

Feb 26, 2024

Querying for Bacterial Pathogen Genomic Data at NCBI

 In 1 collection

DOI

<https://dx.doi.org/10.17504/protocols.io.36wgg3kbbklk5/v1>

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

¹US Food and Drug Administration

GenomeTrakr

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



Maria Balkey

US Food and Drug Administration



Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.36wgg3kbbklk5/v1>

Protocol Citation: Maria Balkey, Ruth Timme, Tina Lusk Pfefer, Candace Bias 2024. Querying for Bacterial Pathogen Genomic Data at NCBI. protocols.io <https://dx.doi.org/10.17504/protocols.io.36wgg3kbbklk5/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

Created: January 16, 2024

Last Modified: February 26, 2024

Protocol Integer ID: 93618

Keywords: genomic pathogen surveillance, INSDC, NCBI, bacterial pathogen genomic data at ncbi purpose, bacterial pathogen genomic data, ncbi pathogen detection, wgs data of bacterial pathogen, bacterial pathogen, ncbi for analysis, pathogen, pathogen detection, contextual information at ncbi, ncbi purpose, ncbi, genometrakr

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

This protocol has three sections:

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

NCBI SRA (Sequence Read Archive) Run Sector

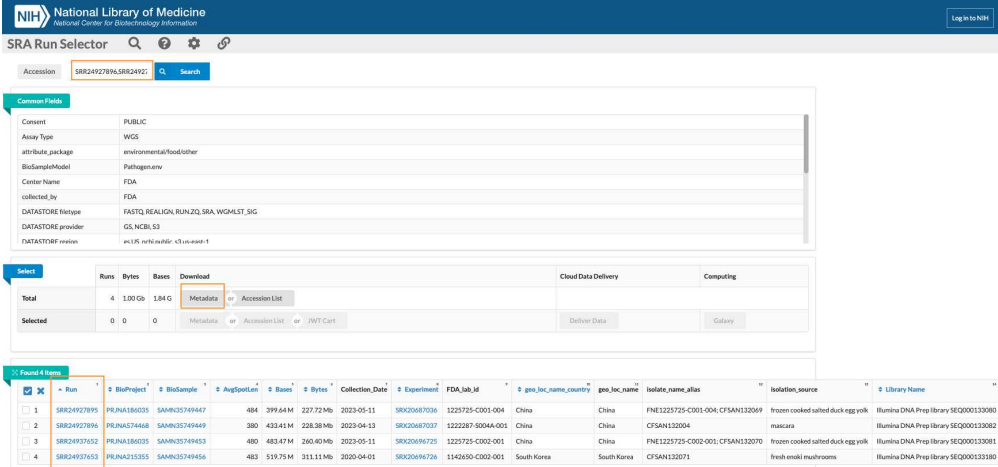
- 1 WGS submissions of pathogen genomes will become public and searchable almost immediately in **SRA Run Selector**.

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's a search bar with the accession SRR24927895 entered. Below the search bar, there's a table of common fields. The 'Download' button is highlighted. Below the download button, there's a table of results with columns: Run, Bytes, Bases, Experiment, SRA Job ID, Accession List, and Library Name. The table shows 4 runs, with the first run highlighted.

Run	Bytes	Bases	Experiment	SRA Job ID	Accession List	Library Name
SRR24927895	484,299,641 M	227,721 M	SRX20687036	1225725-C001-004	China	China
SRR24927896	380,433,411 M	228,381 M	SRX20687037	122287-5044-001	China	China
SRR24937652	480,483,471 M	240,401 M	SRX2068725	1225725-C002-001	China	China
SRR24937653	483,519,751 M	211,111 M	SRX2068726	1142630-C002-001	South Korea	South Korea

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-**

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

SRA

SRR24927896 OR SRR24927895 OR SRR24937653 OR SRR24937652 OR SRR24927939 OR

Create alert Advanced

Log in

Access Public (51)

Source DNA (51)

Type genome (51)

Library Layout paired (51)

Platform Illumina (51)

Strategy Genome (51)

Data in Cloud GS (51) S3 (51)

File Type fastq (51)

[Clear all](#)

[Show additional filters](#)

Summary 20 per page

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

View results as an expanded interactive table using the RunSelector [Send results to Run selector](#)

Search results

Items: 1 to 20 of 51

<< First < Prev Page 1 of 3 Next > Last >>

☐ [Whole genome Illumina MiSeq sequence of Listeria innocua](#)

1. 1 ILLUMINA (Illumina MiSeq) run: 1.1M spots, 519.8M bases, 311.1Mb downloads
Accession: SRX20696726

☐ [Whole genome Illumina MiSeq sequence of Salmonella enterica](#)

