Introduction to NCBI Cloud Computing for Biologists

Cooper J. Park, PhD; Rana Morris, PhD

NIH National Library of Medicine
National Center for Biotechnology Information

Outline

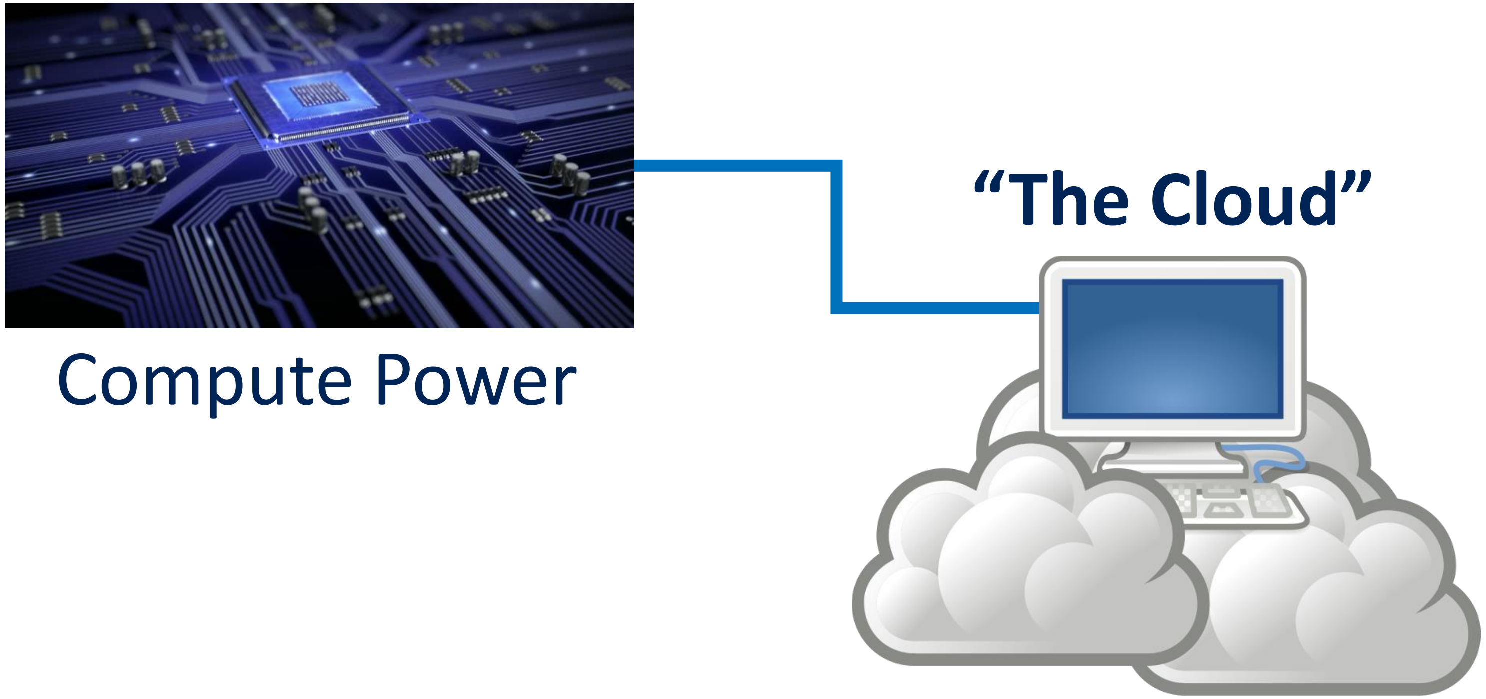
- What is the Cloud
- Objective 0 - Logging In
- Today’s Case Study
- Objective 1 – Navigating the AWS cloud console
- Objective 2 – Mining NCBI’s SRA data
- Objective 3 – Using magicBLAST & Genome Data Viewer in the Cloud
- Wrap up & Billing

What is “The Cloud”

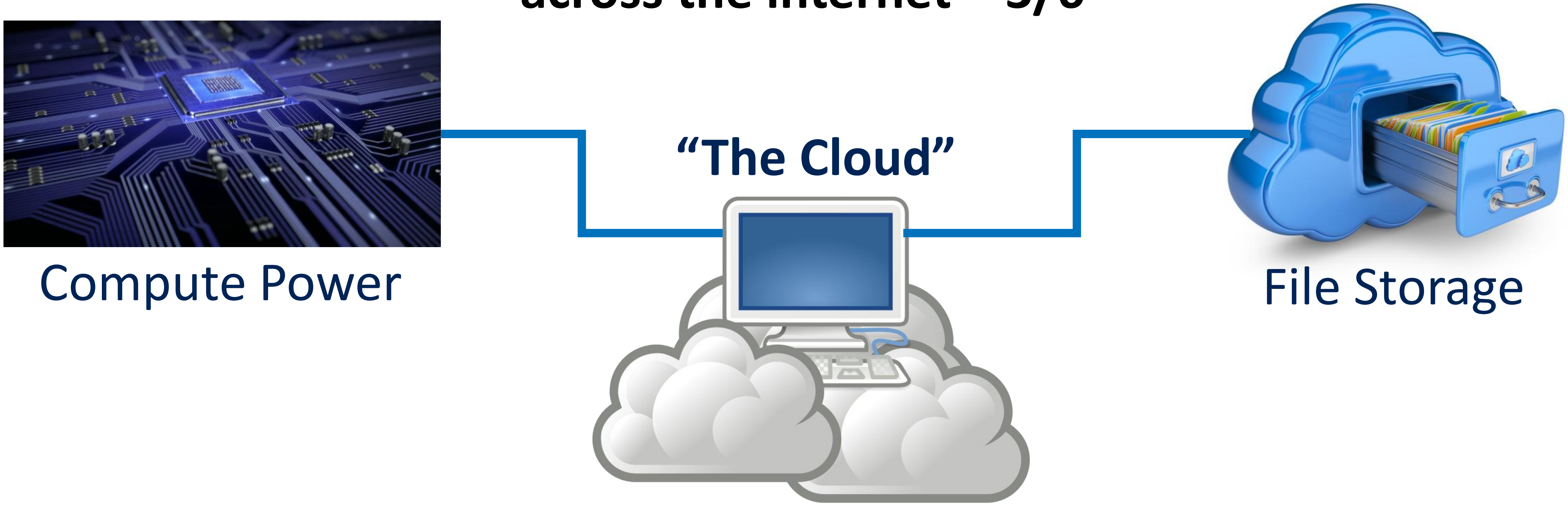
A “one-stop shop” for high-demand computing services delivered across the internet – 1/6

“The Cloud”

A "one-stop shop" for high-demand computing services delivered across the internet – 2/6



A “one-stop shop” for high-demand computing services delivered across the internet – 3/6

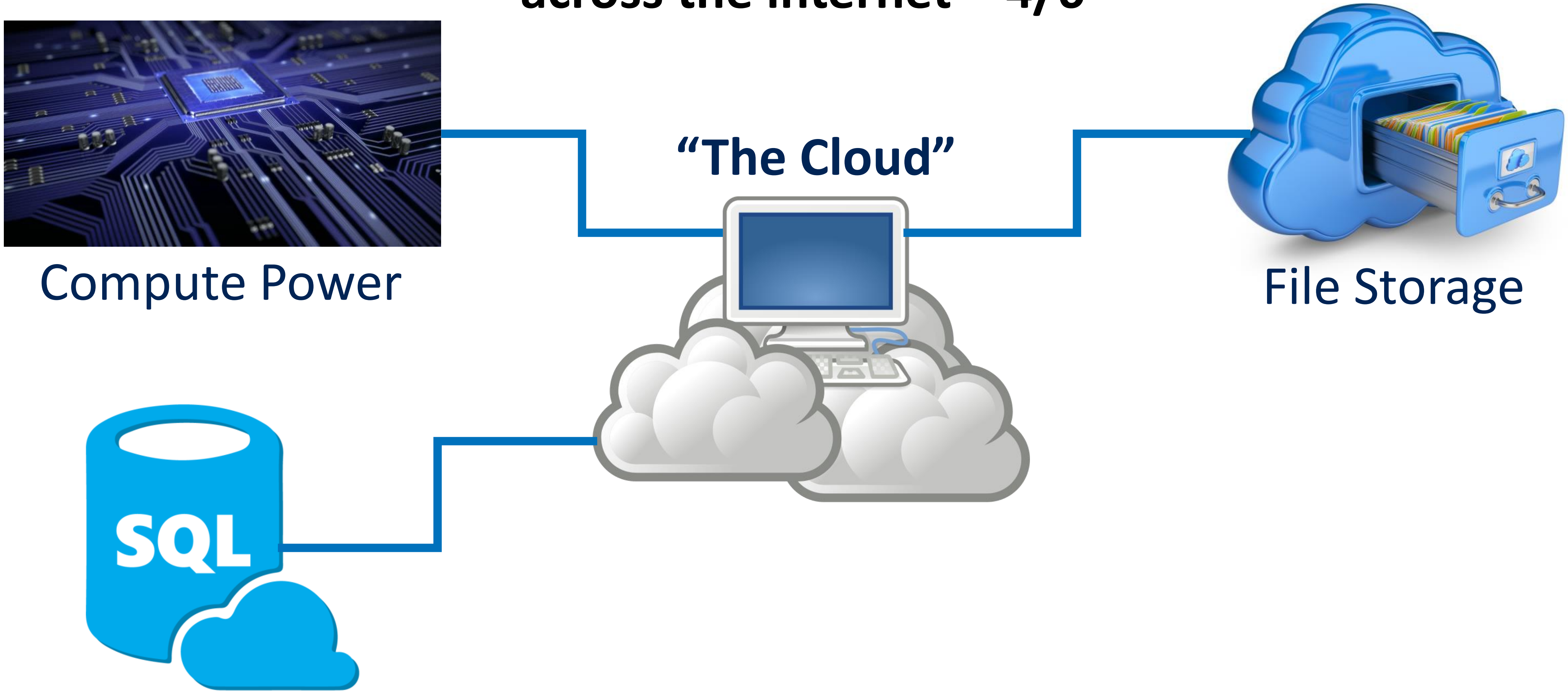


Compute Power

“The Cloud”

