Introduction to NCBI Cloud Computing for Biologists

Cooper J. Park, PhD; Rana Morris, PhD



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"



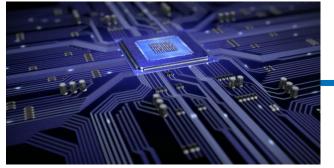
A "one-stop shop" for high-demand computing services delivered across the internet – 1/6

"The Cloud"





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

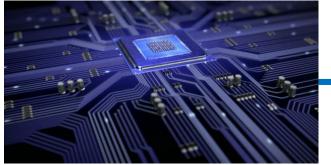


Compute Power





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

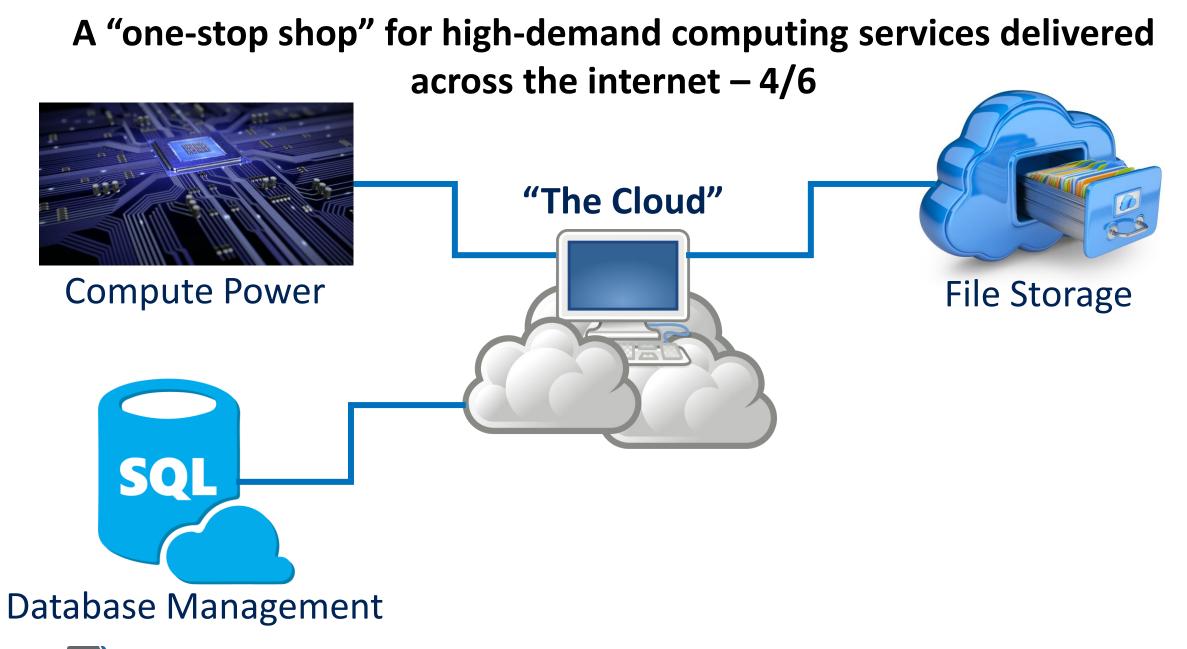


Compute Power

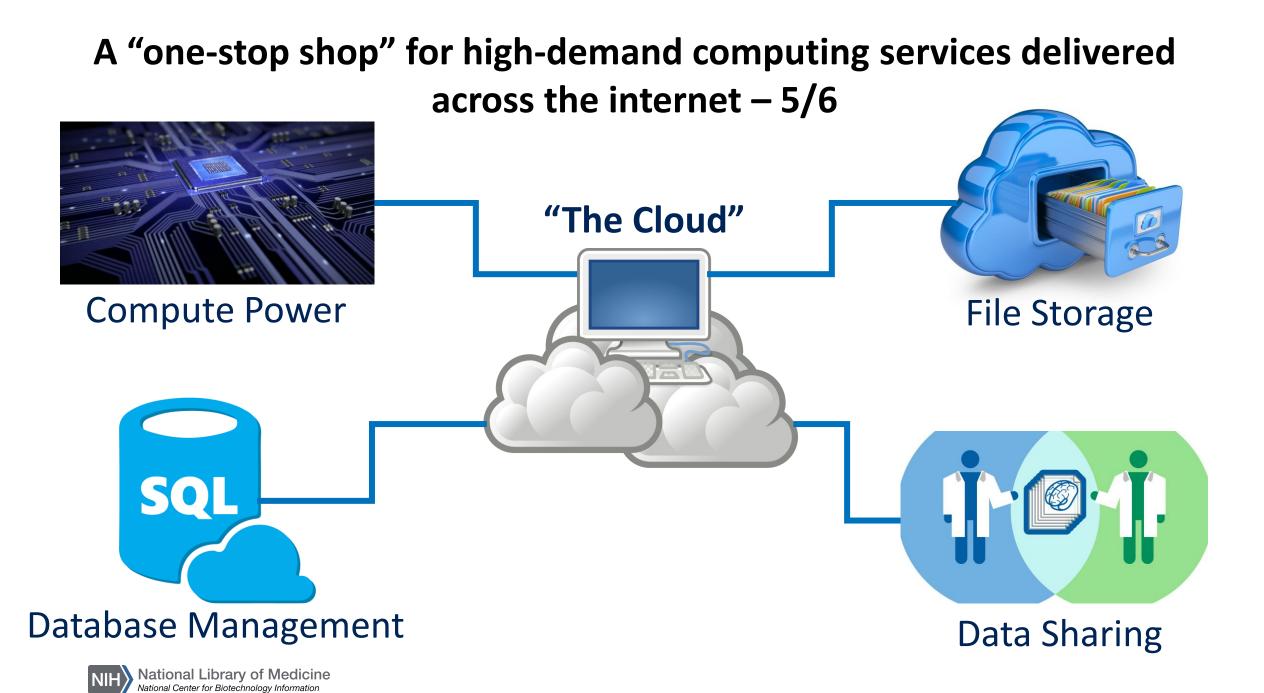


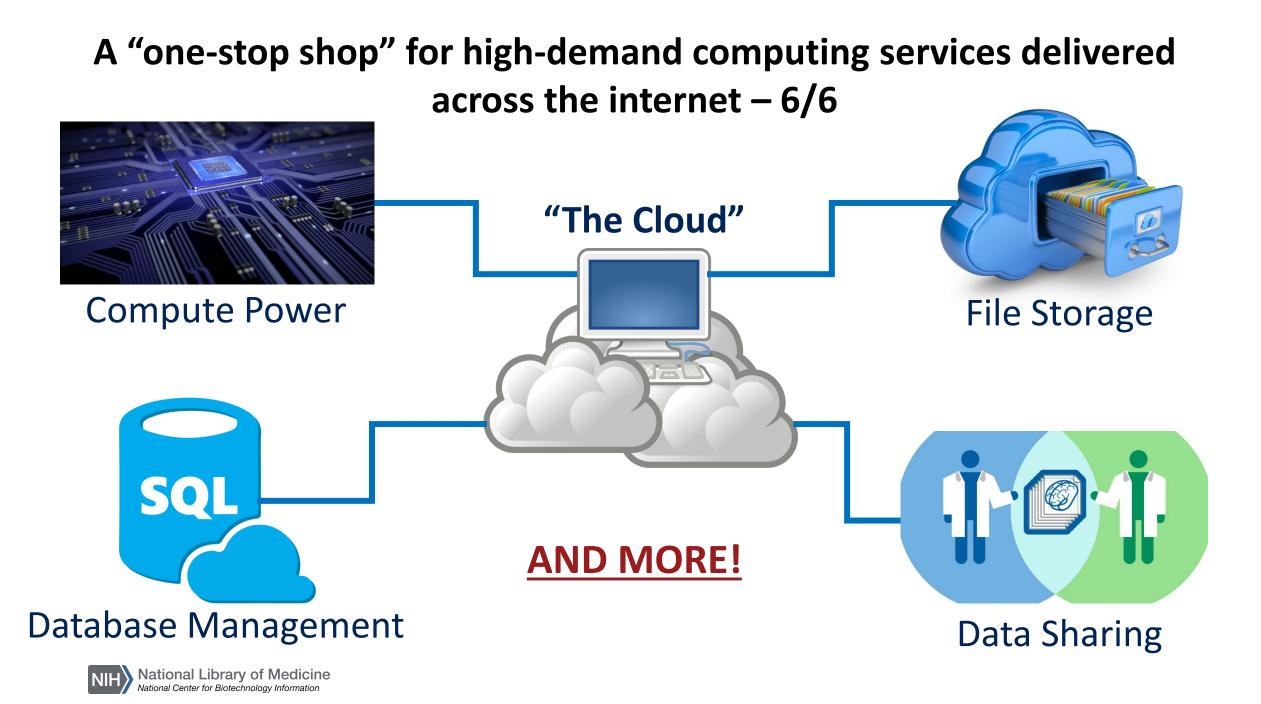


NIH National Library of Medicine National Center for Biotechnology Information



H National Library of Medicine National Center for Biotechnology Information





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

3) Speed and Performance

2) Global Access

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

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

3) Speed and Performance

2) Global Access

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