2. 1 ILLUMINA (Illumina MiSeq) run: 1M spots, 483.5M bases, 260.4Mb downloads
Accession: SRX20696725

☐ [Whole genome Illumina MiSeq sequence of Salmonella enterica](#)

3. 1 ILLUMINA (Illumina MiSeq) run: 1M spots, 476.3M bases, 273Mb downloads
Accession: SRX20687064

☐ [Whole genome Illumina MiSeq sequence of Bacillus cereus](#)

4. 1 ILLUMINA (Illumina MiSeq) run: 1.1M spots, 433.4M bases, 228.4Mb downloads
Accession: SRX20687037

☐ [Whole genome Illumina MiSeq sequence of Salmonella enterica](#)

5. 1 ILLUMINA (Illumina MiSeq) run: 825,629 spots, 399.6M bases, 227.7Mb downloads
Accession: SRX20687036

☐ [Whole genome Illumina MiSeq sequence of Salmonella enterica](#)

6. 1 ILLUMINA (Illumina MiSeq) run: 1.3M spots, 605.7M bases, 354.4Mb downloads
Accession: SRX20687034

☐ [Whole genome Illumina MiSeq sequence of Listeria monocytogenes](#)

7. 1 ILLUMINA (Illumina MiSeq) run: 970,313 spots, 465.4M bases, 238.6Mb downloads
Accession: SRX20687032

Results by taxon

Top Organisms [Tree](#)

Listeria monocytogenes (45)

Salmonella enterica (4)

Listeria innocua (1)

Bacillus cereus (1)

Find related data

Database: [Select](#)

[Find items](#)

Search details

SRR24927896 [All Fields] OR
SRR24927895 [All Fields] OR
SRR24937653 [All Fields] OR
SRR24937652 [All Fields] OR
SRR24927939 [All Fields] OR

[Search](#) [See more...](#)

Recent activity

[Turn Off](#) [Clear](#)

Q SRR24927896 OR SRR24927895 OR SRR24937653 OR SRR24937652 OR SRA

Q SRR24927896 OR SRR24927895 OR SRR24937653 OR SRR24937652 OR SRA

Q SRR24927896 OR SRR24927895 OR SRR24937653 OR SRR24937652 OR SRA

Q SRR16828363 OR SRR16828362 OR

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



NIH

National Library of Medicine

National Center for Biotechnology Information

Log in

BioSample

BioSample

SAMN33598462 OR SAMN36638873 OR SAMN06712285

Search

Create alert Advanced

Help

Organism
Customize ...

Attribute name
collection date
geographic location
host
strain
Customize ...

Sample type
Pathogen.cl (1)
Pathogen.env (2)

Access
Public (3)

Other
Used by SRA (3)

Clear all

Show additional filters

Summary Sort by Has related data

Search results

Items: 3

☐ Pathogen: clinical or host-associated sample from *Cronobacter sakazakii*

1. Identifiers: BioSample: **SAMN06712285**; SRA: SRS20153232; CFSAN: CFSAN062797
Organism: *Cronobacter sakazakii*
strain: MOD1-2151
Package: Pathogen: clinical or host-associated; version 1.0
Accession: SAMN06712285 ID: 6712285
[BioProject](#) [SRA](#) [Nucleotide](#)

☐ Pathogen.env.1.0

2. Identifiers: BioSample: **SAMN36638873**; SRA: SRS18369163; CFSAN: CFSAN131628
Organism: *Escherichia coli*
strain: 23J03FM28-448
Package: Pathogen: environmental/food/other; version 1.0
Accession: SAMN36638873 ID: 36638873
[SRA](#) [Nucleotide](#)

☐ Pathogen.env.1.0

3. Identifiers: BioSample: **SAMN33598462**; SRA: SRS16956222; CFSAN: CFSAN129134
Organism: *Escherichia coli*
strain: 22F-11FS2-208B
Package: Pathogen: environmental/food/other; version 1.0
Accession: SAMN33598462 ID: 33598462
[BioProject](#) [SRA](#) [Nucleotide](#)

Summary Sort by Has related data

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

Find related data

Database: Select

Find items

Search details

SAMN33598462[All Fields] OR
SAMN36638873[All Fields] OR
SAMN06712285[All Fields]

Search See more...

Recent activity

Turn Off Clear

SAMN33598462 OR SAMN36638873 OR
SAMN06712285 (3) BioSample

FDA (130786) BioSample

SRR24927896 OR SRR24927895 OR
SRR24937653 OR SRR24937652 OR SRA

SRR24927896 OR SRR24927895 OR
SRR24937653 OR SRR24937652 OR SRA

SRR24927896 OR SRR24927895 OR
SRR24937653 OR SRR24937652 OR SRA

See more...

