NCBI Pathogen Detection

From next generation sequencing reads to clusters of related bacterial pathogens based on SNPs

https://www.ncbi.nlm.nih.gov/pathogens/

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

Scope

Bacterial pathogens that originate from a food source are a serious concern in the US, with estimates by CDC that each year roughly 1 in 6 Americans (or 48 million people) get sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases. Outbreaks are defined where two or more people become ill when ingesting the same food. If the bacterial pathogens that cause the illness are genetically related, it becomes easier to trace the source of the illness when samples from both patients and food source are compared. US Federal agencies involved in public health for foodborne illnesses have agreed to move forward on this by using whole genome sequencing. A pilot project was started in 2013 for *Listeria* with all federal agencies responsible for food safety (CDC, FDA, and USDA) sequencing all *Listeria* isolates, whether in clinical patients or identified during inspection of food products or processing facilities, and depositing the data in the public archives at NCBI. This project demonstrated that whole genome sequencing could be used to increase the number of clusters detected, with a decrease in median cluster size, and with more outbreaks solved [1]. The goal is to extend this to all 90,000 foodborne bacterial pathogens (*Campylobacter*, *Listeria*, *Escherichia coli* and *Shigella* (STECs), and *Salmonella*) that are collected in the US every year and sequenced in real time by the end of 2018.

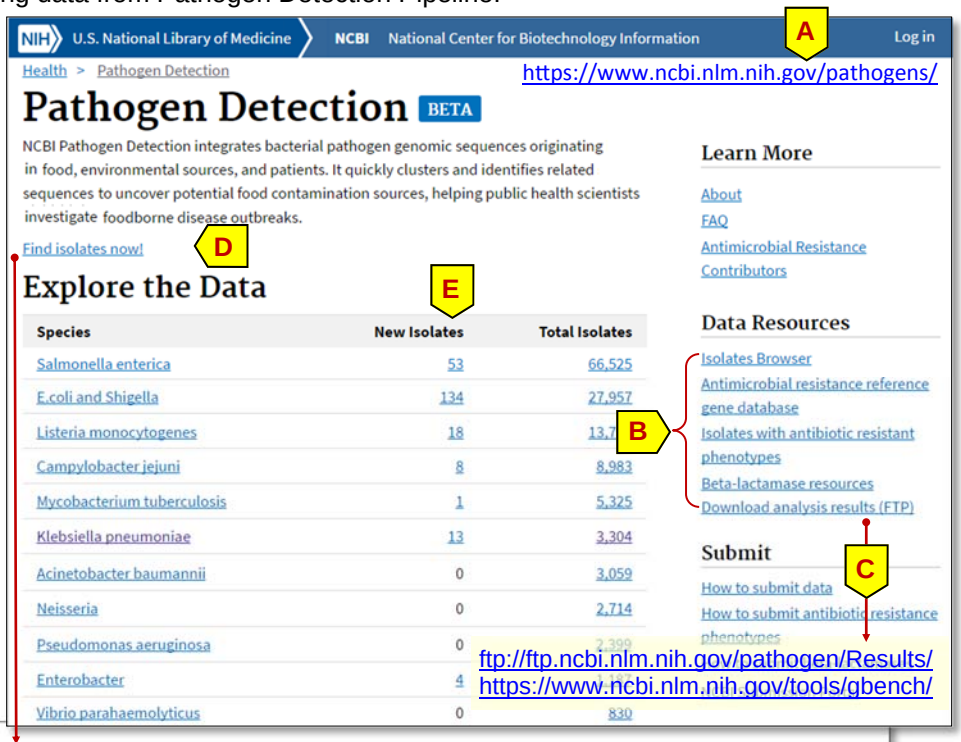
The NCBI Pathogen Detection pipeline takes the incoming sequencing data and assembles, annotates, and clusters the genomes together to facilitate the analysis of genetically related strains to aid outbreak and traceback investigations. Phylogenetic trees are constructed for each SNP cluster using maximum compatibility [2]. Currently clusters are calculated daily for each organism group if new data are submitted and released publicly. The scope of the project has now expanded to include non-foodborne pathogens, especially those that are becoming increasingly antimicrobial resistant (AMR). The workflow uses a reference set of acquired resistance genes/proteins to report for each isolate the AMR genes that are encoded in the genome sequence. This database of antimicrobial resistance pathogens is part of the National Action Plan for Combating Antimicrobial-Resistant Bacteria by providing a resource for the research community on resistant organisms that have genomic sequence data [3].