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

3) Speed and Performance

2) Global Access

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

4) Reproducibility, Security, and Reliability

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

- 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

3) Speed and Performance

2) Global Access

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

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





Meet your commercial cloud providers – 2/2





Google Cloud



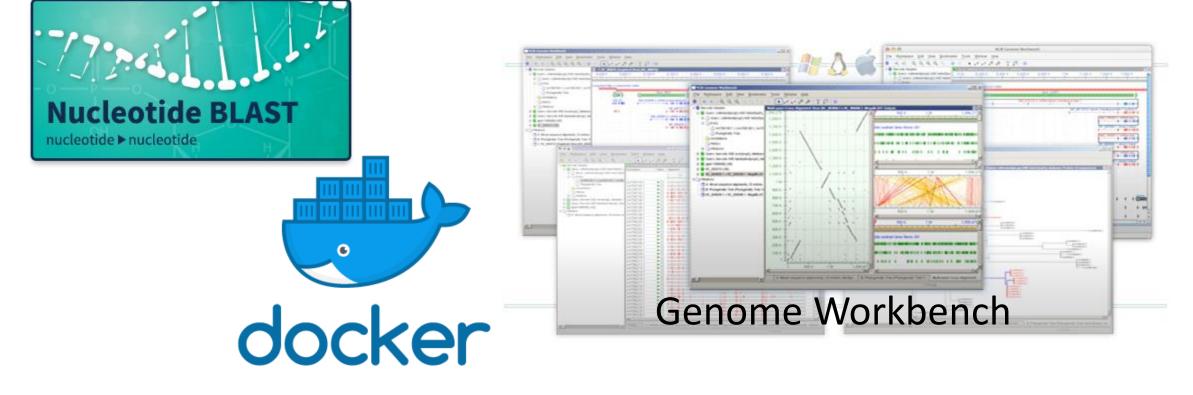


NCBI and the Cloud



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.







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

<u>Username</u>: "Email Prefix" (everything after the "@") <u>Password</u>: <See the chatbox>

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



National Library of Medicine National Center for Biotechnology Information

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).



Sardet 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.



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).



