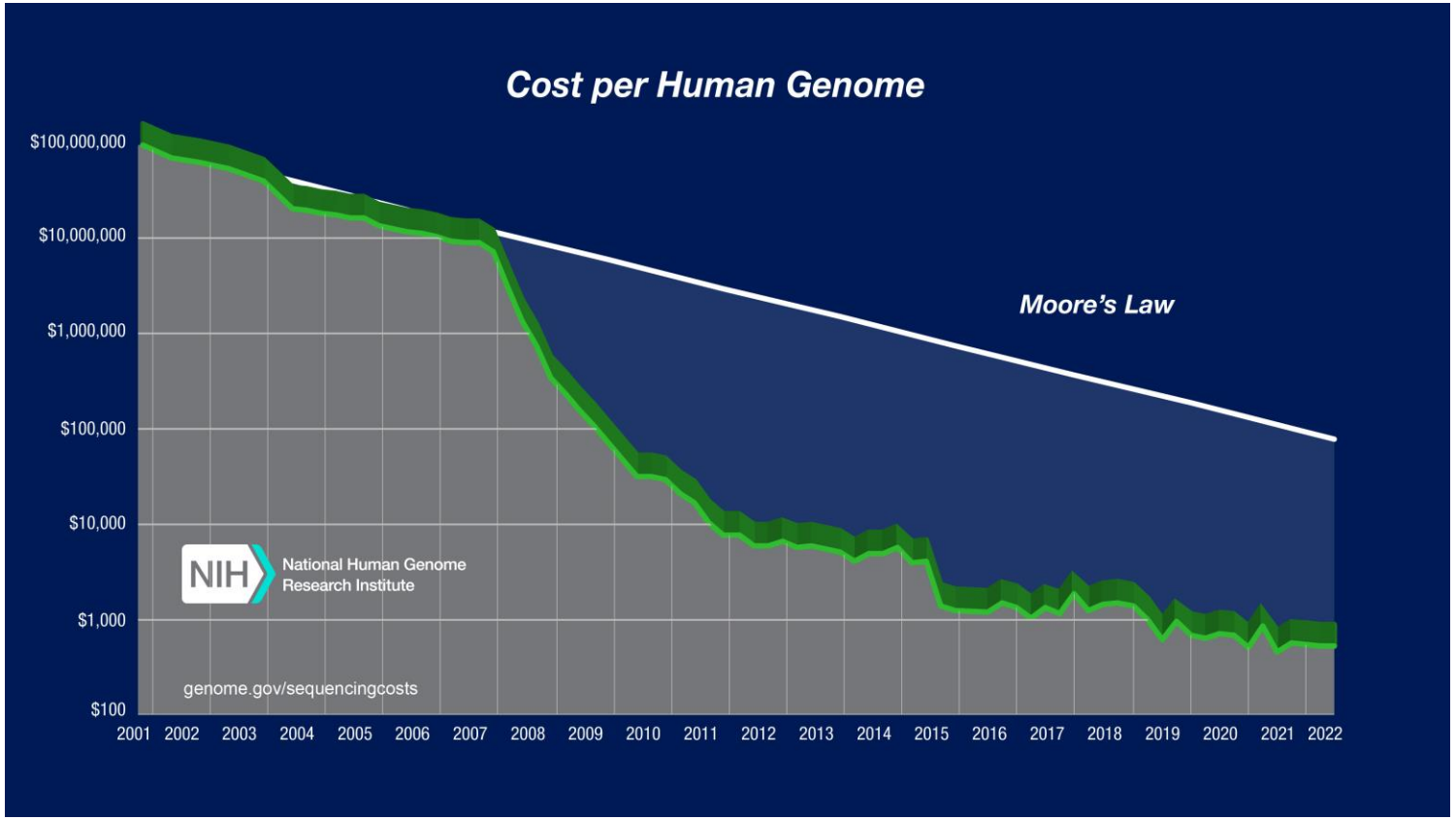
Session 3

Genomes, reads, assemblies, and data repositories

Sequencing

Sequencing costs have fallen dramatically over the past couple of decades

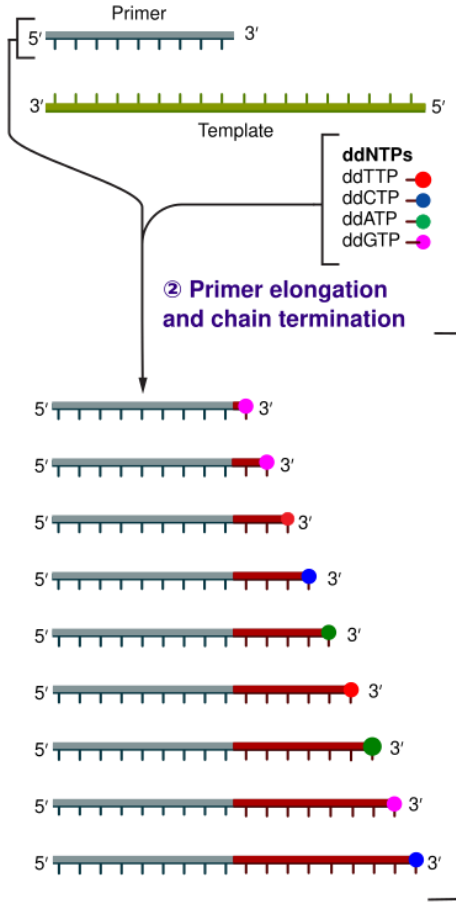


- The cost of sequencing varies by organism, lab, method, etc.
- The cost of a human genome is a common metric

