Antimicrobial Resistance Informatics at NCBI

An NCBI project to identify antimicrobial resistance genes

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

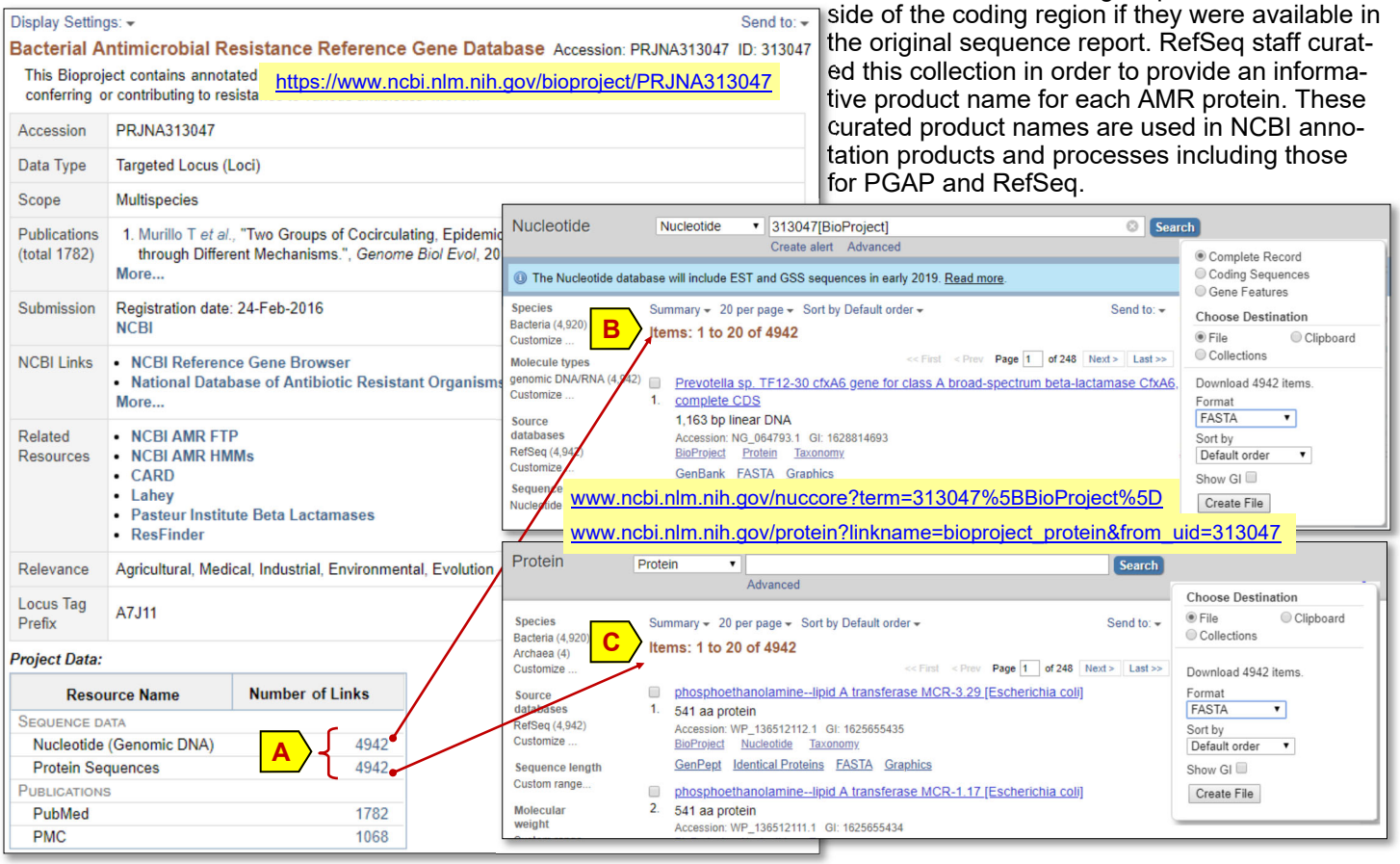
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope

Genome sequencing not only provides critical, fine-scale genetic information that can assist in tracking the spread of infectious organisms, but it also enables the identification of antibiotic resistance genes, which is a critical step in understanding the relationship between resistance phenotype and genotype. As part of the National Action Plan For Combating Antimicrobial-Resistant Bacteria, NCBI is establishing the National Sequence Database of Resistant Pathogens. The database will serve as a central location for collecting and presenting antibiograms (which show the sensitivity or resistance of a bacterial strain to different antibiotic compounds) associated with submitted bacterial genomes. Phenotypes are now displayed in the NCBI Pathogen Detection Isolates Browser for those isolates that have submitted phenotypic tests and are for the organisms that are actively being analyzed for cluster detection to support outbreak detection and traceback investigations.

Data Access

NCBI has developed the Bacterial Antimicrobial Resistance Reference Gene Database, a collection of acquired antimicrobial resistance genes. This collection aggregates collections from multiple sources, curates conflicts, and adds sequences found by review of the literature. Major contributors to NCBI reference set of antimicrobial resistance (AMR) genes include the Lahey group [1], Resfinder [2], the Center for Veterinary Medicine at the Food and Drug Administration, the Comprehensive Antimicrobial Resistance Database (CARD) [3], the Institut Pasteur [4], Dr. Marilyn Roberts [5], and Dr. Derrick Crook. Curation was aided by the use of hidden Markov models (HMMs) from ResFams [6], Pfam [7], and TIGRFAMs collections. Nucleotide records were constructed to include 100 bases of flanking sequences of each side of the coding region if they were available in the original sequence report. RefSeq staff curated this collection in order to provide an informative product name for each AMR protein. These curated product names are used in NCBI annotation products and processes including those for PGAP and RefSeq.



The sequences from this project are collected in the Bioproject record PRJNA313047, readily retrievable using links in the Project Data table (**A**), and can be downloaded to local disk through the "Send to" menu options available in the Nucleotide and Protein databases (**B**, **C**). We can also use the Linux command lines provided by EDirect package [8] to retrieve the sequences:

esearch -db nucleotide -query "313047[bioproject]" | efetch -db nucleotide -format fasta >AMR_nt.fa
elink -db bioproject -id 313047 -target protein | efetch -db protein -format fasta >AMR_aa.fa

NCBI also makes available a data table available through FTP, https://go.usa.gov/xQpvQ, which lists protein and nucleotide accessions, gene symbols, and product names. Additionally, NCBI has developed a set of Hidden Markov Models (HMMs) for use with HMMER (available from http://hmmer.org/) to detect AMR genes. The latest HMM dataset is available from NCBI at: https://ftp.ncbi.nlm.nih.gov/hmm/NCBIfam-AMR/latest/.

AMRFinder, A Tool For Detecting AMR Genes

NCBI has developed AMRFinder, to detect AMR genes in genomic data. This tool is used to generate the resistance genes output in the Isolates Browser described above. AMRFinder uses the AMR protein sequences or genomic nucleotide sequences, HMMs, the hierarchy of gene families and a set of custom rules to generate a report of the names, symbols, and coordinates of acquired AMR genes along with descriptions of the evidence used to identify the sequence (https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/).

AMRFinder runs within a Docker framework to enable use of the latest software and database versions. With a one-line command, a list of annotated proteins, and an optional gff file, AMRFinder can generate a list of AMR proteins. For example, with the following command, we can assess the AMR gene content of Klebsiella pneumoniae isolate KPNIH24 represented by the genome assembly GCF_000714675.1:

amrfinder.sh -n GCF_000714675.1_ASM71467v1_protein.faa [available at https://go.usa.gov/xQv4a]

The above command yields a tab-delimited text file output. For clarity, only the first 15 rows are shown in two sections to cover all columns. The first six columns are:

A. Target identifier: showing the accession for the protein sequence.
B. Contig id: showing the accession of the contig that contains the AMR protein:
C. Start, Stop, and Strand: listing the coordinates and the orientation of the AMR gene coding sequence
D. Gene symbol: showing the most accurate gene symbol possible. If a specific allele can not be provided, AMRFinder will not overcall the allele. For example, WP_004152783.1 is called as *aac(6')-Ib* rather than a more specific allele as it is not 100% identical to known *aac(6')-Ib* alleles.
E. Protein name: displaying the NCBI curated name for the AMR protein.
F. Method: describing the method used to identify the protein. EXACT means there is a 100% identity and length match to a known resistance gene. ALLELE a 100% identity and length match to a known resistance allele (e.g., *blaTEM-1*). BLAST indicates a high identity match to a known AMR protein. PARTIAL indicates a partial length AMR protein sequence. HMM (not shown) indicates a sequence that was only identified through HMM.

