NCBI has **developed the Isolate Browser to allow researchers to identify bacterial genomes with AMR genes**.

Pathogen Detection Portal

results presented in an **easy-to-use web-based interface** and updated <24 hours after new data are submitted

321 newly added *Salmonella* isolates

Species	New Isolates	Total Isolates
Salmonella enterica	321	199,084
E.coli and Shigella	28	73,640
Campylobacter jejuni	73	30,270
Listeria monocytogenes	6	27,061
See more organisms...		

28 different taxa including all four major foodborne bacterial pathogens - more than 390,000 isolates

Top Organisms [\[Tree\]](#)

- Salmonella enterica (225257)
- Escherichia coli (102719)
- ★ Streptococcus pneumoniae (57097)
- Mycobacterium tuberculosis (56753)
- ★ Staphylococcus aureus (56555)
- Listeria monocytogenes (30568)
- Campylobacter jejuni (28219)
- ★ Streptococcus pyogenes (26907)
- Neisseria meningitidis (20454)
- Klebsiella pneumoniae (17141)
- Neisseria gonorrhoeae (14075)
- ★ Enterococcus faecium (12427)
- Clostridioides difficile (11264)
- ★ Streptococcus agalactiae (11090)
- Campylobacter coli (10217)
- Shigella sonnei (9816)
- Campylobacter sp. (8075)
- Pseudomonas aeruginosa (7951)
- Vibrio cholerae (6981)
- Shigella flexneri (6223)
- All other taxa (121943)

coming soon!

S. aureus

Streptococcus

Enterococcus

SRA Search: Bacteria[orgn] AND Illumina[platform] AND wgs[strategy]

