



Feb 17, 2022

# 🔍 Querying the NCBI database for GenomeTrakr data

Maria Balkey<sup>1</sup>, Julie Haendiges<sup>2</sup>, Ruth Timme<sup>2</sup>, Candace.Bias<sup>2</sup>

<sup>1</sup>US F; <sup>2</sup>US Food and Drug Administration

1



[dx.doi.org/10.17504/protocols.io.bznup5ew](https://dx.doi.org/10.17504/protocols.io.bznup5ew)

**GenomeTrakr**

Tech. support email: [genomeTrakr@fda.hhs.gov](mailto:genomeTrakr@fda.hhs.gov)

 Maria Balkey

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

This protocol describes methods to query GenomeTrakr sequencing records and metadata across multiple NCBI resources: BioSample, BioProject, Sequencing Read Archive, Pathogen Detection and Assembly databases.

DOI

[dx.doi.org/10.17504/protocols.io.bznup5ew](https://dx.doi.org/10.17504/protocols.io.bznup5ew)

Maria Balkey, Julie Haendiges, Ruth Timme, Candace.Bias 2022. Querying the NCBI database for GenomeTrakr data. **protocols.io**  
<https://dx.doi.org/10.17504/protocols.io.bznup5ew>



protocol ,

Nov 01, 2021

Feb 17, 2022

54708

:

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

## NCBI Resources

- 1 Whole Genome Sequencing data submitted to NCBI is processed in multiple databases. If your laboratory or collaborator has submitted WGS data for foodborne pathogens to NCBI, you can locate the data at the NCBI resources: BioProject, BioSample, Sequencing Read Archive, Pathogen Detection and Assembly.

You can access NCBI databases at: <https://www.ncbi.nlm.nih.gov/guide/all/>

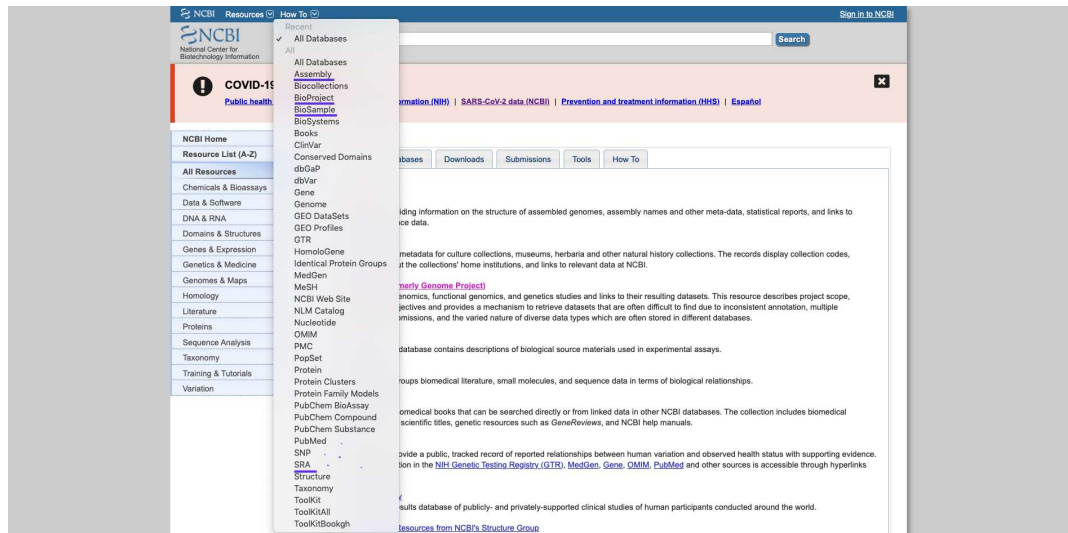


Figure 1: NCBI resources

## BioProject

- 2 BioProject is a collection of biological data related to a surveillance or research effort. Umbrella bioprojects contain several data-level projects. PRJNA593772 (<https://www.ncbi.nlm.nih.gov/bioproject/593772>) comprises a set of umbrella bioprojects, each established for a pathogen being sequenced by the GenomeTrakr network. If you need to find a BioProjects for an specific organism processed by your lab, click at the corresponding organism umbrella Bioproject.