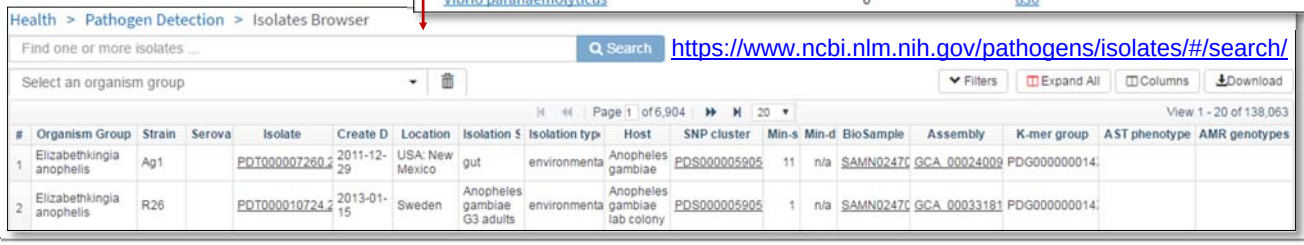
Data Access

Several venues are available for accessing data from Pathogen Detection Pipeline:

1. The Pathogen Detection homepage (**A**) provides links both to the Isolates Browser and FTP results (**B**).
2. The FTP site (**C**) offers access to all of the phylogenetic trees and metadata tables. The tree files can be opened in NCBI Genome Workbench tool and explored locally. See the readme file for more information about the FTP.
3. The beta version of the Isolates Browser (**D**, linked off the Pathogen Detection home page) allows search and browse. In its display, each row is an assembled genome (either from SRA data by the Pathogen Detection Pipeline or genomes deposited in GenBank).

The table (**E**) on the main page specifies specific searches subcategorized by organism group and for the subset of isolates that have been newly deposited since the last cluster calculation.

