

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



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



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/>

Registry of Open Data on AWS

Basic Local Alignment Sequences Tool (BLAST) Databases

[bioinformatics](#) [biology](#) [genetic](#) [genomic](#) [health](#) [life sciences](#) [protein](#) [reference index](#) [transcriptomics](#)

Description
A centralized repository of pre-formatted BLAST databases created by the National Center for Biotechnology Information (NCBI).

Update Frequency
Periodically

License
"NIH Genomic Data Sharing Policy"

Documentation
https://github.com/ncbi/blast_plus_docs

Managed By
National Library of Medicine (NLM)
See all datasets managed by National Library of Medicine (NLM).

Contact
<https://support.nlm.nih.gov/support/create-case/>

Usage Examples

- [BLAST+ Docker by NCBI BLAST](#)

Publications

- [BLAST+: Architecture and Applications](#) by Christian Camacho 1, George

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

Registry of Open Data on AWS

NIH NCBI Sequence Read Archive (SRA) on AWS

[bam](#) [fastq](#) [genetic](#) [genomic](#) [life sciences](#) [STRIDES](#) [transcriptomics](#) [whole genome sequencing](#)

Description
The Sequence Read Archive (SRA), produced by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) at the National Institutes of Health (NIH), stores raw DNA sequencing data and alignment information from high-throughput sequencing platforms. The SRA provides open access to these biological sequence data to support the research community's efforts to enhance reproducibility and make new discoveries by comparing data sets. This bucket contains public SRA data in original format from select high value and newly-released studies.

Update Frequency
Daily

License
U.S. Government work

Documentation
<https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud/>

Managed By
National Library of Medicine (NLM)
See all datasets managed by National Library of Medicine (NLM).

Contact
sra@ncbi.nlm.nih.gov

Usage Examples

Resources on AWS

Description
.bam, .cram, .fastq, .run, .sra, .vcf
Resource type
S3 Bucket
Amazon Resource Name (ARN)
arn:aws:s3:::ncbi-blast-databases
AWS Region
us-east-1
AWS CLI Access (No AWS account required)
aws s3 ls s3://ncbi-blast-databases/ --no-sign-request

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

Registry of Open Data on AWS

COVID-19 Genome Sequence Dataset

[bam](#) [bioinformatics](#) [biology](#) [coronavirus](#) [COVID-19](#) [cram](#) [fastq](#) [genetic](#) [genomic](#) [health](#) [life sciences](#) [MERS](#) [SARS](#) [STRIDES](#) [transcriptomics](#) [virus](#)

Description
A centralized sequence repository for all strains of novel corona virus (SARS-CoV-2) submitted to the National Center for Biotechnology Information (NCBI). Included are both the original sequences submitted by the principal investigator as well as SRA-processed sequences that require the SRA Toolkit for analysis.

Update Frequency
Hourly

License
NIH Genomic Data Sharing Policy

Documentation
<https://www.ncbi.nlm.nih.gov/sra/docs/sra-aws-download/>

Managed By
National Library of Medicine (NLM)
See all datasets managed by National Library of Medicine (NLM).

Contact
<https://support.nlm.nih.gov/support/create-case/>

Usage Examples

Tools & Applications

- Download SRA sequence data using Amazon Web Services (AWS) by NCBI SRA

Resources on AWS

Description
Genomic sequence reads of SARS-CoV-2 and related coronaviridae, organized by NCBI accession. Files in the [sra-src](#) folder are in FASTQ, BAM, or CRAM format (original submission); files in the [run](#) folder are in .sra format and require the [SRA Toolkit](#)

Resource type
S3 Bucket

Amazon Resource Name (ARN)
arn:aws:s3:::sra-pub-sars-cov2

AWS Region
us-east-1

AWS CLI Access (No AWS account required)
aws s3 ls s3://sra-pub-sars-cov2/ --no-sign-request

Description
Metadata for sra-pub-sars-cov2 in an Athena-queryable format

Resource type
S3 Bucket

Amazon Resource Name (ARN)
arn:aws:s3:::sra-pub-sars-cov2-metadata-us-east-1

AWS Region
us-east-1

AWS CLI Access (No AWS account required)
aws s3 ls s3://sra-pub-sars-cov2-metadata-us-east-1/ --no-sign-request