File Storage

A “one-stop shop” for high-demand computing services delivered across the internet – 4/6



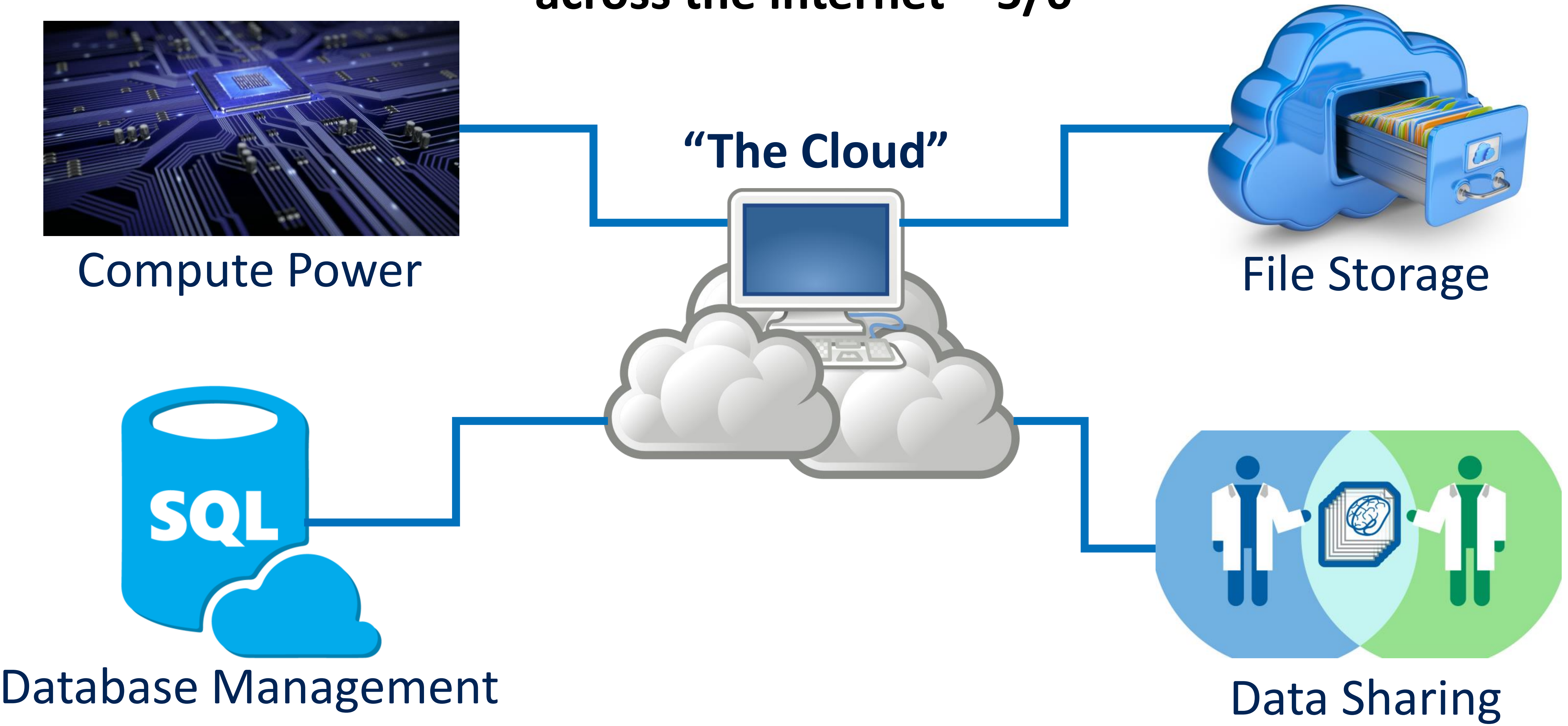
“The Cloud”

Compute Power

File Storage

Database Management

A “one-stop shop” for high-demand computing services delivered across the internet – 5/6



“The Cloud”

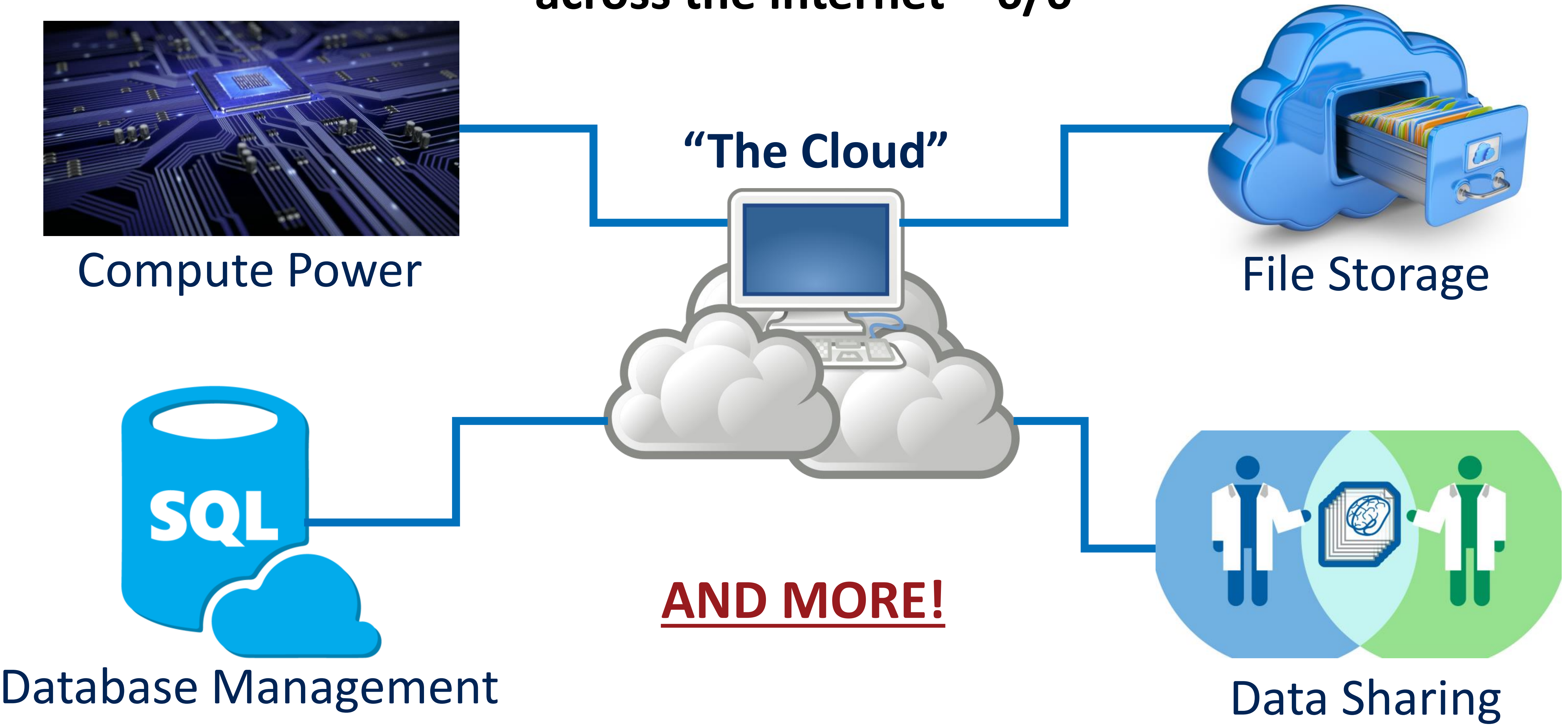
Compute Power

File Storage

Database Management

Data Sharing

A “one-stop shop” for high-demand computing services delivered across the internet – 6/6



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 – 1/5

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

Reasons to use the cloud – 2/5

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

Reasons to use the cloud – 3/5

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

Reasons to use the cloud – 4/5

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

Reasons to use the cloud – 5/5

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 – 1/2

Google Cloud

Microsoft Azure

amazon web services

Meet your commercial cloud providers – 2/2

Google Cloud

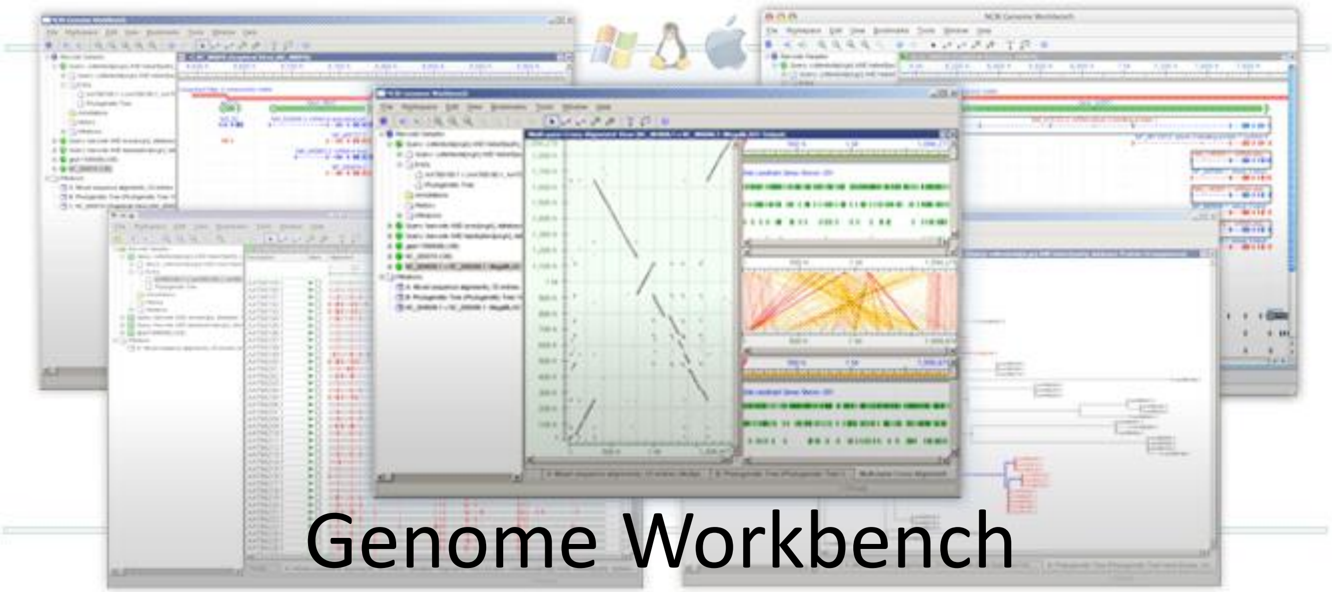
Microsoft Azure

amazon web services

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.



