

These slides were part of an [AWS webinar in 2021](#)

Here are some webinar links to help with setting up AWS Athena

[NCBI Minute: SRA in AWS Athena for SARS-CoV-2 Research and More](#)

[NCBI Minute: Accelerate Genomics Discovery with SRA in the Cloud](#)

Leveraging NCBI's SRA data & tools in AWS ODP and Athena for SARS-CoV-2 search and analyses

Originally Presented: Ryan Connor, Ph.D.

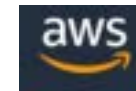
Feb. 25, 2021



NCBI

Extended team comprising of:

- NCBI:NLM STRIDES Program Management
- Sequence Read Archive
- NCBI Virus
- Customer Engagement



Extended team at AWS Life Sciences, Open Data

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine and the NIH STRIDES Initiative.

Do you wish..

1. it was easier to find SRA data based on organismal content?
2. searching SRA based on BioProject and BioSample data was easier?
3. you could get a sense of what could be assembled out of an SRA dataset?

AWS's Open Data Sponsorship Program (ODP)

1. What is this program and why you should care?
2. Open access...what does that mean in terms of cost?

NCBI's data on AWS ODP

Available now:

- COVID-19 Genome Sequence Dataset: <https://registry.opendata.aws/ncbi-covid-19/>
 - SARS-CoV-2 SRA data.
 - SARS-CoV-2 SRA metadata
 - SARS-CoV-2 detection tool
 - SRA Aligned Read Format
- NCBI's Blast Databases: <https://registry.opendata.aws/ncbi-blast-databases/>
- Public SRA data in *original format* from select high value and newly-released studies: <https://registry.opendata.aws/ncbi-sra/>

Coming soon!

- All of the public and controlled-access SRA normalized format data is being migrated to AWS ODP
- Estimated completion in April 2021.

NCBI's open data sets on AWS ODP

URL- <https://registry.opendata.aws/ncbi-blast-databases/>

The screenshot shows the AWS Registry of Open Data page for the Basic Local Alignment Sequences Tool (BLAST) Databases. The page includes a description of the tool, its update frequency, license, and documentation. It also lists resources on AWS, including the Amazon S3 bucket and the AWS CLI command to access the data.

URL- <https://registry.opendata.aws/ncbi-sra/>

The screenshot shows the AWS Registry of Open Data page for the NIH NCBI Sequence Read Archive (SRA) on AWS. The page includes a description of the SRA, its update frequency, license, and documentation. It also lists resources on AWS, including the Amazon S3 bucket and the AWS CLI command to access the data.

URL- <https://registry.opendata.aws/ncbi-covid-19/>

The screenshot shows the AWS Registry of Open Data page for the COVID-19 Genome Sequence Dataset. The page includes a description of the dataset, its update frequency, license, and documentation. It also lists resources on AWS, including the Amazon S3 bucket and the AWS CLI command to access the data.



U.S. National Library of Medicine
National Center for Biotechnology Information

Helpful tips

- **AWS ODP v. commercial buckets**

- Egress from ODP is *free* from anywhere (i.e. cloud and local machines)
 - 2 kinds of SARS-CoV-2 ODP buckets, data and metadata, metadata contains parquet format files for use with Athena

- **Athena**

- Support SQL-like querying of data
- Queries cost money, typically < \$0.10 per query for the datasets being discussed

- **SRA v. GenBank, BioProject, BioSample, experiments, etc.**

- The source data, though not the analytical products being described today, are still available directly from NCBI

- **Metadata** – what do we mean by it here?

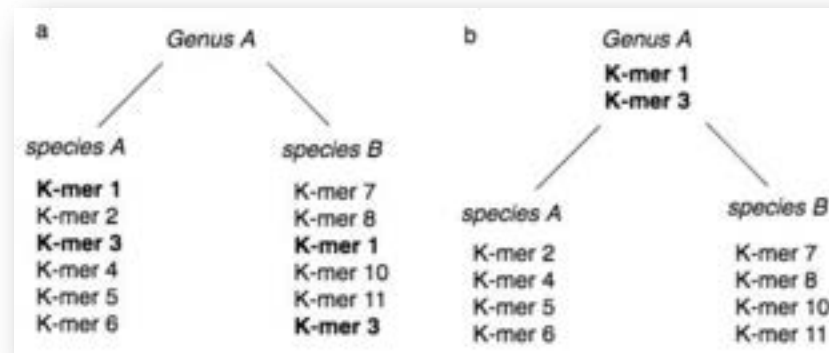
- Anything not the sequence data

What data is in scope for this talk?

- Public, not controlled-access SRA data that contains SARS-CoV-2 sequence
 - Illumina platform only
 - Stay tuned for long-read data
- How do we determine which runs contain SARS-CoV-2 data?
 - SRA Taxonomy Analysis Tool (STAT)

How does the STAT tool work?

- K-mer based taxonomic classification, fast & scalable



- Preprint:

- <https://www.biorxiv.org/content/10.1101/2021.02.16.431451v1.full.pdf>

Other Flavors of STAT

- Coronaviridae Detection Tool
 - Docker - <https://hub.docker.com/r/ncbi/sars-cov-2-detection-tool>
 - Documentation - <https://www.ncbi.nlm.nih.gov/sra/docs/sra-detection-tool/>
- Human Scrubber
 - Docker - <https://hub.docker.com/r/ncbi/sra-human-scrubber>
 - Github - <https://github.com/ncbi/sra-human-scrubber>

Learning objectives:

1. Getting set-up, real fast.
2. How to search against user submitted metadata.
3. How to search against NCBI calculated metadata.

