



FEB 26, 2024

OPEN  ACCESS**DOI:**

[dx.doi.org/10.17504/protocols.io.
36wgq3kbklk5/v1](https://dx.doi.org/10.17504/protocols.io.36wgq3kbklk5/v1)

Protocol Citation: Maria Balkey, Ruth Timme, Tina Lusk Pfefer, Candace Hope Bias 2024. Querying for Bacterial Pathogen Genomic Data at NCBI.

protocols.io
<https://dx.doi.org/10.17504/protocols.io.36wgq3kbklk5/v1>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
We use this protocol and it's working

Querying for Bacterial Pathogen Genomic Data at NCBI

Maria Balkey¹, Ruth Timme¹, Tina Lusk Pfefer¹, Candace Hope Bias¹

¹US Food and Drug Administration

GenomeTrakr

Tech. support email: genomeTrakr@fda.hhs.gov



Maria Balkey

ABSTRACT

PURPOSE: This document provides detailed instructions on how to find bacterial pathogen genomic data and associated contextual information at NCBI, specifically at the BioSample, SRA, and Pathogen Detection databases.

SCOPE: This protocol is intended for use by any laboratory submitting WGS data of bacterial pathogens to NCBI for analysis within NCBI Pathogen Detection. This includes US labs connected to GenomeTrakr, NARMS, Vet-LIRN, NAHLN, and other international networks and submitters.

BEFORE START INSTRUCTIONS

This protocol has three sections:

- **Section 1:** NCBI-SRA Run Selector
- **Section 2:** NCBI-BioSample
- **Section 3:** NCBI-SRA
- **Section 4:** NCBI-Pathogen Detection

Created: Jan 16, 2024

Last Modified: Feb 26, 2024

PROTOCOL integer ID: 93618

Keywords: genomic pathogen surveillance, INSDC, NCBI

NCBI SRA (Sequence Read Archive) Run Sector

- 1 WGS submissions of pathogen genomes will become public and searchable almost immediately in [SRA Run Selector](#).

NCBI's **SRA Run Selector** serves as a platform for querying contextual data, or metadata, from *both* BioSample and SRA. This is a good first place to check for a recent submission, obtaining a single table containing a suite of NCBI accessions and metadata for both samples and sequence data. For our application, users could enter a single or comma-separated list of IDs.

Example query with SRA run accessions: SRR24927895,SRR24927896,SRR24937652,SRR24937653

Additionally, a BioProject accession (or multiple BioProject accessions) can be entered to retrieve a complete table of submissions linked to that BioProject(s).

The screenshot shows the NCBI SRA Run Selector interface. At the top, there is a search bar with the query "SRR24927895,SRR24927896,SRR24937652,SRR24937653". Below the search bar, there are sections for "Common Fields" and "Select". The "Common Fields" section displays various metadata fields like Consent (PUBLIC), Assay Type (WGS), and Center Name (FDA). The "Select" section shows a table of results with columns for Run, BioProject, BioSample, ArgPoint, Bases, Bytes, Collection Date, Experiment, FDA Lab ID, Geo Loc, Isolate Name, Isolation Source, and Library Name. The results table contains four rows corresponding to the submitted SRA accessions. A sidebar on the right provides links for Cloud Data Delivery and Computing.

	Run	BioProject	BioSample	ArgPoint	Bases	Bytes	Collection Date	Experiment	FDA Lab ID	Geo Loc	Isolate Name	Isolation Source	Library Name
1	SRR24927895	PRJNA186035	SAMN35749447	484	399.64M	227.72Mb	2023-05-11	SRX02687036	1225725-C001-004	China	China	FNE1225725-C001-004; CFSAN132069	Illumina DNA Prep library SEQ000133080
2	SRR24927896	PRJNA574468	SAMN35749448	380	431.41M	228.38Mb	2023-04-13	SRX02687037	1222287-9004-001	China	China	CFSAN132004	Illumina DNA Prep library SEQ000133082
3	SRR24937652	PRJNA186035	SAMN35749453	480	483.47M	260.40Mb	2023-05-11	SRX02696725	1225725-C002-001	China	China	FNE1225725-C002-001; CFSAN132070	Illumina DNA Prep library SEQ000133081
4	SRR24937653	PRJNA215355	SAMN35749456	483	519.75M	311.11Mb	2020-04-01	SRX02696726	1142650-C002-001	South Korea	South Korea	CFSAN132071	Illumina DNA Prep library SEQ000133180

Isolate and experimental metadata displayed in tabular format.

Download table: Click on the -Metadata- button to download the tab-delimited file.

