

For today's exercises & your own copy of the
slides, please visit:

https://bit.ly/Feb8_Cloud4Virologists

Introduction to NCBI Cloud Computing for Virologists

Cooper J. Park, PhD



National Library of Medicine
National Center for Biotechnology Information

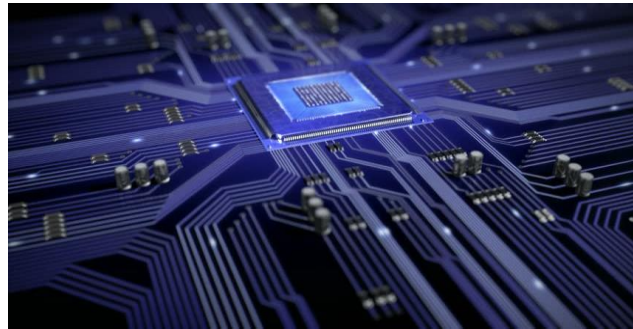
Outline

- **What is the Cloud**
- **Objective 0** - Logging In & Creating an S3 bucket
- **Today's Story**
- **Objective 1** – Consensus Sequences from SRA reads using EC2 instances
- **Objective 2** – Search SRA metadata using Athena
- **Objective 3** – Visualize Sequence Alignments using the NCBI Sequence Viewer
- **Wrap up & Billing**



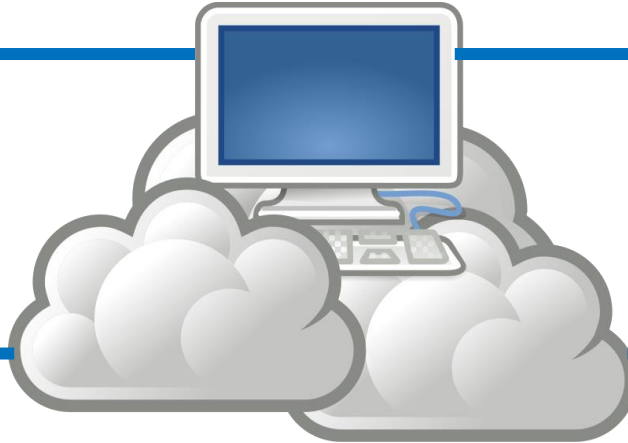
What is “The Cloud”

A “one-stop shop” for high-demand computing services delivered across the internet



Compute Power

“The Cloud”

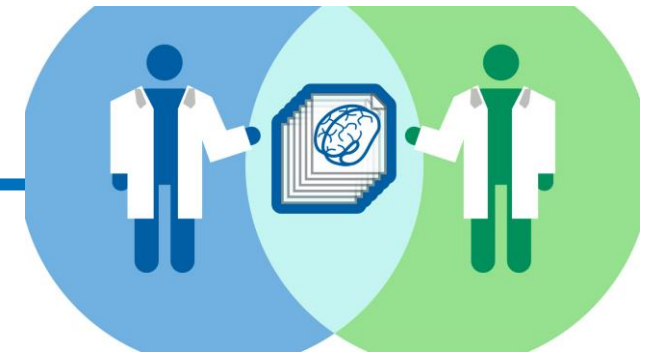


File Storage



Database Management

AND MORE!



Data Sharing

POLL!

Which aspect of your own computational research slows your progress down

Reasons to use the cloud

1) Cost

- Pay only for what you use
- Often cheaper than managing your own infrastructure

2) Global Access

- Data can be shared and accessed seamlessly on a global scale

3) Speed and Performance

- Resources can be optimized for specific needs
- Workflows can be scaled to meet demand
- New technologies/services constantly developed and immediately available

4) Reproducibility, Security, and Reliability

- Easily back-up, protect, version control and recover crucial data
- Computing environments can be saved with 3rd party tools to replicate workflows