Setting up..Step 1

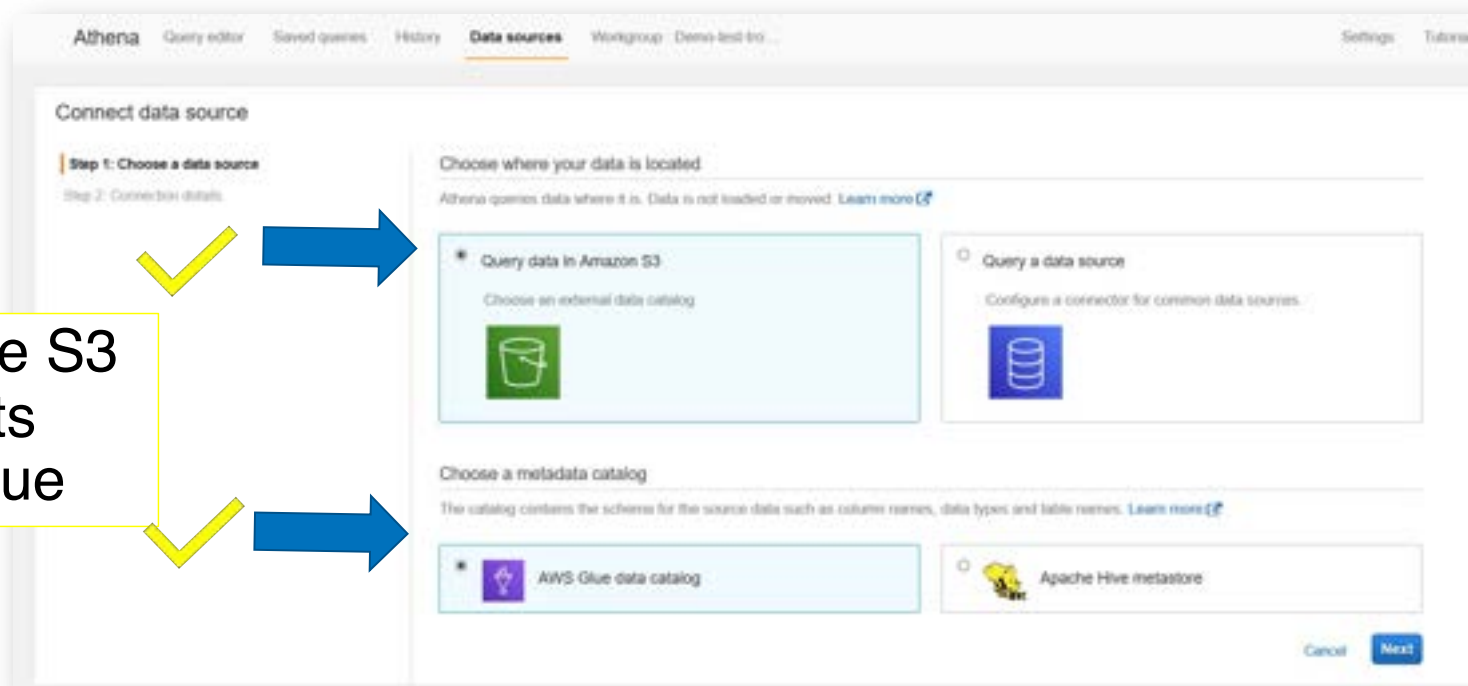
⌄ Connect to Data Source



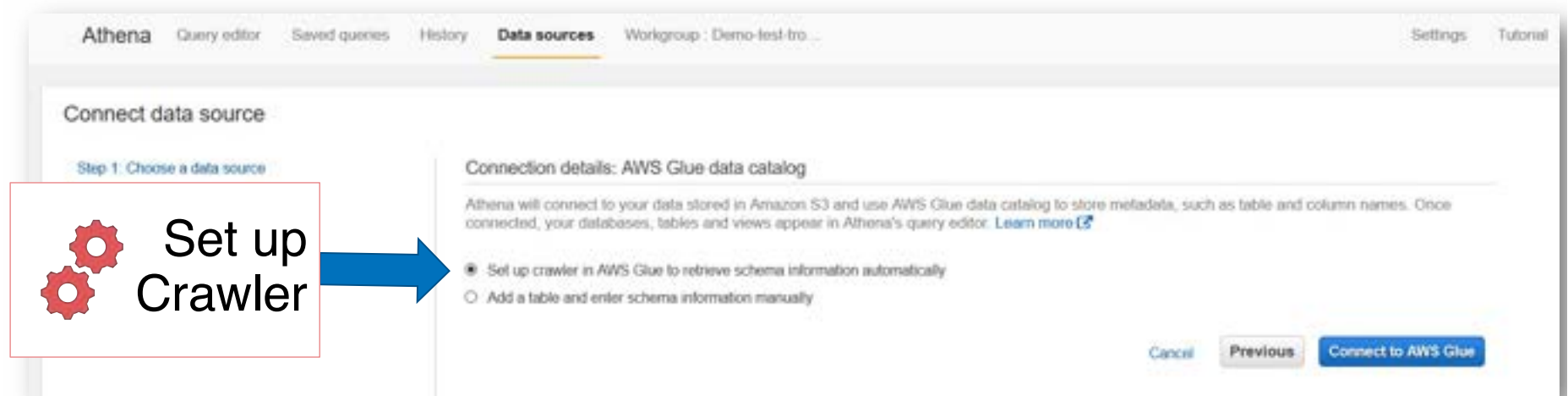
Details are available here- https://www.youtube.com/watch?v=_F4FhcDWSJg