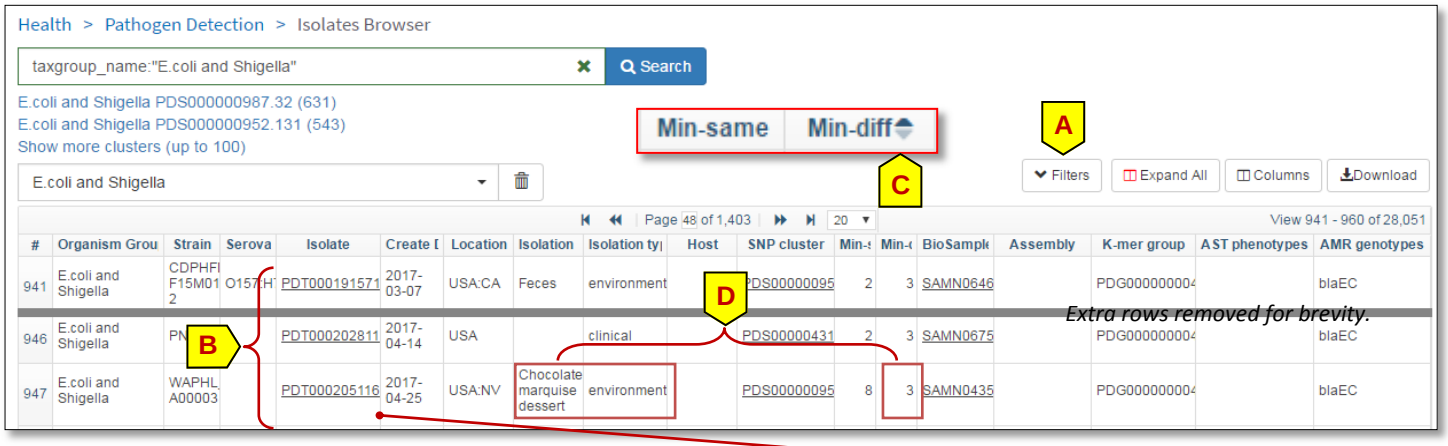




Usage Examples

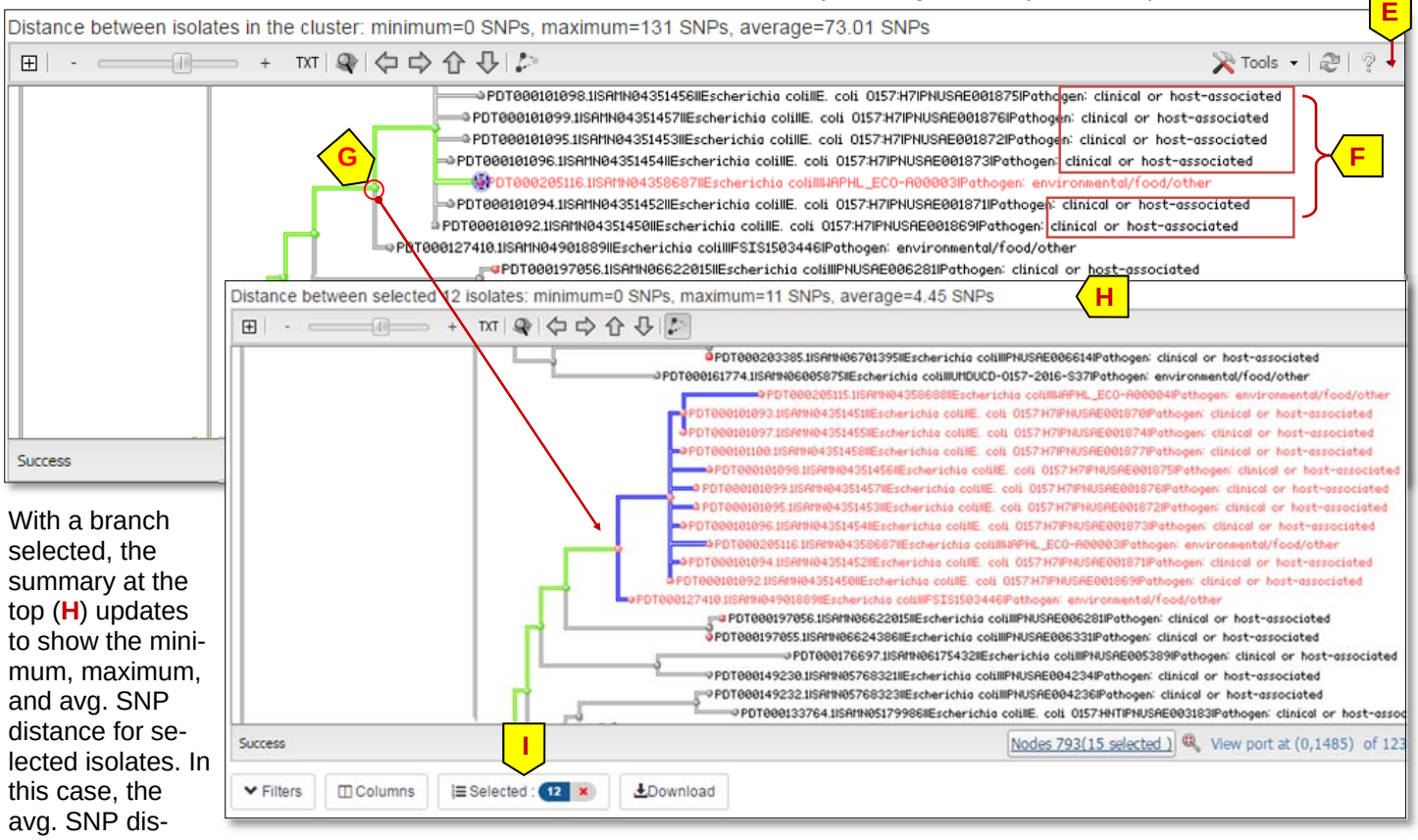
A. Isolation Browser: Basic navigation

From the Pathogen Detection homepage's table (p1, **E**), you can click on the number in the 'New Isolates' column for an organism group to locate new clusters for that group. You can search directly in the Isolate Browser using structured terms. For *E. coli* and *Shigella* group, the structured query terms are: *taxgroup_name:"E.coli and Shigella" AND new:1* (http://bit.ly/ncbi-pathogen-1). The search results will change every time new data is deposited. You can toggle open the Filters section (**A**, more on this in p 3) to narrow down isolates displayed. In the Isolate Browser, the **Min-same** and **Min-diff** columns (insert) report on the SNP differences for each isolate (**B**) based on Isolation type (clinical vs. environmental). Sorting on the **Min-diff** column (**C**) in ascending order brings isolates with the shortest SNP distance between a clinical and environmental samples to the top. This similarity is critically important for public health labs.



B. Use Case One: To identify potential clusters of interest for public health

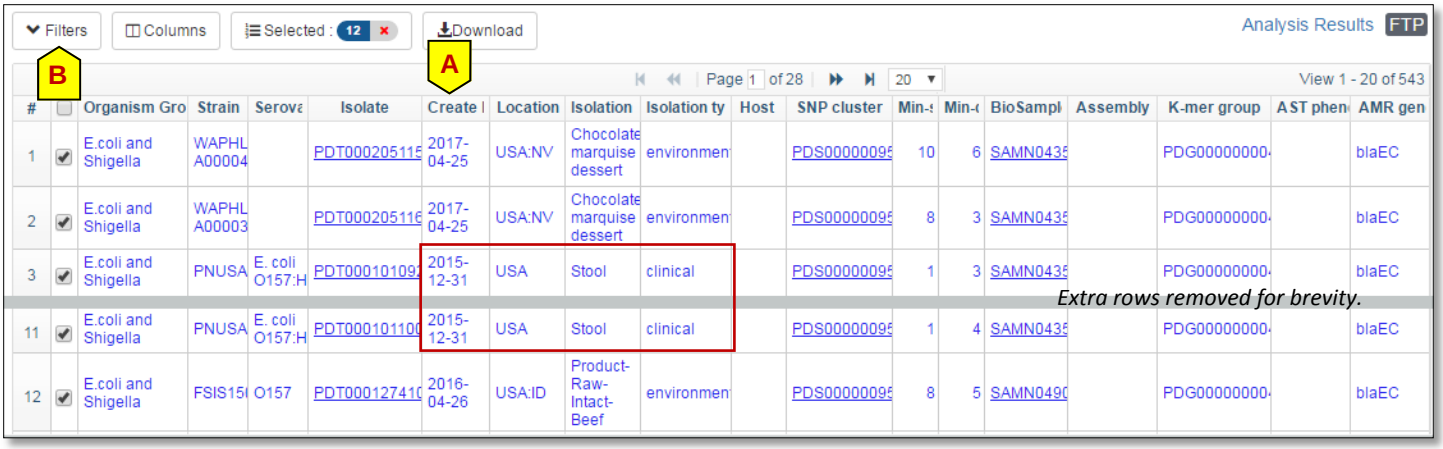
We will use an environmental isolate PDT000205116.1 (row 947) in this exercise, which is 3 SNPs away from a clinical isolate (**D**). Clicking the isolate identifier opens the SNP Tree Viewer (**E**), which zooms to the particular branch of the phylogenetic tree where that isolate is located. Our environmental isolate is clustered with several clinical isolates (**F**). We can examination the details of a branch in the SNP distance tree by clicking a node (such as **G**).