NCBI Sequencing Read Archive (SRA)

- 2** Submissions should be available in SRA within 1-3 days.

Strain identifiers and/or NCBI accessions can be used as queries. Multiple isolate identifiers can be included in the search box by using the " OR " separator.

Example with SRA run accessions: SRR24927896 OR SRR24927895 OR SRR24937653 OR SRR24937652.

Users can click on the SRA records themselves, or they can send the results to SRA run selector for a tabular output.

Click **-Send results to Run selector-**

The screenshot shows the NCBI SRA search interface. The search bar at the top contains the query: "SRR24927896 OR SRR24927895 OR SRR24937653 OR SRR24937652 OR SRR24927939 OR". Below the search bar, there are filters for Access (Public 51), Source (DNA 51), Type (genome 51), Library Layout (paired 51), Platform (Illumina 51), Strategy (Genome 51), Data in Cloud (GS 51), File Type (fastq 51), and a 'Clear all' link. The main results table shows 20 items per page, with the first item being a whole genome Illumina MiSeq sequence of Listeria innocua. To the right of the results, there are sections for 'Results by taxon', 'Find related data', 'Search details' (listing the search terms used), and 'Recent activity'. A 'Send results to Run selector' button is highlighted in red.

NCBI BioSample

- 3** Submissions should be available in BioSample within 1-3 days.

Strain identifiers and/or NCBI accessions can be used to query NCBI BioSample. Multiple isolate identifiers can be included in the search box using the "OR" separator.

Example with BioSample accessions: SAMN33598462 OR SAMN36638873 OR SAMN06712285

Users can click to view each BioSample query result to review the metadata submitted for that sample (or isolate for our use case). Each BioSample record also includes links to the connected [BioProject](#), and sequenced data generated from that BioSample, for example, raw reads in [SRA](#), and/or genome assemblies in the Nucleotide database, or [GenBank](#).

National Library of Medicine
National Center for Biotechnology Information

BioSample BioSample SAMN20176111 Search Box Create alert Advanced Log in Help

Full + Pathogen: environmental/food/other sample from *Salmonella enterica*

Identifiers BioSample SAMN20176111; SRA: SR9458997; CFSAN: CFSAN096271 NCBI accessions and isolate identifiers

Organism *Salmonella enterica* cellular organisms; Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Salmonellales

Package Pathogen: environmental/food/other; version 1.0

Attributes

- strain RA1_157
- collection date 2015-12-02
- geographic location Ecuador
- isolate name alias CFSAN096271
- collected by UNIETAR
- latitude and longitude missing
- source type animal
- isolation source skin chiller (*Gallus gallus domesticus*)
- project name GenomeTrakr
- sequenced by FDA Center for Food Safety and Applied Nutrition
- PublicAccession CFSAN096271
- attribute_package environmental/food/other
- Genus *Salmonella*
- Species *enterica*
- ProjectAccession PRJNA377900
- ontological term chicken:FOODON_03411457, zone of skin:UBERON_0000014; CURATION
- IFSAAC+ Category veterinary clinical/research|chicken

Link to BioProject PRJNA377900 *Salmonella enterica*
Retrieve all samples from this project

Submission CFSAN: 2021-07-12

Accession: SAMN20176111 ID: 20176111 BioProject SRA Nucleotide Links to NCBI databases

Send to: Related information
BioProject SRA Nucleotide Assembly Taxonomy

Links for accessing or exporting data in different formats

Search details SAMN20176111[All Fields] See more...

Recent activity Turn Off Clear Your browsing activity is empty.

BioSample query results can be exported in different formats by clicking on **-Send To-**.

Send to: Filters: [Manage Filters](#)

Choose Destination

File Clipboard

Collections

Download 340 items.

Format

Summary (text) Full (text) Full XML (text) BioSample ID list Accessions List

NCBI Pathogen Detection

- 4** NCBI Pathogen Detection performs cluster analysis and genotyping screening for antimicrobial resistance, stress response and virulence genes. Visit NCBI-Pathogen Detection [HowTo](#) for extensive documentation on how to access analysis results.

Analysis results expected in <1 day for all organisms except *Salmonella*, which requires about two days for cluster results.



Health > Pathogen Detection > Help > How To

Search page ▲ ▼

How To

These are links to demonstrate how to access the NCBI Pathogen Detection data for some of the analysis tasks you might want to do. For more details see [our documentation](#). The How To's are meant to demonstrate techniques to access the data available and should be generalizable to answer other questions. If you have suggestions for changes or other How To's you think would be useful please set us know at pd_help@ncbi.nlm.nih.gov.