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 Search Box

Full ▾ Links for accessing or exporting data in different formats

Pathogen: environmental/food/other sample from Salmonella enterica

Identifiers BioSample: SAMN20176111; SRA: SRS9458997; CFSAN: CFSAN096271 NCBI accessions and isolate identifiers

Organism [Salmonella enterica](#)
cellular organisms; Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella

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
IFSAC+ 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

Related information

- BioProject
- SRA
- Nucleotide
- Assembly
- Taxonomy

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.

NIH National Library of Medicine
National Center for Biotechnology Information

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.

[Health](#) > Pathogen Detection

BETA

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.

Search isolates:

SAMN32574131 SAMN32574129 SAMN32768065 SAMN32710378 SAMN33139339 SAMN3257

Search

Learn More

[About](#)

Success Stories

FAQ

[Browser Factsheet](#)

Antimicrobial Resistance Factsheet

Antimicrobial Resistance (NDARO)

Antimicrobial Resistance Tools and Resources

Contributors

[Help](#)

How To

Data Resources

[Isolates Browser](#)

[Microbial Browser for Identification of Genetic and Genomic Elements \(MicroBIGG-E\)](#)

1000

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

Table 1. Cluster-level results

Table 2. Isolate-level results

NIH

National Library of Medicine

National Center for Biotechnology Information

Search

Go

Saved Searches

Watched Isolates

Health > Pathogen Detection > Isolates Browser

Help

SARIN2574131 SARIN2574129 SARIN2768065 SARIN32710378 SARIN3119379 SARIN32570945 SARIN32614137 SARIN32661475 SARIN32661475

X

Share

Save

Saved Searches

Watched Isolates

Filters

Matched Clusters

#	Organism groups	SNP cluster	Matched isolates	Matched clinical isolates	Matched environmental isolates	Total isolates	Minimal min-diff	Minimal min-name	Latest update
1	Salmoneila enterica	PGSD0000000000.90	3	0	3	551	0	0	2024-05-04
2	Salmoneila enterica	PGSD0000000000.26.362	13	0	13	335	2	0	2024-05-07
3	Salmoneila enterica	PGSD0000000000.76.764	4	0	4	215	2	1	2014-01-09
4	Salmoneila enterica	PGSD0000000000.54.54	1	0	1	73	2	2	2023-12-22
5	Salmoneila enterica	PGSD0000000000.762.172	1	0	1	397	3	3	2024-05-07
6	Salmoneila enterica	PGSD0000000000.26	8	0	8	32	0	0	2023-09-11
7	Salmoneila enterica	PGSD0000000000.90	7	0	7	1393	5	0	2024-02-27

Matched Isolates

Records per page20

Choose columnsDownloadExpand allCross-browser selection

Displaying 1 - 20 of 28

#	Organism group	Scientific name	Run	Sever	WGS accession	Read origin	ARR category	ISFAC case	Source type	Host	K-mer group	AST phenotype	AMRfxn	PD Ref Gen.	AMRfxnIndex	BiProject	Collected by	Lab/Lin	N50	Length	Cov.	Library ver
1	Salmoneila enterica	Salmoneila enterica	SR123456135		ABRQIQ0000000000.1	Complete (v)	AMR3	naaT3	REF	Show all 4 genes	PGD0000000000.2890	3.11.26	2023-08-08.2	COMPLETED	PRJNA314276	Teves Tech. U.		K29217	5089597	42	PAIRED	
2	Salmoneila enterica	Salmoneila enterica	SR123456132		ABRQIQ0000000000.1	Complete (v)	AMR3	naaT3	REF	Show all 4 genes	PGD0000000000.2890	3.11.26	2023-08-08.2	COMPLETED	PRJNA314276	Teves Tech. U.		462203	4994500	39	PAIRED	
3	Salmoneila enterica	Salmoneila enterica	SR123456129		ABRQIQ0000000000.1	NONE					PGD0000000000.2890	3.11.26	2023-08-08.2	COMPLETED	PRJNA314276	Teves Tech. U.		322988	4743271	31	PAIRED	

Feedback



Protocol references

NCBI-Pathogen Detection Documentation: <https://www.ncbi.nlm.nih.gov/pathogens/docs/HowTo/>

NCBI-Pathogen Detection Documentation: https://www.ncbi.nlm.nih.gov/pathogens/pathogens_help/

SRA Documentation: <https://www.ncbi.nlm.nih.gov/sra/docs/submitupdate/>

BioSample Documentation: <https://www.ncbi.nlm.nih.gov/biosample/docs/>