AMRFinder also provides alignment details in additional columns:

Target identifier	Contig id	Start	Stop	Strand	Gene symbo	Protein name	Method
WP_000027057.1	NZ_CP008798.1	46209	47069	+	blaTEM-1	class A broad-spectrum beta-lactamase TEM-1	ALLELE
WP_000034420.1	NZ_CP008798.1	16469	17260	-	sul3	sulfonamide-resistant dihydropteroate synthase Sul3	EXACT
WP_000095725.1	NZ_CP008800.1	55144	56403	+	cmlA1	chloramphenicol efflux MFS transporter CmlA1	EXACT
WP_000219391.1	NZ_CP008798.1	5110	6015	-	mph(A)	Mph(A) family macrolide 2'-phosphotransferase	EXACT
WP_000259031.1	NZ_CP008797.1	4E+06	3564548	-	sul1	sulfonamide-resistant dihydropteroate synthase Sul1	EXACT
WP_000259031.1	NZ_CP008798.1	84433	85272	+	sul1	sulfonamide-resistant dihydropteroate synthase Sul1	EXACT
WP_000347934.1	NZ_CP008797.1	2E+06	1763628	-	oqxB	multidrug efflux RND transporter permease subunit OqxB	EXACT
WP_000679427.1	NZ_CP008797.1	4E+06	3564889	-	qacEdelta1	quaternary ammonium compound efflux SMR transporter QacE delta 1	ALLELE
WP_0 9427.1	NZ_C 798.1	84092	8 9	+	qa lta1	q nary ammonium compound efflux SMR transporter QacE delta 1	LE
WP_0 2814.1	NZ_C 798.1	21156	1	+	ap a	a glycoside O-phosphotransferase APH(4)-Ia	T
WP_000800531.1	NZ_CP008800.1	57457	57789	+	qacL	quaternary ammonium compound efflux SMR transporter QacL	EXACT
WP_001083725.1	NZ_CP008798.1	82232	82729	+	dfrA12	trimethoprim-resistant dihydrofolate reductase DfrA12	EXACT
WP_001199192.1	NZ_CP008798.1	20151	20927	+	aac(3)-IV	aminoglycoside N-acetyltransferase AAC(3)-IV	EXACT
WP_001206316.1	NZ_CP008800.1	56496	57287	+	aadA1	ANT(3'')-Ia family aminoglycoside nucleotidyltransferase AadA1	EXACT
WP_001206356.1	NZ_CP008798.1	83137	83928	+	aadA2	ANT(3'')-Ia family aminoglycoside nucleotidyltransferase AadA2	EXACT