Objective 0 – Logging in & Navigating the AWS Console page

National Library of Medicine
National Center for Biotechnology Information

Login Walkthrough

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

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

Password: <See the chatbox>

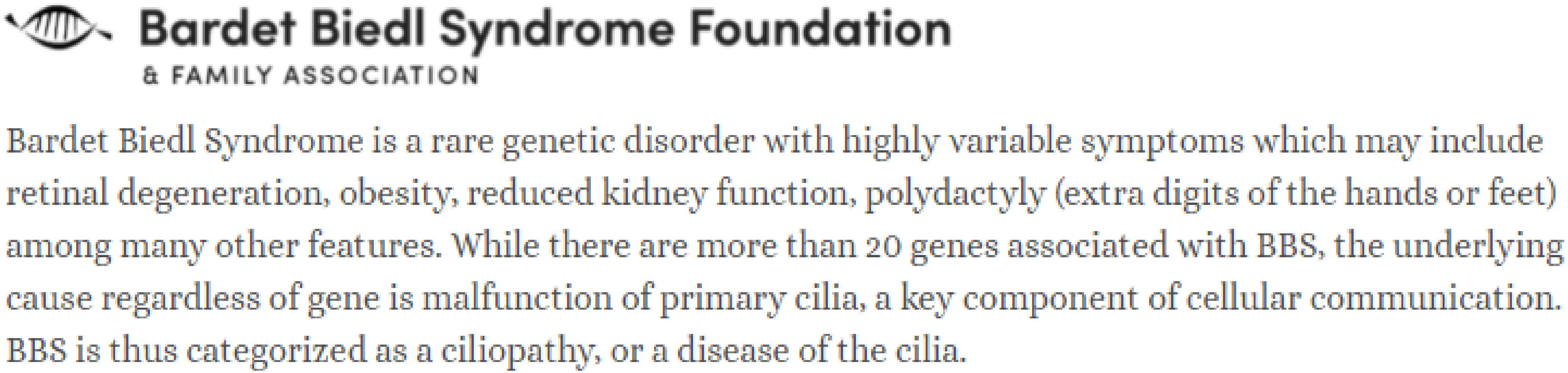
Full Documentation at: https://parkcoj.github.io/Intro-to-NCBI-Cloud-Computing/

Outline

- About NCBI
- What is the Cloud
- Objective 0 - Logging In
- Today’s Case Study
- Objective 1 – Mining SRA metadata using AWS Athena
- Objective 2 – Aligning sequence reads using AWS EC2 & MagicBLAST
- Objective 3 – Visualize read alignment in Genome Data Viewer
- Wrap up & Billing

Case Study – Clinical background – 1/2

- Through years of clinical tests and evaluations, a 3-year-old Guyanese child is diagnosed with Bardet-Biedl syndrome (BBS).



Bardet Biedl Syndrome Foundation
& FAMILY ASSOCIATION

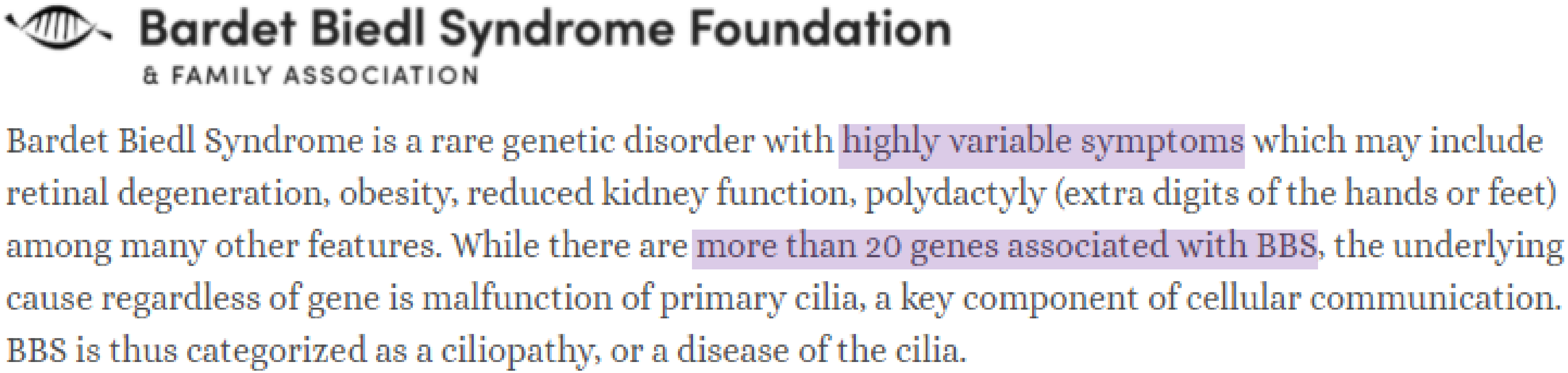
Bardet Biedl Syndrome is a rare genetic disorder with highly variable symptoms which may include retinal degeneration, obesity, reduced kidney function, polydactyly (extra digits of the hands or feet) among many other features. While there are more than 20 genes associated with BBS, the underlying cause regardless of gene is malfunction of primary cilia, a key component of cellular communication. BBS is thus categorized as a ciliopathy, or a disease of the cilia.

NIH National Library of Medicine
National Center for Biotechnology Information

Paper - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778042/

Case Study – Clinical background – 2/2

- Through years of clinical tests and evaluations, a 3-year-old Guyanese child is diagnosed with Bardet-Biedl syndrome (BBS).



Bardet Biedl Syndrome Foundation
& FAMILY ASSOCIATION

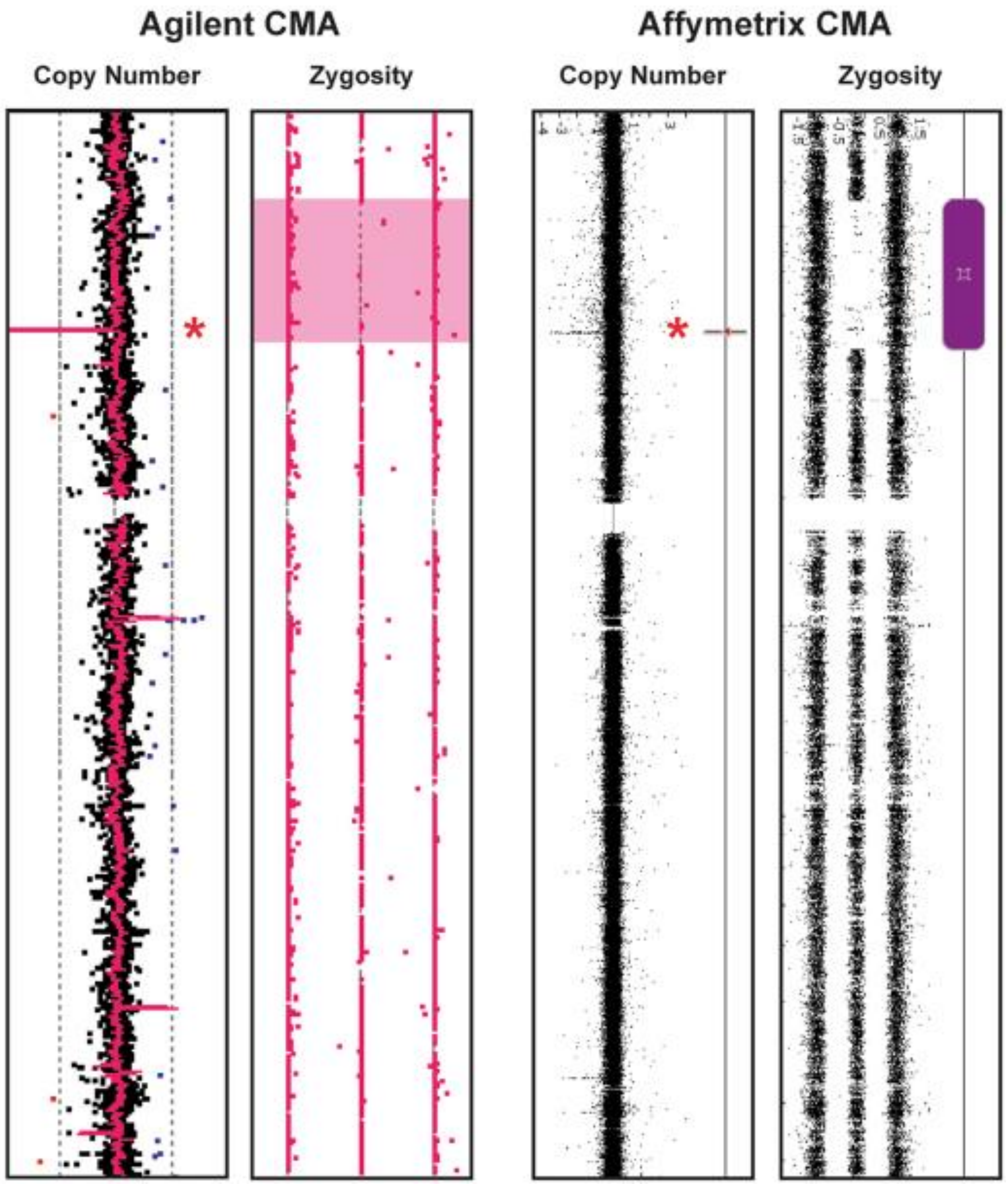
Bardet Biedl Syndrome is a rare genetic disorder with highly variable symptoms which may include retinal degeneration, obesity, reduced kidney function, polydactyly (extra digits of the hands or feet) among many other features. While there are more than 20 genes associated with BBS, the underlying cause regardless of gene is malfunction of primary cilia, a key component of cellular communication. BBS is thus categorized as a ciliopathy, or a disease of the cilia.