Helpful tips

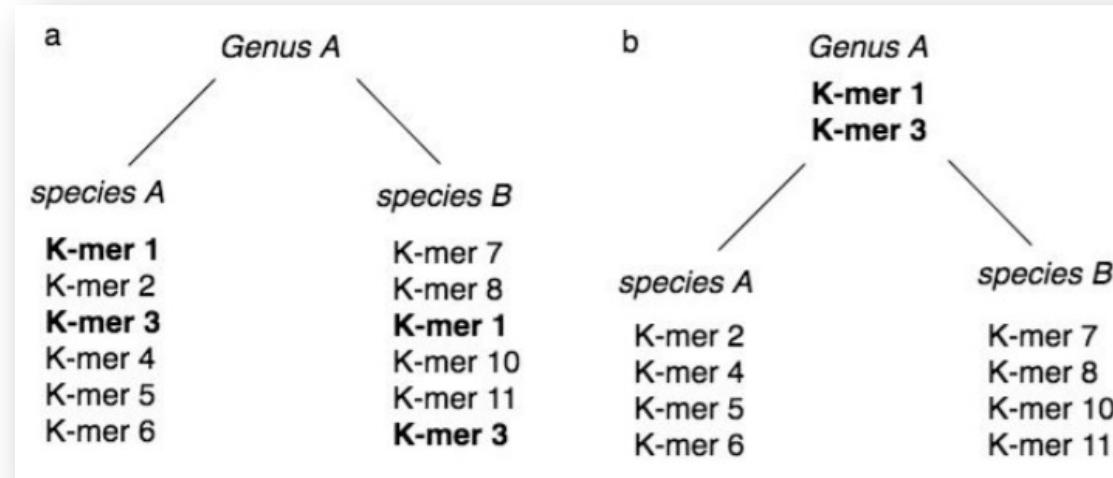
- **AWS ODP v. commercial buckets**
 - Egress from ODP is *free* from anywhere (i.e. cloud and local machines)
 - 2 kinds of SARS-CoV-2 ODP buckets, data and metadata, metadata contains parquet format files for use with Athena
- **Athena**
 - Support SQL-like querying of data
 - Queries cost money, typically < \$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

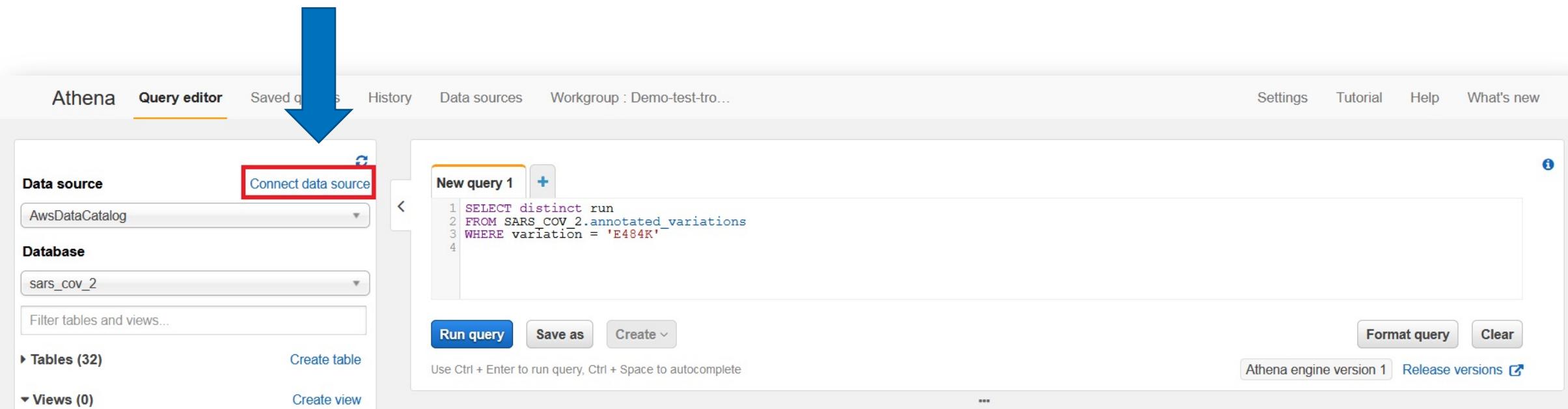
- Coronaviridae Detection Tool
 - Docker - <https://hub.docker.com/r/ncbi/sars-cov-2-detection-tool>
 - Documentation - <https://www.ncbi.nlm.nih.gov/sra/docs/sra-detection-tool/>
- Human Scrubber
 - Docker - <https://hub.docker.com/r/ncbi/sra-human-scrubber>
 - Github - <https://github.com/ncbi/sra-human-scrubber>