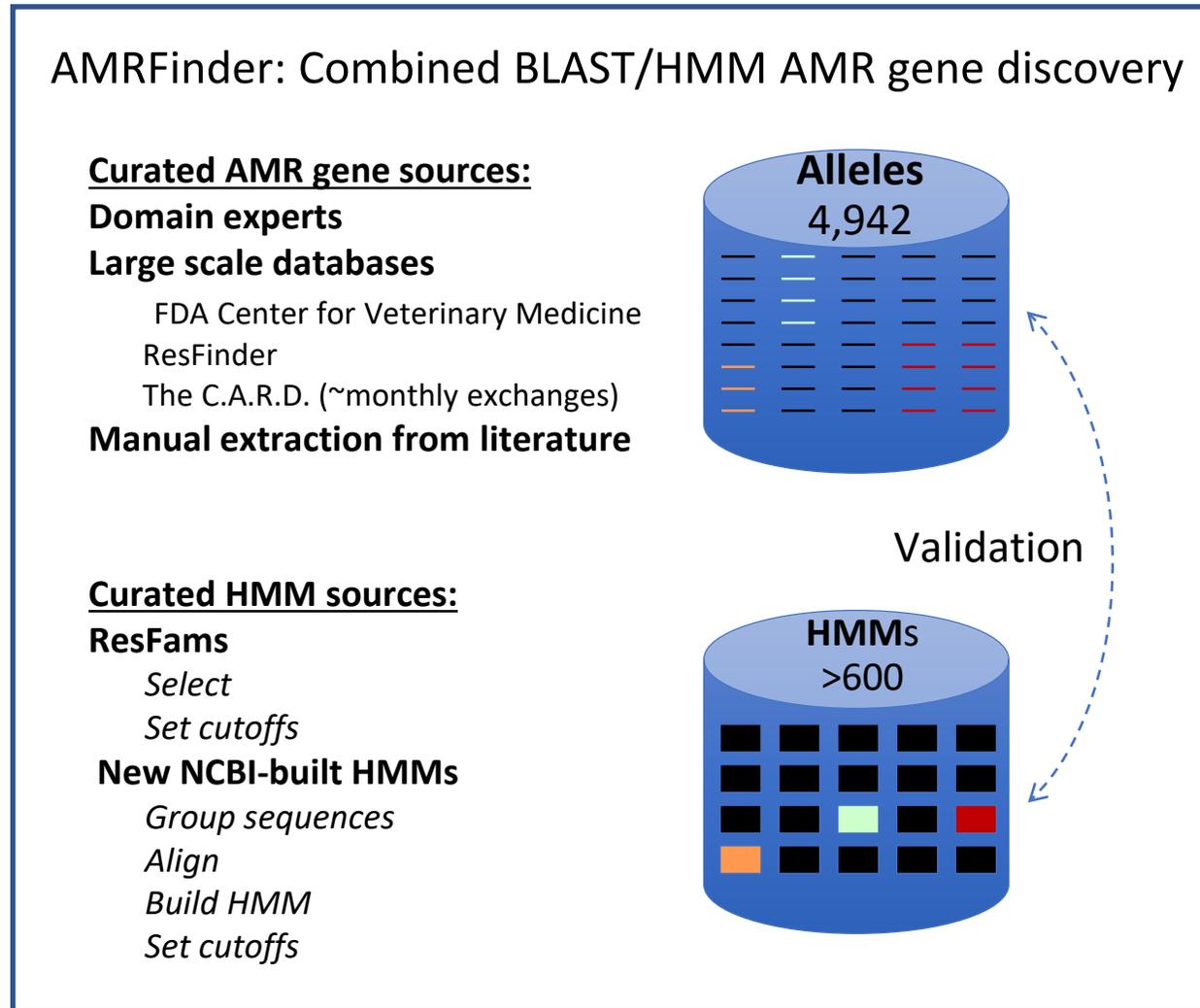
Automating AMR Gene Detection

Pathogen Detection Pipeline (SRA):

GenomeTrakr
PulseNet
PHE
FDA/CDC Antimicrobial Resistant Isolate Bank
State laboratories
Clinical laboratories

Genbank assemblies:

General submissions



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

Isolates Browser landing page

Health > Pathogen Detection > Isolates Browser

<https://www.ncbi.nlm.nih.gov/pathogens/isolates/>

Find one or more isolates ...

Select an organism group

Page 1 of 19,186
View 1 - 20 of 383,716

#	Organism Group	Strain	Serovar	Isolate	Create Date	Location	Isolation Sou	Isolation type	Host	SNP cluster	Min-sar	Min-diff	BioSample	Assembly	AMR genotypes
1	Salmonella enterica	PNUSAS06		PDT000518134.1	2019-06-05	USA		clinical			n/a	n/a	SAMN11962413		
2	Salmonella enterica	PNUSAS07		PDT000518135.1	2019-06-05	USA		clinical		PDS000026800.35	12	20	SAMN11962417		
3	Salmonella enterica	PNUSAS07		PDT000518130.1	2019-06-05	USA		clinical			n/a	n/a	SAMN11961811		
4	Salmonella enterica	PNUSAS07		PDT000518133.1	2019-06-05	USA		clinical		PDS000026769.7	0	n/a	SAMN11961799		
5	Salmonella enterica	PNUSAS07		PDT000518132.1	2019-06-05	USA		clinical		PDS000042752.95	15	15	SAMN11961803		aac(3)-II aac(6)-IIc aadA2 Show all 16 genes