- Authors sought to confirm this clinical diagnosis using some newer “long-read” sequencing technologies.

NIH National Library of Medicine
National Center for Biotechnology Information

Paper - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778042/

Case Study cont.



Authors first perform a chromosomal microarray test to identify which of the 20 BBS genes could be affected and **identify a deletion in *BBS9!***

Is this deletion the cause of BBS in the child? To answer this, we need to:

a) confirm whether this deletion is truly present

b) Identify any previously known clinical associations between mutations in this gene and the BBS disorder using NCBI resources

Paper - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778042/

Case Study – Our goals

Objective 1 – Search for the child's sequencing reads from deposited into NCBI's SRA database

Objective 2 - Align the DNA sequences against a template (aka: *Reference*) genome sequence for comparison

Objective 3 - Visualize the read alignment to confirm the deletion and investigate any known clinical relevance

NIH National Library of Medicine
National Center for Biotechnology Information

Paper - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778042/

Objective 1 – Search for the sequencing reads deposited into NCBI’s SRA database with AWS Athena

NIH National Library of Medicine
National Center for Biotechnology Information

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

NIH National Library of Medicine
National Center for Biotechnology Information

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

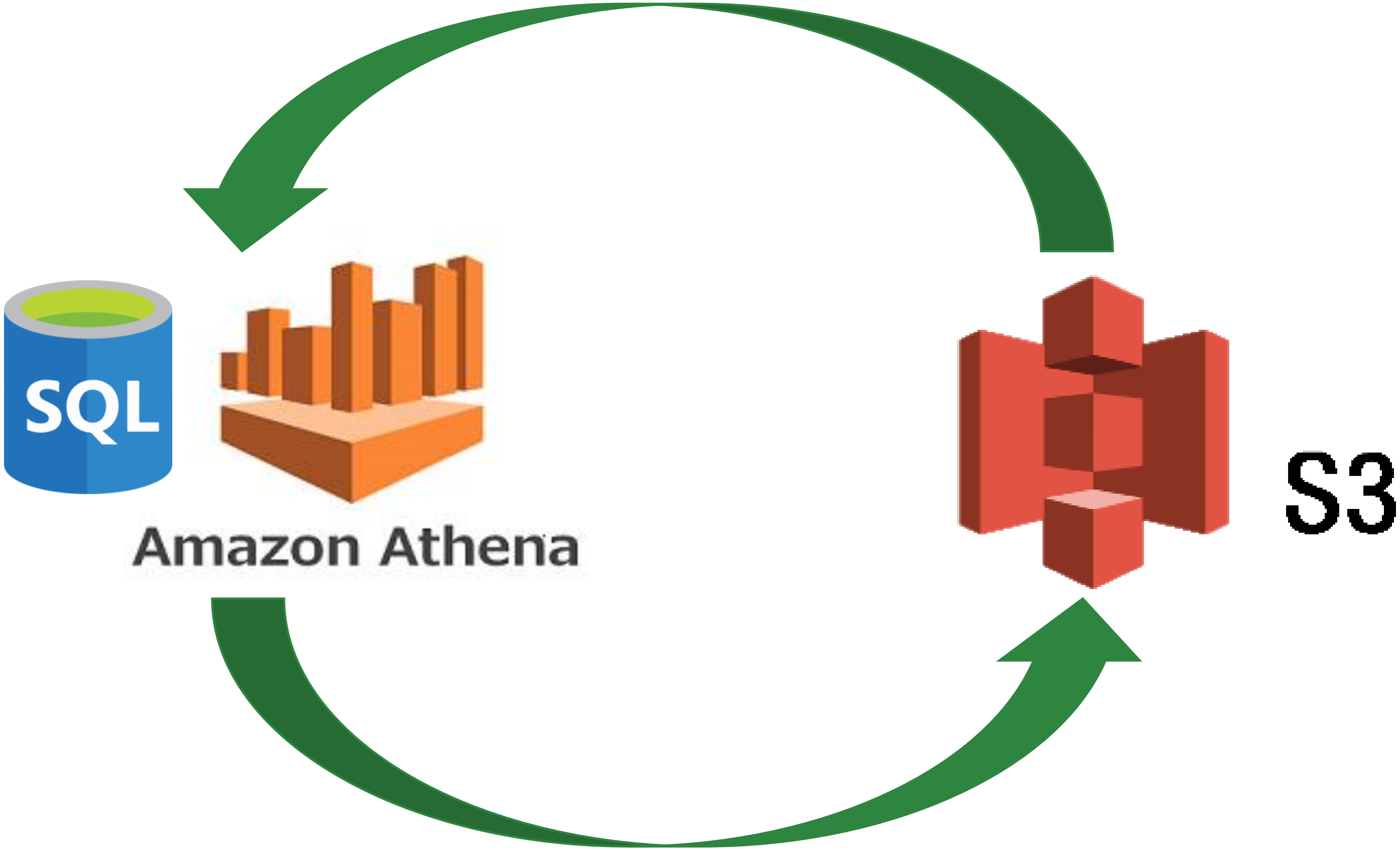
- AWS data-table querying platform designed to rapidly query large tables of data using the SQL language
- NCBI offers all SRA read metadata as a table we can import into Athena
    - We can query the metadata with Athena to pull out only useful sequence data to use in our own research
- Results can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

Amazon Athena

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!**

S3



Import results and mine data in table format

Store data mining results and save useful queries

Objective 1 - Goals

Computational

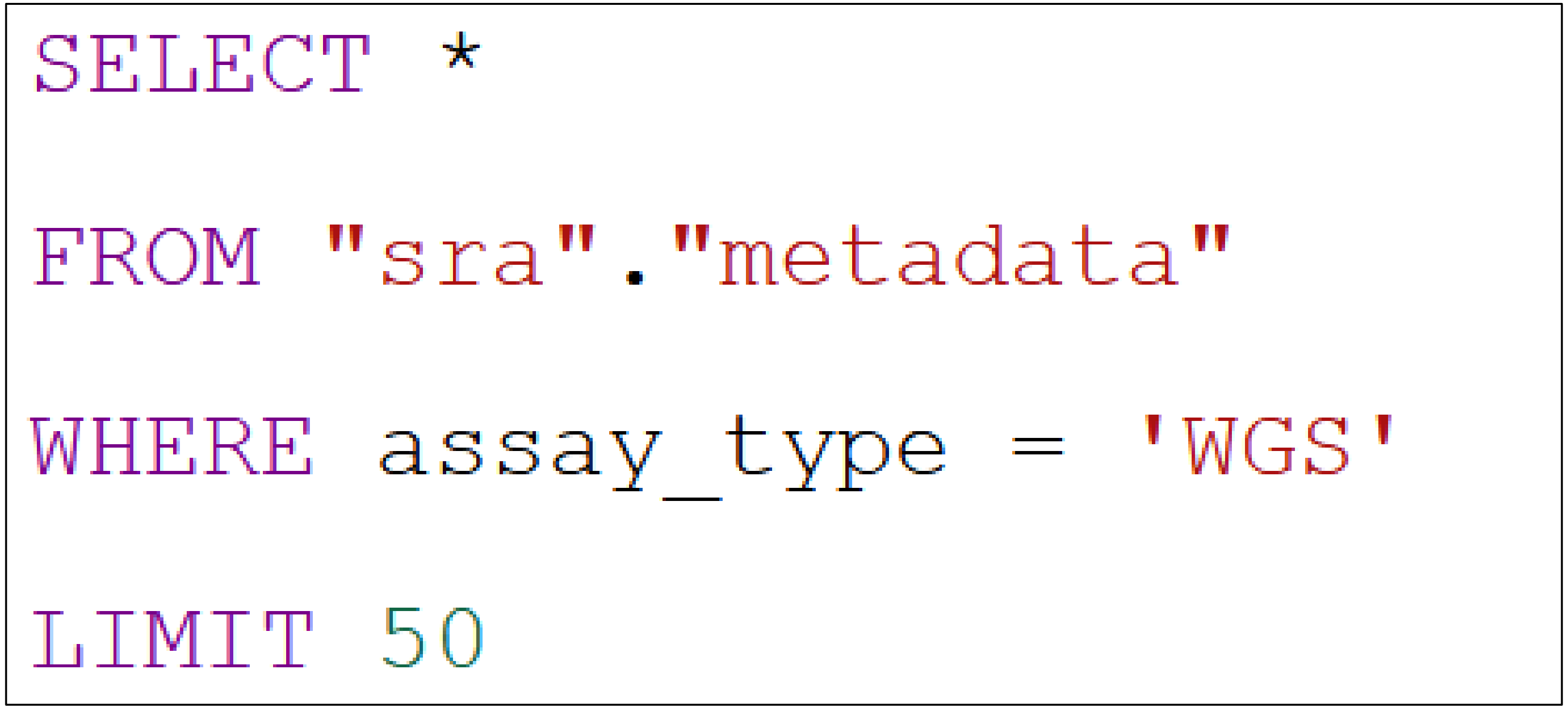
- Create an S3 bucket to store results and files
- Use basic SQL commands to query Athena data tables
- Save query results to personal computer and an S3 bucket