Setting up.. Step 2

Choose S3
Buckets
and Glue



Setting up.. Step 3



Athena Query editor Saved queries History **Data sources** Workgroup: Demo-test-fo... Settings Tutorial

Connect data source


Step 1: Choose a data source

Connection details: AWS Glue data catalog

Athena will connect to your data stored in Amazon S3 and use AWS Glue data catalog to store metadata, such as table and column names. Once connected, your databases, tables and views appear in Athena's query editor. [Learn more](#)

- ☒ Set up crawler in AWS Glue to retrieve schema information automatically
- ☐ Add a table and enter schema information manually

Cancel Previous **Connect to AWS Glue**



Set up Crawler

Setting up.. Step 4

👉 Specify 'Data stores' and 'Crawl all folders'



The screenshot shows a dialog box titled "Add crawler" with a close button (X) in the top right corner. On the left is a sidebar with a list of steps: "Crawler info" (checked), "Crawler-1", "Crawler source type" (highlighted in green), "Data store", "S3 data store", "Include table", "Include", "Include", and "Include all steps". The main area is titled "Specify crawler source type" and contains the following text: "Choose Existing catalog tables to specify catalog tables as the crawler source. The selected tables specify the data stores to crawl. This option doesn't support JDBC data stores." Below this text are two sections. The first, "Crawler source type", has two radio buttons: "Data stores" (selected) and "Existing catalog tables". The second, "Repeat crawls of S3 data stores", has two radio buttons: "Crawl all folders" (selected) and "Crawl new folders only". At the bottom right are "Back" and "Next" buttons. Two large blue arrows point from the left towards the "Data stores" and "Crawl all folders" options.

Setting up.. Step 5

+ Add S3 Location for Metadata

Add crawler

Add a data store

Choose a data store
S3

Connection
Select a connection

Optionally, include a network connection to use with this S3 target. Note that each crawler is added to one network connection so any future S3 targets will also use the same connection (or none, if left blank).

Add connection

Crawl data in
☒ Specified path in my account
☐ Specified path in another account

Include path
s3://ra-pub-sars-cov2-metadata-us-east-1

All folders and files contained in the include path are crawled. For example, type s3://mybucket/myFolder to crawl all objects in myFolder within myBucket.

Exclude patterns (optional)

Back Next

Setting up.. Step 6

✓ Specify database



The screenshot shows the 'Add crawler' console window with the title 'Configure the crawler's output'. On the left, a sidebar lists steps: 'Crawler info', 'Crawler 1', 'Crawler source type', 'Data stores', 'Data store', 'IAM Role', 'Schedule', 'Output', and 'Choose an output'. The 'Database' field is highlighted with a red box and a blue arrow. Below it is an 'Add database' button. The 'Prefix added to tables (optional)' field is empty. At the bottom, there are 'Back' and 'Next' buttons.

Add crawler

Configure the crawler's output

Database ⓘ

s3://pub-sar

Add database

Prefix added to tables (optional) ⓘ

Type a prefix added to table names

- Grouping behavior for S3 data (optional)
- Configuration options (optional)

Back Next

Setting up.. Step 7

▶ Start the Crawler



The screenshot shows the AWS Glue console interface. On the left, the navigation menu includes 'Data catalog', 'Databases', 'Tables', 'Connections', 'Crawlers' (highlighted with a blue arrow), 'Classifiers', and 'Schema registries'. The main panel is titled 'Crawlers' and contains a description: 'A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.' Below this, there are buttons for 'Add crawler', 'Run crawler' (highlighted with a red box and a blue arrow), and 'Action'. A search bar is also present. A table lists three crawlers:

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	AlignedRoads		Ready	Logs	3 mins	3 mins	0	7
<input checked="" type="checkbox"/>	Sars-Cov-2		Ready	Logs	2 mins	2 mins	19	1
<input type="checkbox"/>	TotalSRA		Ready	Logs	4 mins	4 mins	0	7

... Ready to go! 🚀

Searching Submitter-Provided Metadata

❖ Some popular tasks include:

- I want Amplicon Sequencing data ([assay type](#))
- I want samples submitted as SARS-CoV-2 ([Org.](#))
- I want sample from the USA ([geog. location](#))

❖ Other questions you can address:

- I want Illumina platform data
- I want data released since the start of 2021
- I want data collected before 2020
- I want data submitted by Quest Diagnostics



This information originates in NCBI BioProject and BioSample



Table Definitions -

<https://www.ncbi.nlm.nih.gov/sra/docs/aligned-metadata-tables/>

Search Against Assay Type

```
SELECT run
FROM SARS_COV_2.metadata
WHERE assay_type = 'AMPLICON'
```