Sardet 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 "longread" sequencing technologies.

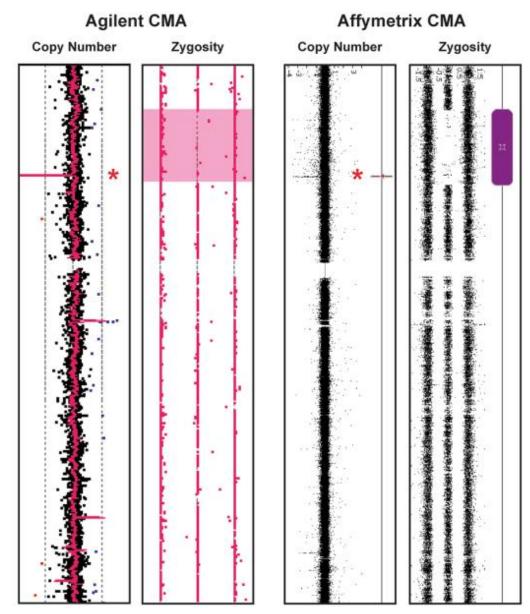
Paper - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778042/

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



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

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



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



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





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 <u>all</u> 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

Amazon Athena

Results can be saved to 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 <u>free!</u>





Import results and mine data in table format





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



National Library of Medicine National Center for Biotechnology Information

SQL programming language basics – 1/6

SELECT *

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



SQL programming language basics – 2/6

SELECT *

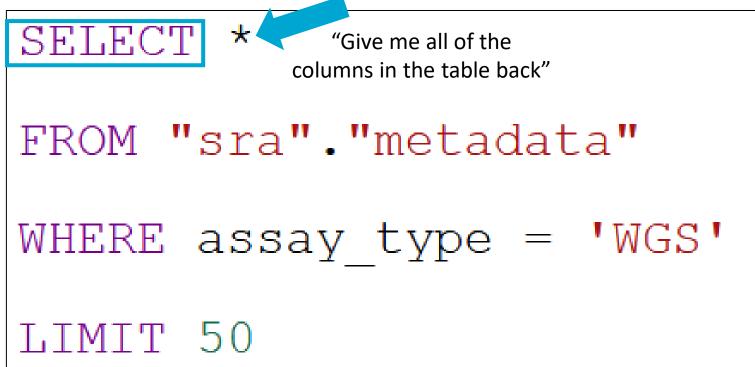
FROM "sra"."metadata"