Case Study

- Find sequence data associated with case study publication

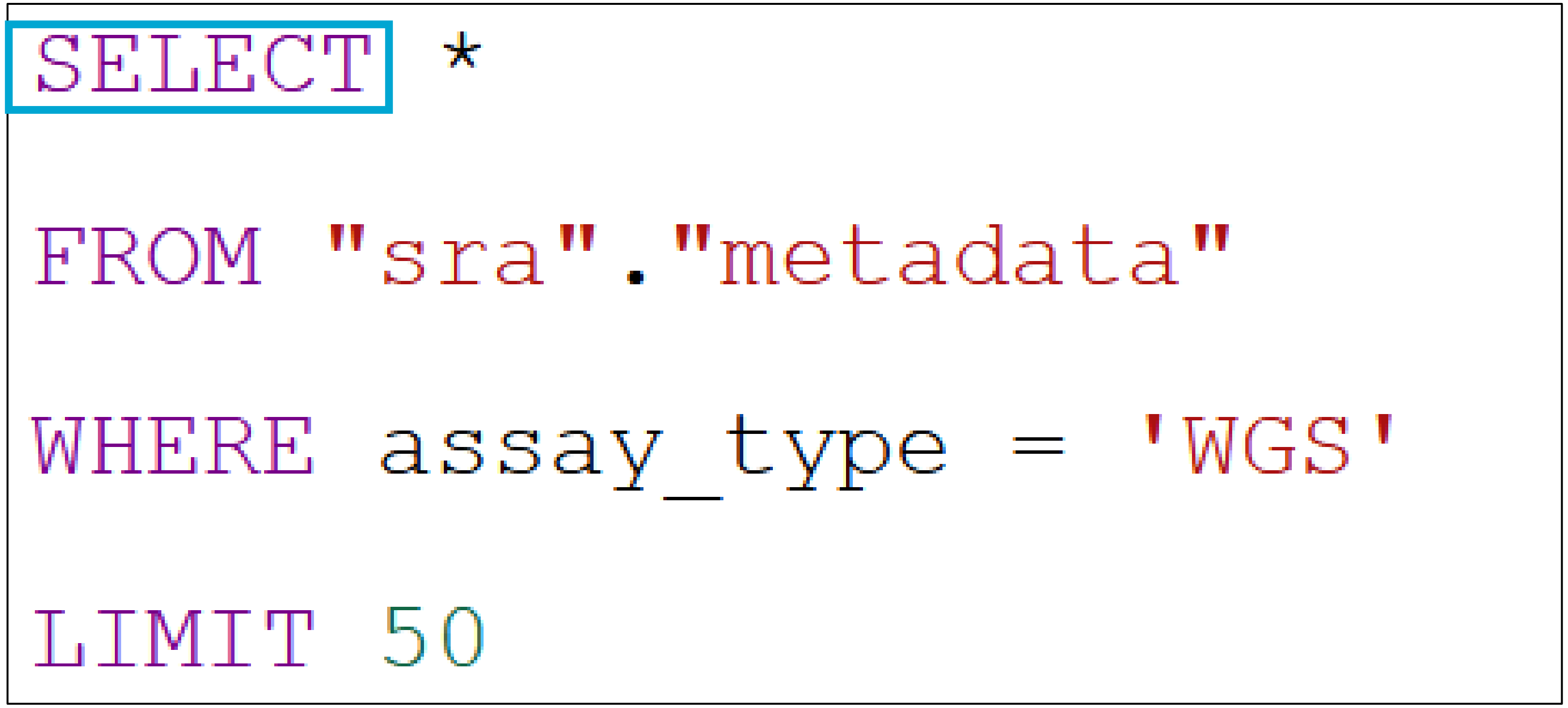
S3 & Athena Walkthrough

SQL programming language basics – 1/6



```
SELECT *

FROM "sra"."metadata"

WHERE assay_type = 'WGS'

LIMIT 50
```

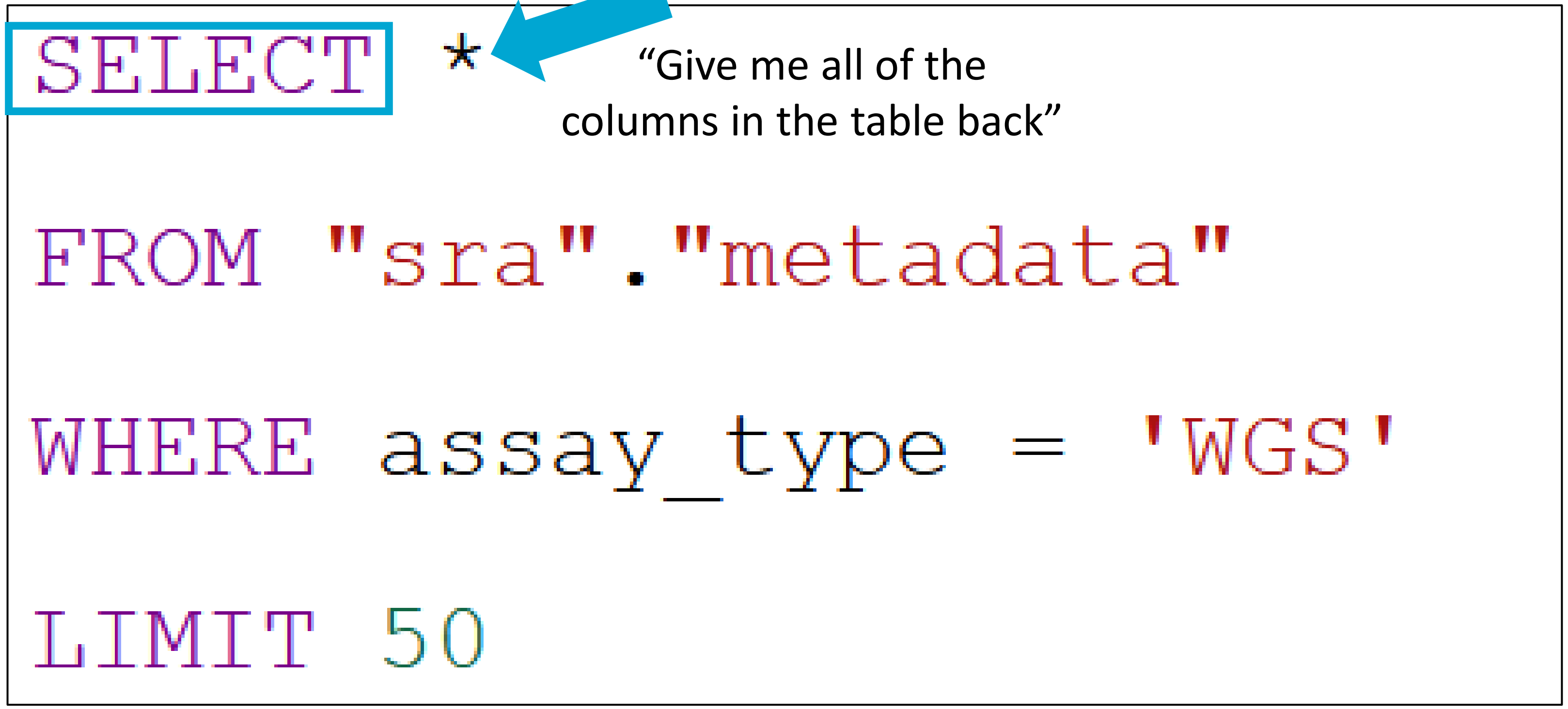
SQL programming language basics – 2/6



```
SELECT *

FROM "sra"."metadata"

WHERE assay_type = 'WGS'

LIMIT 50
```

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

SQL programming language basics – 3/6

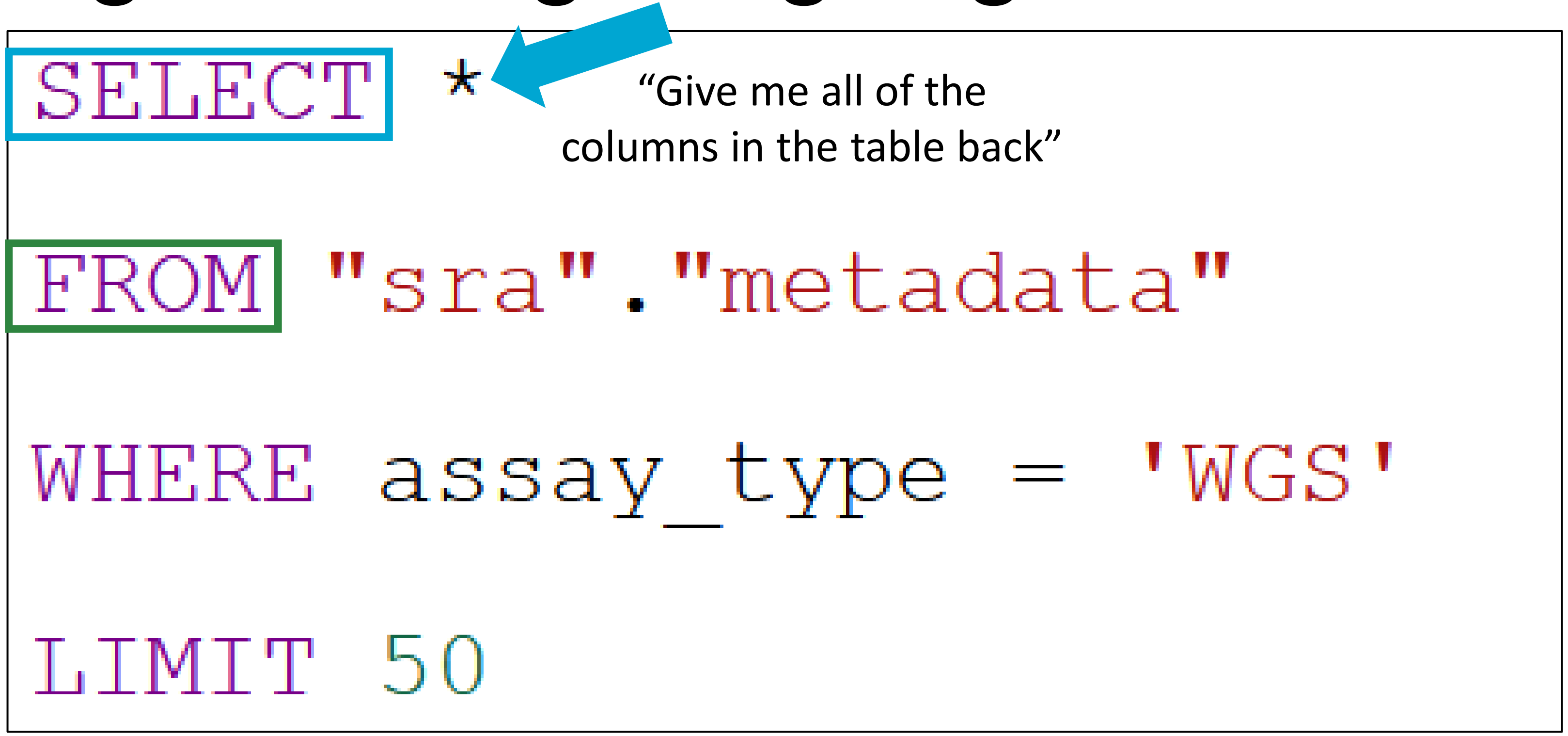


```
SELECT *
FROM "sra"."metadata"
WHERE assay_type = 'WGS'
LIMIT 50
```