The screenshot shows the NCBI SRA Explorer interface. At the top, a query editor displays the SQL query: `SELECT run FROM SARS_COV_2.metadata WHERE assay_type = 'AMPLICON'`. Below the query editor, there are buttons for 'Run query', 'Save as', and 'Create'. The status bar indicates '(Run time: 1.23 seconds, Data scanned: 381.24 KB)'. Below the query editor, there are buttons for 'Format query' and 'Clear'. The results section shows a table with the following data:

	run
1	SRR13060144
2	SRR13060143
3	SRR13060142
4	SRR13060141
5	SRR13060140
6	SRR13060338
7	SRR13060337

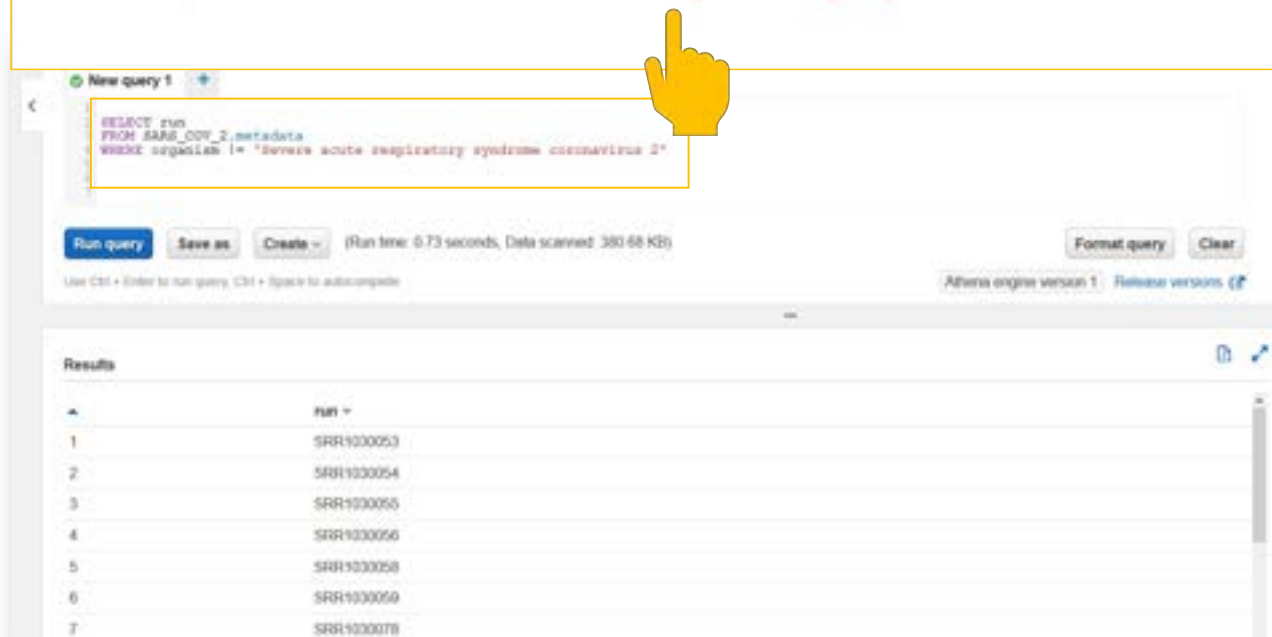
Available Assay Types



- RNA-Seq
- OTHER
- ChIP-Seq
- MeDIP-Seq
- WGS
- AMPLICON
- Targeted-Capture
- FL-cDNA
- WGA
- WXS

Search Against Submitted Organism

```
SELECT run
FROM SARS_COV_2.metadata
WHERE organism != 'Severe acute respiratory syndrome coronavirus 2'
```



The screenshot shows a web-based query interface. At the top, a text box contains the SQL query: `SELECT run FROM SARS_COV_2.metadata WHERE organism != 'Severe acute respiratory syndrome coronavirus 2'`. A yellow hand icon points to the `!=` operator. Below the text box are buttons for 'Run query', 'Save as', and 'Create'. Below these buttons, it says '(Run time: 0.73 seconds, Data scanned: 380.68 KB)'. To the right are buttons for 'Format query' and 'Clear'. Below the buttons, it says 'Use Ctrl + Enter to run query, Ctrl + Space to autocomplete'. Below this is a section labeled 'Results' with a table showing 7 rows of results. The first column is 'run' and the second column is 'run'.

