These slides were part of an **AWS Webinar in 2021**

NCBI Minute: SRA in AWS Athena for SARS-CoV-2 Research and More
https://www.youtube.com/watch?v=_F4FhcDWSJg&t=1080s

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!

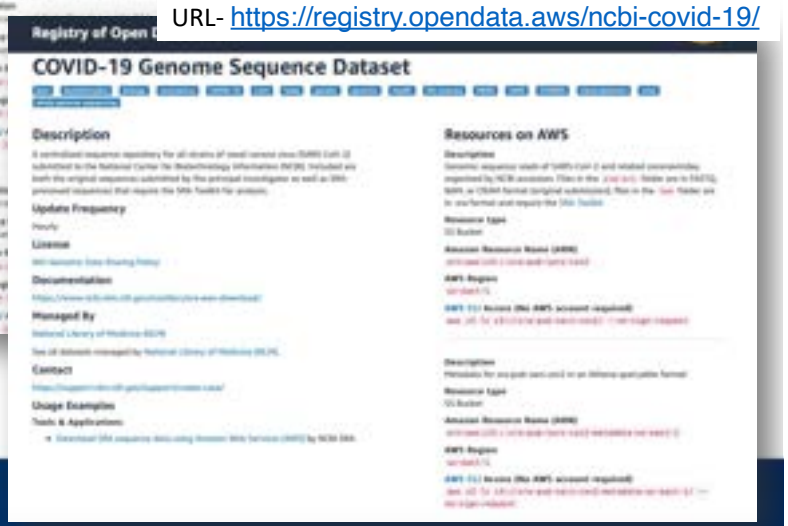
- All of the public and controlled-access SRA normalized format data is being migrated to AWS ODP
- Estimated completion in April 2021.

NCBI's open data sets on AWS ODP

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

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

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



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

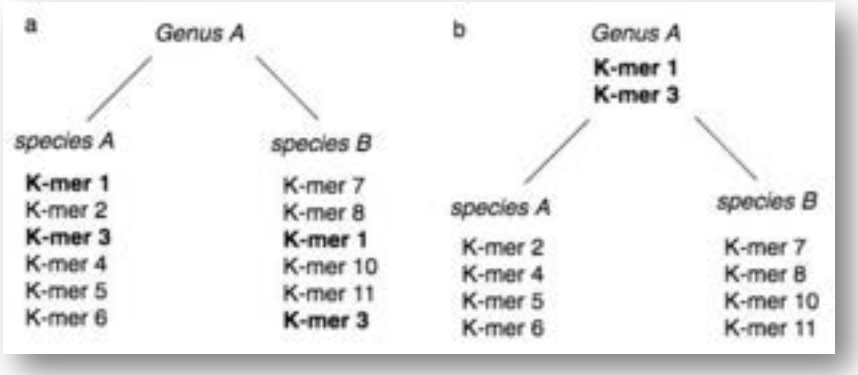
What data is in scope for this talk?

- 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.