"Give me all of the columns in the table back"

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

SQL programming language basics – 4/6



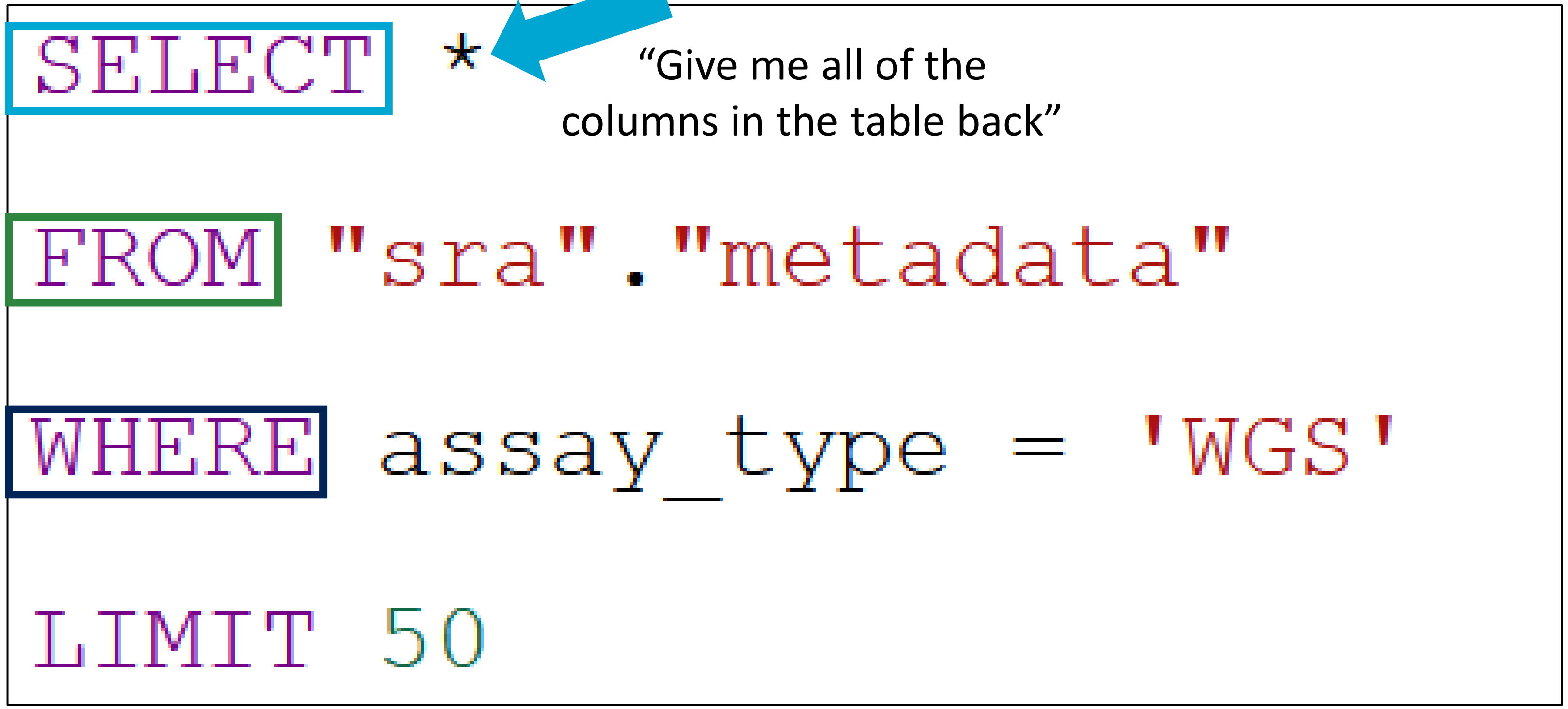
```
SELECT *

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

SQL programming language basics – 5/6



```
SELECT *

FROM "sra"."metadata"

WHERE assay_type = 'WGS'

LIMIT 50
```

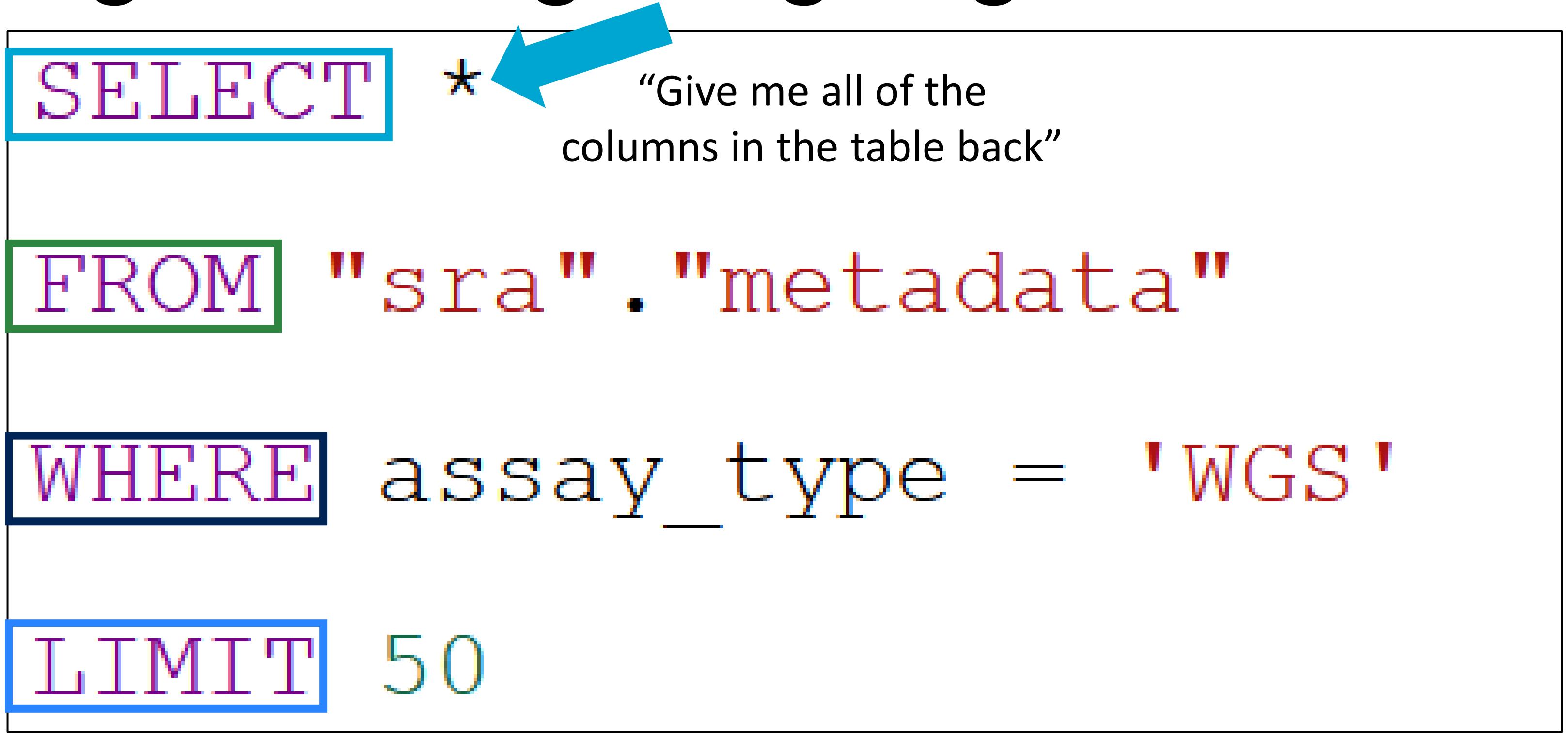
"Give me all of the columns in the table back"

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

SQL programming language basics – 6/6