	run
1	SRR1030053
2	SRR1030054
3	SRR1030055
4	SRR1030056
5	SRR1030058
6	SRR1030059
7	SRR1030078

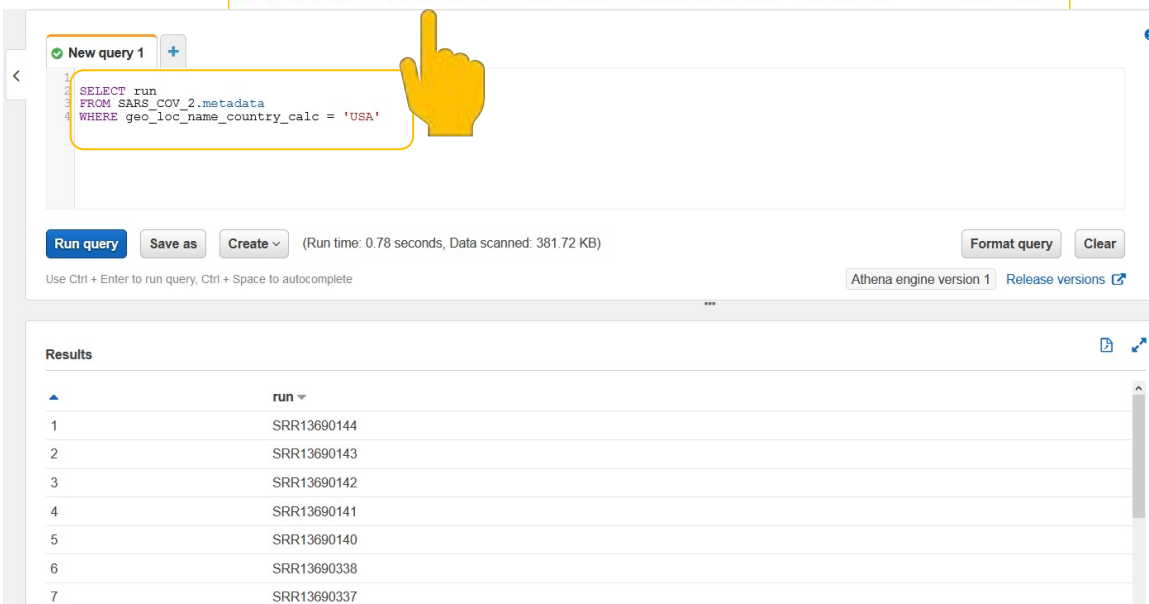
💡 Note '!=' means not equal

💡 Top 3 Organisms:

- *Severe acute respiratory syndrome coronavirus 2*
- *Homo sapiens*
- *Mus musculus*

Search Against Geographic Location

```
SELECT run
FROM SARS_COV_2.metadata
WHERE geo_loc_name_country_calc = 'USA'
```



The screenshot shows a query interface with a text input field containing the SQL query. Below the input field are buttons for 'Run query', 'Save as', and 'Create'. To the right of these buttons, it shows '(Run time: 0.78 seconds, Data scanned: 381.72 KB)'. Further right are 'Format query' and 'Clear' buttons. Below the input field, there is a 'Results' section with a table. The table has a header row with 'run' and a dropdown arrow. The table contains 7 rows of data, each with a number in the first column and a run ID in the second column.

	run
1	SRR13690144
2	SRR13690143
3	SRR13690142
4	SRR13690141
5	SRR13690140
6	SRR13690338
7	SRR13690337

Top 3 Countries Currently

- *United Kingdom*
- *USA*
- *Australia*

Names follow INSDC
Specifications -

<https://www.ncbi.nlm.nih.gov/genbank/collab/country/>

Searching Against NCBI Metadata:

A. STAT Results

- Taxonomy
 - Each run includes rows only for taxids with at least 1 kmer hit
- Self vs total hits
 - Self hits – hits directly to the associated taxa
 - Keep in mind how kmers are mapped up the tax hierarchy by STAT
 - Total hits – hits directly to that associated taxa plus hits to child taxa

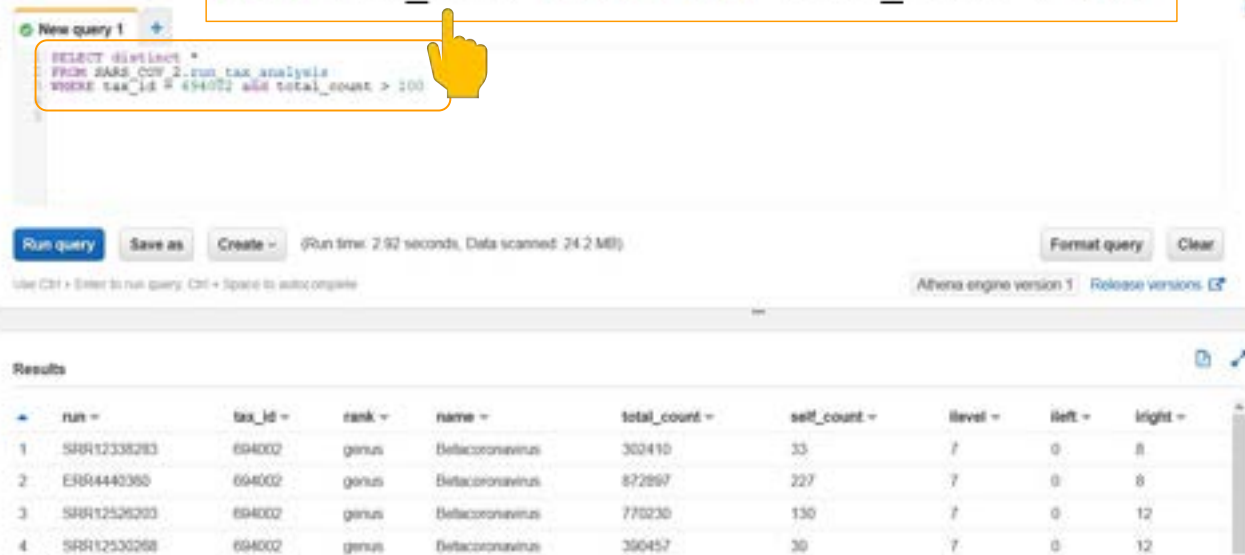
Search Using STAT Results

```
SELECT distinct *  
FROM SARS_COV_2.run tax_analysis  
WHERE tax_id = 694002 and total_count > 100
```