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



The screenshot shows the AWS Athena Query Editor. At the top, there's a navigation bar with tabs: 'Athena' (selected), 'Query editor' (highlighted in orange), 'Saved queries', 'History', 'Data sources', and 'Workgroup : Demo-test-tro...'. To the right of the workgroup are links for 'Settings', 'Tutorial', 'Help', and 'What's new'. Below the navigation bar, there's a sidebar on the left with sections for 'Data source' (set to 'AwsDataCatalog'), 'Database' (set to 'sars_cov_2'), and buttons for 'Tables (32)', 'Views (0)', 'Create table', and 'Create view'. The main area is titled 'New query 1' and contains a code editor with the following SQL query:

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

Below the code editor are buttons for 'Run query', 'Save as', and 'Create'. To the right are buttons for 'Format query' and 'Clear'. At the bottom of the main area, it says 'Use Ctrl + Enter to run query, Ctrl + Space to autocomplete'. In the bottom right corner, there are links for 'Athena engine version 1' and 'Release versions'.

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

Setting up.. Step 2

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

Connect data source

Step 1: Choose a data source

Step 2: Connection details

Choose where your data is located

Athena queries data where it is. Data is not loaded or moved. [Learn more](#)

Query data in Amazon S3
Choose an external data catalog.


Query a data source
Configure a connector for common data sources.


Choose a metadata catalog

The catalog contains the schema for the source data such as column names, data types and table names. [Learn more](#)

AWS Glue data catalog
 AWS Glue data catalog

Apache Hive metastore
 Apache Hive metastore

[Cancel](#) **Next**

Choose S3 Buckets and Glue



Setting up.. Step 3

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

Connect data source

Step 1: Choose a data source



Set up Crawler

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**

Setting up.. Step 4



Specify 'Data stores' and 'Crawl all folders'

Add crawler X

Crawler info
Crawler-1

Crawler source type

Data store

IAM Role

Schedule

Output

Review all steps

Specify crawler source type

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.

Crawler source type

Data stores
 Existing catalog tables

Repeat crawls of S3 data stores

Crawl all folders
 Crawl new folders only

Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.

[Back](#) [Next](#)

Setting up.. Step 5

+ Add S3 Location for Metadata

Add crawler

Add a data store

Crawler info
Crawler-1

Crawler source type
Data stores

Data store

IAM Role

Schedule

Output

Review all steps

Choose a data store
S3

Connection
Select a connection

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

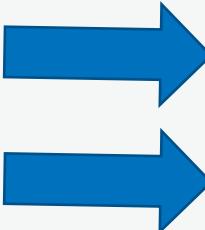
Crawl data in
 Specified path in my account
 Specified path in another account

Include path
s3://sra-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

Add crawler X

Crawler info
Crawler-1

Crawler source type
Data stores

Data store
S3: s3://sra-pub-sar...