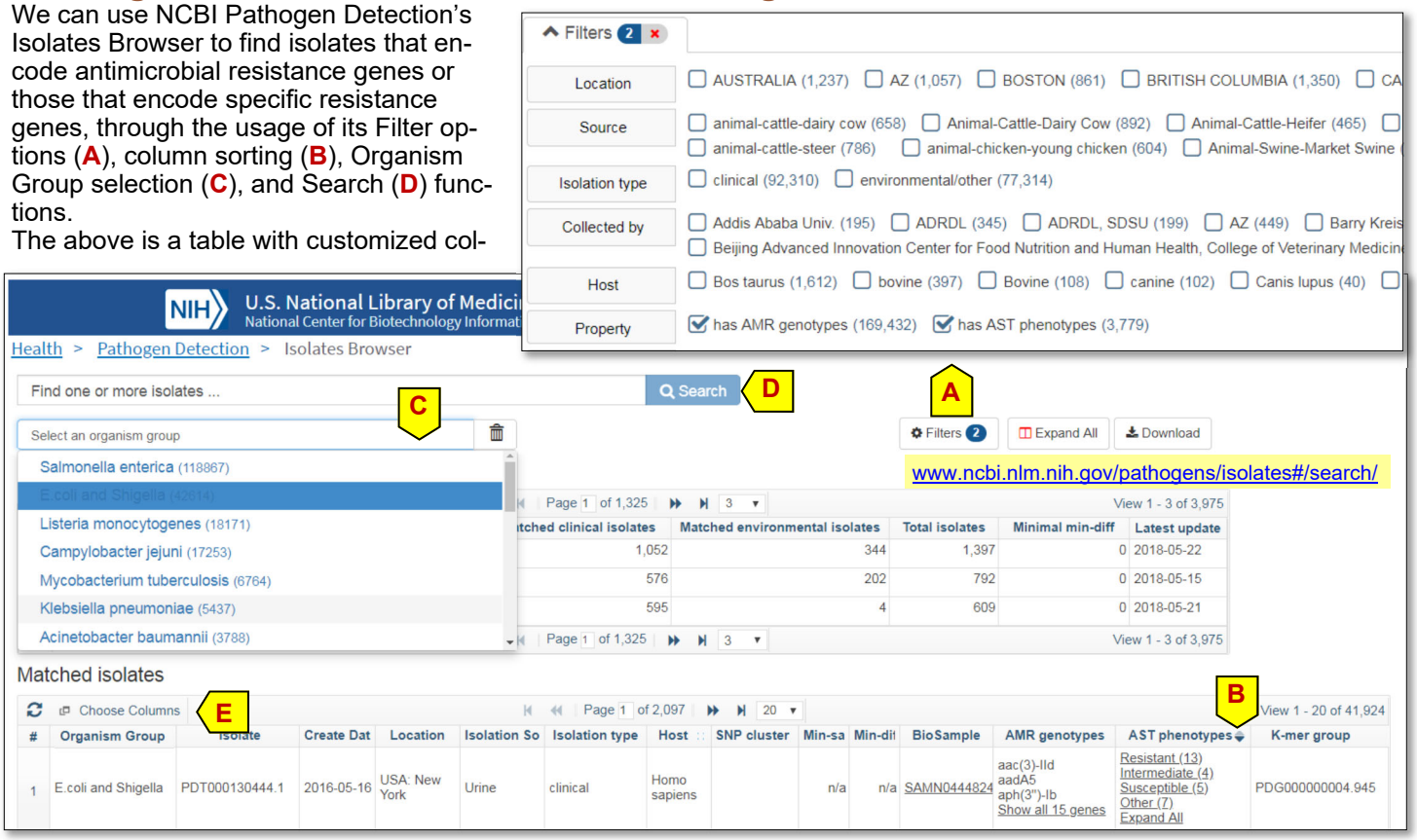
G. Target length: the lengths of the target protein.
H. Reference protein length: the reference protein.
I. % Coverage of … & % Identity to … : the coverage of and identity to the reference protein.
J. Alignment length: the length of the alignment.
K. Accession & Name of closest protein: the closest protein's accession and name.

Target length	Reference protein length	% Coverage of reference protein	% Identity to reference protein	Alignment length	Accession of closest protein	Name of closest protein
286	286	100	100	286	WP_000027057.1	class A broad-spectrum beta-lactamase TEM-1
263	263	100	100	263	WP_000034420.1	sulfonamide-resistant dihydropteroate synthase Sul3
419	419	100	100	419	WP_000095725.1	chloramphenicol efflux MFS transporter CmlA1
301	301	100	100	301	WP_000219391.1	Mph(A) family macrolide 2'-phosphotransferase
279	279	100	100	279	WP_000259031.1	sulfonamide-resistant dihydropteroate synthase Sul1
2	2	100	100	279	WP_0002590 1	sulfonamide-resistant dihydropteroate synthase Sul1
1	1	100	100	105	WP_0003479	multidrug efflux RND transporter permease subunit OqxB
115	115	100	100	115	WP_000679427.1	quaternary ammonium compound efflux SMR transporter QacE delta 1
115	115	100	100	115	WP_000679427.1	quaternary ammonium compound efflux SMR transporter QacE delta 1
341	341	100	100	341	WP_000742814.1	aminoglycoside O-phosphotransferase APH(4)-Ia
110	110	100	100	110	WP_000800531.1	quaternary ammonium compound efflux SMR transporter QacL
165	165	100	100	165	WP_001083725.1	trimethoprim-resistant dihydrofolate reductase DfrA12
258	258	100	100	258	WP_001199192.1	aminoglycoside N-acetyltransferase AAC(3)-IV
263	263	100	100	263	WP_001206316.1	ANT(3'')-Ia family aminoglycoside nucleotidyltransferase AadA1
263	263	100	100	263	WP_001206356.1	ANT(3'')-Ia family aminoglycoside nucleotidyltransferase AadA2