**GenomeTrakr umbrella project**

This project comprises a set of umbrella biojects, each established for a pathogen being sequenced by the GenomeTrakr network. These genomes, along with those collected by other networks, like PulseNet, are analyzed in real-time by public health officials who are detecting and responding to foodborne outbreaks and contamination events. Automated clustering and genotyping screens for these data can be found at NCBI Pathogen Detection (linked below).

Accession: PRJNA593772 ID: 593772

**GenomeTrakr umbrella project**

Project Type	Number of Projects
Genome sequencing Highest level of assembly : SRA or Trace	3

BioProject accession	Assembly level	Organism	Title
PRJNA741099	SRA or Trace	Listeria monocytogenes	Listeria monocytogenes Genome sequencing and assembly (Institute of Environmental...)
PRJNA720150	SRA or Trace	Salmonella enterica	Salmonella enterica genome sequencing and assembly (Institute of Environmental...)
PRJNA741100	SRA or Trace	Shigella	New Zealand Shigella Genome sequencing and assembly (Institute of Environmental...)

**Umbrella project**

BioProject accession	Name	Title
PRJNA692474	Bacillus cereus	GenomeTrakr umbrella project for Bacillus cereus (FDA)
PRJNA692474	Campylobacter sp.	GenomeTrakr umbrella project for Campylobacter jejuni and Campylobacter coli (FDA/CFSAN)
PRJNA200488	Clostridium botulinum	GenomeTrakr umbrella project for Clostridium botulinum (US Food and Drug Administration)
PRJNA357477	Cyclospora cayentanensis	CycloTrakr (Cyclospora cayentanensis GenomeTrakr) (CFSAN)
PRJNA200919	Escherichia coli	GenomeTrakr umbrella project for Escherichia coli and Shigella sp. (FDA/CFSAN)
PRJNA208402	GenomeTrakr umbrella project for Chronobacter sp.	GenomeTrakr umbrella project for Chronobacter sp. (FDA/CFSAN)
PRJNA706995	Klebsiella oxytoca	Umbrella for Klebsiella oxytoca (FDA Center for Food Safety...)
PRJNA706995	Klebsiella pneumoniae	Umbrella for Klebsiella pneumoniae (FDA Center for Food Safety...)
PRJNA514048	Listeria monocytogenes	GenomeTrakr umbrella project for Listeria monocytogenes (FDA Center for Food Safety...)
PRJNA706994	Multispecies	GenomeTrakr umbrella for research organisms (FDA Center for Food Safety...)
PRJNA153844	Salmonella enterica	GenomeTrakr umbrella project for Salmonella enterica (Center for Food Safety and...)
PRJNA484553	Staphylococcus	GenomeTrakr Umbrella Project for Staphylococcus sp. (FDA Center for Food Safety...)
PRJNA245885	Vibrio sp.	GenomeTrakr umbrella project for Vibrio parahaemolyticus (FDA/CFSAN)
PRJNA745491	Vibrio vulnificus	GenomeTrakr Umbrella Project for Vibrio vulnificus (US Food and Drug Administration)
PRJNA745494	Yersinia enterocolitica	GenomeTrakr umbrella project for Yersinia enterocolitica (FDA Center for Food Safety...)

**Figure 2:** GenomeTrakr Umbrella BioProjects

2.1 If you are interested in searching for specific type of data, you can click on **Browse by Project attributes** and narrow your search by using filters such as: Project, Data Type, Scope, Property , Kingdom, Group, Subgroup.

## BioSample

- 3 BioSample (<https://www.ncbi.nlm.nih.gov/biosample/>) is the database for the isolate or sample metadata. Users access biosample records at using the **search box** and typing laboratory identifiers ( strain, isolate name alias, FDA\_Lab\_ID, BioProjects) or specific attributes separated by " OR " e.g. "CFSAN0001 OR CFSAN0002 OR CFSAN0003" "Salmonella enterica".