Meet your commercial cloud providers



Google Cloud



NCBI and the Cloud

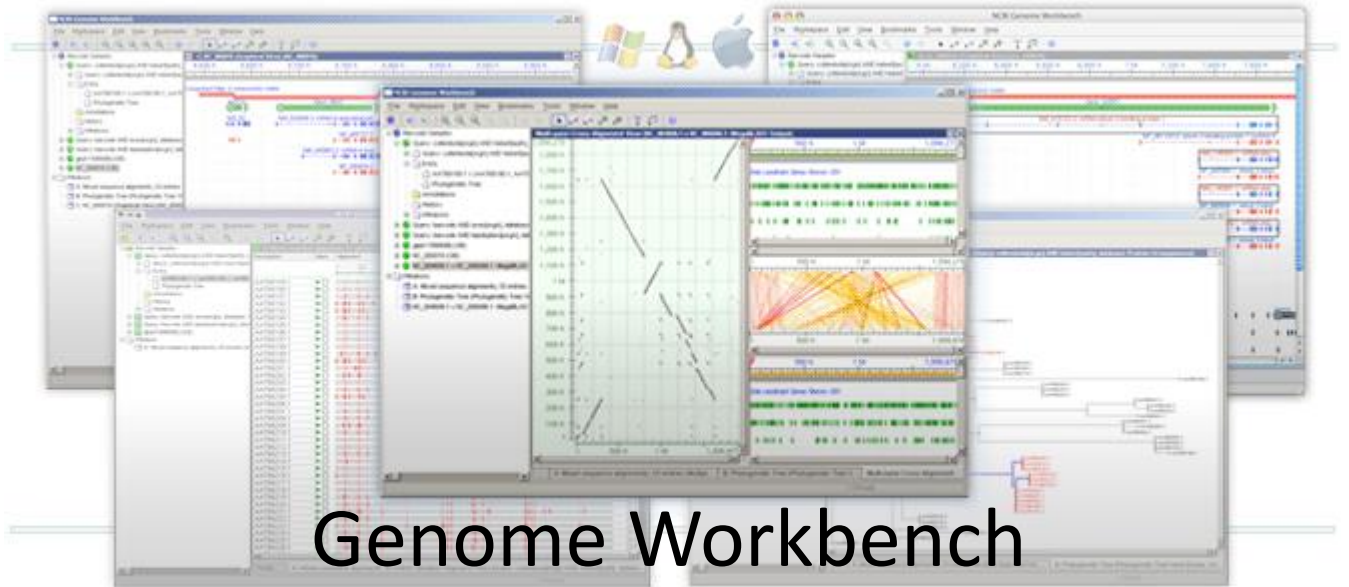


SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.



docker



Genome Workbench

Objective 0 – Logging in to the AWS Console Page & Creating an S3 bucket

S3 Bucket (aka: “Storage”)

- S3 buckets are the “hard drive” of your cloud computer
- Designed for long term storage of files and easy sharing
- Pay for what you use
 - Price increases with storage size/duration and data transfer rates
 - Today’s S3 is **free!**



Login Walkthrough

<https://codeathon.ncbi.nlm.nih.gov>

Username: “Email Prefix” (everything after the “@”)

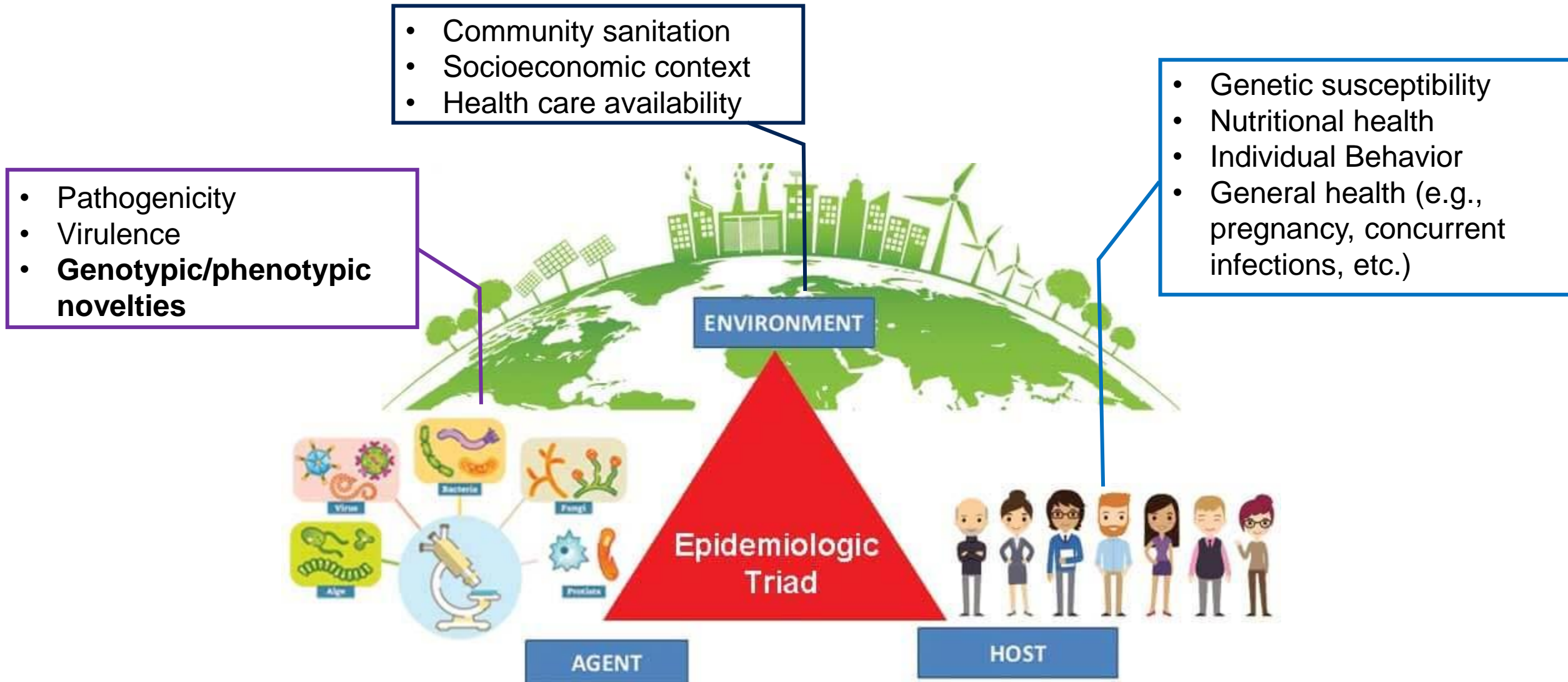
Password: <See the chatbox>

Full Documentation at: https://bit.ly/Feb8_Cloud4Virologists

Outline

- What is the Cloud
- Objective 0 - Logging In & Creating an S3 bucket
- **Today's Story**
- **Objective 1** – Consensus Sequences from SRA reads using EC2 instances
- **Objective 2** – Search SRA metadata using Athena
- **Objective 3** – Visualize Sequence Alignments using the NCBI Sequence Viewer
- **Wrap up & Billing**