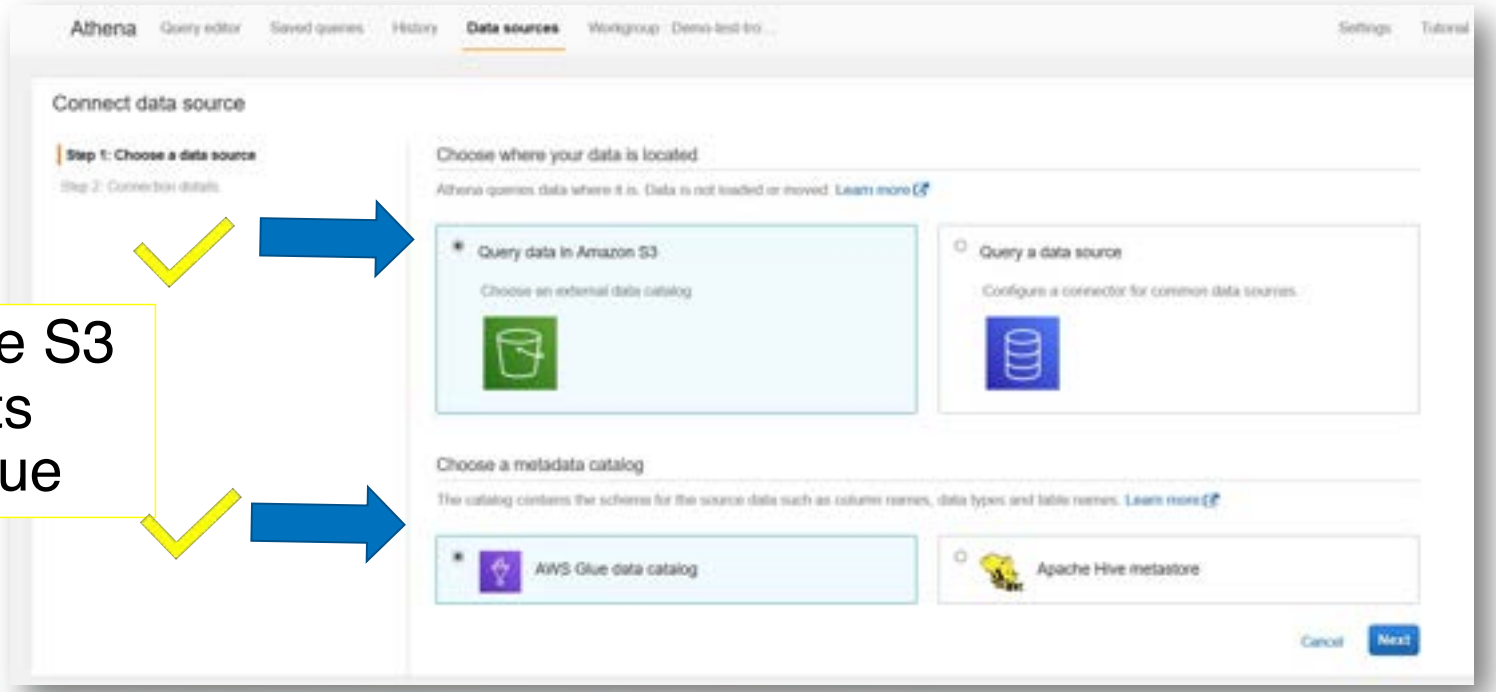
Setting up..Step 1

Connect to Data Source

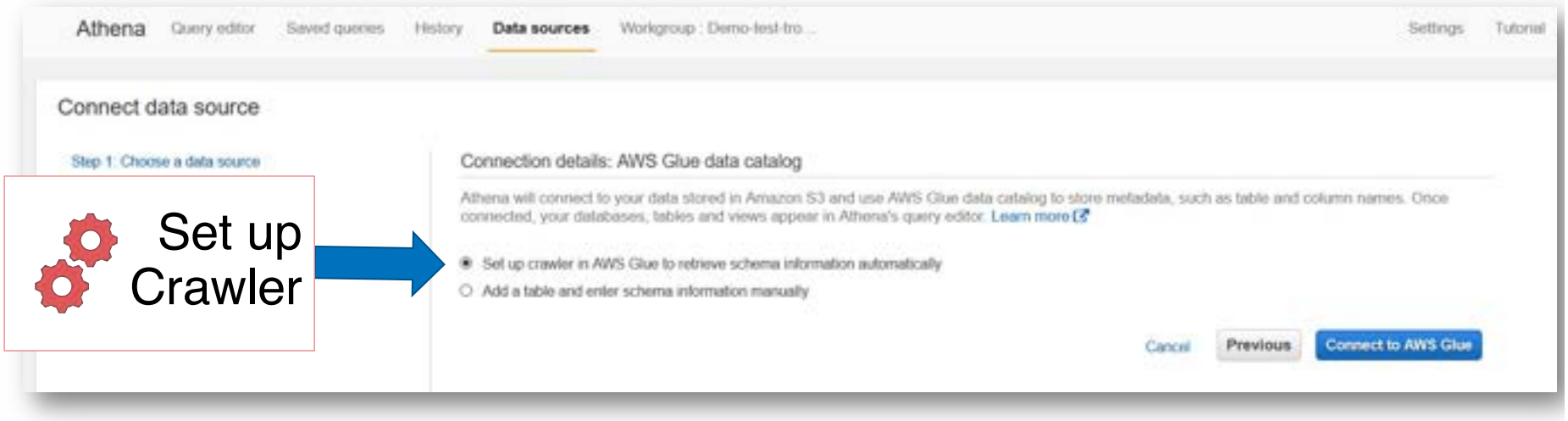
Details are available here- https://www.youtube.com/watch?v=_F4FhcDWSJg