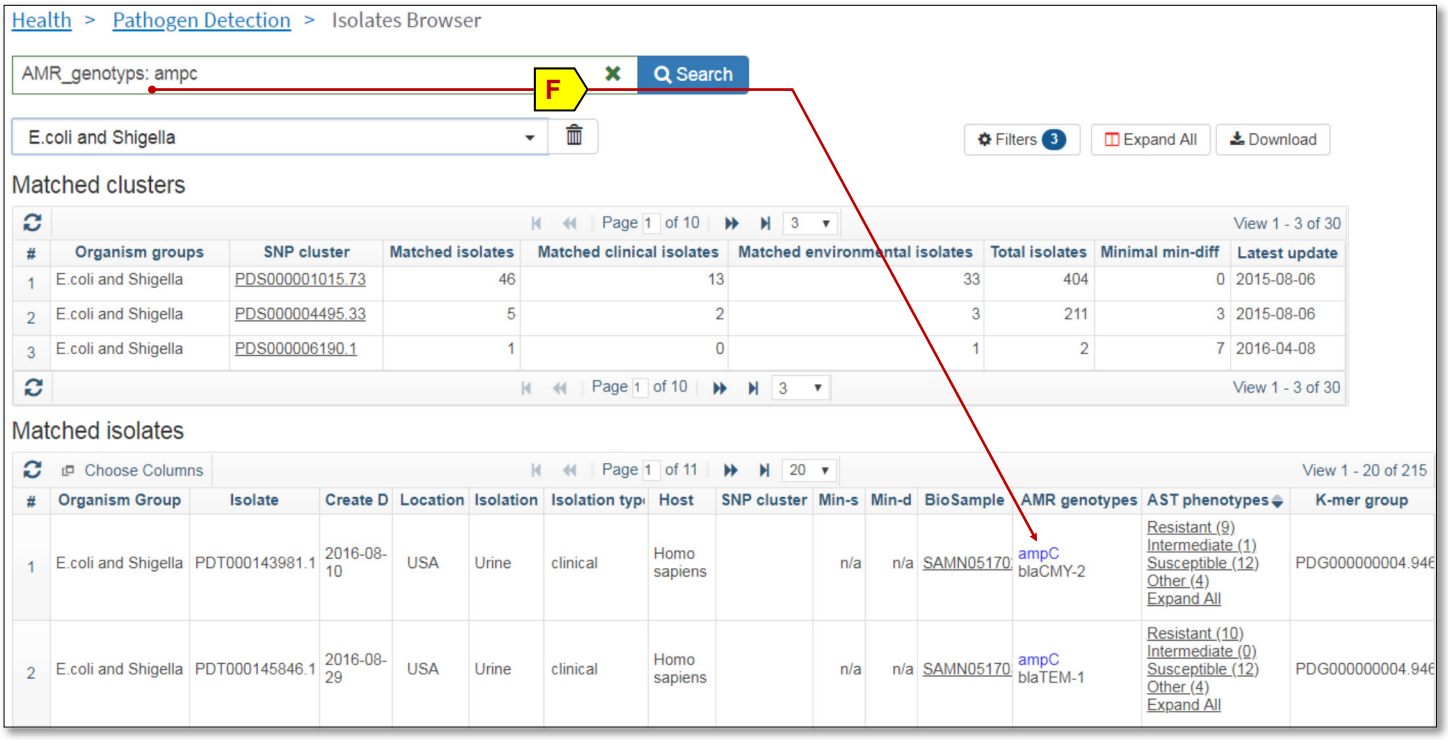
Interrogate Antimicrobial Resistance Through Isolates Browser

We can use NCBI Pathogen Detection's Isolates Browser to find isolates that encode antimicrobial resistance genes or those that encode specific resistance genes, through the usage of its Filter options (**A**), column sorting (**B**), Organism Group selection (**C**), and Search (**D**) functions.