Case Study: Genomic Epidemiology



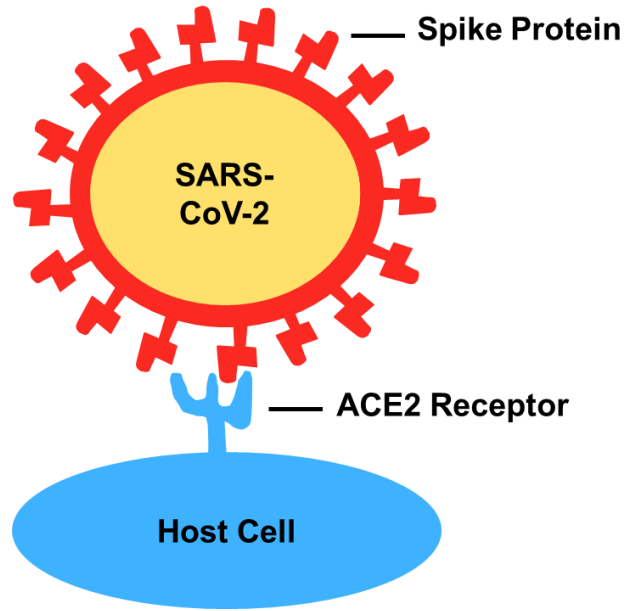
Case Study: Sars-CoV-2 Pandemic

- Daily releases of new sequences to:
 - SRA (raw reads)
 - Genbank (assembled)
- RefSeq Reference Sequence of Wuhan strain
- Easy access via web, command line & cloud

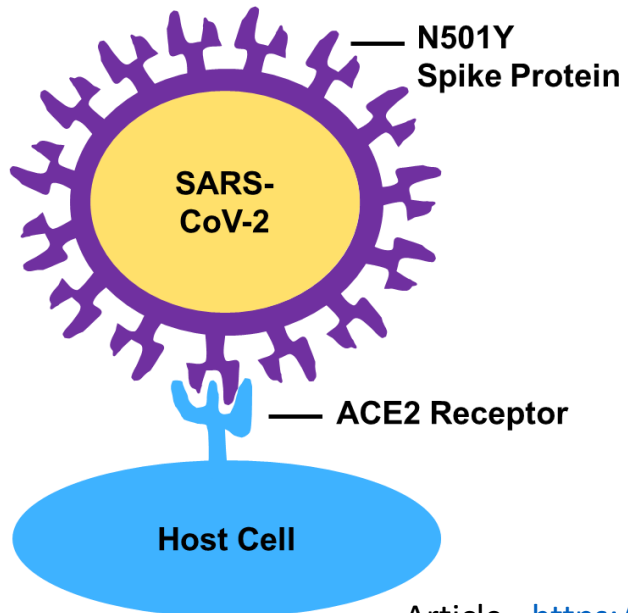


Today's Ultimate Goal: Identify novel mutations in a modern Sars-CoV-2 infection compared to the traditional "Wuhan" reference strain

Case Study: N501Y Mutation



WUHAN



N501Y



Case Study: Our Objectives

Objective 1 - Build a consensus sequence from SRA reads and align to reference genome using AWS EC2 instances

Objective 2 - Search SRA metadata using Athena

Objective 3 - Visualize Sequence Alignments using the NCBI Sequence Viewer

Objective 1 – Consensus Sequences from SRA reads using EC2 instances



National Library of Medicine
National Center for Biotechnology Information

EC2 instance (aka: “Remote Computer”)

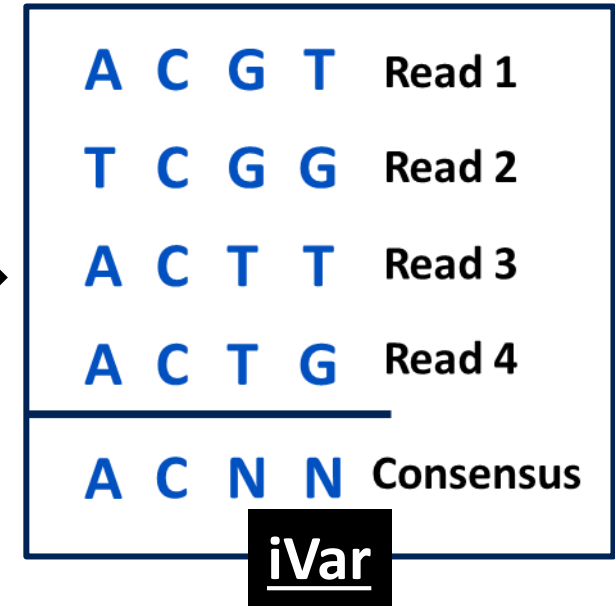
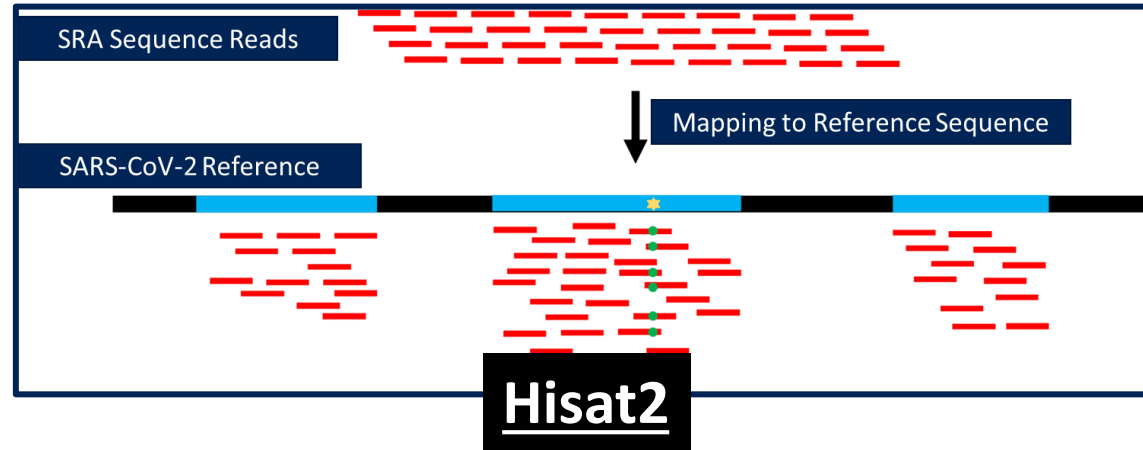
- EC2 instances basically “remote computers”
 - Install software, perform data analyses, manage other AWS services using AWS CLI
- Lots of different customization options including OS, hard drive space, and memory
- Pay for what you use
 - Price increases with larger hardware needs and longer runtime
 - Today’s EC2 is roughly **\$0.20/hour/person**
 - Turn it off when not in use!