```
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

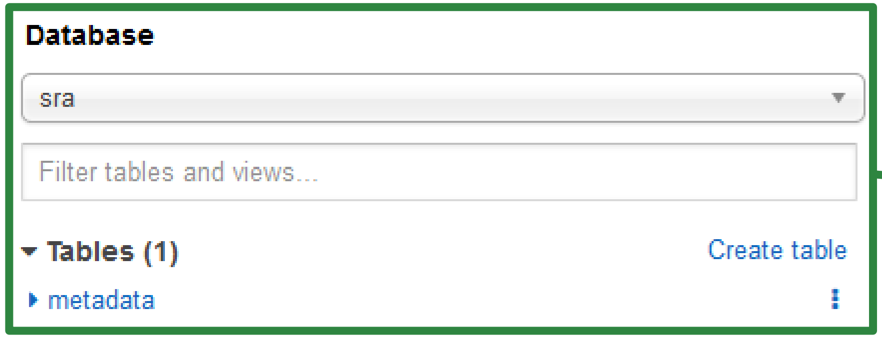
Restrict the results to a given number of rows

```sql
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

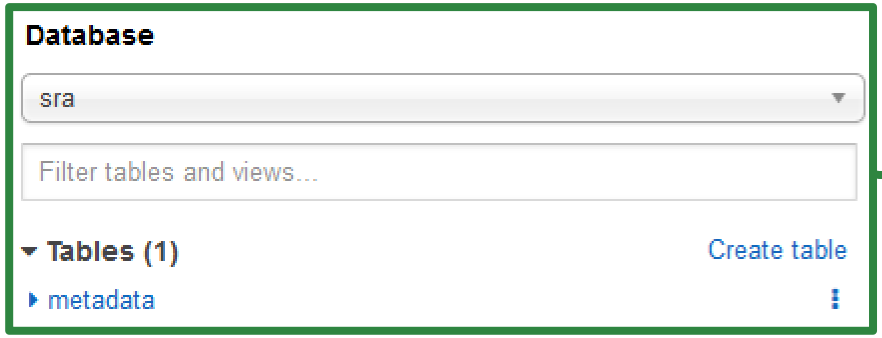
```
SELECT *
FROM "sra"."metadata"
WHERE assay_type = 'WGS'
```

	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



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

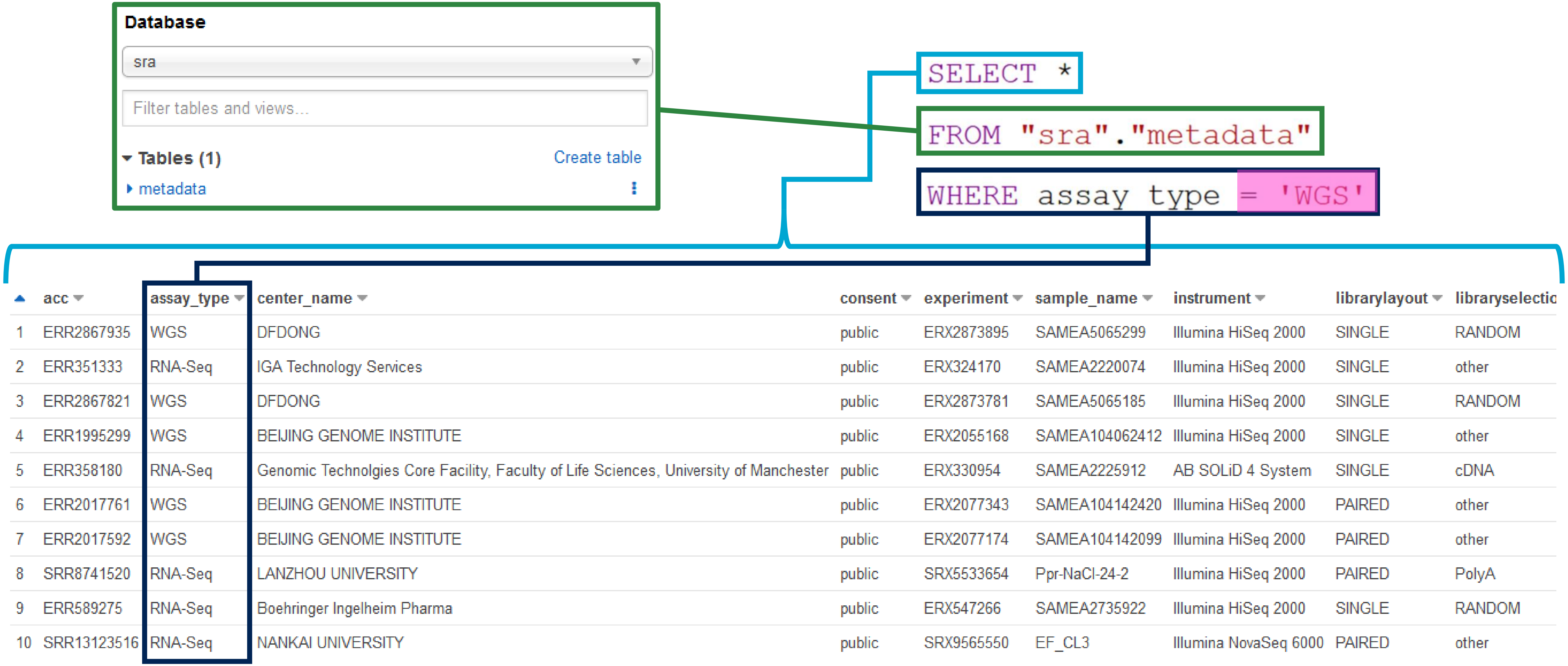
	acc	assay_type	center_name	consent	experiment	sample_name	instrument	librarylayout	libraryselectic
1	ERR2867935	WGS	DFDONG	public	ERX2873895	SAMEA5065299	Illumina HiSeq 2000	SINGLE	RANDOM
2	ERR351333	RNA-Seq	IGA Technology Services	public	ERX324170	SAMEA2220074	Illumina HiSeq 2000	SINGLE	other
3	ERR2867821	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



Database

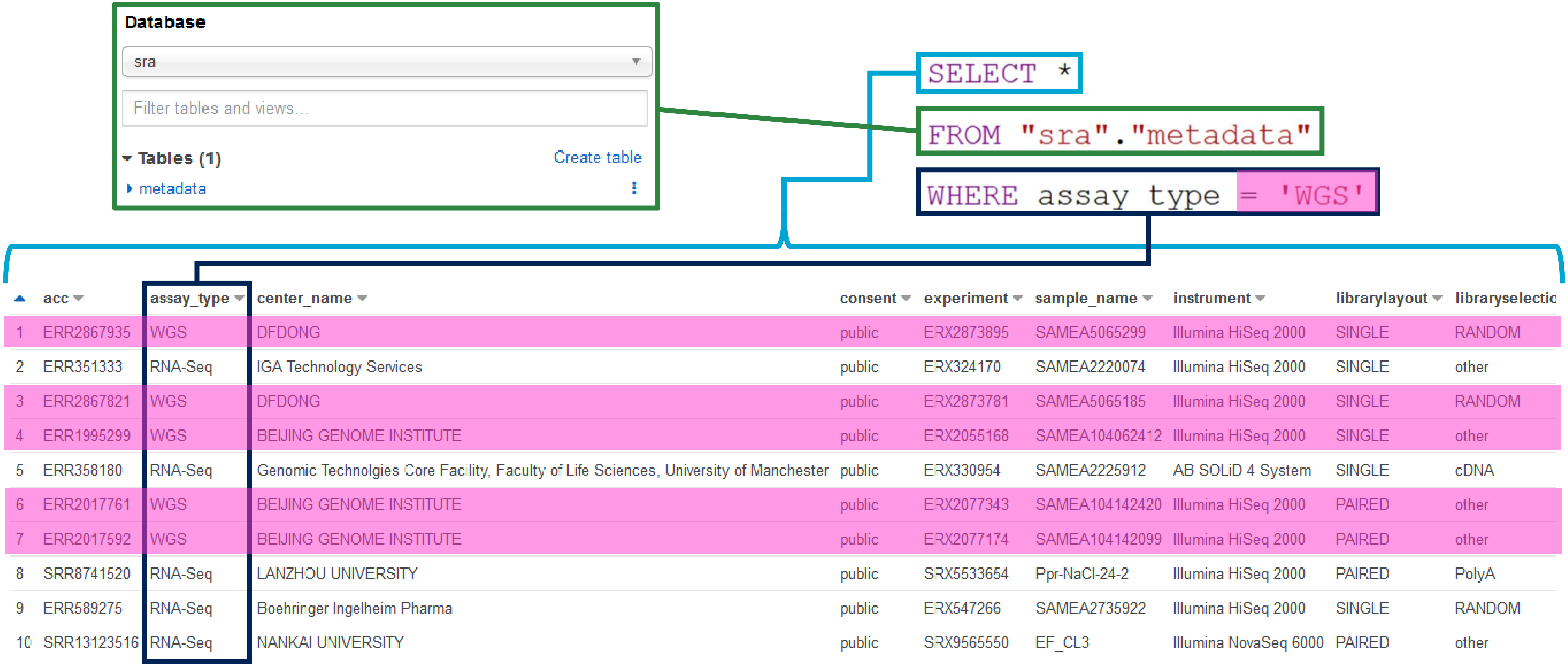
```
SELECT *
FROM "sra"."metadata"
WHERE assay_type = 'WGS'
```

	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