WHERE assay_type = 'WGS' LIMIT 50

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



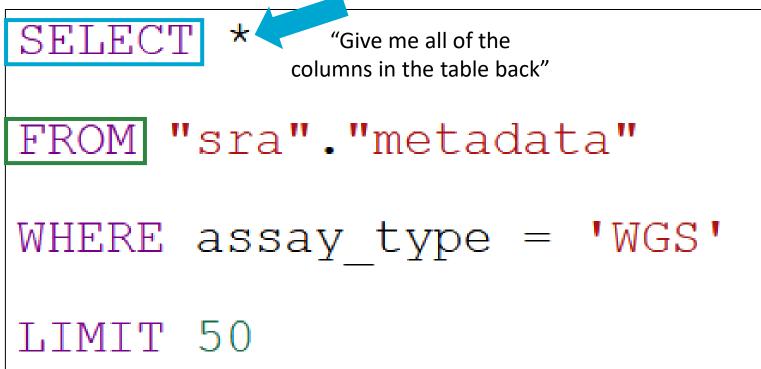
SQL programming language basics – 3/6



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



SQL programming language basics – 4/6

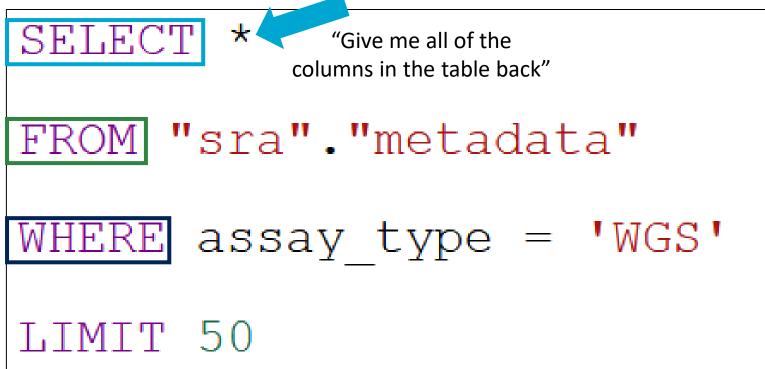


Choose the table columns you want to see for each hit from the table

Choose which table of data you are querying against



SQL programming language basics – 5/6

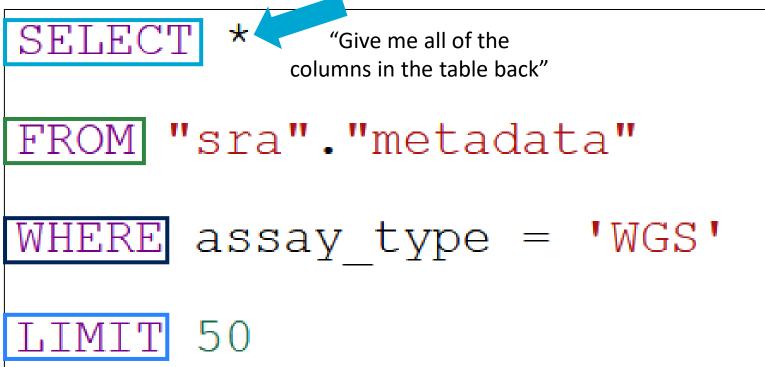


Choose the table columns you want to see for each hit from the table

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



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



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-NaCI-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 |



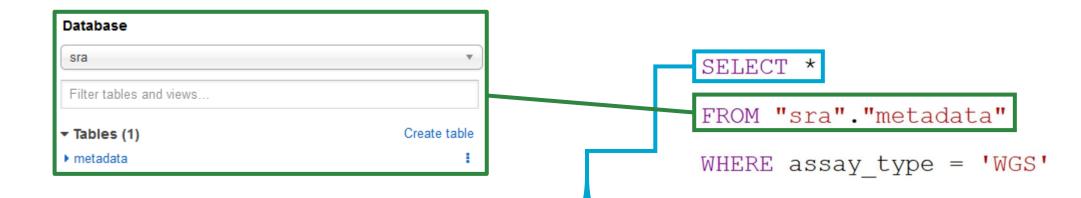


FROM "sra"."metadata"

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-NaCI-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 |





| • | 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-NaCI-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 |