Building an EC2 Instance Walkthrough

Analysis Pipeline

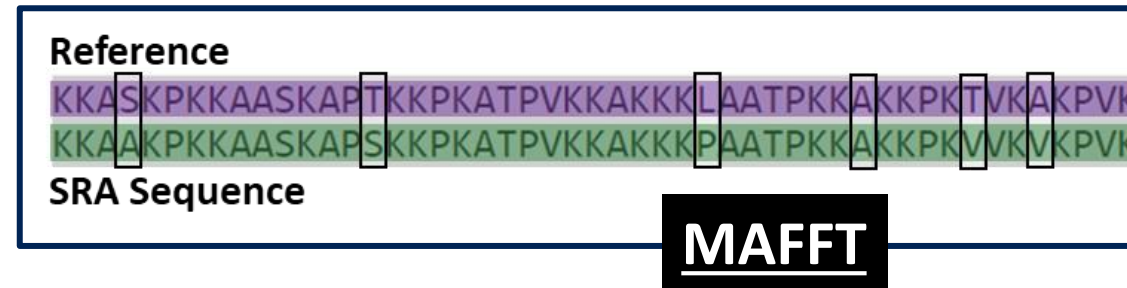
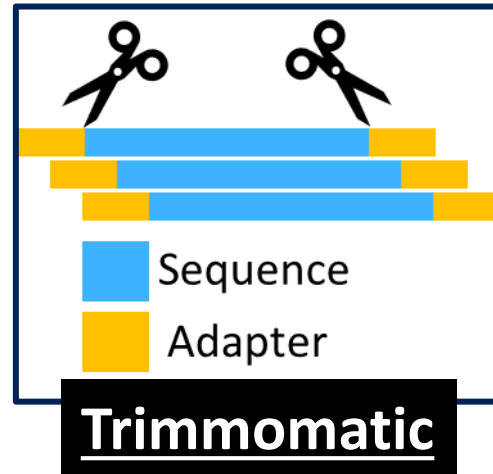
NCBI
Reference
Sequence



SRA
Toolkit



SRA
Sequence
Reads



Supporting Software

- Samtools
 - <http://www.htslib.org/doc/>
 - Manipulate Hisat2 files into formats usable by iVar
- A mazon Web Service Command Line Interface
 - <https://docs.aws.amazon.com/cli/index.html>
 - Moving data between EC2 and S3



Objective 2 - Goals

Computational:

- Create, customize, and manage an EC2 instance
- Align sequence reads, generate a consensus sequence, and align genomes
- Upload files from your remote instance to your S3 bucket

Case Study:

- Identify novel mutations in our recently sequenced genome compared to the traditional Wuhan strain.

POLL!

How familiar are you with using a Unix
command line (aka: terminal)?

EC2 Data Analysis Walkthrough

Objective 2 – Search SRA metadata using Athena

What is the Sequence Read Archive

<https://www.ncbi.nlm.nih.gov/sra>

- Collection of user-submitted nucleotide sequencing reads, most of which are publicly available to download
 - Current size = >10 petabytes
- You can search the data online using the URL above, or by using AWS Athena