Setting up.. Step 2

Choose S3 Buckets and Glue



Setting up.. Step 3

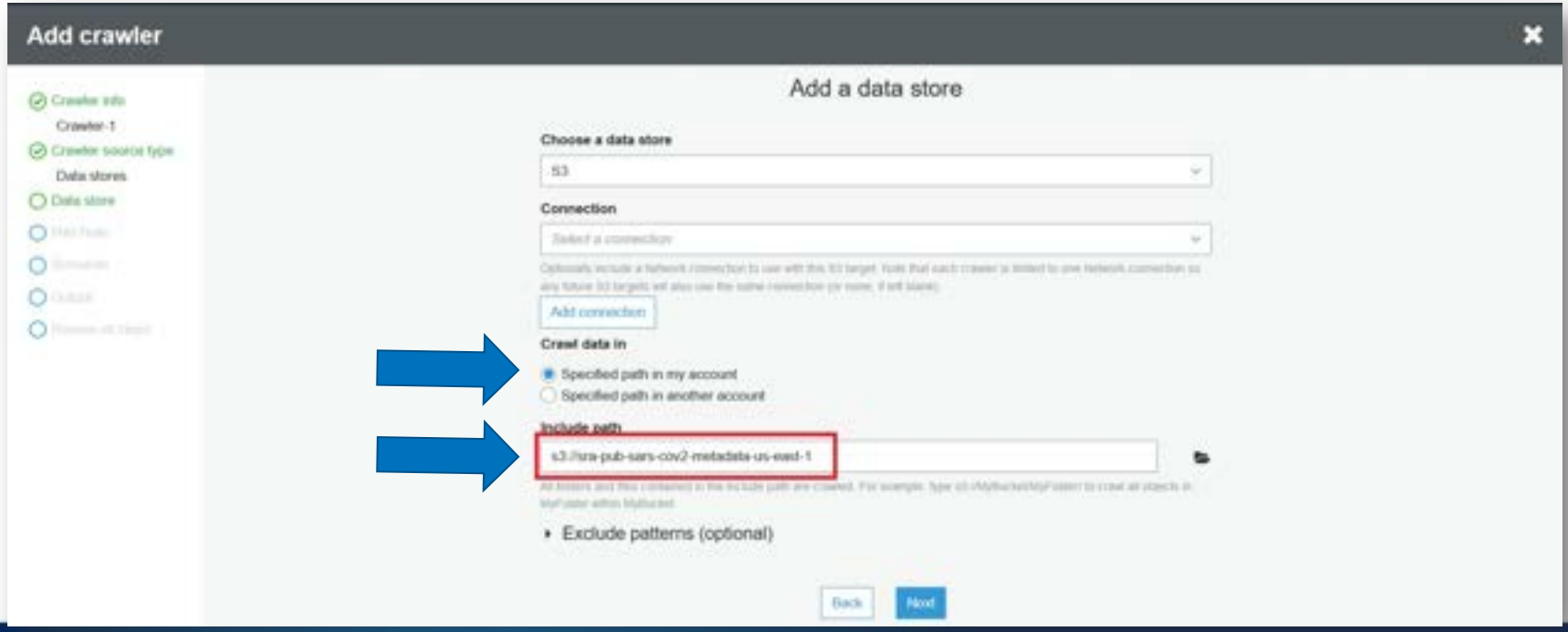


Setting up.. Step 4

Specify 'Data stores' and 'Crawl all folders'

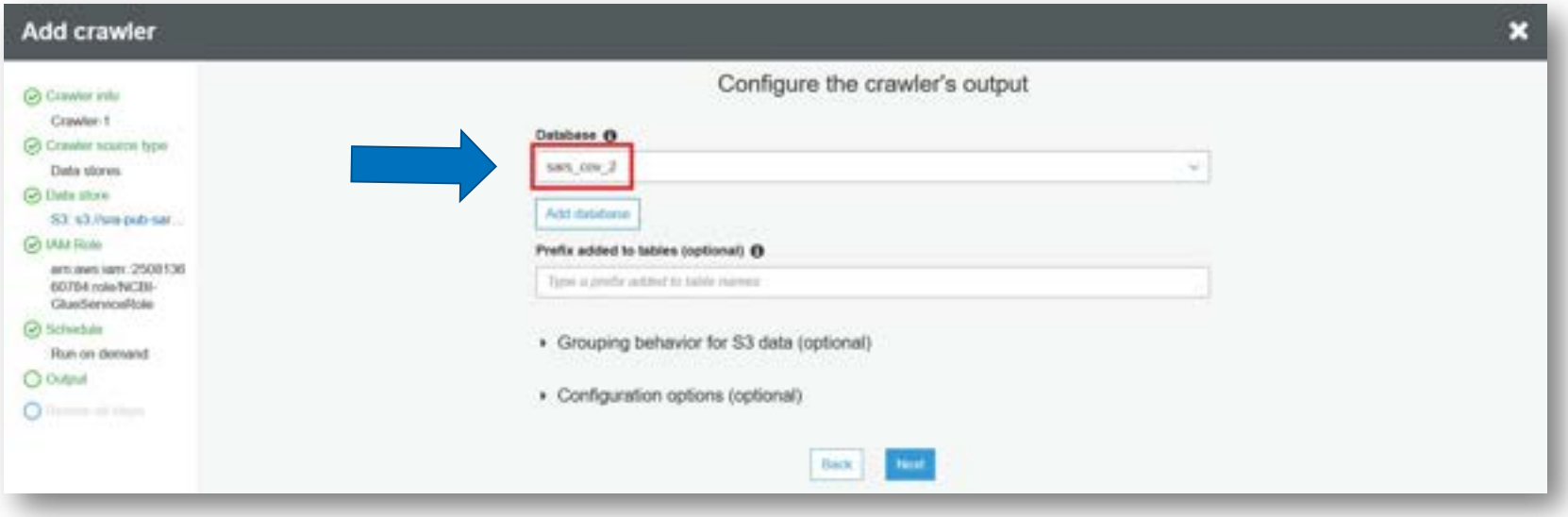
Setting up.. Step 5

- Add S3 Location for Metadata



Setting up.. Step 6

- Specify database



Setting up.. Step 7

Start the Crawler

… 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)