💡 Note the use of 'and' to indicate 2 'where' clauses

💡 We think **filtering based on STAT results is very powerful**, but we also realize it can be a little confusing

💡 **We'd love to hear from you** about any issues you have in using this data, or ideas you have about how to make using it easier at:
sra@ncbi.nlm.nih.gov



The screenshot shows the NCBI SRA query interface. At the top, a query is entered: `SELECT distinct * FROM SARS_COV_2.run tax_analysis WHERE tax_id = 694002 and total_count > 100`. Below the query, there are buttons for 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'. The 'Run query' button is highlighted. Below the buttons, the results are displayed in a table. The table has columns: run, tax_id, rank, name, total_count, self_count, level, left, and right. The results show four rows of data for the tax_id 694002.

run	tax_id	rank	name	total_count	self_count	level	left	right
SRR12335283	694002	genus	Betacoronavirus	302410	33	7	0	8
ERR4440360	694002	genus	Betacoronavirus	872897	227	7	0	8
SRR12526200	694002	genus	Betacoronavirus	770230	130	7	0	12
SRR12530268	694002	genus	Betacoronavirus	260457	30	7	0	12

Searching Against NCBI Metadata:

B. Assembled Sequences

- Contigs
 - ✓ Assembled using SAUTE using the SARS-CoV-2 RefSeq as a guide
 - Conservative assembly
 - SAUTE [GitHub](#) site.
 - ✓ Checked against (nucleotide) [nt Blast](#) database
 - ✓ Checked using STAT
 - ✓ Annotated using VIGOR3
 - We still recommend VADR for GenBank submission
 - VADR Github - <https://github.com/ncbi/vadr>

Search Against Contig Length

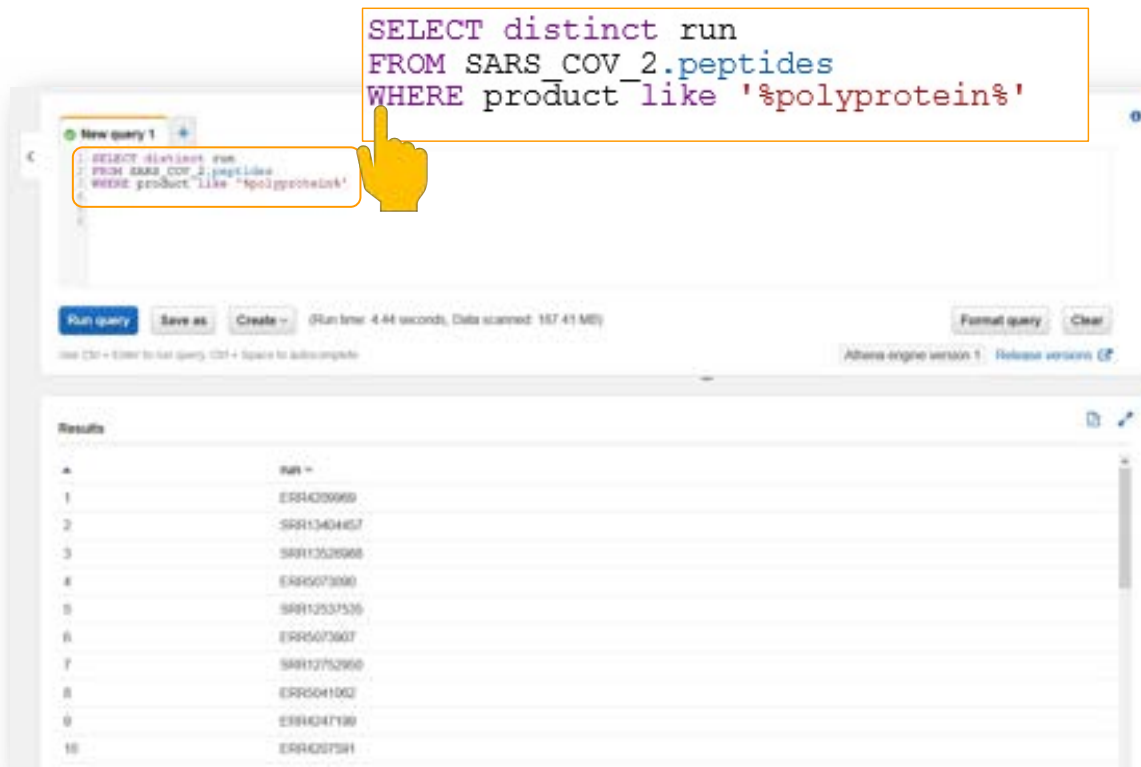
```
SELECT distinct run
FROM SARS_COV_2.contigs
WHERE length > 28000
```

Use 'min' and 'max' to find the range of values in your selection

The screenshot shows the NCBI SRA Explorer interface. A query is entered in the 'New query 1' field: `SELECT distinct run FROM SARS_COV_2.contigs WHERE length > 28000`. A yellow hand icon points to the query input area. Below the query input, there are buttons for 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'. The 'Run query' button is highlighted. Below the buttons, the status bar shows '(Run time: 2.35 seconds, Data scanned: 42.81 MB)'. The 'Results' section displays a table with two columns: 'run' and 'contigs'. The table contains six rows of results, each with a 'run' ID and a 'contigs' ID.