The above is a table with customized column (set through "Choose Columns," **E**), which lists all isolates with AMR genotyping or AST phenotyping data, sorted by AST phenotypes.

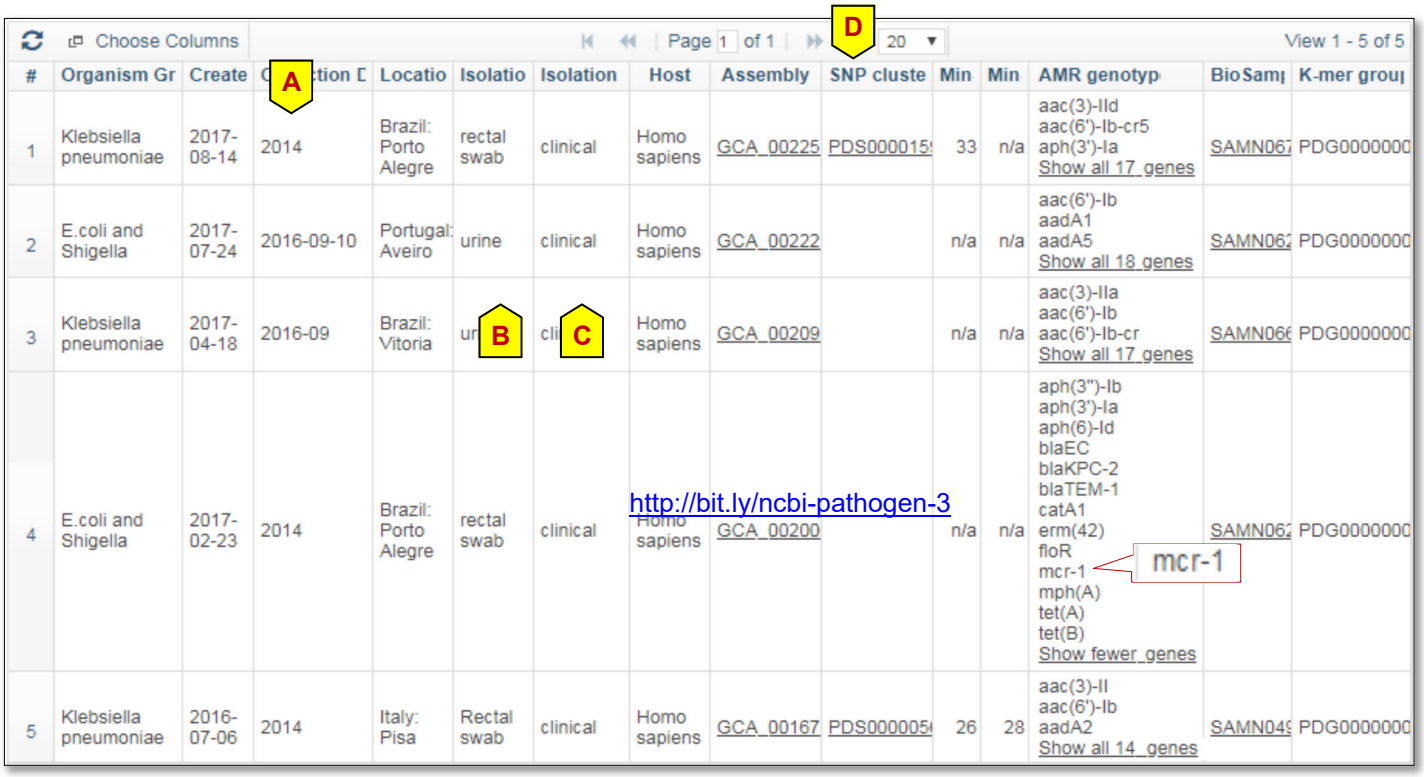


We can combine Filter selection, organism group limit, and the search box to refine the look up. For example, we can look for *E. coli* and *Shigella* isolates, reported in 2015 and 2016, that have both antimicrobial genotypic and phenotypic information, and that also have an *ampC* gene (**F**).