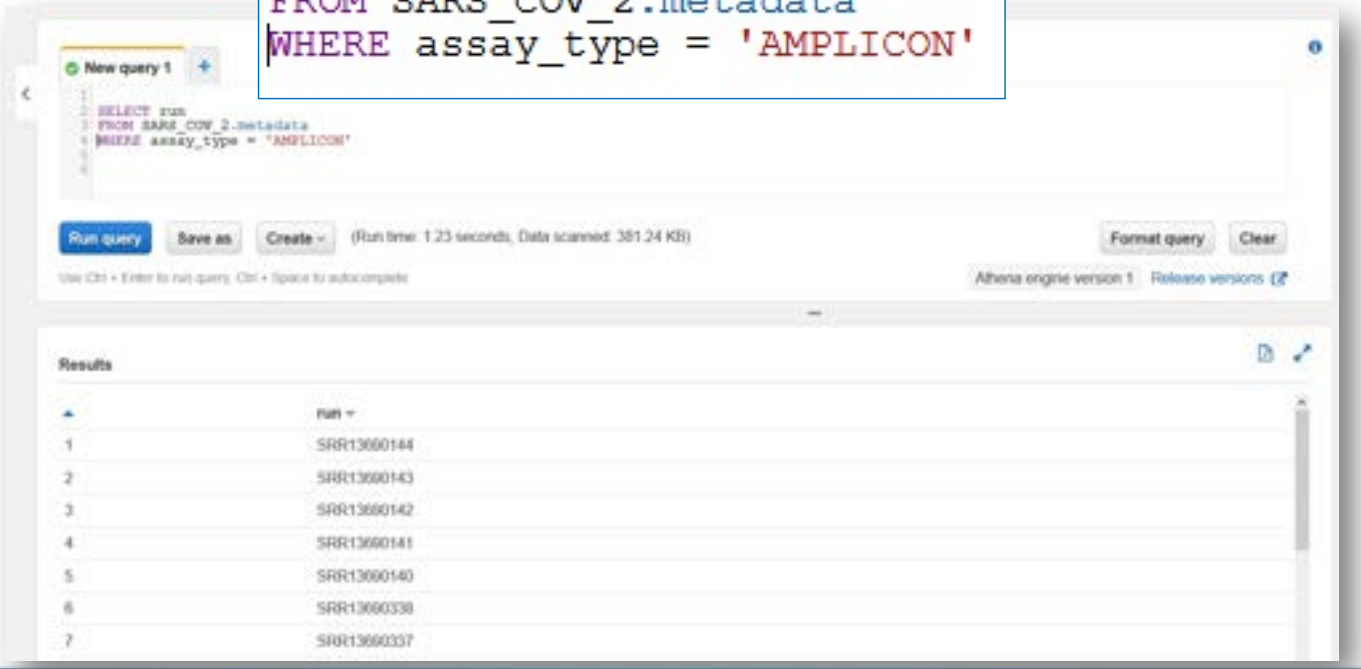
❖Other questions you can address:
- I want Illumina platform data
- I want data released since the start of 2021
- I want data collected before 2020
- I want data submitted by Quest Diagnostics

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'
```

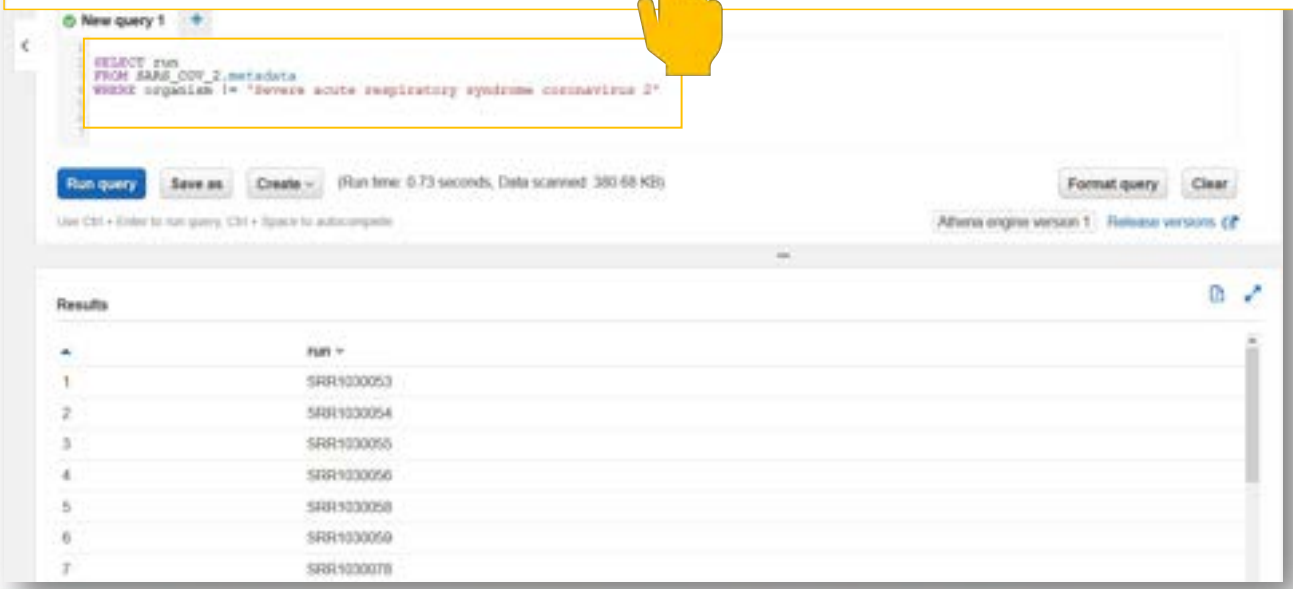


Available Assay Types

- *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'
```