With a branch selected, the summary at the top (**H**) updates to show the minimum, maximum, and avg. SNP distance for selected isolates. In this case, the avg. SNP distance of 4.45 indicates a very closely related set of 12 isolates. We can click "Selected ..." button (**I**) to bring the selected isolates to the top in the datatable below the tree to analyze the metadata associated with these isolates.

B. Use Case One: To identify potential clusters of interest for public health (cont.)

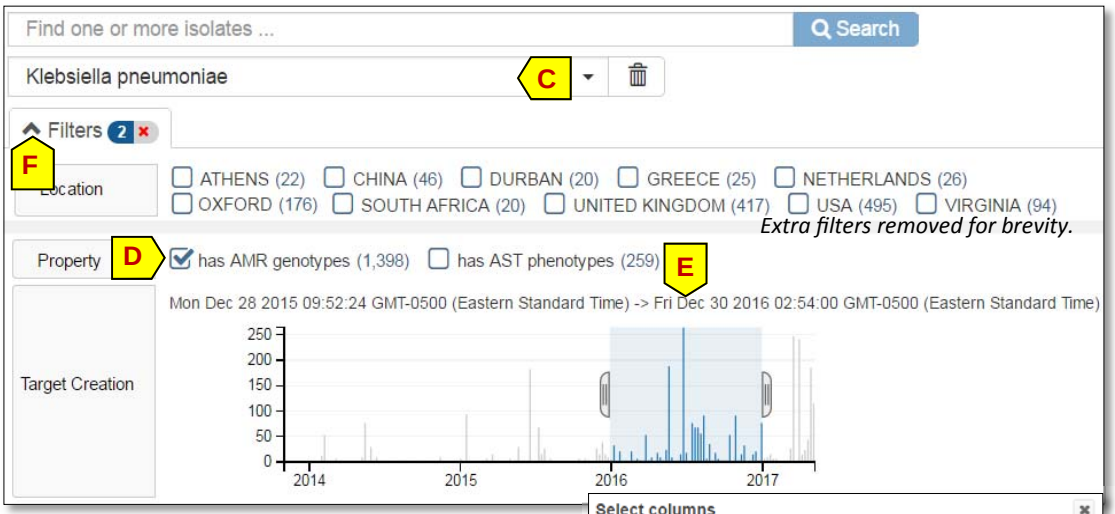
In the sorted table (**A**), the clinical isolates (boxed) are from 2015, while the environmental isolates are from raw beef in 2016 and desserts in 2017. Any public health lab reviewing this data would need to make a determination based on additional metadata if further investigation is warranted.



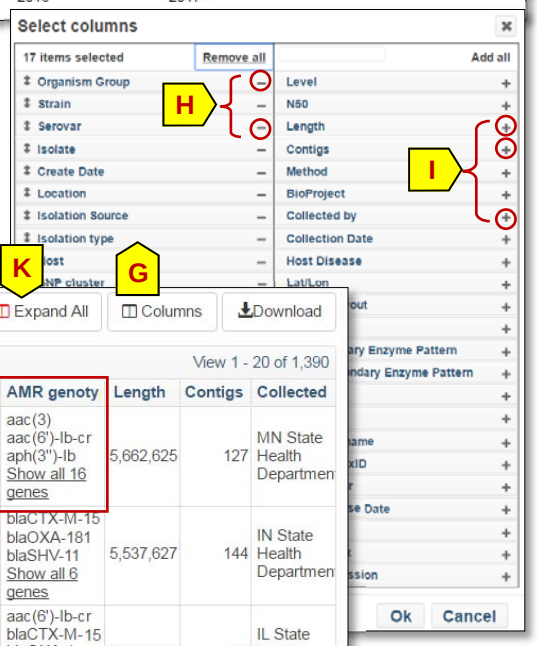
#	Organism Gro	Strain	Serova	Isolate	Create	Location	Isolation	Isolation ty	Host	SNP cluster	Min-s	Min-c	BioSampl	Assembly	K-mer group	AST phen	AMR gen
1	E.coli and Shigella	WAPHL A00004		PDT000205115	2017-04-25	USA:NV	Chocolate marquise dessert	environmen		PDS00000095	10	6	SAMN0435		PDG00000000		blaEC
2	E.coli and Shigella	WAPHL A00003		PDT000205116	2017-04-25	USA:NV	Chocolate marquise dessert	environmen		PDS00000095	8	3	SAMN0435		PDG00000000		blaEC
3	E.coli and Shigella	PNUSA	E. coli O157:H	PDT000101092	2015-12-31	USA	Stool	clinical		PDS00000095	1	3	SAMN0435		PDG00000000		blaEC
11	E.coli and Shigella	PNUSA	E. coli O157:H	PDT000101100	2015-12-31	USA	Stool	clinical		PDS00000095	1	4	SAMN0435		PDG00000000		blaEC
12	E.coli and Shigella	FSIS15(	O157	PDT000127410	2016-04-26	USA:ID	Product-Raw-Intact-Beef	environmen		PDS00000095	8	5	SAMN0490		PDG00000000		blaEC