Visual How Tos

- [How To: Find an isolate you submitted \(.pptx\)](#)
- [How To: Find the latest *Salmonella* in the Isolates Browser \(.pptx\)](#)
- [How To: Download a list of human, clinical *E. faecalis* isolates \(.pptx\)](#)
- [How To: Identify isolates in the same SNP cluster that share a set of genes \(.pptx\)](#)
- [How To: Download a list of all carbapenem resistance genes and point mutations from the Reference Gene Catalog \(.pptx\)](#)
- [How To: Download all the reference sequences for a set of proteins \(.pptx\)](#)
- [How To: Find all the known resistance mechanisms to a given drug \(.pptx\)](#)
- [How To: Download the nucleotide sequence of all MCR-1 alleles \(.pptx\)](#)
- [How To: Identify all the contigs that share a set of genes \(.pptx\)](#)
- [How To: Identify isolates that have a pair of genes on the same contig \(.pptx\)](#)
- [How To: Download a MicroBIGG-E table > 100,000 rows \(all CARBAPENEM genes found in Pathogen Detection isolate assemblies\) \(.pptx\)](#)
- [How To: Cite the Pathogen Detection Resource and the Data Contained Within \(.pptx\)](#)
- [How To: Identify carbapenem-resistant isolates without common acquired carbapenem resistance genes \(.pptx\)](#)
- [How To: Find, using the MicroBIGG-E Map, *blaKPC*-containing *Klebsiella pneumoniae* isolates from China and the U.S. in the Isolates Browser \(.pptx\)](#)
- [How To: Download, using the MicroBIGG-E Map, *blaKPC* nucleotide sequences from *Klebsiella pneumoniae* isolates from China and the U.S. \(.pptx\)](#)

Strain identifiers and/or NCBI accessions can be used to query NCBI Pathogen Detection. Delimiters are **not** needed for querying this database.

Example query with BioSample accessions: SAMN33598462 SAMN36638873 SAMN06712285. Direct cut/paste from an excel table also works here.

National Library of Medicine
National Center for Biotechnology Information

[Health](#) > Pathogen Detection

Pathogen Detection BETA

i NCBI Pathogen Detection integrates [bacterial and fungal pathogen genomic sequences](#) from numerous ongoing surveillance and research efforts whose sources include food, environmental sources such as water or production facilities, and patient samples. Foodborne, hospital-acquired, and other clinically infectious pathogens are included.

The system provides two major automated real-time analyses: 1) it quickly clusters related pathogen genome sequences to identify potential transmission chains, helping public health scientists investigate disease outbreaks, and 2) as part of the National Database of Antibiotic Resistant Organisms (NDARO), NCBI screens genomic sequences using AMRFinderPlus to identify the antimicrobial resistance, stress response, and virulence genes found in bacterial genomic sequences, which enables scientists to track the spread of resistance genes and to understand the relationships among antimicrobial resistance, stress response, and virulence.

A The new [MicroBIGG-E Map](#) interface shows the geographical distribution of acquired resistance alleles/genes and point mutations.

Search isolates: SAMN32574131 SAMN32574129 SAMN32768065 SAMN32710378 SAMN33139339 SAMN3257

[Learn More](#)

About

Success Stories

FAC

Browser Factsheet

Antimicrobial Resistance Factsheet

Antimicrobial Resistance (NDARO)

Antimicrobia

Con

[Help](#)

Data Resources

Isolates Browser

Microbial Browser for Identification of Genetic and Genomic Elements (MicroBIGG-E)

For each query, users will see results summarized within two tables:

Table 1. Cluster-level results

Table 2. Isolate-level results

Health > Pathogen Detection > Isolates Browser										
Search: <input type="text"/> <input type="button" value="Search"/>										
Share Save Saved Searches Watched Isolates										
Matched clusters	Cluster ID	Organism group	QNP cluster	Matched isolates	Matched clinical isolates	Matched environmental isolates	Total isolates	Mineral min-off	Mineral min-score	Last update
4	S	Organism group								
1	S	Salmonella enterica	FDG00001200005.90	3	0	3	551	0	2024-01-04	
2	S	Salmonella enterica	FDG000000120_202	13	0	13	315	2	2024-01-17	
3	S	Salmonella enterica	FDG000000120_203	4	0	4	2123	2	2024-01-17	
4	S	Salmonella enterica	FDG000000120_204	1	0	1	13	4	2024-01-17	
5	S	Salmonella enterica	FDG000000120_202_172	1	0	1	397	5	2024-01-17	
6	S	Salmonella enterica	FDG0000004105.96	8	0	8	31	5	2023-09-11	
7	S	Salmonella enterica	FDG00001705488.93	7	0	7	10931	5	2024-01-17	