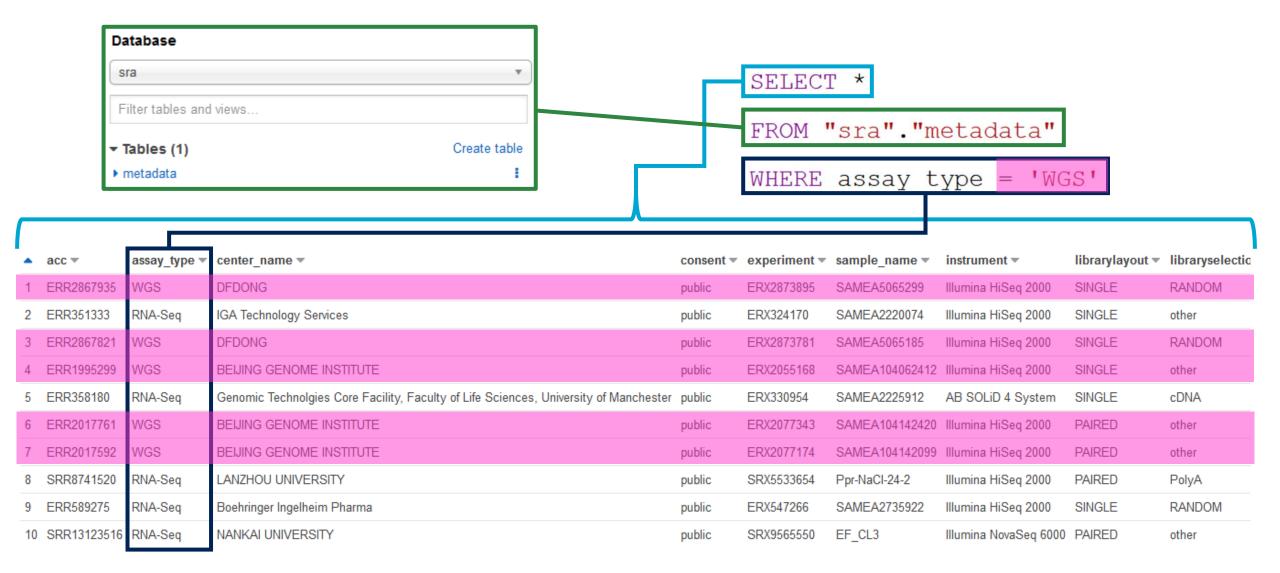


| Da | atabase | | | | | | | | | | | | |
|---|---------------------|--|---------------------------|--------------------------|----------------------------------|--|------------------|-----------------|--|--|--|--|--|
| s | ra | ▼ | _ | SELECT | [* | | | | | | | | |
| F | ilter tables and | views | | | | | | | | | | | |
| - 1 | Tables (1) | Create table | | FROM "sra"."metadata" | | | | | | | | | |
| ≯r | metadata | 1 | | WHERE | S' | | | | | | | | |
| | | | | | | | | | | | | | |
| 200 = | assau tupo = | contor namo = | consent 💌 | experiment - | sample name 🔻 | instrument 💌 | librarylayout 🔻 | librarysoloctic | | | | | |
| | assay_type ▼ WGS | Center_name ▼ DFDONG | public | ERX2873895 | SAMEA5065299 | Illumina HiSeq 2000 | SINGLE | RANDOM | | | | | |
| | RNA-Seq | IGA Technology Services | public | ERX324170 | SAMEA2220074 | Illumina HiSeq 2000 | SINGLE | other | | | | | |
| | WGS | DFDONG | public | ERX2873781 | SAMEA5065185 | Illumina HiSeq 2000 | SINGLE | RANDOM | | | | | |
| ERR1995299 | WGS | BEIJING GENOME INSTITUTE | public | ERX2055168 | SAMEA104062412 | Illumina HiSeq 2000 | SINGLE | other | | | | | |
| CIVICID00200 | | | | | | 40.001/0.40 | SINGLE | cDNA | | | | | |
| | RNA-Seq | Genomic Technolgies Core Facility, Faculty of Life Sciences, University of Man | nchester public | ERX330954 | SAMEA2225912 | AB SOLiD 4 System | SINGLE | CDINA | | | | | |
| ERR358180 | RNA-Seq WGS | Genomic Technolgies Core Facility, Faculty of Life Sciences, University of Man BEIJING GENOME INSTITUTE | nchester public public | ERX330954 ERX2077343 | | AB SOLID 4 System Illumina HiSeq 2000 | PAIRED | other | | | | | |
| ERR358180 ERR2017761 | | | | | SAMEA104142420 | · · · · · | | | | | | | |
| ERR358180 ERR2017761 ERR2017592 | WGS | BEIJING GENOME INSTITUTE | public | ERX2077343 | SAMEA104142420 | Illumina HiSeq 2000 | PAIRED | other | | | | | |
| ERR358180 ERR2017761 ERR2017592 SRR8741520 | wgs wgs | BEIJING GENOME INSTITUTE BEIJING GENOME INSTITUTE | public | ERX2077343 ERX2077174 | SAMEA104142420 SAMEA104142099 | Illumina HiSeq 2000 Illumina HiSeq 2000 | PAIRED PAIRED | other other | | | | | |