6	Salmonella enterica	PNUSAS07		PDT000518131.1	2019-06-05	USA		clinical		PDS000002504.307	10	10	SAMN11961796		aph(3'')-Ib aph(6)-Id blaCMY-2 Show all 6 genes

Isolates Browser landing page

View 1 - 20 of 383,716

AMR genotypes



U.S. National Library of Medicine
National Center for Biotechnology Information

Log in

[Health](#) > [Pathogen Detection](#) > Isolates Browser

Find one or more isolates ...

Search

Select an organism group

Filters

Expand All

Download

#	Organism Group	Strain	Serovar	Isolate	Create Date	Location	Isolation Sou	Isolation type	Host	SNP cluster	Min-sar	Min-difi	BioSample	Assembly	AMR genotypes
1	Salmonella enterica	PNUSAS06		PDT000518134.1	2019-06-05	USA		clinical			n/a	n/a	SAMN11962413		
2	Salmonella enterica	PNUSAS07		PDT000518135.1	2019-06-05	USA		clinical		PDS000026800.35	12	20	SAMN11962417		
3	Salmonella enterica	PNUSAS07		PDT000518130.1	2019-06-05	USA		clinical			n/a	n/a	SAMN11961811		
4	Salmonella enterica	PNUSAS07		PDT000518133.1	2019-06-05	USA		clinical		PDS000026769.7	0	n/a	SAMN11961799		
5	Salmonella enterica	PNUSAS07		PDT000518132.1	2019-06-05	USA		clinical		PDS000042752.95	15	15	SAMN11961803		aac(3)-II aac(6)-IIc aadA2 Show all 16 genes
6	Salmonella enterica	PNUSAS07		PDT000518131.1	2019-06-05	USA		clinical		PDS000002504.307	10	10	SAMN11961796		aph(3'')-Ib aph(6)-Id blaCMY-2 Show all 6 genes