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

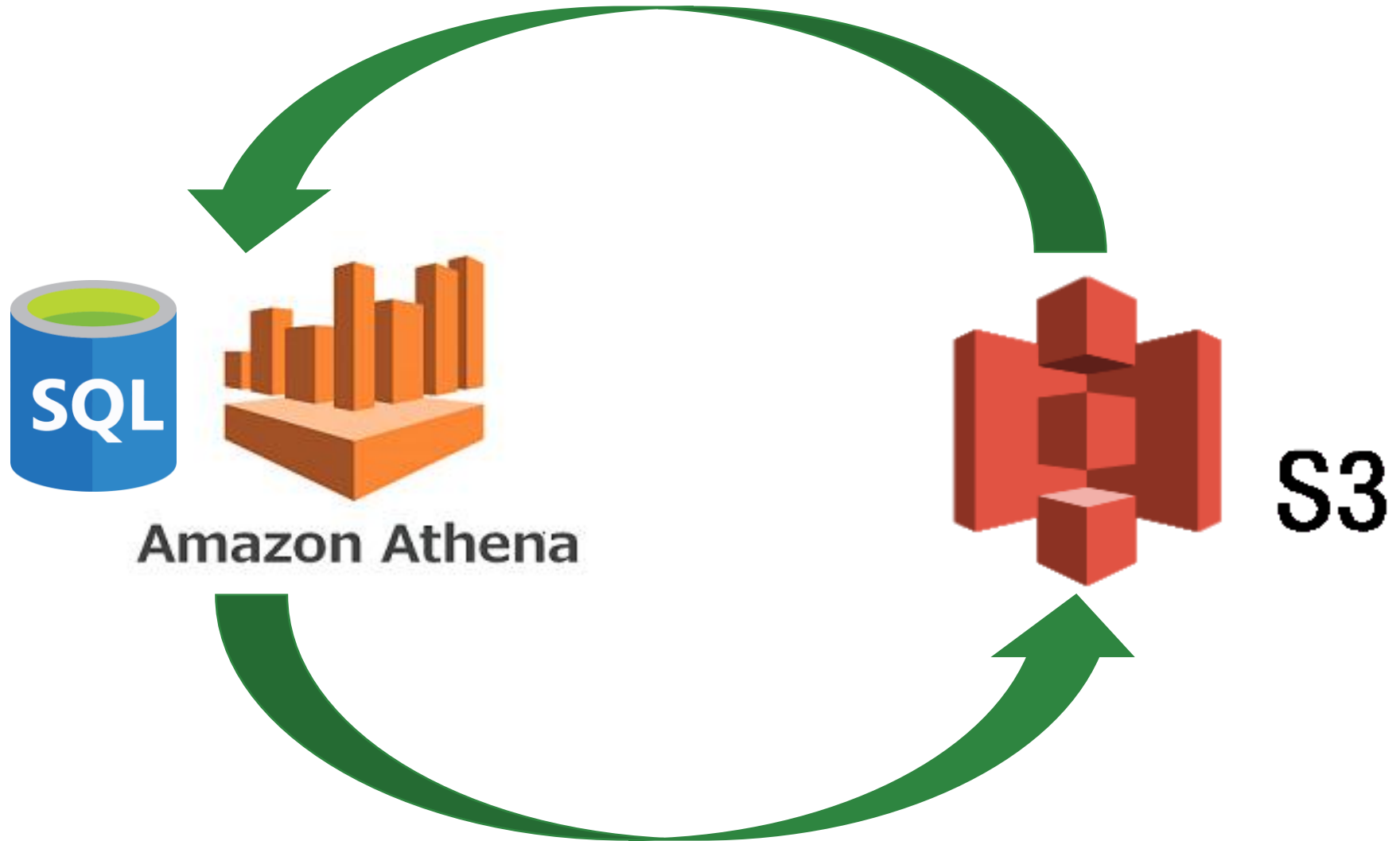
AWS Athena

- AWS data-table querying platform designed to rapidly query large tables of data using the SQL language
- NCBI offers all SRA read metadata as a table we can import into Athena
 - We can query the metadata with Athena to pull out only useful sequence data to use in our own research
- Results can be saved to an S3 bucket



Amazon Athena

**Import results and mine
data in table format**



**Store data mining results
and save useful queries**

Objective 2 - Goals

Computational

- Use basic SQL commands to query Athena data tables
- Save query results to personal computer and an S3 bucket

Case Study

- Find sequence data & metadata associated with our sequence reads

Athena Setup Walkthrough

SQL programming language basics

```
SELECT *  
FROM "sra"."metadata"  
WHERE assay_type = 'WGS'  
LIMIT 50
```

“Give me all of the columns in the table back”

Choose the table columns you want to see for each hit from the table

Choose which table of data you are querying against

Choose the columns you want to filter the data by

Restrict the results to a given number of rows

Database

sra

Filter tables and views...

▼ **Tables (1)** Create table

▶ metadata

SELECT *

FROM "sra"."metadata"

WHERE assay_type = 'WGS'

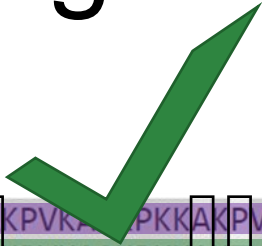
acc	assay_type	center_name	consent	experiment	sample_name	instrument	librarylayout	libraryselectic
1 ERR2867935	WGS	DFDONG	public	ERX2873895	SAMEA5065299	Illumina HiSeq 2000	SINGLE	RANDOM
2 ERR351333	RNA-Seq	IGA Technology Services	public	ERX324170	SAMEA2220074	Illumina HiSeq 2000	SINGLE	other
3 ERR2867821	WGS	DFDONG	public	ERX2873781	SAMEA5065185	Illumina HiSeq 2000	SINGLE	RANDOM
4 ERR1995299	WGS	BEIJING GENOME INSTITUTE	public	ERX2055168	SAMEA104062412	Illumina HiSeq 2000	SINGLE	other
5 ERR358180	RNA-Seq	Genomic Technolgies Core Facility, Faculty of Life Sciences, University of Manchester	public	ERX330954	SAMEA2225912	AB SOLiD 4 System	SINGLE	cDNA
6 ERR2017761	WGS	BEIJING GENOME INSTITUTE	public	ERX2077343	SAMEA104142420	Illumina HiSeq 2000	PAIRED	other
7 ERR2017592	WGS	BEIJING GENOME INSTITUTE	public	ERX2077174	SAMEA104142099	Illumina HiSeq 2000	PAIRED	other
8 SRR8741520	RNA-Seq	LANZHOU UNIVERSITY	public	SRX5533654	Ppr-NaCl-24-2	Illumina HiSeq 2000	PAIRED	PolyA
9 ERR589275	RNA-Seq	Boehringer Ingelheim Pharma	public	ERX547266	SAMEA2735922	Illumina HiSeq 2000	SINGLE	RANDOM
10 SRR13123516	RNA-Seq	NANKAI UNIVERSITY	public	SRX9565550	EF_CL3	Illumina NovaSeq 6000	PAIRED	other