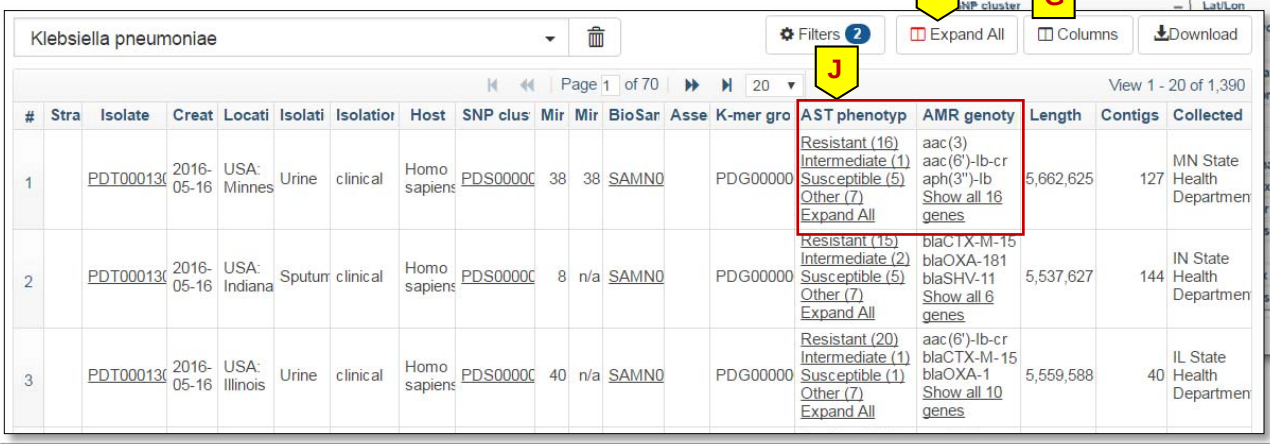
C. Use Case Two: To identify isolates encoding antimicrobial resistance genes

The Filter control (**B**) allows filtering by location, isolation source, sample collecting lab, host, whether the isolate has antimicrobial resistance genes or has antibiotic susceptibility test phenotypic data, a date range control, and by the scientific name of the isolate. In this example we set the organism group control above the Filter to 'Klebsiella pneumoniae,' check 'has AMR genotypes,' and set date range control to 2016 (**C**, **D**, **E**). This setting combination selects all *K. pneumoniae* released in 2016 that encode an antimicrobial resistance gene, such as chromosomal encoded ampC beta lactamases. Click the up arrow (**F**) to close the Filter control.



You can customized the columns shown in the metadata table using the Column control (**G**). The example has 'Organism group' and 'serovar' column removed (**H**) since we are dealing only with *Klebsiella* isolates (through filter), and 'Collected by', 'Length' plus 'Contigs' add (**I**) to provide information on isolate collectors and the assembly quality. Clicking the "AST Phenotypes" column header sorts to the top those isolates tested for antibiotic susceptibility (**J**), many of which are multi-drug resistant. Clicking "Expand All" button (**K**) expands the rows in the table to reveal details (p4, **A**).