U.S. National Library of Medicine
National Center for Biotechnology Information

<https://www.ncbi.nlm.nih.gov/pathogens/isolates/>



Large Scale Requires Concise Information

- hundreds of genomes per day
- can't be 'artisanal'; flipping through multiple columns/rows/tables will not work
- Need *concise, discrete signifier* that conveys appropriate information about genotype (and possibly phenotype)

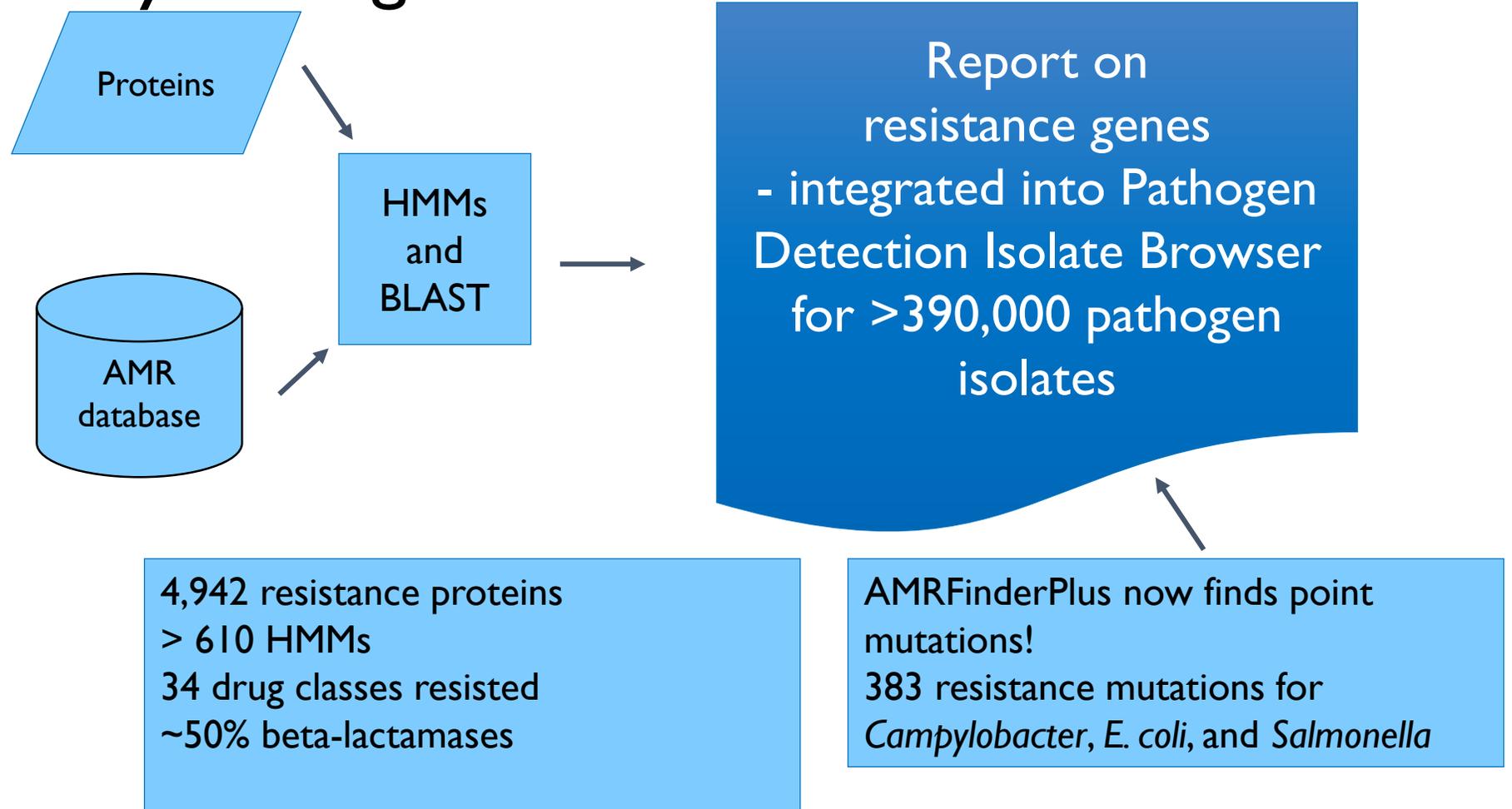
That signifier is the gene symbol

- *blaKPC-2* (KPC-2) is a shorthand for gene function:
 - “carbapenem-hydrolyzing class A beta-lactamase KPC-2”
- Need to specify precisely what we do *and* do **not** know:
 - *blaKPC-2*, *blaKPC*, and *bla* do not mean the same thing
- We do not want to:
 - *overspecify*: claim KPC-2 when only KPC-family
 - *underspecify*: claim KPC-family when KPC-28
 - *misclassify*: call KPC-2 as KPC-28
- For AMR genes, we have developed **AMRFinderPlus** to identify AMR genes in genomic data

AMRFinderPlus Has a Hierarchical Structure

	Protein name	
Exact match	KPC-2	<i>Resistance to carbapenems and other beta-lactam antibiotics.</i>
HMM score > cutoff of KPC family	KPC family	<i>Likely resistance to carbapenems and other beta-lactam antibiotics.</i>
HMM score > cutoff	class A b-lactamase	<i>Class A beta-lactamase of unknown specificity.</i>
HMM score < cutoff	not beta-lactamase	Prevents false-positive identification as a beta-lactamase. Not reported.

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



Isolates Browser landing page

View 1 - 20 of 383,716

AMR genotypes



U.S. National Library of Medicine
National Center for Biotechnology Information

Log in

[Health](#) > [Pathogen Detection](#) > Isolates Browser

Find one or more isolates ...

Search

Select an organism group

Filters

Expand All

Download

#	Organism Group	Strain	Serovar	Isolate	Create Date	Location	Isolation Sou	Isolation type	Host	SNP cluster	Min-sar	Min-difi	BioSample	Assembly	AMR genotypes
1	Salmonella enterica	PNUSAS06		PDT000518134.1	2019-06-05	USA		clinical			n/a	n/a	SAMN11962413		
2	Salmonella enterica	PNUSAS07		PDT000518135.1	2019-06-05	USA		clinical		PDS000026800.35	12	20	SAMN11962417		
3	Salmonella enterica	PNUSAS07		PDT000518130.1	2019-06-05	USA		clinical			n/a	n/a	SAMN11961811		
4	Salmonella enterica	PNUSAS07		PDT000518133.1	2019-06-05	USA		clinical		PDS000026769.7	0	n/a	SAMN11961799		
5	Salmonella enterica	PNUSAS07		PDT000518132.1	2019-06-05	USA		clinical		PDS000042752.95	15	15	SAMN11961803		aac(3)-II aac(6)-IIc aadA2 Show all 16 genes
6	Salmonella enterica	PNUSAS07		PDT000518131.1	2019-06-05	USA		clinical		PDS000002504.307	10	10	SAMN11961796		aph(3'')-Ib aph(6)-Id blaCMY-2 Show all 6 genes