Athena Queries Walkthrough

Objective 3 – Visualize Sequence Alignments using the NCBI Sequence Viewer

Case Study - Using the sequences

Align Sequences



Reference `KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKAPKKAAPKVK`
Our Sequence `KKAAKPKKAASKAPSKKPKATPVKKAKKKLPAATPKKAKKPKVVKV KPVKASKPKKAKTVK`



S3

Visualize Alignment

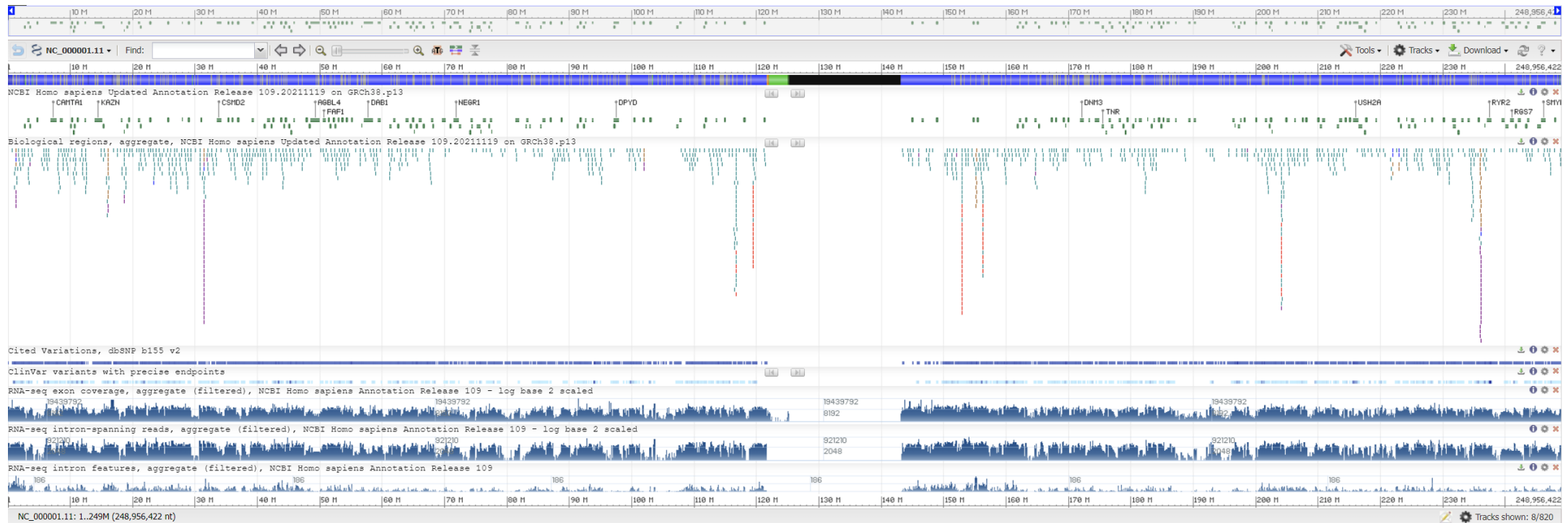


NCBI
Sequence
Viewer



NCBI Sequence Viewer - 1

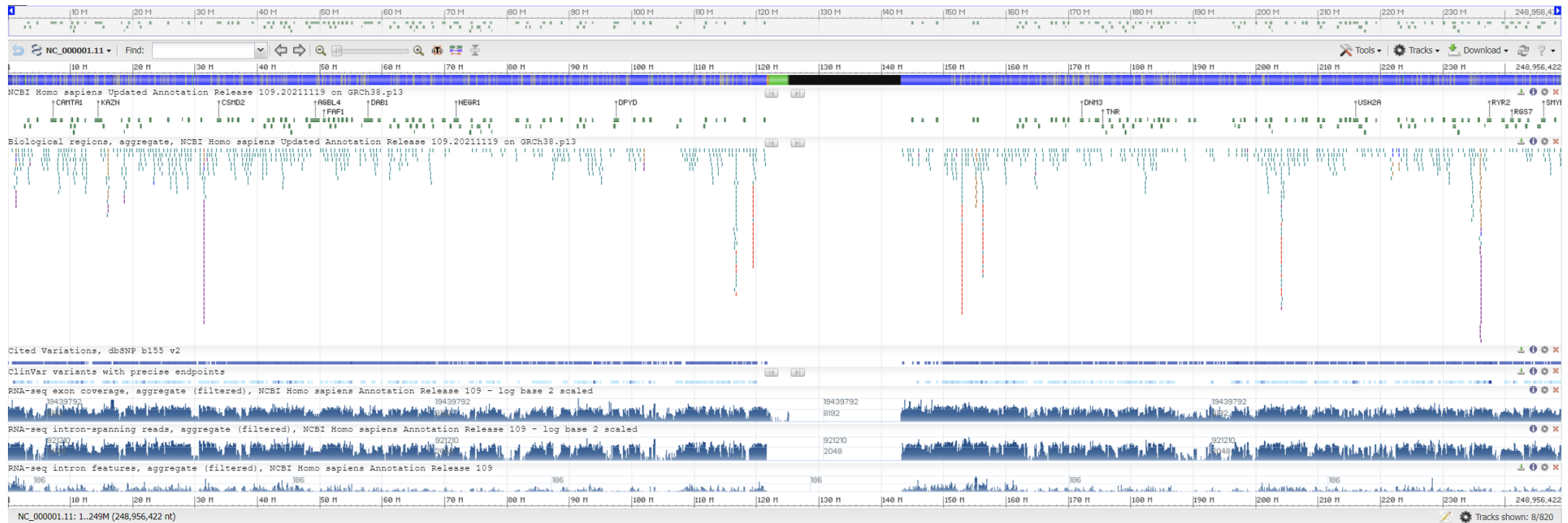
Sequence Viewer is a web tool which supports the visualization of genetic data mapped against any genetic sequence.