	run	contigs
1	ERR4147587	
2	ERR4157000	
3	ERR4238317	
4	ERR4296962	
5	ERR4296964	
6	ERR4296968	

Search Against Peptide Products



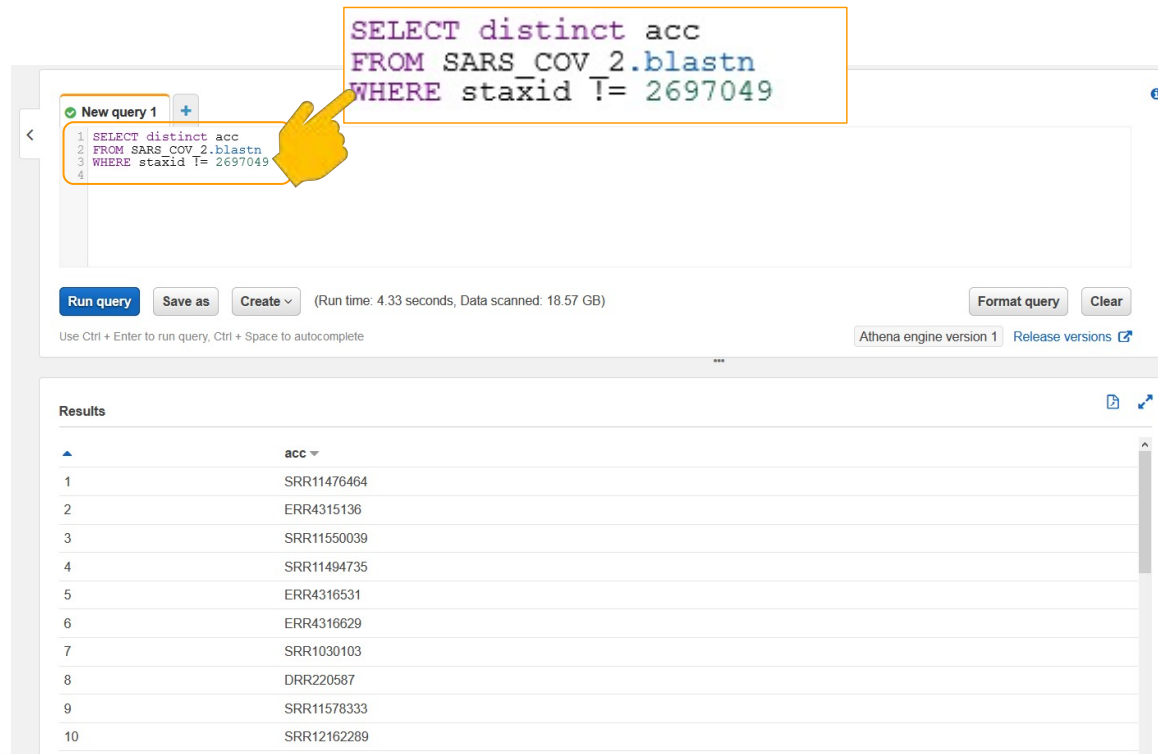
The screenshot shows a web-based query interface. At the top, a text box contains the following SQL query:

```
SELECT distinct run
FROM SARS_COV_2.peptides
WHERE product like '%polyprotein%'
```

A yellow hand icon points to the query text. Below the query text, a smaller box shows the same query. The interface includes buttons for 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'. Below the query, a 'Results' section displays a table with two columns: 'run' and 'product'. The table contains 10 rows of data, with the 'run' column values being: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. The 'product' column values are: ESR4259999, SRR113404657, SRR113526968, ESR45073000, SRR112537536, ESR45073607, SRR112752960, ESR45041002, ESR450417190, and ESR4507594.

- 💡 The '%' acts as a wild card - effectively, any string that includes the value between the two '%' will be found
- 💡 Names come from VIGOR3
- 💡 If you have a complete genome you would like to submit to GenBank we recommend using VADR to ensure no errors during the submission process
<https://github.com/ncbi/vadr>

Search Against Top Blast Hit Taxa



The screenshot shows the NCBI Athena query interface. A yellow box highlights the SQL query: `SELECT distinct acc
FROM SARS_COV_2.blastn
WHERE staxid != 2697049`. A yellow hand icon points to the query. Below the query, there are buttons for 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'. The 'Run query' button is highlighted. Below the buttons, the text '(Run time: 4.33 seconds, Data scanned: 18.57 GB)' is displayed. The 'Results' section shows a table with 10 rows and 2 columns: 'acc' and 'staxid'. The 'acc' column contains accession numbers, and the 'staxid' column contains the staxid values.

	acc
1	SRR11476464
2	ERR4315136
3	SRR11550039
4	SRR11494735
5	ERR4316531
6	ERR4316629
7	SRR1030103
8	DRR220587
9	SRR11578333
10	SRR12162289