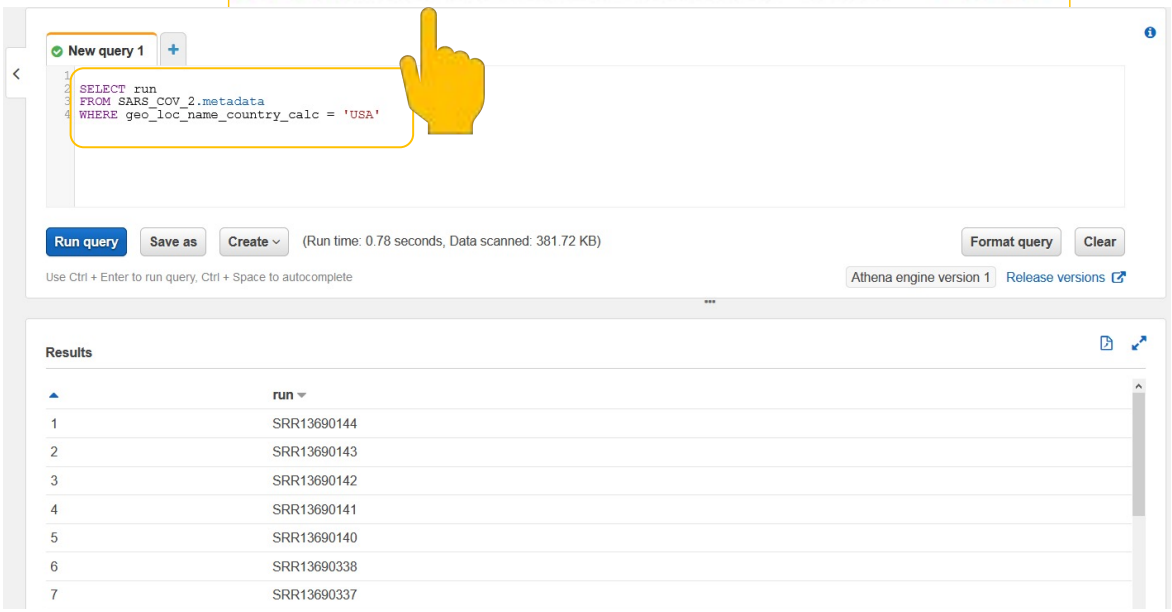


Note '!=' means not equal

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'
```



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:

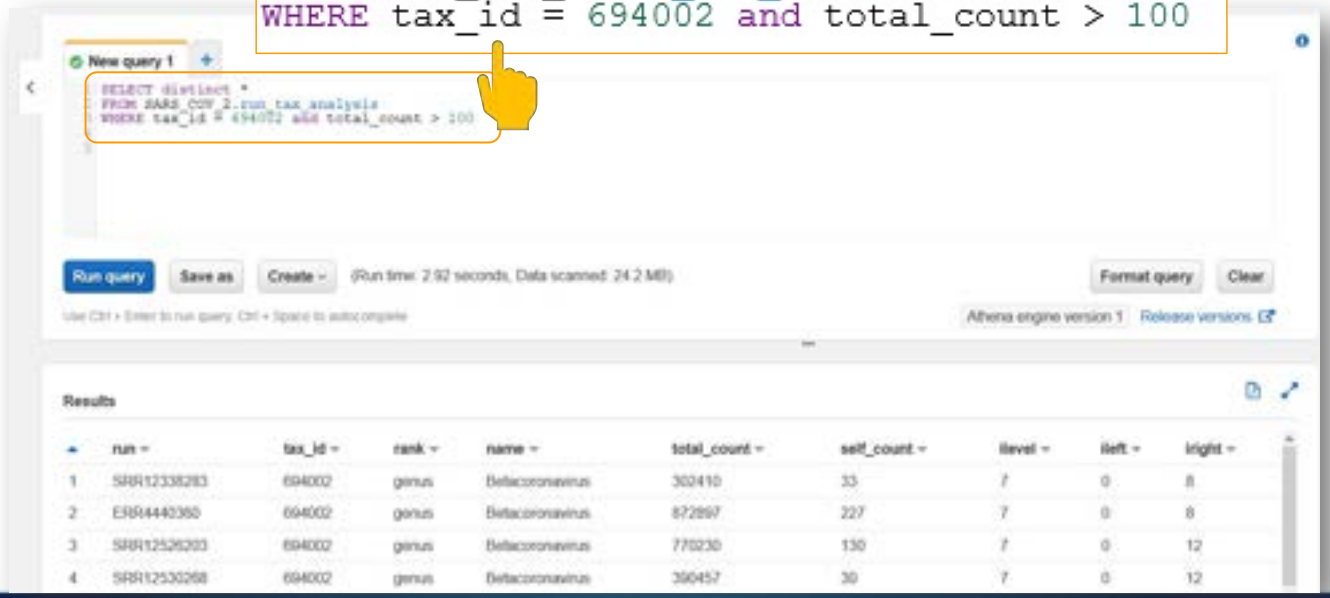
A. STAT Results

- 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

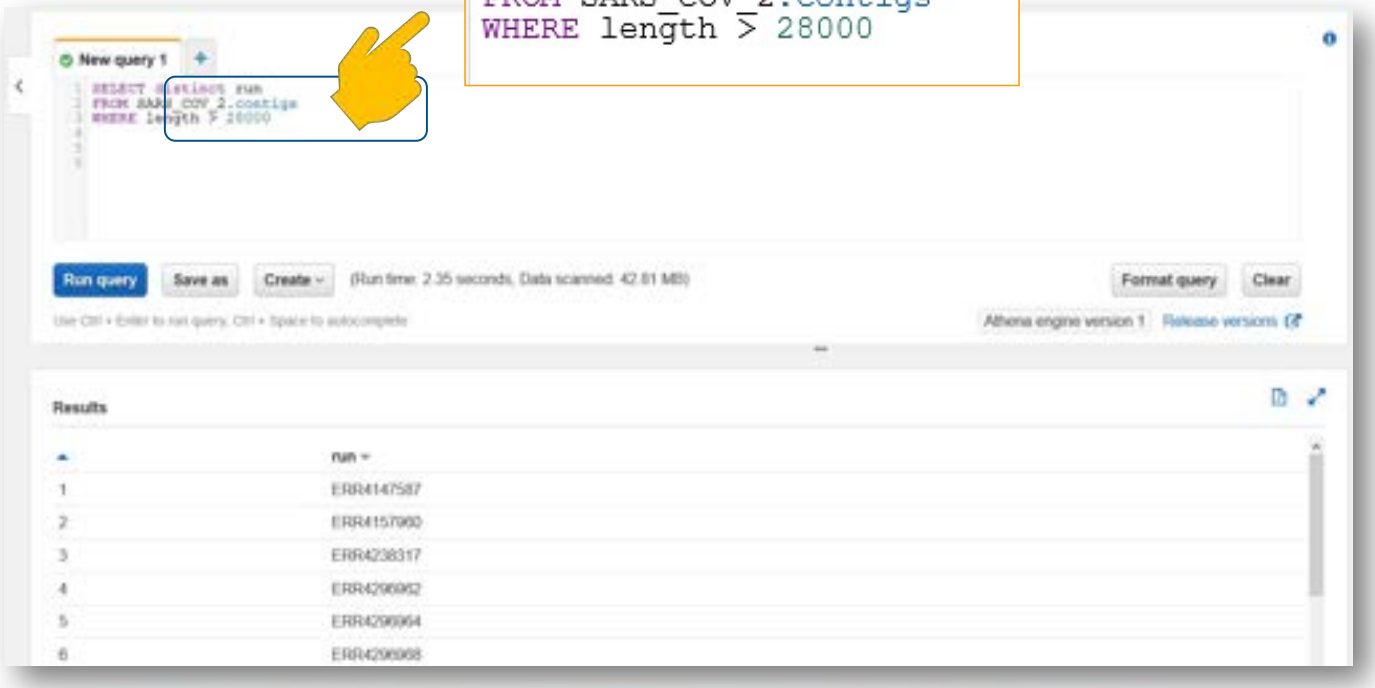
Searching Against NCBI Metadata:

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

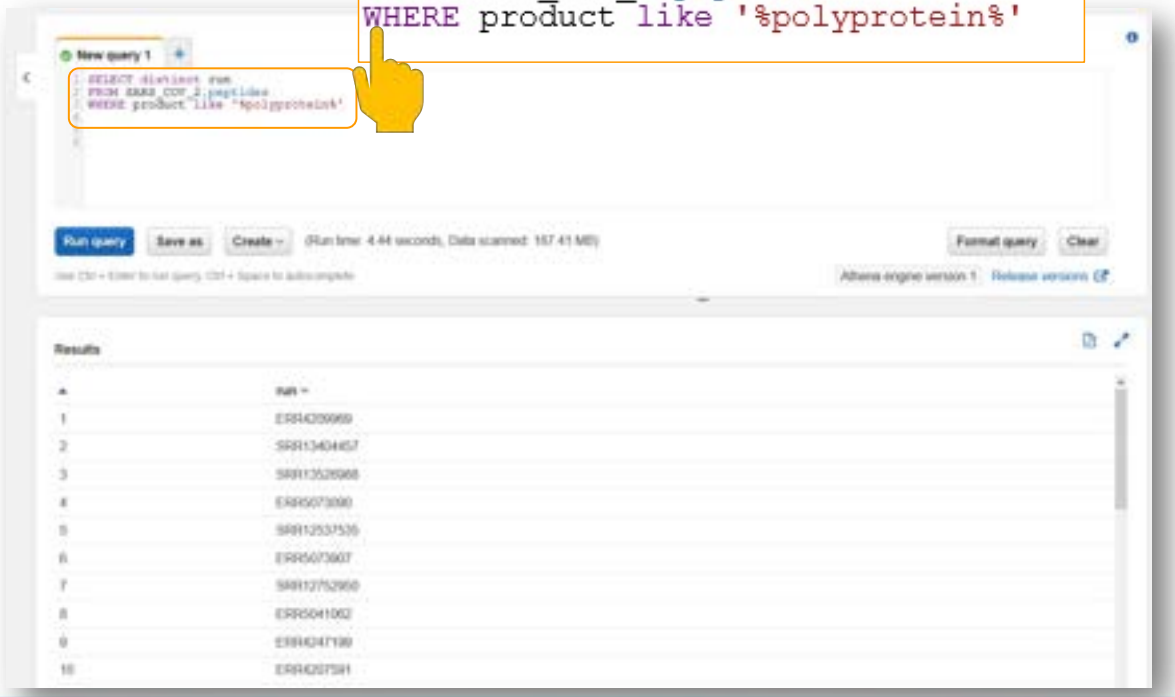
```
SELECT distinct run
FROM SARS_COV_2.contigs
WHERE length > 28000
```