NCBI Sequence Viewer - 2

Data is visualized in “Tracks”

- Can include sequence annotations, coverage graphs, GWAS data, alignment data, and more!



Objective 3 - Goals

Computational:

- Access and navigate Sequence Viewer
- Upload custom data track to SV
- Parse biological meaning from alignment results

Case Study:

- Identify variations between the “Wuhan” strain and our own sequence
- Compare identified variations to known variations to identify potential pathogenic benefits

Sequence Viewer Walkthrough

Billing

- The most important question in cloud computing...

“How Much Will This Cost Me?”



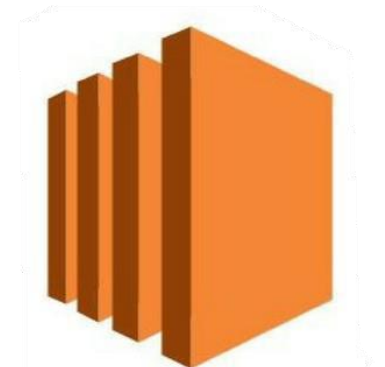
Amazon Athena



Amazon Glue



S3



Amazon EC2

POLL!

How much do you think today's workshop
cost per person?

Billing

- The most important question in cloud computing...

“How Much Will This Cost Me?”

Everything you did in this workshop cost ~\$0.50



Amazon Athena



Amazon Glue



S3



Amazon EC2

Billing

- AWS strives to be transparent about costs
 - <https://calculator.aws/#/estimate> - Build a price estimate based on anticipated service usage
 - <https://aws.amazon.com/free/> - View free-tier uses on most AWS services
- Several tools such as Cost Explorer can help you break down usage across a group

From Introduction to Intermediate & Beyond!



AWS Batch can automate the distribution of work across multiple EC2 instances



AWS Lambda can automate code execution without managing hardware needs



AWS CLI can manage all AWS products from a computer terminal, automating any step of a process

How do I get an account to try this stuff out?

Initial free accounts can be gotten from:

- AWS: <https://aws.amazon.com/free/>
- GCP: <https://cloud.google.com/free/>
- Microsoft Azure: <https://azure.microsoft.com/en-us/free/>

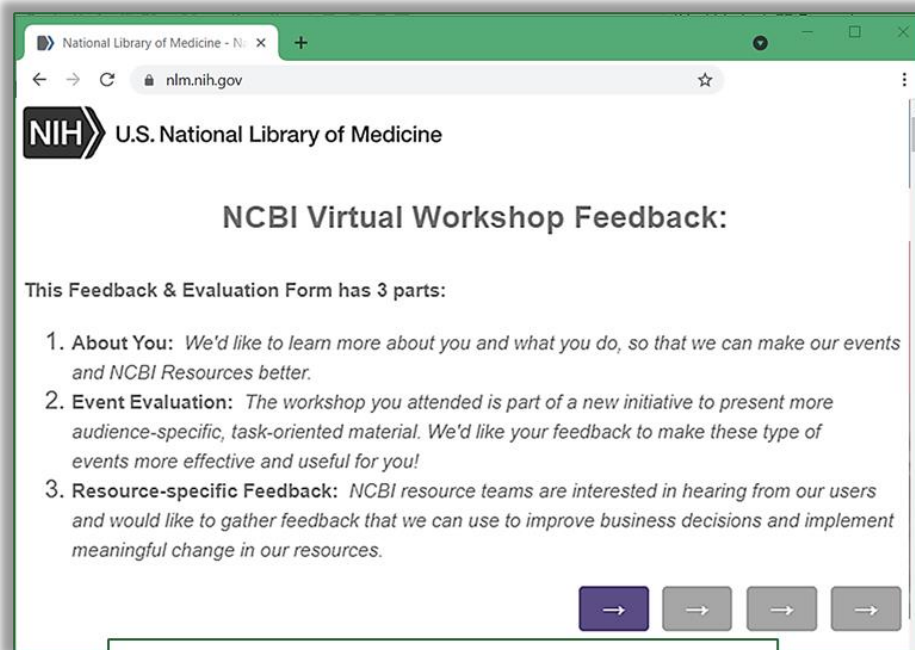
For NIH-funded research teams, you can also get help from the NIH Strides program: <https://datascience.nih.gov/strides/>

What's next?

At the end of the workshop, you'll see a Feedback pop-up to let us know what you thought of today's event.

You'll receive a **follow-up Email** with a link to the **Feedback survey** as well as a **workshop webpage** with information about this event and links to materials and the video recording.

This page will also be linked from the **NCBI Outreach Events page for this workshop**, so you'll be able to find it later.