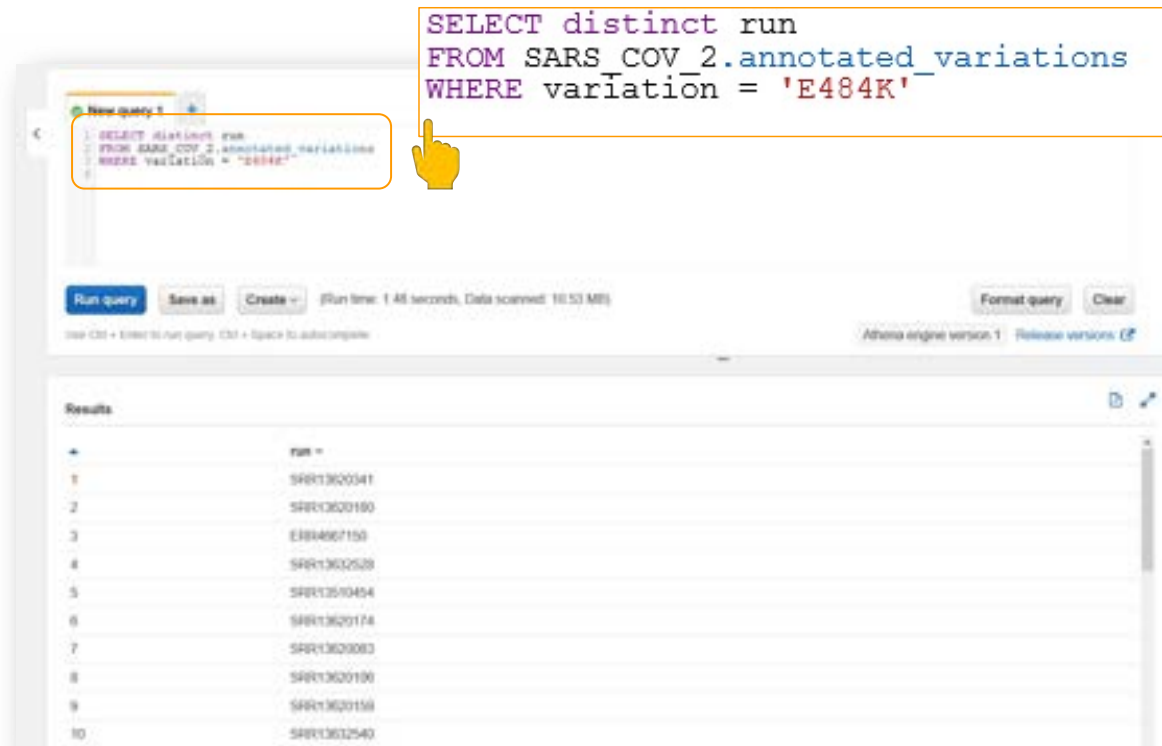
- 💡 Tax IDs are from the NCBI Taxonomy database - <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>
- 💡 Only the top hit is reported
- 💡 Result from megablast against nt BLAST database

Search Against NCBI Metadata:

C. VCF Results:

- Documentation
 - <https://www.ncbi.nlm.nih.gov/sra/docs/sars-cov-2-variant-calling>
- Outline
 - Trimming via Trimmomatic
 - Hisat2 for alignment to SARS-CoV-2
 - Samtools for bam conversion
 - Bcftools for pileup and VCF generation

Search Against Protein Variation



The screenshot shows a web interface for searching protein variations. A SQL query is entered in a text box, and the results are displayed in a table below.

```
SELECT distinct run
FROM SARS_COV_2.annotated_variations
WHERE variation = 'E484K'
```

Run query Save as Create (Run time: 1.45 seconds, Data scanned: 10.53 MB) Format query Clear

Results

	run =
1	SRR13620341
2	SRR13620160
3	E884867150
4	SRR13622528
5	SRR13510454
6	SRR13620174
7	SRR13620883
8	SRR13620190
9	SRR13620158
10	SRR13612540



Also include variations listed by position, reference and alternate alleles, protein name, protein position, reference, and alternate amino acid

How to get SRA runs for your analyses?

- You can download your Athena results as a CSV!
- AWS Commands
 - `aws s3 cp --recursive s3://sra-pub-sars-cov2/RA0/ERR4145453 ./`
- SRA Toolkit
 - https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc
- May I interest you in a SARF (SRA Aligned Read Format)
 - Compatible with SRA Toolkit
 - Reads aligned to contigs
 - Can extract reads, reads aligned to contigs, or contigs

I have my runs, what now?

- Whatever you want!
 - BLAST DBs
 - Assembly
 - Variant Calling
 - Something else? Let us know: sra@ncbi.nlm.nih.gov
- Stay tuned for future webinars on NCBI Cloud-based tools!
 - to our blog - [NCBI Insights](#).
 - Follow us on



DIY resources

- AWS Docs – Does AWS have any links?
- NCBI Help Docs
 - Getting Started - <https://www.ncbi.nlm.nih.gov/sra/docs/sra-aws-download/>
 - Athena Set-up - <https://www.ncbi.nlm.nih.gov/sra/docs/sra-athena/>
 - Athena Use - <https://www.ncbi.nlm.nih.gov/sra/docs/sra-athena-examples/>
 - Table Definitions - <https://www.ncbi.nlm.nih.gov/sra/docs/aligned-metadata-tables/>

DIY Resources contd..

- NCBI Cloud Data & tools YouTube playlist (User:NCBINLM):
 - <https://tinyurl.com/SRAonthecloud>
- NCBI's COVID-19 resources: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>

Keeping the dialog going..

- Here is how to reach us, Email: sra@ncbi.nlm.nih.gov
- Share your ideas on improving our existent documentation: <https://tinyurl.com/SRAcloudDoc>.
- Send us your questions or input on new functionality. (e.g. API for Athena)
- Let us know how we can better serve you!