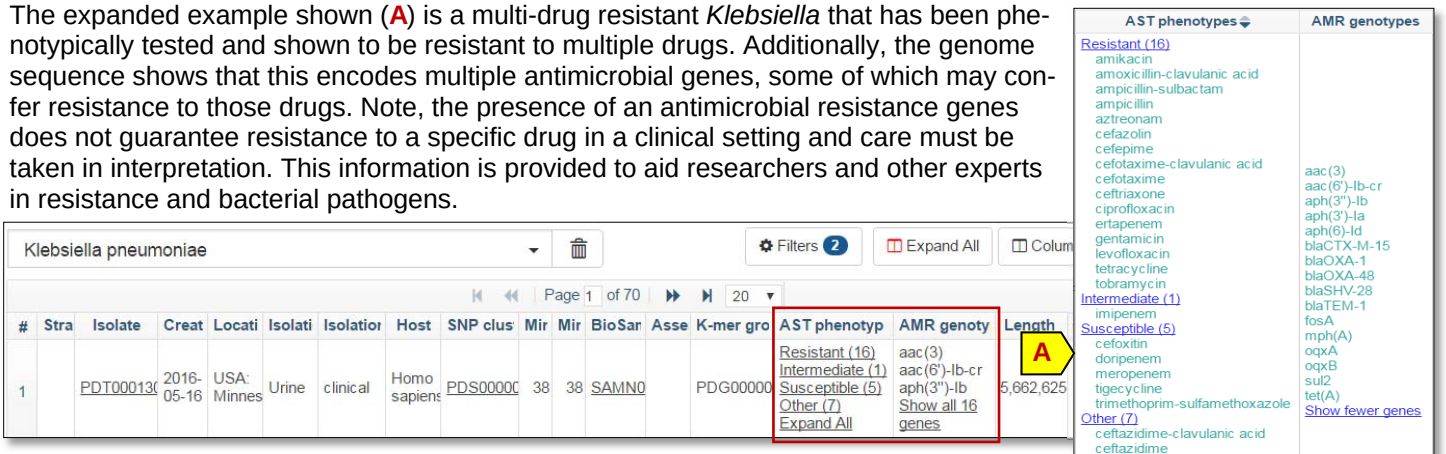




#	Stra	Isolate	Creat	Locati	Isolati	Isolatior	Host	SNP clus	Mir	Mir	BioSar	Asse	K-mer gro	AST phenotyp	AMR genoty	Length	Contigs	Collected
1		PDT00013(	2016-05-16	USA: Minnes	Urine	clinical	Homo sapiens	PDS00000(	38	38	SAMN0		PDG00000	Resistant (16) Intermediate (1) Susceptible (5) Other (7) Expand All	aac(3) aac(6')-Ib-cr aph(3'')-Ib Show all 16 genes	5,662,625	127	MN State Health Departmen
2		PDT00013(	2016-05-16	USA: Indiana	Sputum	clinical	Homo sapiens	PDS00000(	8	n/a	SAMN0		PDG00000	Resistant (15) Intermediate (2) Susceptible (5) Other (7) Expand All	blaCTX-M-15 blaOXA-181 blaSHV-11 Show all 6 genes	5,537,627	144	IN State Health Departmen
3		PDT00013(	2016-05-16	USA: Illinois	Urine	clinical	Homo sapiens	PDS00000(	40	n/a	SAMN0		PDG00000	Resistant (20) Intermediate (1) Susceptible (1) Other (7) Expand All	aac(6')-Ib-cr blaCTX-M-15 blaOXA-1 Show all 10 genes	5,559,588	40	IL State Health Departmen