```
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



```
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 Walkthrough

National Library of Medicine
National Center for Biotechnology Information

Objective 2 – Aligning sequence reads using AWS EC2 & MagicBLAST

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

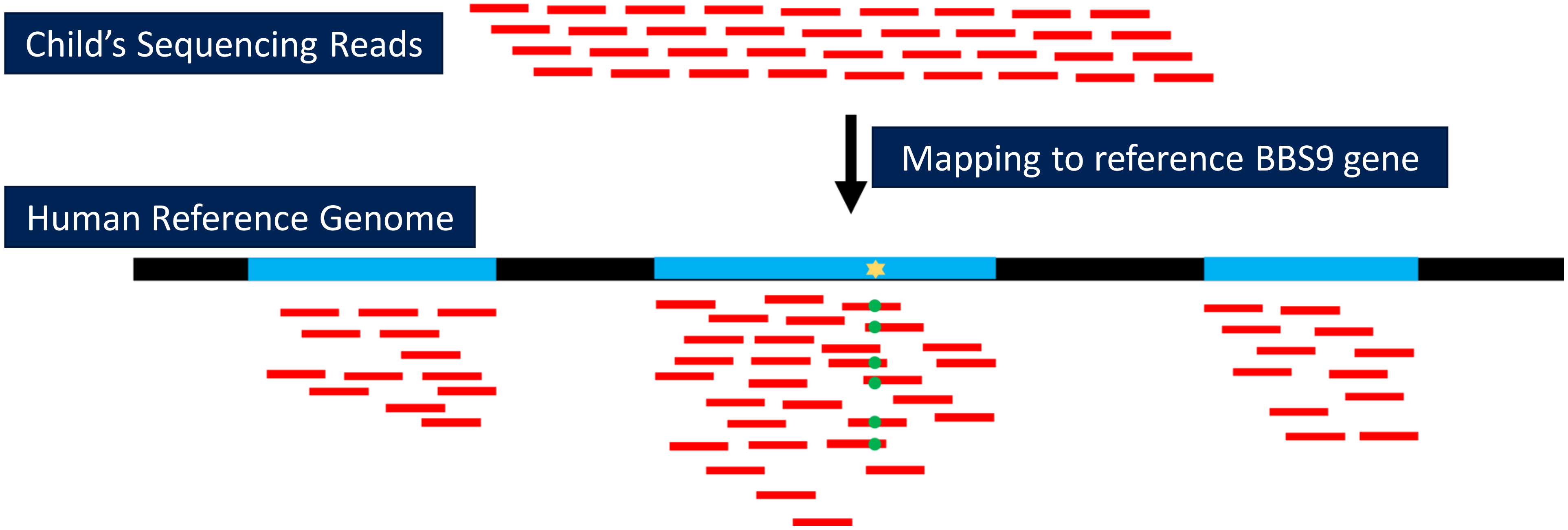
- Pay for what you use
    - Price increases with larger hardware needs and longer runtime
    - Today’s EC2 is roughly **<u>$0.20/hour/person</u>**
    - <u>Turn it off when not in use!</u>

EC2 Walkthrough

National Library of Medicine
National Center for Biotechnology Information

MagicBLAST

- A “flavor” of BLAST which aligns next-generation RNA or DNA sequencing reads against BLAST databases
  - Can use user-created custom databases OR NCBI maintained ones



Documentation - https://ncbi.github.io/magicblast/

Supporting Software

- Samtools
  - http://www.htslib.org/doc/
  - Manipulate MagicBLAST files into formats usable by Genome Data Viewer

- **A**mazon **W**eb **S**ervice **C**ommand **L**ine **I**nterface
  - 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
- Run MagicBLAST and format output files with Samtools
- Upload files from your remote instance to your S3 bucket

Case Study:

- Align child's DNA to human reference genome for compare against "expected" sequence

magicBLAST Walkthrough

National Library of Medicine
National Center for Biotechnology Information

Objective 3 – Visualize the read alignment to confirm the deletion and investigate any known clinical relevance

National Library of Medicine
National Center for Biotechnology Information

Case Study - Using the sequences

Align Sequences

Reference KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAK ASKPKKAKPVK
Child KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK

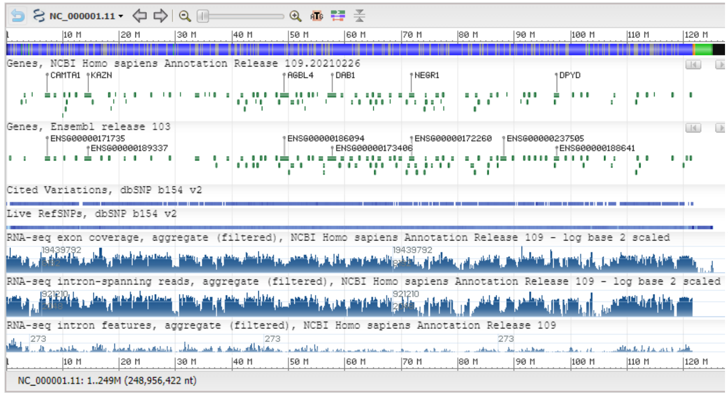
NCBI Magic-BLAST RNA-seq mapping tool

Amazon EC2

S3

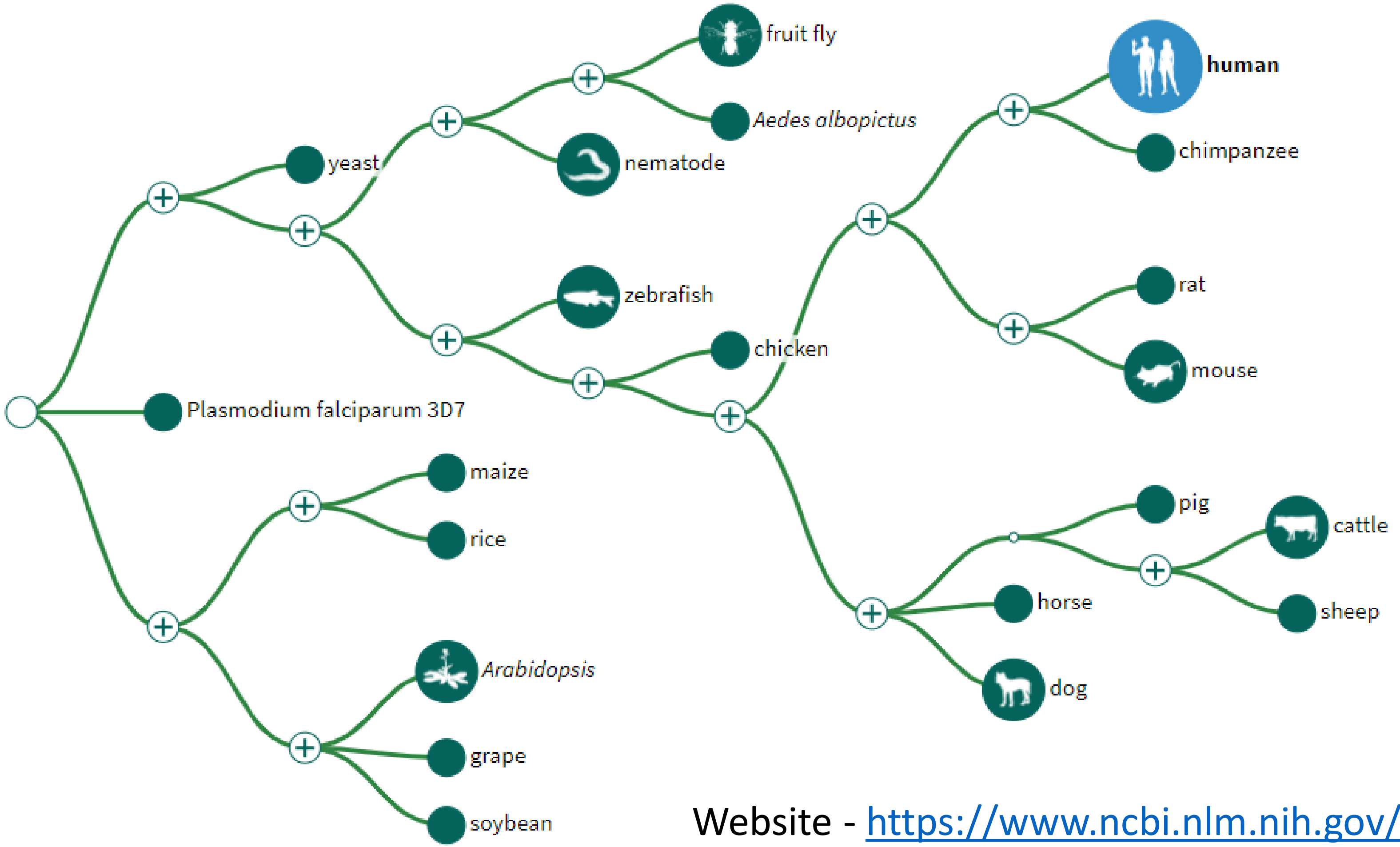
Visualize Alignment

Genome Data Viewer



Genome Data Viewer

GDV is a genome “browser” which supports the visualization of genetic data mapped against >1000 NCBI curated/annotated eukaryotic reference genomes



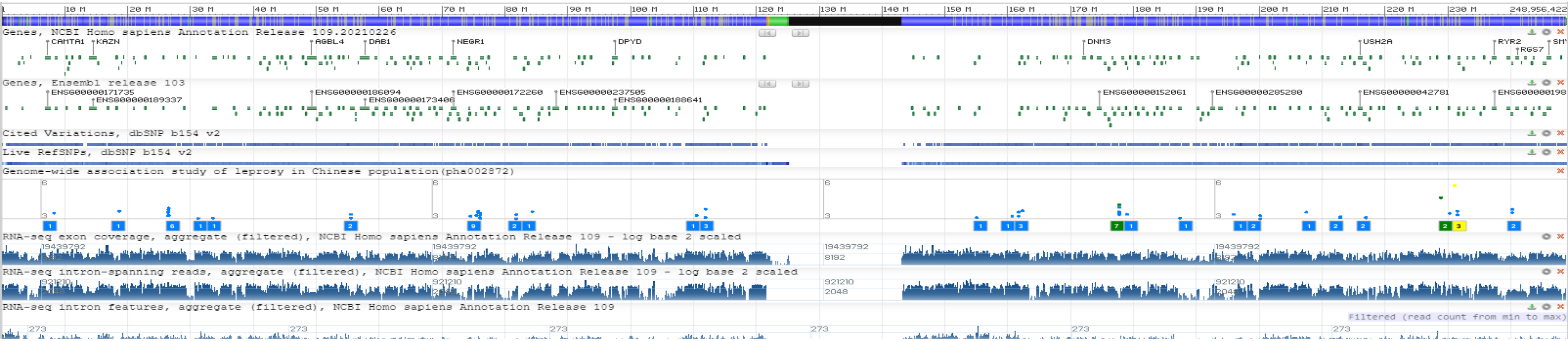
NIH National Library of Medicine
National Center for Biotechnology Information

Website - https://www.ncbi.nlm.nih.gov/genome/gdv/

Genome Data Viewer cont.

Data is visualized in “tracks”

- Can include gene/feature annotations, sequence coverage, GWAS data, and more!
- Users can mix/match between their own tracks and access NCBI/partner provided ones



Website - https://www.ncbi.nlm.nih.gov/genome/gdv/

Objective 3 - Goals

Computational:

- Access and navigate Genome Data Viewer
- Upload custom data tracks to GDV
- Parse biological meaning from alignment results
- Use NCBI track data to find known clinical relevance

Case Study:

- Identify structural changes between patient DNA and reference sequence to identify possible deletions in BBS related gene
- Use NCBI dbVar data to match results to known structural variants

GDV Walkthrough

Billing – 1/4

- The most important question in cloud computing...

Billing – 2/4

- The most important question in cloud computing...

“How Much Will This Cost Me?”

Amazon Athena

Amazon Glue

S3

Amazon EC2

NIH National Library of Medicine
National Center for Biotechnology Information

Website - https://www.ncbi.nlm.nih.gov/genome/gdv/

POLL!

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

Billing – 3/4

- The most important question in cloud computing...

“How Much Will This Cost Me?”

Billing – 4/4

- 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 cont.

- 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

Thank you!