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

Search Against Peptide Products

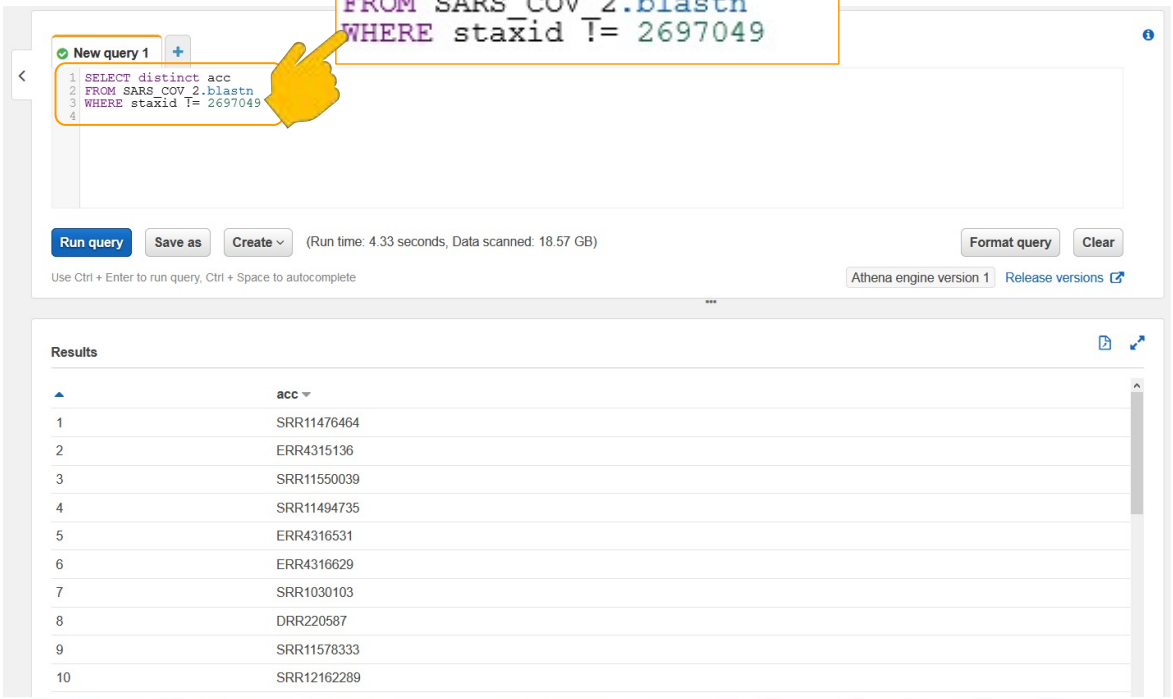
```
SELECT distinct run
FROM SARS_COV_2.peptides
WHERE product like '%polyprotein%'
```



- 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

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



- 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

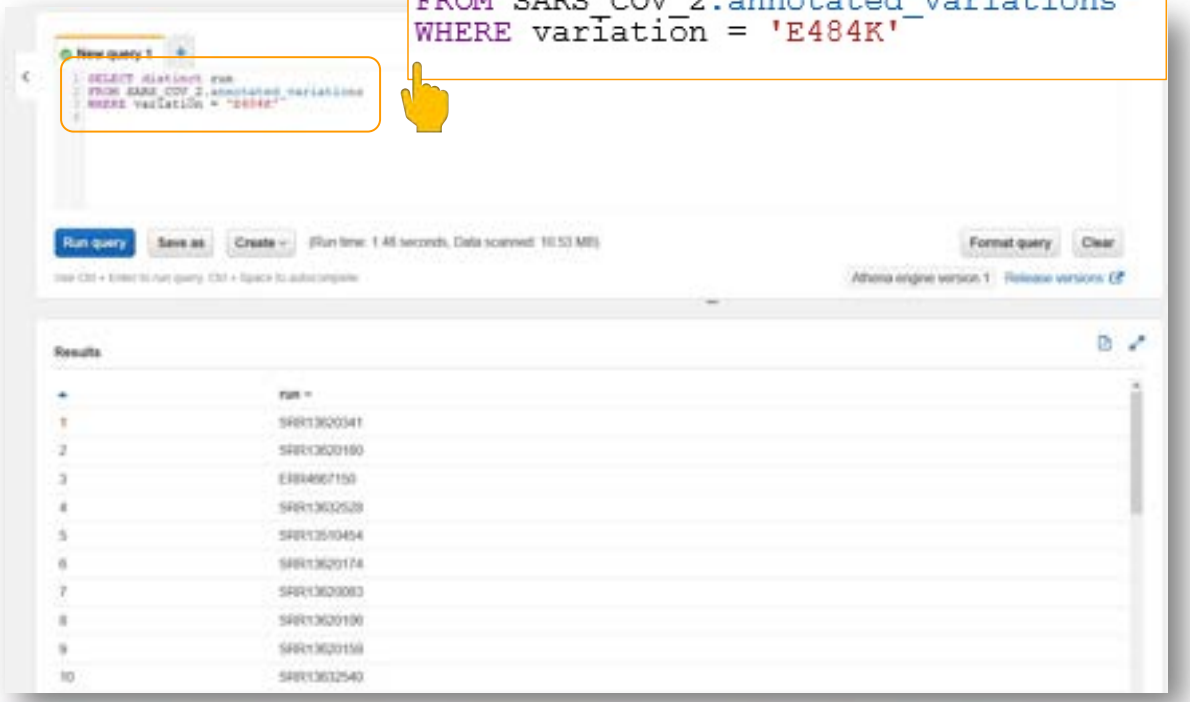
Search Against NCBI Metadata:

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

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