**BioSample**

Search: SAMN15406445 OR SAMN15406444 OR SAMN15406446 OR SAMN15406447 OR SAMN15406448

**COVID-19 Information**

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Organism: Summary | 20 per page | Sort by Has related data | Send to: Filters: Manage Filters

**Search results**

Items: 6

1. Pathogen: environmental/food/other sample from Listeria monocytogenes  
Identifiers: BioSample: SAMN15406451; Sample name: TB0696; SRA: SRS6926403  
Organism: Listeria monocytogenes  
Package: Pathogen: environmental/food/other; version 1.0  
Accession: SAMN15406451 ID: 15406451  
BioProject SRA Nucleotide

2. Pathogen: environmental/food/other sample from Listeria monocytogenes  
Identifiers: BioSample: SAMN15406449; Sample name: TB0694; SRA: SRS6926401  
Organism: Listeria monocytogenes  
Package: Pathogen: environmental/food/other; version 1.0  
Accession: SAMN15406449 ID: 15406449  
BioProject SRA Nucleotide

3. Pathogen: environmental/food/other sample from Listeria monocytogenes  
Identifiers: BioSample: SAMN15406447; Sample name: TB0692; SRA: SRS6926399  
Organism: Listeria monocytogenes  
Package: Pathogen: environmental/food/other; version 1.0  
Accession: SAMN15406447 ID: 15406447  
BioProject SRA Nucleotide

**Search details**

SAMN15406445[All Fields] OR  
SAMN15406444[All Fields] OR  
SAMN15406446[All Fields] OR  
SAMN15406447[All Fields] OR  
SAMN15406448[All Fields] OR