| Da | atabase | | | | | | | |
|---------------|------------------|---|--------------|--------------|----------------|------------------------|-----------------|-----------------|
| sr | ra | • | | SELECT | [* | | | |
| Fi | ilter tables and | views | | | | | | |
| - T | Tables (1) | Create table | | FROM ' | 'sra"."m | etadata" | | |
| In € | metadata | | | WHERE | assay t | ype <mark>= 'WG</mark> | S' | |
| | | l | | | | | | |
| acc 🔻 | assay_type ▼ | center_name 💌 | consent 🔻 | experiment 🔻 | sample name 🔻 | instrument 🔻 | librarylayout 🔻 | libraryselectic |
| | | DFDONG | public | ERX2873895 | SAMEA5065299 | Illumina HiSeq 2000 | SINGLE | RANDOM |
| | | IGA Technology Services | public | ERX324170 | SAMEA2220074 | Illumina HiSeq 2000 | SINGLE | other |
| | | DFDONG | public | ERX2873781 | SAMEA5065185 | Illumina HiSeq 2000 | SINGLE | RANDOM |
| ERR1995299 | WGS | BEIJING GENOME INSTITUTE | public | ERX2055168 | SAMEA104062412 | Illumina HiSeq 2000 | SINGLE | other |
| ERR358180 | RNA-Seq | Genomic Technolgies Core Facility, Faculty of Life Sciences, University of Manche | ester public | ERX330954 | SAMEA2225912 | AB SOLiD 4 System | SINGLE | cDNA |
| ERR2017761 | WGS | BEIJING GENOME INSTITUTE | public | ERX2077343 | SAMEA104142420 | Illumina HiSeq 2000 | PAIRED | other |
| ERR2017592 | WGS | BEIJING GENOME INSTITUTE | public | ERX2077174 | SAMEA104142099 | Illumina HiSeq 2000 | PAIRED | other |
| SRR8741520 | RNA-Seq | LANZHOU UNIVERSITY | public | SRX5533654 | Ppr-NaCI-24-2 | Illumina HiSeq 2000 | PAIRED | PolyA |
| ERR589275 | RNA-Seq | Boehringer Ingelheim Pharma | public | ERX547266 | SAMEA2735922 | Illumina HiSeq 2000 | SINGLE | RANDOM |
| 0 SRR13123516 | RNA-Seg | NANKAI UNIVERSITY | public | SRX9565550 | EF CL3 | Illumina NovaSeq 6000 | PAIRED | other |







Athena Walkthrough





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>**
 - Turn it off when not in use!



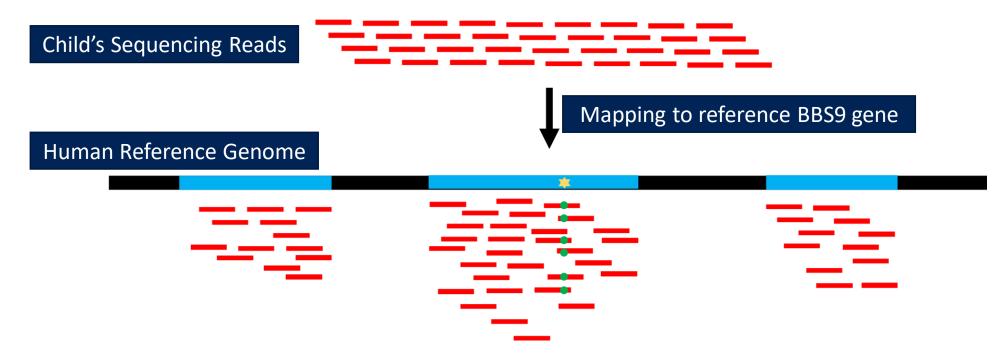


EC2 Walkthrough



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
- <u>A</u>mazon <u>W</u>eb <u>Service</u> <u>Command</u> <u>Line</u> 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
- 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 KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKA Child KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKV/KVKPVKASKPKKAKTVK