U.S. National Library of Medicine
National Center for Biotechnology Information



Searching for KPC and MCR positive organisms



Searching for carbapenem sensitive, *blaKPC*+ isolates



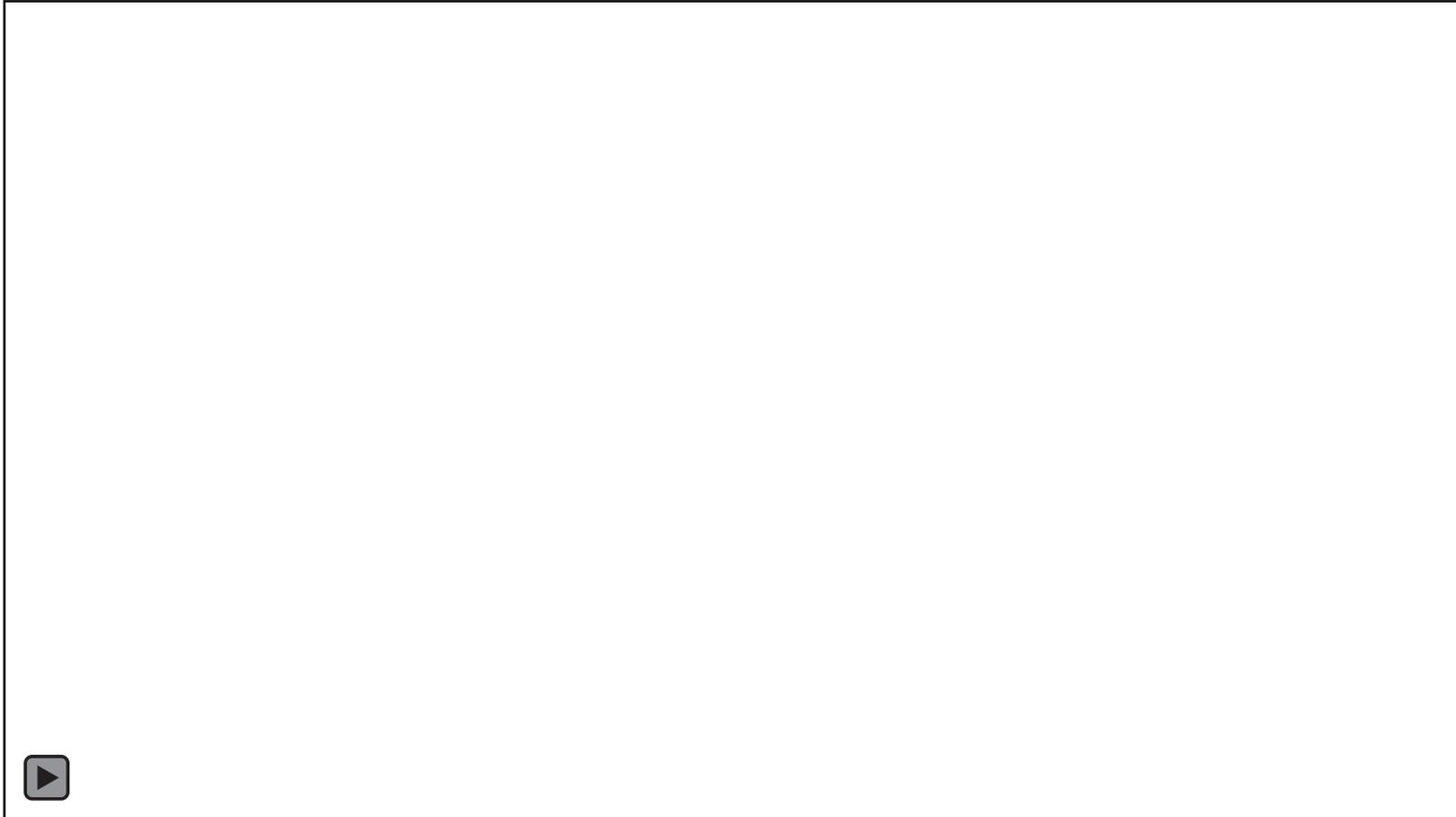
Reference Gene Catalog



Reference Gene Catalog

- what it enables:
 - search for genes or alleles
 - provides accessions
 - standardized names
 - link out to every isolate in Isolates Browser w/that gene
 - downloadable

Reference Gene Catalog



Amikacin resistance genes

#	gene family	product name	type	subtype	class	subclass	refseq protein	refseq	genbank	genbank nucle	curated refseq start
1	aphA16	APH(3') family aminoglycoside O-phosphotransferase AphA16	AMR	AMR	AMINOGLYCOSIDE	AMIKACIN/KANAMYCIN	WP_004206941.1	NG_05	AKL037	CP011593.1	No
2	aph(3')-XV	aminoglycoside O-phosphotransferase APH(3')-XV	AMR	AMR	AMINOGLYCOSIDE	AMIKACIN/KANAMYCIN	WP_034067921.1	NG_05	ABY489	EU165039.1	No
3	aph(3')-XV	aminoglycoside O-phosphotransferase APH(3')-XV	AMR	AMR	AMINOGLYCOSIDE	AMIKACIN/KANAMYCIN	WP_032492579.1	NG_05	ACX940	GQ926879.1	No
4	aph(3')-XV	aminoglycoside O-phosphotransferase APH(3')-XV	AMR	AMR	AMINOGLYCOSIDE	AMIKACIN/KANAMYCIN	WP_019407932.1	NG_04	CAD913	Y18050.2	No
5	aph(3')-VIb	aminoglycoside O-phosphotransferase APH(3')-VIb	AMR	AMR	AMINOGLYCOSIDE	AMIKACIN/KANAMYCIN	WP_000422633.1	NG_04	CAF294	AJ627643.4	No
6	aph(3')-VIa	aminoglycoside O-phosphotransferase APH(3')-VIa	AMR	AMR	AMINOGLYCOSIDE	AMIKACIN/KANAMYCIN	WP_000422636.1	NG_04	CAA305	X07753.1	No
7	aph(3')-VIIa	aminoglycoside O-phosphotransferase APH(3')-VIIa	AMR	AMR	AMINOGLYCOSIDE	AMIKACIN/KANAMYCIN	WP_063842175.1	NG_04	AAA768	M29953.1	No