**Recent activity**

Turn Off Clear

Q SAMN15406445 OR SAMN15406444 OR SAMN15406446 OR SAMN15406447 BioSample

Q SAMN1542141 OR SAMN15406445 OR SAMN15406447 OR SAMN15406448 BioSample

**Figure 3:** Searching records in BioSample

The data from biosample can be downloaded by clicking the send to icon and

### 3.1 choosing the destination of the file (summary, full text, full XML, biosample ID list or Accessions list).

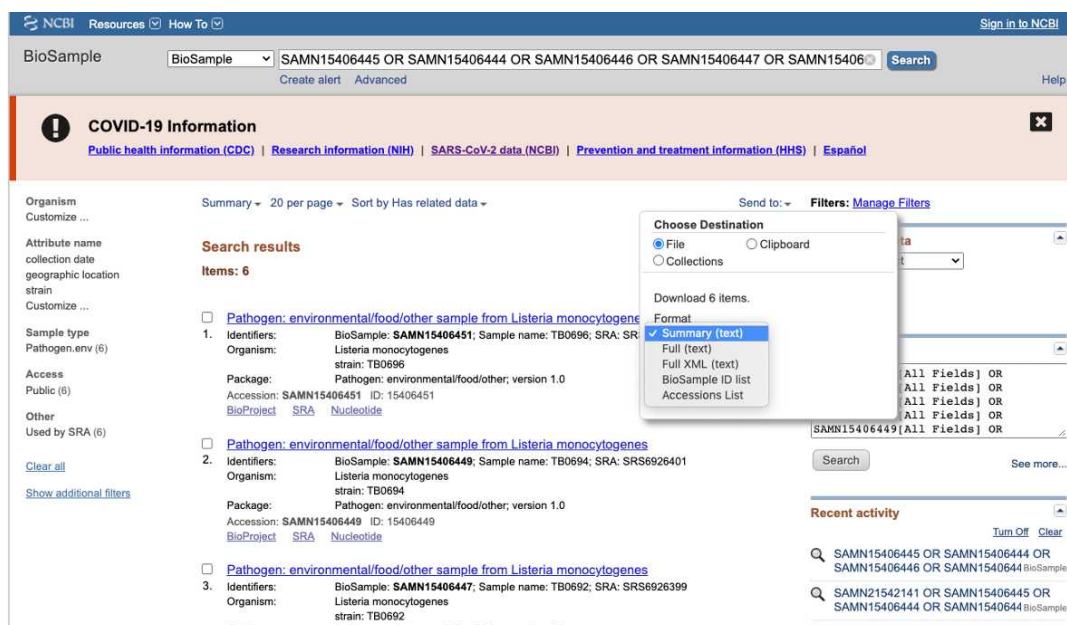


Figure 4: Downloading records from NCBI BioSample

### 3.2 SRA and nucleotide links are available if sequencing and assembly data were submitted to NCBI.

### Pathogen: environmental/food/other sample from Salmonella enterica

Identifiers	BioSample: SAMN08640061; SRA: SRS3667400; CFSAN: CFSAN076999																													
Organism	<a href="#">Salmonella enterica</a> cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella																													
Package	<a href="#">Pathogen: environmental/food/other; version 1.0</a>																													
Attributes	<table><tr><td><b>isolation source</b></td><td>non tidal fresh water creek</td></tr><tr><td><b>geographic location</b></td><td><a href="#">USA:MD</a></td></tr><tr><td><b>latitude and longitude</b></td><td>missing</td></tr><tr><td><b>strain</b></td><td>CNSV-T15-MD05-10-RV-C</td></tr><tr><td><b>isolate name alias</b></td><td>CFSAN076999</td></tr><tr><td><b>collection date</b></td><td>2017</td></tr><tr><td><b>collected by</b></td><td>USDA-ARS</td></tr><tr><td><b>attribute_package</b></td><td>environmental/food/other</td></tr><tr><td><b>IFSAC+ Category</b></td><td>environmental-water</td></tr><tr><td><b>source type</b></td><td>Environmental</td></tr><tr><td><b>PublicAccession</b></td><td>CFSAN076999</td></tr><tr><td><b>ProjectAccession</b></td><td>PRJNA271470</td></tr><tr><td><b>Species</b></td><td>enterica</td></tr><tr><td><b>Genus</b></td><td>Salmonella</td></tr></table>		<b>isolation source</b>	non tidal fresh water creek	<b>geographic location</b>	<a href="#">USA:MD</a>	<b>latitude and longitude</b>	missing	<b>strain</b>	CNSV-T15-MD05-10-RV-C	<b>isolate name alias</b>	CFSAN076999	<b>collection date</b>	2017	<b>collected by</b>	USDA-ARS	<b>attribute_package</b>	environmental/food/other	<b>IFSAC+ Category</b>	environmental-water	<b>source type</b>	Environmental	<b>PublicAccession</b>	CFSAN076999	<b>ProjectAccession</b>	PRJNA271470	<b>Species</b>	enterica	<b>Genus</b>	Salmonella
<b>isolation source</b>	non tidal fresh water creek																													
<b>geographic location</b>	<a href="#">USA:MD</a>																													
<b>latitude and longitude</b>	missing																													
<b>strain</b>	CNSV-T15-MD05-10-RV-C																													
<b>isolate name alias</b>	CFSAN076999																													
<b>collection date</b>	2017																													
<b>collected by</b>	USDA-ARS																													
<b>attribute_package</b>	environmental/food/other																													
<b>IFSAC+ Category</b>	environmental-water																													
<b>source type</b>	Environmental																													
<b>PublicAccession</b>	CFSAN076999																													
<b>ProjectAccession</b>	PRJNA271470																													
<b>Species</b>	enterica																													
<b>Genus</b>	Salmonella																													
Links																														
BioProject	<a href="#">PRJNA271470</a> Salmonella enterica Retrieve <a href="#">all samples</a> from this project																													
Submission	<a href="#">CFSAN</a> ; 2018-03-06																													
<a href="#">Accession: SAMN08640061</a> <a href="#">ID: 8640061</a>																														
<a href="#">BioProject</a> <a href="#">SRA</a> <a href="#">Nucleotide</a>																														