Interrogate Antimicrobial Resistance Through Isolates Browser (cont.)

The colistin resistance gene *mcr* has been found to occur on mobile elements and encodes resistance to one of the last line drugs that can be used. Isolates that also encode carbapenemases are of significant interest to the community. We can use structured search terms to find isolates that encode *mcr* and any allele of the KPC family of beta lactamases: AMR_genotypes:mcr* AND AMR_genotypes:blaKPC*.



#	Organism Gr	Create	Collection D	Locatio	Isolatio	Isolation	Host	Assembly	SNP cluste	Min	Min	AMR genotyp	BioSamp	K-mer grou
1	Klebsiella pneumoniae	2017-08-14	2014	Brazil: Porto Alegre	rectal swab	clinical	Homo sapiens	GCA_00225	PDS000015	33	n/a	aac(3)-IId aac(6')-Ib-cr5 aph(3')-Ia Show all 17 genes	SAMN067	PDG0000000
2	E.coli and Shigella	2017-07-24	2016-09-10	Portugal: Aveiro	urine	clinical	Homo sapiens	GCA_00222		n/a	n/a	aac(6')-Ib aadA1 aadA5 Show all 18 genes	SAMN062	PDG0000000
3	Klebsiella pneumoniae	2017-04-18	2016-09	Brazil: Vitoria	ur	cli	Homo sapiens	GCA_00209		n/a	n/a	aac(3)-IIa aac(6')-Ib aac(6')-Ib-cr Show all 17 genes	SAMN066	PDG0000000
4	E.coli and Shigella	2017-02-23	2014	Brazil: Porto Alegre	rectal swab	clinical	Homo sapiens	GCA_00200		n/a	n/a	aph(3")-Ib aph(3')-Ia aph(6)-Id blaEC blaKPC-2 blaTEM-1 catA1 erm(42) floR mcr-1 mph(A) tet(A) tet(B) Show fewer genes	SAMN062	PDG0000000
5	Klebsiella pneumoniae	2016-07-06	2014	Italy: Pisa	Rectal swab	clinical	Homo sapiens	GCA_00167	PDS000005	26	28	aac(3)-II aac(6')-Ib aadA2 Show all 14 genes	SAMN049	PDG0000000

http://bit.ly/ncbi-pathogen-3

Currently, only five isolates that encode both genes are in the database (**A**). The 'Collected by' and 'Location' columns (**B**, **C**) show that they were collected in different locations in 2014 and 2016. The genomes of all isolates are available in GenBank (**D**). The expanded view for the fourth isolate shows that it encodes *mcr-1* (insert).

Future Work

The NCBI Pathogen Detection team is working on a number of enhancements to the interface, which include:

1) genomic locations of resistance genes,
2) ability to search using specific MIC values, and
3) adding metal and biocide resistance genes to the genes that are screened.

References

1. Bush K, Jacoby GA. 2010. Updated functional classification of beta-lactamases. Antimicrob Agents Chemother 54:969-76.
2. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640-4.
3. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. 2013. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 57:3348-57.
4. Klebsiella Sequence Typing at Pasteur Institute: http://bigsdb.pasteur.fr/klebsiella/klebsiella.html
5. Chopra I, Roberts M. 2001. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. Microbiol Mol Biol Rev 65:232-60.
6. Gibson MK, Forsberg KJ, Dantas G. 2015. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 9:207-16.
7. Finn, RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy S, Heger A, Heterington K, Mistry J, Sonnhammer EL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res. 42: D222-30.
8. Kans J. Entrez Direct: E-utilities on the UNIX Command Line. 2013 Apr 23 [Updated 2018 May 29]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK179288/