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/RAO/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!

Thank You
Merci
Obrigado
Dyakuyu
Dank Je
Kiitos
Grazie
Salamat
Arigato
Spasibo
Vinaka
Asante
Multumesc
Mochchakkeram
Dankon
Terma Kasih
Kia Ora
Chokrane
Maake
Cam on ban
Raibh Maith Agat
Matur Nuwun
Welalin
Ua Tsaug Rau Koj
Dank Je
Obrigado
Mamana
Juspaxar
Kiitos
Spasibo
Kiitos
Grazie
Asante
Merci
Maake
Mochchakkeram
Obrigado
Dank Je
Kiitos
Multumesc
Kiitos
Grazie
Asante
Matur Nuwun
Raibh Maith Agat
Dank Je
Spasibo
Mochchakkeram
Obrigado
Kiitos
Maake
Kiitos
Obrigado
Kia Ora
Chokrane
Nirringrazzjak
Asante
Maake
Mamana
Multumesc
Dank Je
Mochchakkeram
Dank Je
Spasibo
Raibh Maith Agat
Kiitos
Grazie
Kiitos
Obrigado
Dankon
Dank Je
Mochchakkeram
Grazie
Spasibo
Obrigado
Multumesc
Kiitos
Chokrane
Welalin
Asante
Spasibo
Mamana
Kiitos
Terma Kasih
Obrigado
Kiitos
Obrigado
Grazie
Kia Ora
Chokrane
Dank Je
Kiitos
Merci
Cam on ban
Raibh Maith Agat
Asante
Merci
Mochchakkeram
Obrigado
Dank Je
Raibh Maith Agat
Spasibo
Matur Nuwun
Chokrane
Spasibo
Salamat
Obrigado
Dank Je
Mochchakkeram
Obrigado
Kiitos
Grazie
Kiitos
Dank Je
Mamana
Mochchakkeram
Dank Je
Salamat