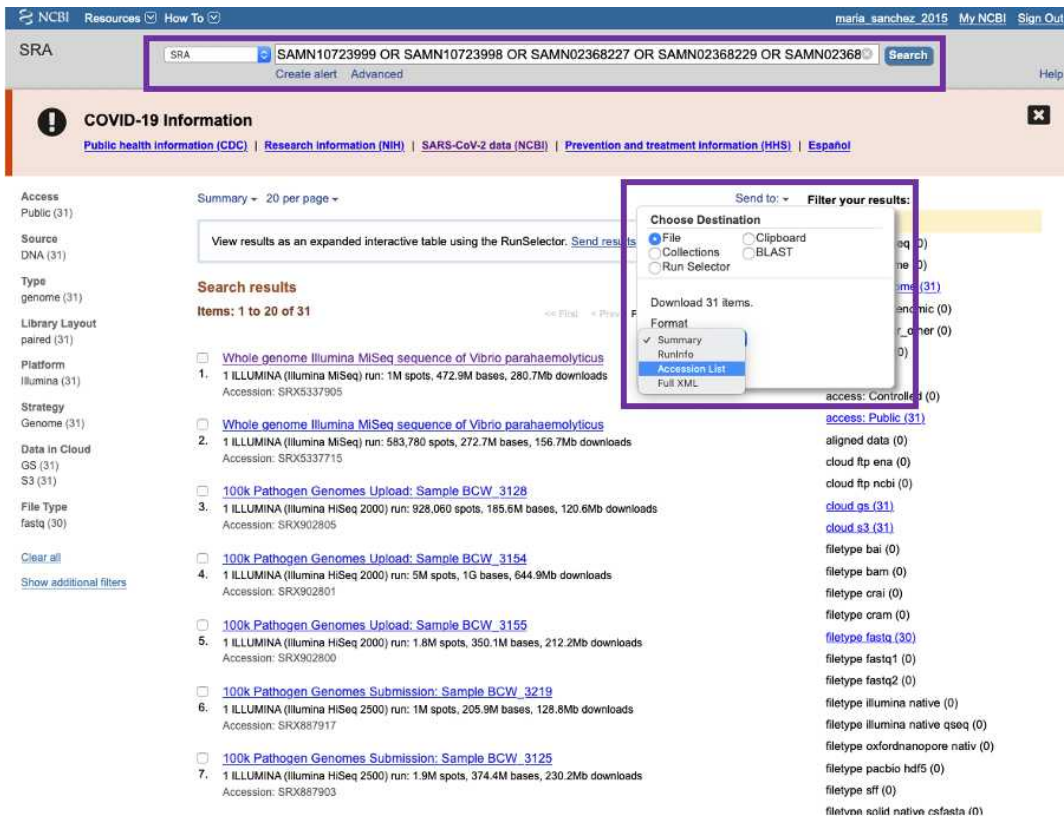
**Figure 5:** SRA and Nucleotide (assembly) access from BioSample.

## Sequencing Read Archive

### 4

The Sequencing Read Archive (SRA) <https://www.ncbi.nlm.nih.gov/sra/> is the primary repository of raw whole genome sequencing data. You can access records at SRA by typing laboratory identifiers ( strain, isolate name alias, FDA\_Lab\_ID, BioProjects) or specific attributes in the search box. Identifiers might need to be separated by " OR " e.g. -CFSAN0001 OR CFSAN0002 OR CFSAN0003.

You can export SRA accessions by clicking at the **Send to** bottom and choosing **file** and the **accession list** format.



**Figure 6:** Downloading SRA accessions from NCBI SRA.

#### 4.1 You can download sequencing data files from NCBI using SRA Toolkit, Run Browser and the cloud.

- [Download sequence data files using SRA Toolkit](#)
- [Download sequence data from the Run Browser](#)
- [Download SRA sequence data from the Cloud](#)

#### Run Selector

5 You can download a combination of sample/isolate and sequencing metadata in a tab-delimited file using Run Selector (<https://www.ncbi.nlm.nih.gov/Traces/study/>).

- Enter the accessions in the search box at the SRA browser, click **Search**, the output will include all the found records.
- Click **Send to** on the top of the SRA page, check the **Run Selector** radio button, and click the button **Go**.
- If necessary, refine your results by using various filters provided by the **Run Selector's** interface.
- Click the **Metadata** button. This will generate a tabular file with metadata available for each Run.