IAM Role
arn:aws:iam::2508136
60784:role/NCBI-GlueServiceRole

Schedule
Run on demand

Output

Review all steps

Configure the crawler's output

Database i
 Add database

Prefix added to tables (optional) i

▶ Grouping behavior for S3 data (optional)

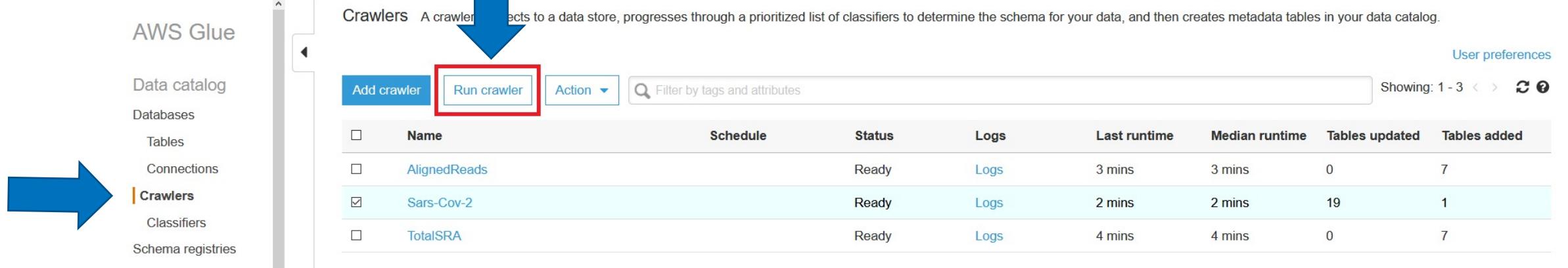
▶ Configuration options (optional)

Back Next



Setting up.. Step 7

▶ Start the Crawler



AWS Glue

Crawlers 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.

Add crawler Run crawler Action ▾ Filter by tags and attributes Showing: 1 - 3 < > User preferences

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
AlignedReads		Ready	Logs	3 mins	3 mins	0	7
Sars-CoV-2		Ready	Logs	2 mins	2 mins	19	1
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](#))

 This information originates in NCBI BioProject and BioSample

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

❖ 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

Search Against Assay Type

The screenshot shows an Athena query interface. In the top-left, a code editor contains the following SQL query:

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

Below the code editor, the query history shows the same query with line numbers 1 through 6. To the right of the code editor are buttons for "Run query", "Save as", "Create", "Format query", and "Clear". Below these buttons, it says "(Run time: 1.23 seconds, Data scanned: 381.24 KB)". Further down, there's a note "Use Ctrl + Enter to run query, Ctrl + Space to autocomplete". On the far right, it says "Athena engine version 1" and "Release versions".

The bottom section is titled "Results" and displays a table with the following data:

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

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'
```

A screenshot of the AWS Athena console. A yellow hand icon points to the query text in the top-left pane. The query is:

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

The results pane below shows a table with one column, "run", containing 7 entries:

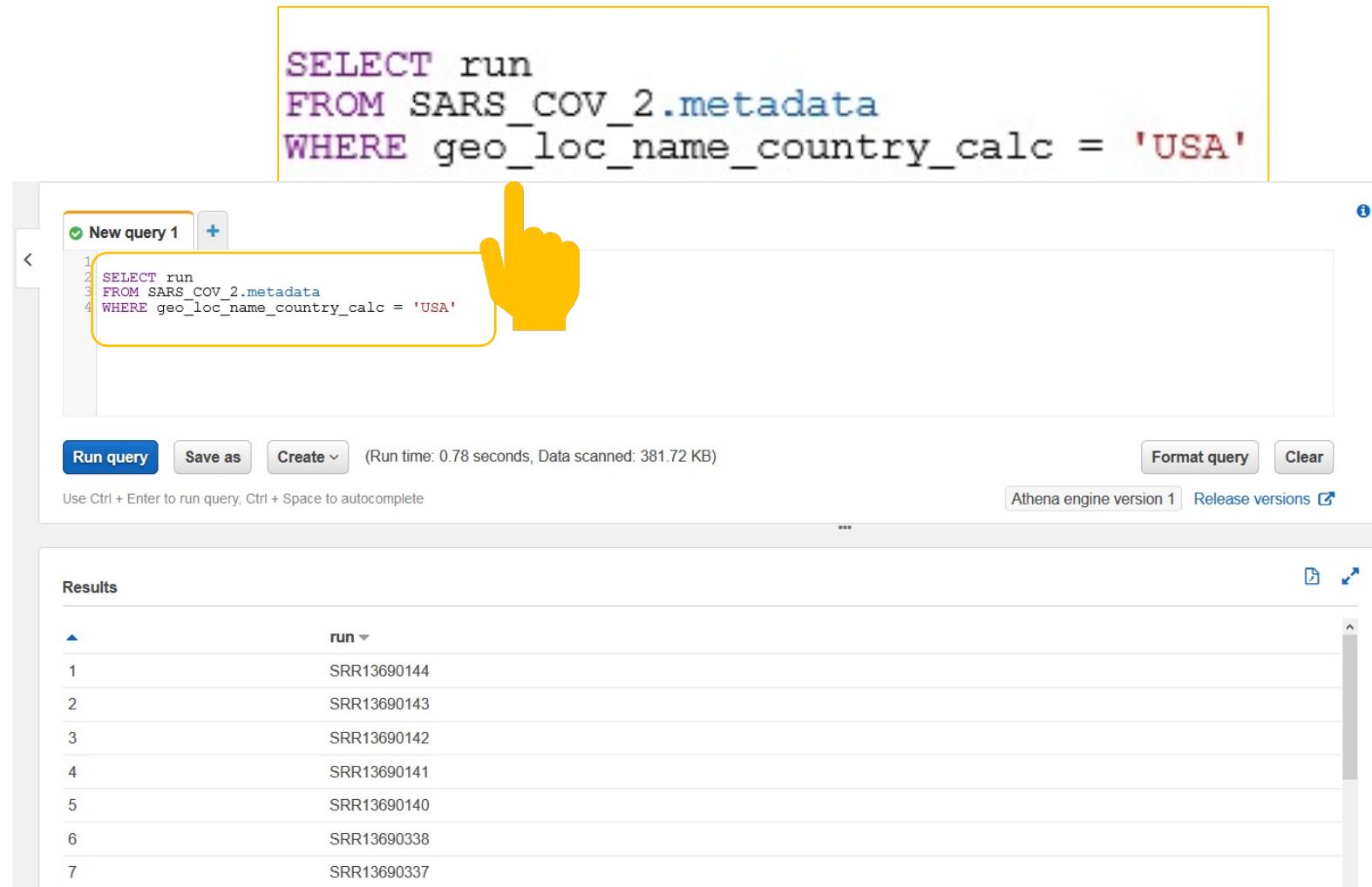
	run
1	SRR1030053
2	SRR1030054
3	SRR1030055
4	SRR1030056
5	SRR1030058
6	SRR1030059
7	SRR1030078

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'
```

New query 1

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

Run query Save as Create (Run time: 0.78 seconds, Data scanned: 381.72 KB) Format query Clear Athena engine version 1 Release versions

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

	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 -

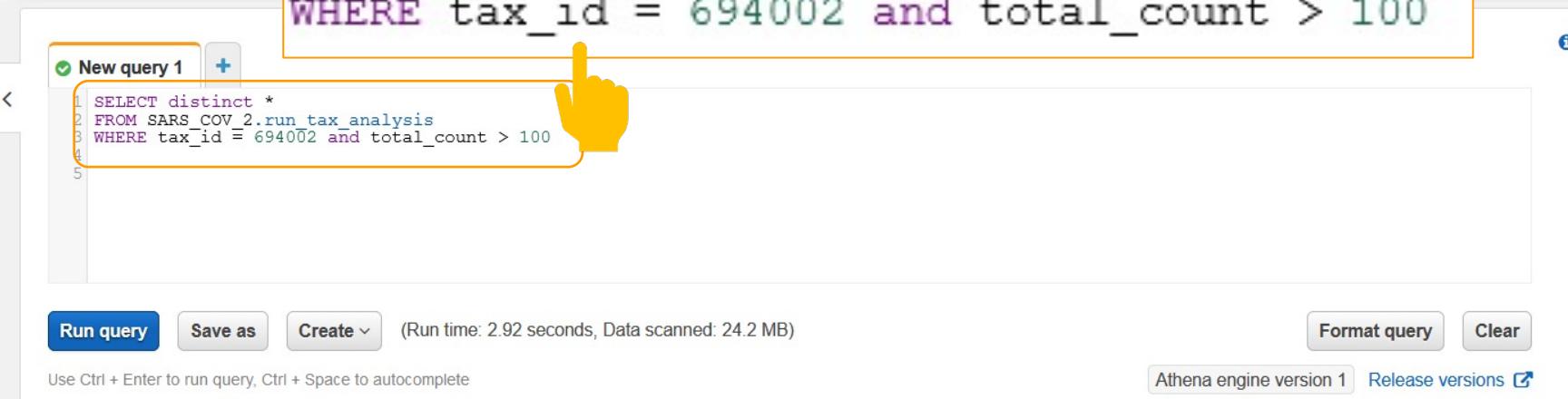
<https://www.ncbi.nlm.nih.gov/genbank/collab/country/>