NCBI Magic-BLAST RNA-seq mapping tool

Visualize Alignment

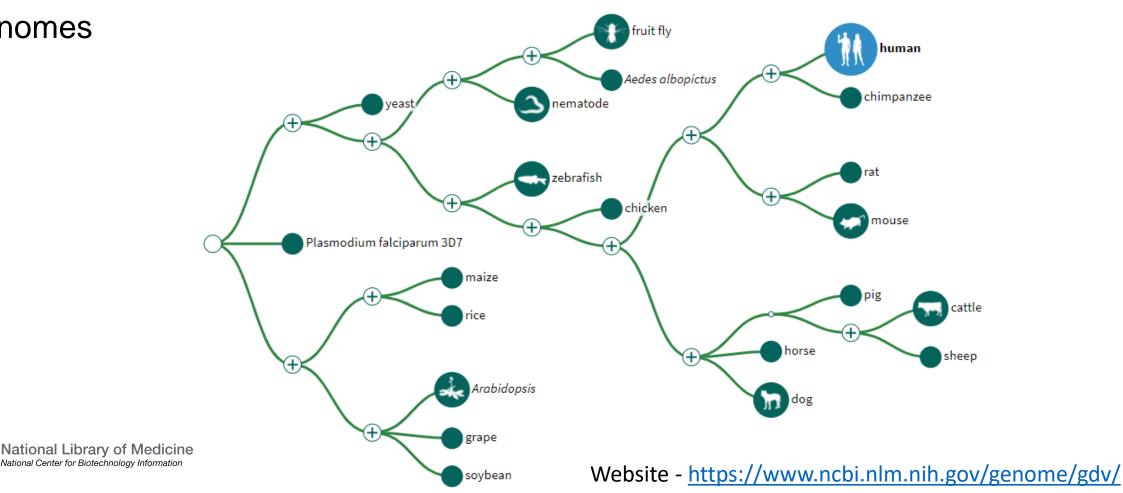
| | 581 | - | 1• 🗇 🖒 | | | • 🔍 🌆 🖡 | | | | | | | |
|--------|-------------------|--------------------------------|----------------------------------|--|----------------------------|-------------|-------------------------|----------------------------|---------------------|--------------|-----------------------|-------------------|----------|
| | | 10 M | 20 M | 30 M | 40 M | 50 H | 60 M | 70 M | 80 M | 98 H | 100 M | 110 M | 120 M |
| | | NCBI Homo CAMTA1 KAZH | | Annotation | Release | | 0226 DAB1 | NEGR1 | ., | | DPYD | 111 | |
| Genome | | ENS6888888171 | elease 10: 735 00000189337 | | | | 186094 ENSG00000173 | 1 ENS6000 | 00172260 | ENS6000002 | 37505 ENS600000188 | 3641 | |
| Data | | | , dbSNP bi SNP b154 v | | | | | | | | | | |
| Viewer | 197 197 197 | exon cov 39792 intron-sj | ليعتم فرام م | Production of the second s | والالإسلام | لعراقتهم | no sapiens NCBI Homo | 9439792 3434 | WIN ^{NY} U | e philosof p | un datam | hantan da | 2 scales |
| | с n N | | and the second | (^{prov} int | اللار العالي الافتار مع | and the sta | o population (| 921210 30 Martin (1947) | to Mili | i palintagi | WHEEL COM | enter he | 1 |
| | RNA-seq | intron f | eatures, a | | 273 | .), NODI | Homo sapie | ens Annot | 27 at 1011 Kei | 3 3 | | alfa fa la far an | |



Genome Data Viewer

genomes

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



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

| L | 10 M | | 20 M | 30 M | 40 M | 50 M | 60 M | 70 M | 80 M | 90 M | 100 M | | 110 M | 120 M | 1 | 130 M | 140 M | 150 M | 160 M | | 170 M | 180 M | 190 M | 200 M | 210 M | 220 M | 230 M | 248,956,422 |
|---------|------------------|----------------|-----------------------|-------------|-------------------------------|----------------|-------------------------|-----------------------------|--------------------|----------------------------|--------------------|---------------------|--|--|-----|------------------|-----------|------------------|-------------------------|---------------|--|-----------------------|-------------------|-------------------------------|-------------------|-------------|---------------------|---------------------------|
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | NCBI I CAMTA1 | | apiens A | innotatio | on Releas | | | | | | DPYD | | | | | | | | | | римз | | | | | SH2A | | RYR2 SM' |
| | . J | 1 R | 1 | | 14.14 | | 1 I I I I | 11 11 11 11 | 1.1 | 11 | • | | | | | | | • | | | | | | • 14 - 1 | | | · · | |
| Genes, | ENSGOOD | 20017173 | 000189337 | | | ENSGOO | 300186094 † ENS60000 | | 000172260 | ENSGØØØØ | 237505 1ENSG000 | 001886 | 341 | | | | | | | | | 00000152061 | ENSGØ | 0000285280 | , f ^e | NSG00000 | 042781 | ± © × ENS600000198 |
| • • | ÷ | - • • | | · - · | · · · · · · | | - 10 Y | ንግለ እጥ | 1876 | | | 14 | 151 | · · · · | | | 1.77 | | | | | | P | | 16 MA | | | * ***** |
| Cited | Variat: | ions, (| dbSNP b1 | .54 v2 | | | | | | | | | | | | | | | | | | | | | | | | ± 0 × |
| Live H | efSNPs | , dbSNI | P b154 v | 72 | | | | | | | | | | | | | | | | | | | | | | | | ± 0 × |
| Genome | -wide a | associa | ation st | udy of : | leprosy i | n Chines | e populat | ion (pha0028) | 72) | | | | | | | | | | | | | | | | | | | × |
| | B 3 • | | t | | | | | 6 | | | | | .: | | | 6 | | | | | | | 6 | | | | 1 | : |
| RNA-se | 1 q exon | 1 cover: | age, ago | regate | (filtered | l), NCBI | Homo sapi | ens Annotat: | 2 1 ion Releas | se 109 - | - log b | ase 2 | 1 3 scaled | | | | | 1 | 1 3 | | 7 | 1 | | 2 | 2 2 | | 2 3 | 2 |
| 1 | 9439792 | | | | در است _{ار ا} ستار ا | | A discussion | 19439792 2011 - 19439792 | And a state of the | لي دروا بداريا و | n n n hai | glavi ^{te} | an kalan yan | an a | J. | 19439792 8192 | والدادسان | ومعولاتهم | | وريطوا | an a | | 194397 11 3197 | 92 1 | particular states | natur P | الارمحد الأحك | The state of the state of |
| RNA-se | q intro | on-spai | nning re | ads, ag | gregate | filtered |), NCBI H | lomo sapiens | Annotatio | on Relea | ase 109 | - 10 | g base | 2 scal | .ed | | | | | | | | | | | | | 0 × |
| | 2148 I.I. | a land a shirt | and the second second | | فقار ويعلن ويأمل | الشقيقين الراد | A contraction | 921210 2049. | lor, Alto | ر الارمانية (¹ | n de la de | | and the first | | | 921210 2048 | (and a di | a and philipping | ر الط ^ا لكان | n ha na ha ha | nti di tenge | | 921210 | الله والتندية _و ال | al at distan | n king sete | disabilities of the | ing and stilled as |
| RNA-se | q intro | on feat | tures, a | ggregat | e (filter | ed), NCE | I Homo sa | piens Annot | ation Rele | ease 109 | 9 | | | | | | | | | | | | | | Filt | tered (re | ead count fro | ± ⊙ × om min to max) |
| 27 1 | 3 | ndeta das | and come entered | مىر جىلغى د | 2 4 4 4 4 4 4 4 4 4 | 73 4 | | | 27 | 3 ha shi kan | | يلى . | a fa da sa | In | 2 | 273 | | المدارية المالية | in ada | | 273 | and the second second | | والمراجع المراجع والمراجع | 273 | | | and the balance and |