Quinolone resistance point mutations

subtype:POINT AND subclass:QUINOLONE ✕ 🔍 Search

Database version: 2019-04-29.1 ▼ Filters

🔄 Choose Columns Page 1 of 7 20

#	allele	gene family	product name	type	subtype	class	subclass	refseq protein	organism
1	gyrA_A119E	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281271.1	Salmonella
2	gyrA_A119S	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281271.1	Salmonella
3	gyrA_A119V	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281271.1	Salmonella
4	gyrA_A131G	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281271.1	Salmonella
5	gyrA_A196E	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281243.1	Escherichia
6	gyrA_A51V	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281243.1	Escherichia
7	gyrA_A67P	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281271.1	Salmonella
8	gyrA_A67S	gyrA	DNA gyrase subunit A	AMR	POINT	QUINOLONE	QUINOLONE	WP_001281242.1	Escherichia

Automated Alerts: Email Notifications

Dear Colleagues:

New Salmonella enterica results are available.

2019-06-15T21:25:58 ET

Query "hygromycin/apramycin" : AMR_genotypes:aph(4)-I* OR AMR_genotypes:aph(7'')-I*

Summary of found isolates:

Total isolates: 1, clinical: 0, environmental: 1

Clustered (in 1 clusters):1, clinical: 0, environmental:

Unclustered: 0, clinical: 0, environmental: 0

Cluster list:

1. PDS000003955.415 hits:1 clinical:0 environmental:1

<https://www.ncbi.nlm.nih.gov/Structure/tree/#!/tree/Salm>

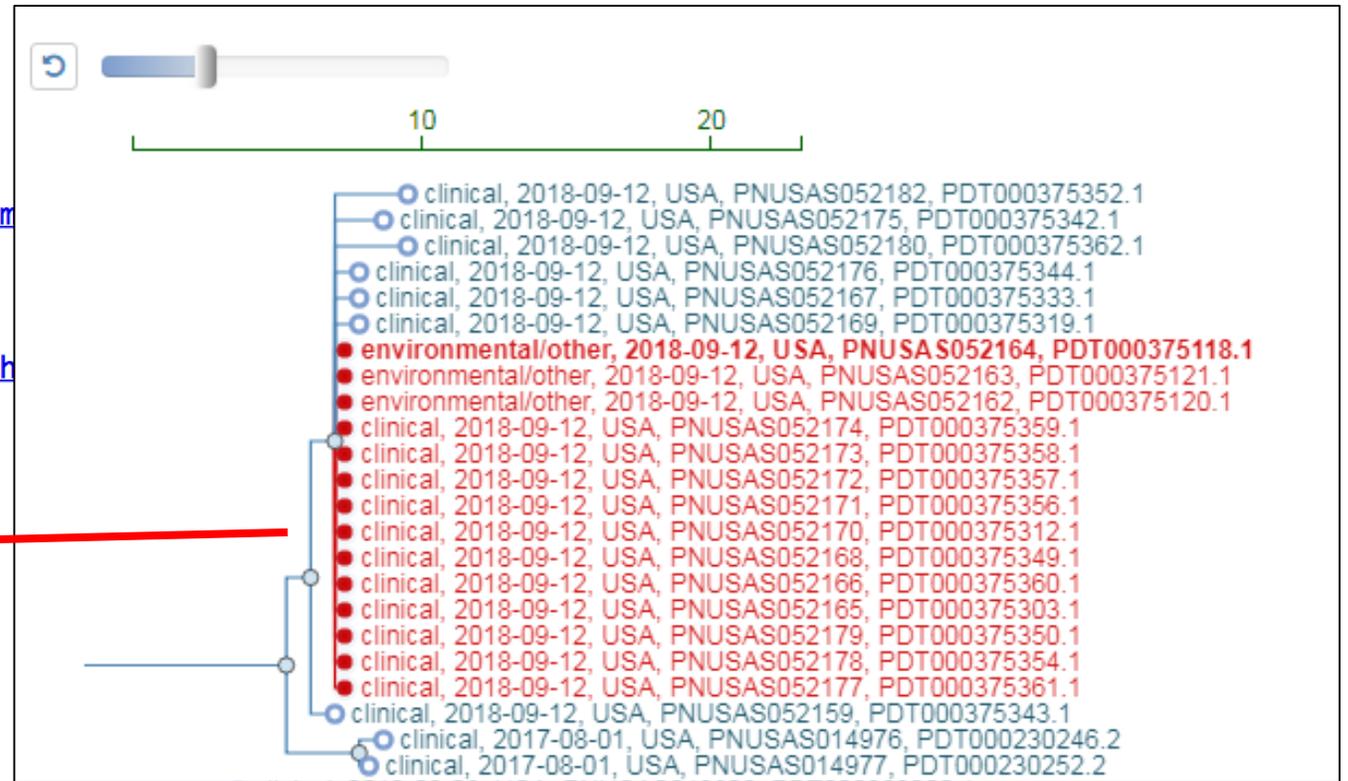
Total list of isolate(s):