National Library of Medicine - NLM

nlm.nih.gov

NIH U.S. National Library of Medicine

NCBI Virtual Workshop Feedback:

This Feedback & Evaluation Form has 3 parts:

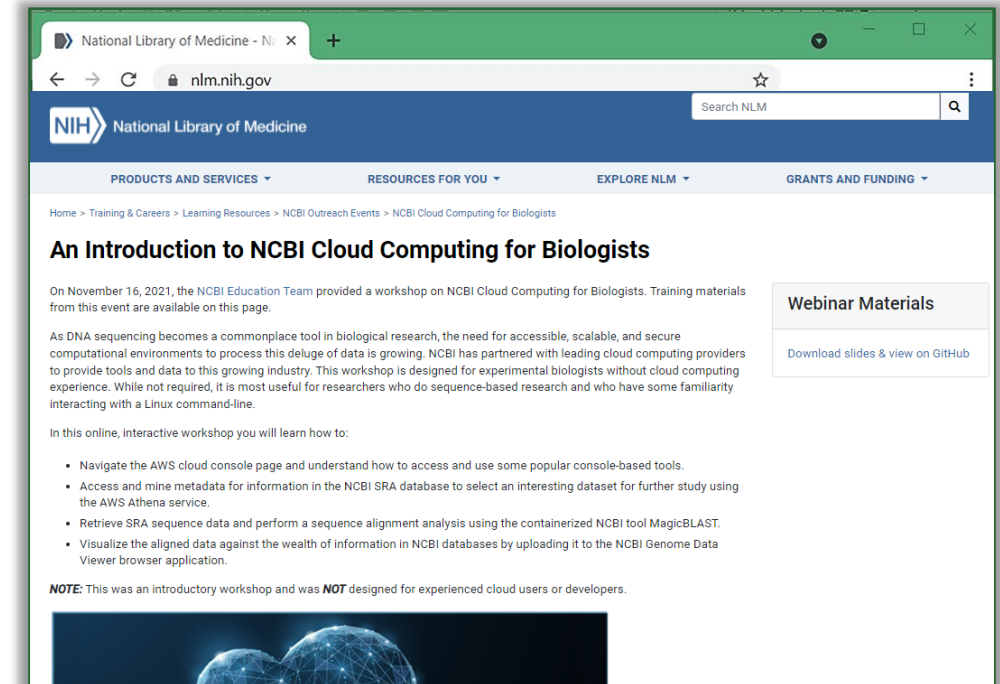
- 1. About You:** We'd like to learn more about you and what you do, so that we can make our events and NCBI Resources better.
- 2. Event Evaluation:** The workshop you attended is part of a new initiative to present more audience-specific, task-oriented material. We'd like your feedback to make these type of events more effective and useful for you!
- 3. Resource-specific Feedback:** NCBI resource teams are interested in hearing from our users and would like to gather feedback that we can use to improve business decisions and implement meaningful change in our resources.

→ → → →

Thank you for providing your feedback!

To get your eCertificate, please click this link and add your name - then you can download and/or print out your certificate:
<https://www.nlm.nih.gov/>

If you would like to provide additional input on NCBI resources you encountered at this event, please write to us at workshops@ncbi.nlm.nih.gov.



National Library of Medicine

Search NLM

PRODUCTS AND SERVICES ▾ RESOURCES FOR YOU ▾ EXPLORE NLM ▾ GRANTS AND FUNDING ▾

Home > Training & Careers > Learning Resources > NCBI Outreach Events > NCBI Cloud Computing for Biologists

An Introduction to NCBI Cloud Computing for Biologists

On November 16, 2021, the NCBI Education Team provided a workshop on NCBI Cloud Computing for Biologists. Training materials from this event are available on this page.

As DNA sequencing becomes a commonplace tool in biological research, the need for accessible, scalable, and secure computational environments to process this deluge of data is growing. NCBI has partnered with leading cloud computing providers to provide tools and data to this growing industry. This workshop is designed for experimental biologists without cloud computing experience. While not required, it is most useful for researchers who do sequence-based research and who have some familiarity interacting with a Linux command-line.

In this online, interactive workshop you will learn how to:

- Navigate the AWS cloud console page and understand how to access and use some popular console-based tools.
- Access and mine metadata for information in the NCBI SRA database to select an interesting dataset for further study using the AWS Athena service.
- Retrieve SRA sequence data and perform a sequence alignment analysis using the containerized NCBI tool MagicBLAST.
- Visualize the aligned data against the wealth of information in NCBI databases by uploading it to the NCBI Genome Data Viewer browser application.

NOTE: This was an introductory workshop and was **NOT** designed for experienced cloud users or developers.

Webinar Materials

Download slides & view on GitHub

HOW CAN YOU KEEP UP WITH IT ALL & LEARN MORE?

@NCBI



NLMNIH



NCBI



NCBIInsights.ncbi.nlm.nih.gov



info@ncbi.nlm.nih.gov

NCBI.NLM

National Center for Biotechnology Information (NCBI)



Using NCBI's Primer-BLAST to design & analyze PCR primers

783

Biological researchers often design specific PCR primers to amplify a single genomic or mRNA template or a set of closely related templates. In addition, PCR amplification with specially designed primers is sometimes used to identify an organism or group of organisms based on targeted RNA or genomic DNA amplification of an isolate. NCBI's Primer-BLAST combines the primer design features of the popular Primer3 package with a specificity

NCBI Outreach Events

Search Upcoming

Keywords Location Select Date Range

Choose a Category Choose an event type

Events

- 14 - 15 OCT Using Web BLAST Effectively
2021-10-14 @ 01:00 PM - 2021-10-14 @ 02:30 PM
Online Event
NCBI Workshop
- 26 - 26 OCT NCBI Resources for Genetic Disease Discovery & Clinical Support
2021-10-26 @ 01:00 PM - 2021-10-26 @ 04:00 PM
Online Event
NCBI Workshop
- 28 - 28 OCT An NCBI Guide to Finding and Analyzing Metagenomic Data
2021-10-28 @ 01:00 PM - 2021-10-28 @ 02:30 PM
Online Event
NCBI Workshop

Thank you!