The screenshot shows the NCBI SRA search results page. The search criteria are SAMN10723999 OR SAMN10723998 OR SAMN02368227 OR SAMN02368229 OR SAMN02368227. The results list includes several entries for *Vibrio parahaemolyticus* and *100k Pathogen Genomes Upload: Sample BCW\_3128*. A 'Send results' dialog box is open, showing options to send data to a file, clipboard, collections, BLAST, or Run Selector. The 'Run Selector' option is selected, and the 'Go' button is highlighted.

**Figure 7:** Sending data to SRA Run Selector.

## Pathogen Detection

- Visit the Pathogen Detection (<https://www.ncbi.nlm.nih.gov/pathogens/>) to access real-time analyses of isolates obtained from ongoing pathogen surveillance activities.

More details on how to navigate the Pathogen Detection can be found at:  
[https://www.ncbi.nlm.nih.gov/pathogens/pathogens\\_help/](https://www.ncbi.nlm.nih.gov/pathogens/pathogens_help/)

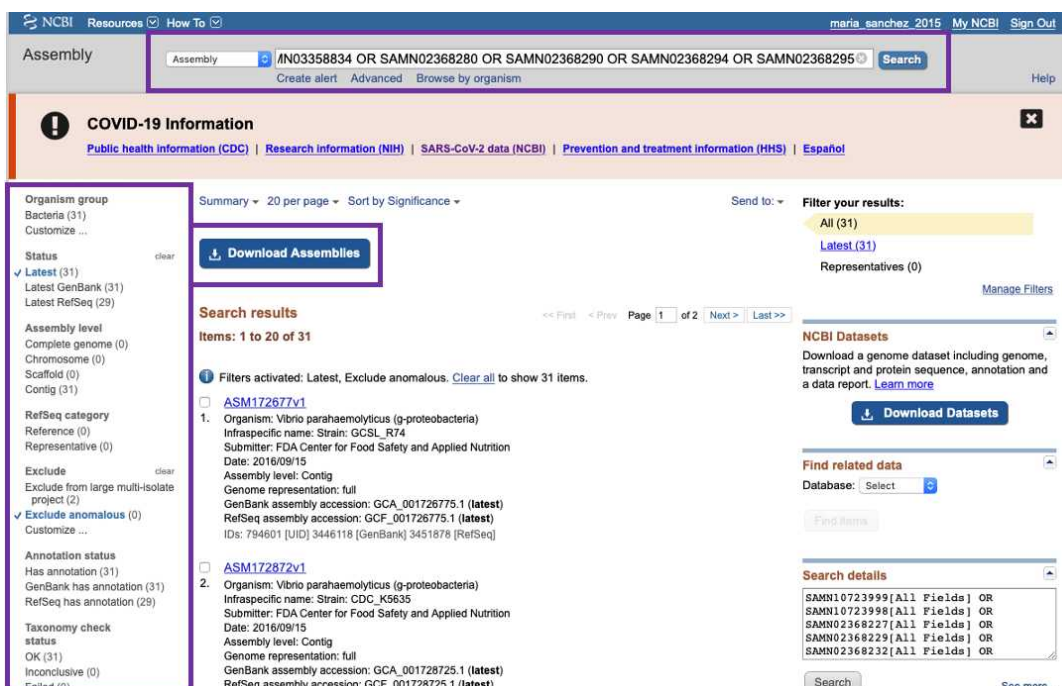
Resources from NCBI Pathogen Detection to address specific research questions.

- [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\(.pptx\)](#)

## Assembly

If you need to download multiple assembled genomes, access the NCBI Assembly resource

- 7 (<https://www.ncbi.nlm.nih.gov/assembly/>). Enter the identifiers in the search box and click in the "Download Assemblies" button. In the left side of this interface, you can refine your search by applying multiple filters. For more details on programatically download genomes from NCBI visit <https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/>.



**Figure 8:** Downloading assemblies from NCBI Assembly

## NCBI Insights

- 8 If you want to keep up with NCBI news, sign up for NCBI insights updates. The *NCBI Insights* Blog offers guidance on the latest NCBI resources.