hits:1 clinical:0 environmental:1

<https://www.ncbi.nlm.nih.gov/pathogens/isolates/#/search>

Requires MyNCBI login

- email includes links highlighting new isolates within SNP trees when opened in browser



Conclusions

- The huge amount of data requires annotation systems that are rapid, accurate, and can convey information concisely
- AMRFinder combines a manually curated AMR gene database, curated HMMs, and BLAST to identify genes and provide specific gene symbols
- The Isolates Browser can be used to track small-scale changes in AMR gene content
- The Reference Gene Catalog interface also allows users to search for isolates with AMR genes of interest

Acknowledgements

NCBI

Richa Agarwala
Azat Badretdin
Slava Brover
Joshua Cherry
Vyacheslav Chetvernin
Robert Cohen
Michael DiCuccio
Daniel Haft
Arjun Prasad
Douglas Slotta
Edward Rice
Kirill Rotmistrovsky
Stephen Sherry
Sergey Shiryev
Martin Shumway
Tatiana Tatusova
Igor Tolstoy
Chunlin Xiao
Leonid Zaslavsky
Alexander Zasytkin
Alejandro A. Schaffer
Lukas Wagner
Aleksandr Morgulis

William Klimke
Kim Pruitt
James Ostell

FDA/CVM

Patrick McDermott
Greg Tyson
Shaohua Zhao

CDC

Jason Folster

USDA

Glenn Tillman
Cesar Morales
Mustafa Simmon
Jamie Wasilenko

NIH/NIAID

US ARMY/MEDCOM
FDA/CFSAN

JCVI

Broad Institute
Brigham & Women's Hospital
Wadsworth Institute

pd-help@ncbi.nlm.nih.gov

ncbi.nlm.nih.gov/pathogens

- 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 Resources

AMRFinder is publicly available for use in your pipeline:

<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/>

Curated AMR gene download:

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047>

AMR HMM download:

<https://ftp.ncbi.nlm.nih.gov/hmm/NCBIfam-AMRFinder/>

Table of AMR gene accessions and names:

<https://www.ncbi.nlm.nih.gov/pathogens/isolates#/refgene/>

Isolate Browser:

<https://www.ncbi.nlm.nih.gov/pathogens/isolates>

Questions: pd-help@ncbi.nlm.nih.gov



Ask Us
for a full schedule
or posters & slides

The NCBI team would like to learn more about workflows and **user needs for large scale analyses of microbial genomes** to support our users better.

We would be especially interested in anyone interested in doing **large-scale data analyses using cloud-based resources**.

If you are interested in participating and would be willing to **share your email address, contact our personnel at booth #443**.