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



A screenshot of the AWS Athena console. The query editor shows a query:

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

The entire query is highlighted with an orange box, and a yellow hand cursor points to the WHERE clause. Below the editor, there are buttons for "Run query", "Save as", "Create", and "Format query". The status bar indicates "(Run time: 2.92 seconds, Data scanned: 24.2 MB)".

The results table has columns: run, tax_id, rank, name, total_count, self_count, ilevel, ileft, iright. The data is as follows:

	run	tax_id	rank	name	total_count	self_count	ilevel	ileft	iright
1	SRR12338283	694002	genus	Betacoronavirus	302410	33	7	0	8
2	ERR4440360	694002	genus	Betacoronavirus	872897	227	7	0	8
3	SRR12526203	694002	genus	Betacoronavirus	770230	130	7	0	12
4	SRR12530268	694002	genus	Betacoronavirus	390457	30	7	0	12

 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

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

The screenshot shows the AWS Athena console interface. A yellow hand icon points to the first line of the query in the left-hand query editor. The query itself is highlighted with an orange border:

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

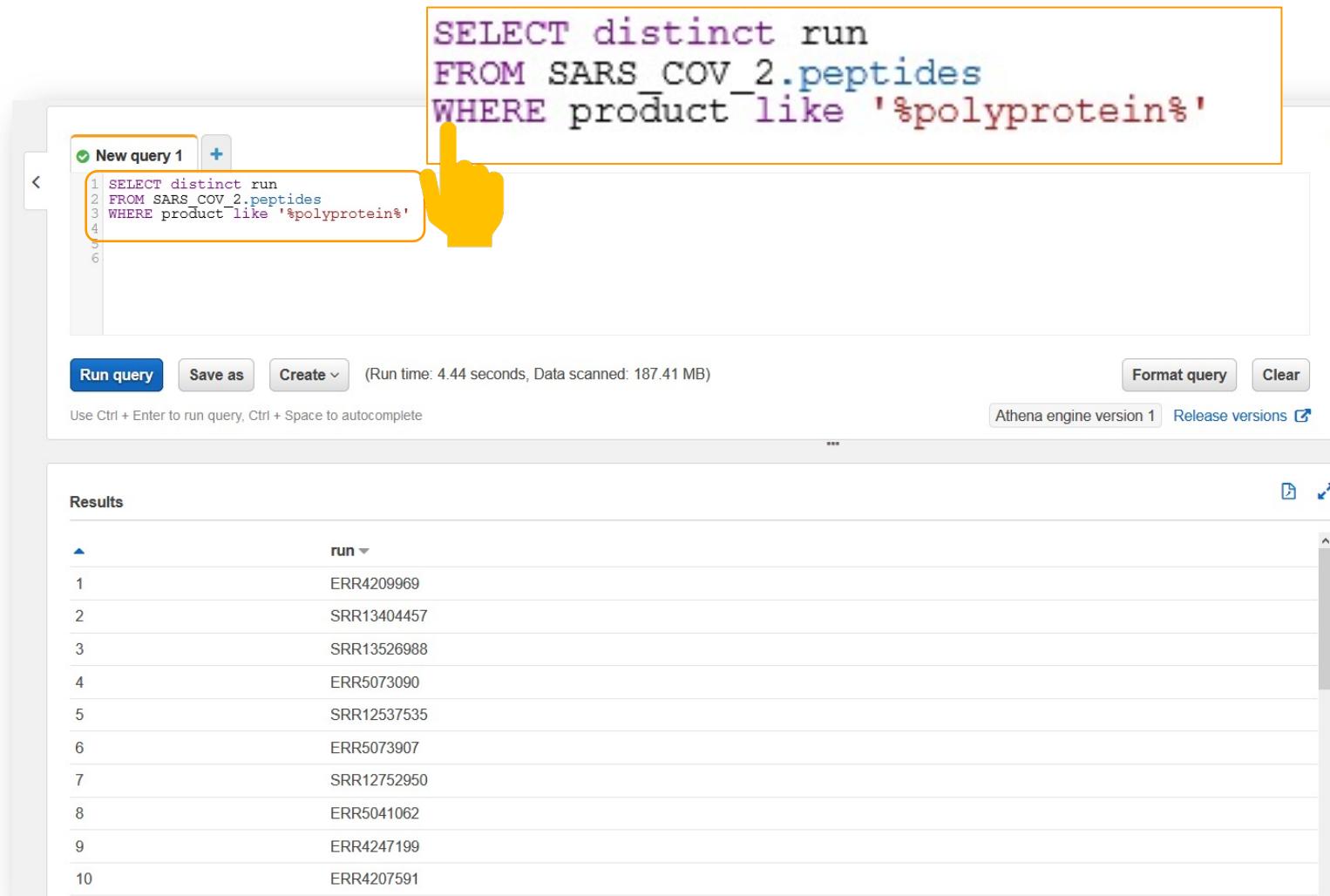
The query editor also displays the full query with line numbers (1-6) and includes buttons for "Run query", "Save as", "Create", "Format query", and "Clear". Below the editor, it says "(Run time: 2.35 seconds, Data scanned: 42.81 MB)". On the right, there's a note about using "min" and "max" to find value ranges.