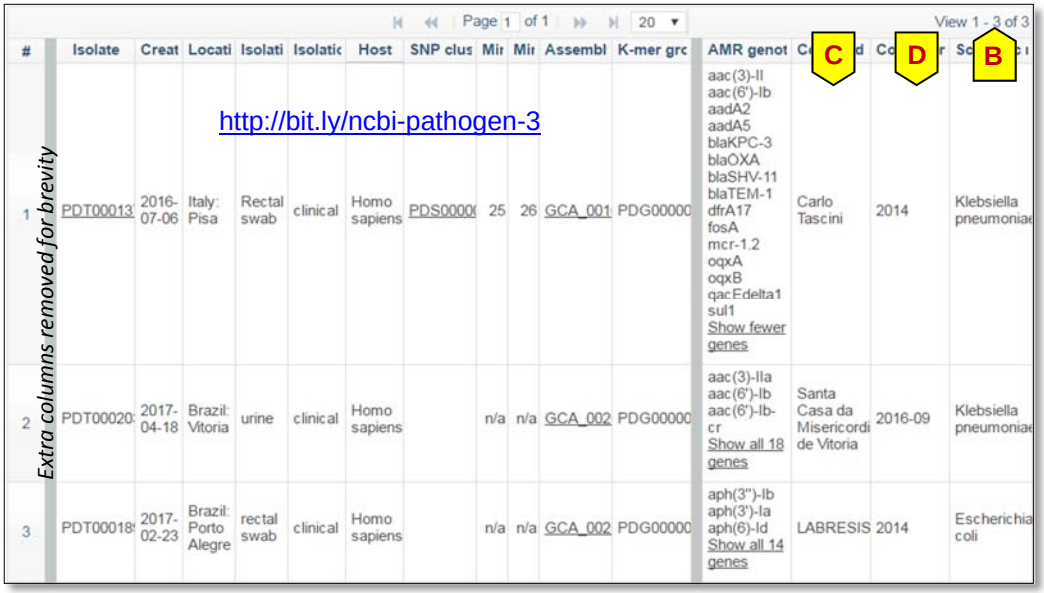
C. Use Case Two: To identify isolates encoding antimicrobial resistance genes (cont.)

The expanded example shown (**A**) is a multi-drug resistant *Klebsiella* that has been phenotypically tested and shown to be resistant to multiple drugs. Additionally, the genome sequence shows that this encodes multiple antimicrobial genes, some of which may confer resistance to those drugs. Note, the presence of an antimicrobial resistance genes does not guarantee resistance to a specific drug in a clinical setting and care must be taken in interpretation. This information is provided to aid researchers and other experts in resistance and bacterial pathogens.



D. Use Case Three: To find isolates that encode specific resistance genes

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*. Currently, only three isolates that encode both genes are in the database (**B**). The 'Collected by' and 'Collection date' columns (**C**, **D**) show that they were collected by different groups in 2014 and 2016. The genomes of all three isolates are available in GenBank. The expanded view for the first isolate shows that it encodes mcr-1.2.



Future Work

The NCBI Pathogen Detection team is working on a number of enhancements to the interface including: 1) genomic locations of resistance genes, 2) ability to search using specific MIC values.

Submissions and Contact

If you wish to submit data to the NCBI Pathogen Detection pipeline, please review the submission page:
https://www.ncbi.nlm.nih.gov/pathogens/submit/

You can address questions on this resource to the NCBI Pathogen Detection team directly:
pd-help@ncbi.nlm.nih.gov

References

1. Allard et al. (2016). The Practical value of Food Pathogen Traceability through BUILDING a Whole-Genome Sequencing Network and database. J Clin Microbiol. 2016 Mar 23. pii: JCM.00081-16.
2. Cherry J. (2017). A practical exact maximum compatibility algorithm for reconstruction of recent evolutionary history. BMC Bioinformatics. 2017 Feb 23;18(1):127. doi: 10.1186/s12859-017-1520-4.
3. Jackson et al. (2016). Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clin Infect Dis. 2016 Apr 18. pii: ciw242.
4. National Action Plan for Combating Antimicrobial-Resistant Bacteria.
https://www.whitehouse.gov/blog/2015/03/27/our-plan-combat-and-prevent-antibiotic-resistant-bacteria
https://obamawhitehouse.archives.gov/sites/default/files/docs/national_action_plan_for_combating_antibotic-resistant_bacteria.pdf