National Library of

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

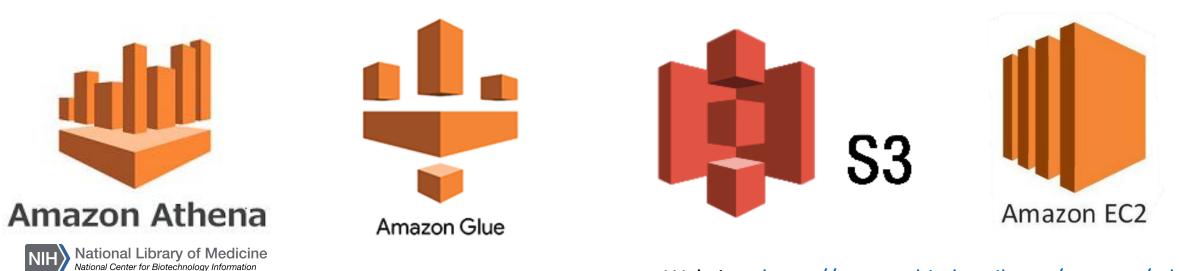
• The most important question in cloud computing...



Billing – 2/4

• The most important question in cloud computing...

"How Much Will This Cost Me?"



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?"

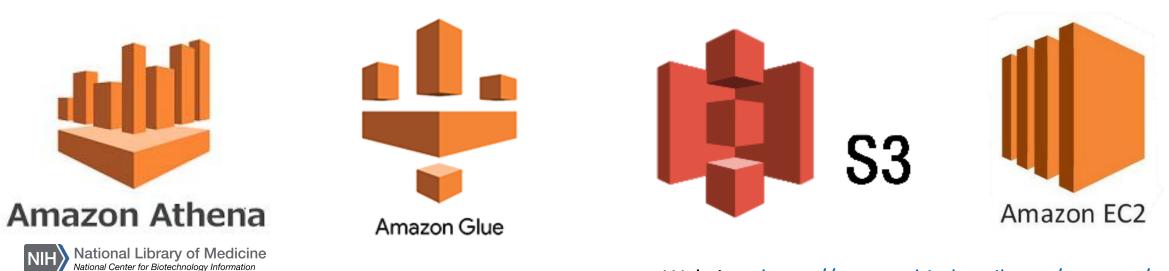


Billing – 4/4

• The most important question in cloud computing...

"How Much Will This Cost Me?"

Everything you did in this workshop cost ~\$0.50



Billing cont.

- AWS strives to be transparent about costs
 - <u>https://calculator.aws/#/estimate</u> Build a price estimate based on anticipated service usage
 - <u>https://aws.amazon.com/free/</u> 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!