Results

run	
1	ERR4147587
2	ERR4157960
3	ERR4238317
4	ERR4296962
5	ERR4296964
6	ERR4296968

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

Search Against Peptide Products



A screenshot of the AWS Athena console. The query editor shows the following SQL code:

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

The code is highlighted with an orange box. A yellow hand cursor is pointing at the WHERE clause. Below the editor, the results section displays a table of 10 runs, each with a unique identifier:

run
ERR4209969
SRR1340447
SRR13526988
ERR5073090
SRR12537535
ERR5073907
SRR12752950
ERR5041062
ERR4247199
ERR4207591

 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 AWS Athena console interface. At the top, a query is entered:

```
SELECT distinct acc  
FROM SARS_COV_2.blastn  
WHERE staxid T= 2697049
```

A yellow hand icon points to the first line of the query. Below the query, there are several buttons: Run query, Save as, Create, Format query, Clear, Athena engine version 1, and Release versions. A status message indicates: (Run time: 4.33 seconds, Data scanned: 18.57 GB). Below the buttons, a note says: Use Ctrl + Enter to run query, Ctrl + Space to autocomplete.

The Results section displays a table with 10 rows, ordered by 'acc' (column header). The data is as follows:

	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 the AWS Athena console interface. A query is being run against the 'SARS_COV_2.annotated_variations' database. The query is:

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

A yellow callout box highlights the WHERE clause of the query. A yellow hand cursor icon is positioned over the highlighted text. Below the query editor, there are buttons for 'Run query', 'Save as', 'Create', and 'Format query'. The results section shows a table with the 'run' column, listing 10 entries from 1 to 10, each corresponding to a sample ID.

	run
1	SRR13620341
2	SRR13620180
3	ERR4667150
4	SRR13632528
5	SRR13510454
6	SRR13620174
7	SRR13620083
8	SRR13620106
9	SRR13620159
10	SRR13632540



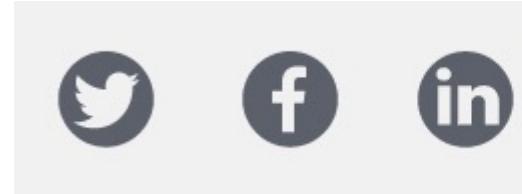
Also include variations listed by position, reference and alternate alleles, protein name, protein position, reference, and alternate amino acid

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!