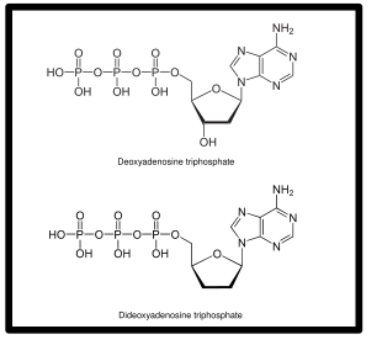
Source: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Sanger sequencing

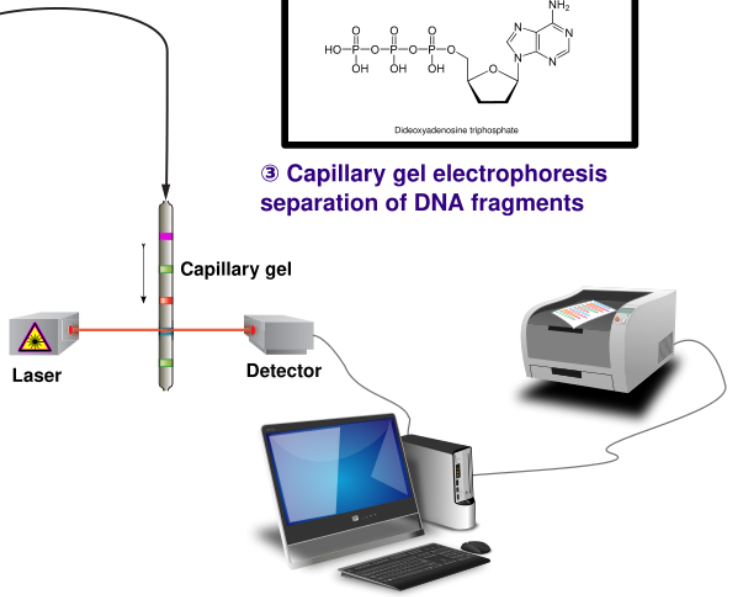
① Reaction mixture
- Primer and DNA template
- ddNTPs with flourochromes
- DNA polymerase
- dNTPs (dATP, dCTP, dGTP, and dTTP)

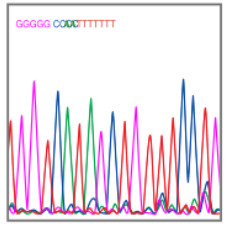


② Primer elongation and chain termination

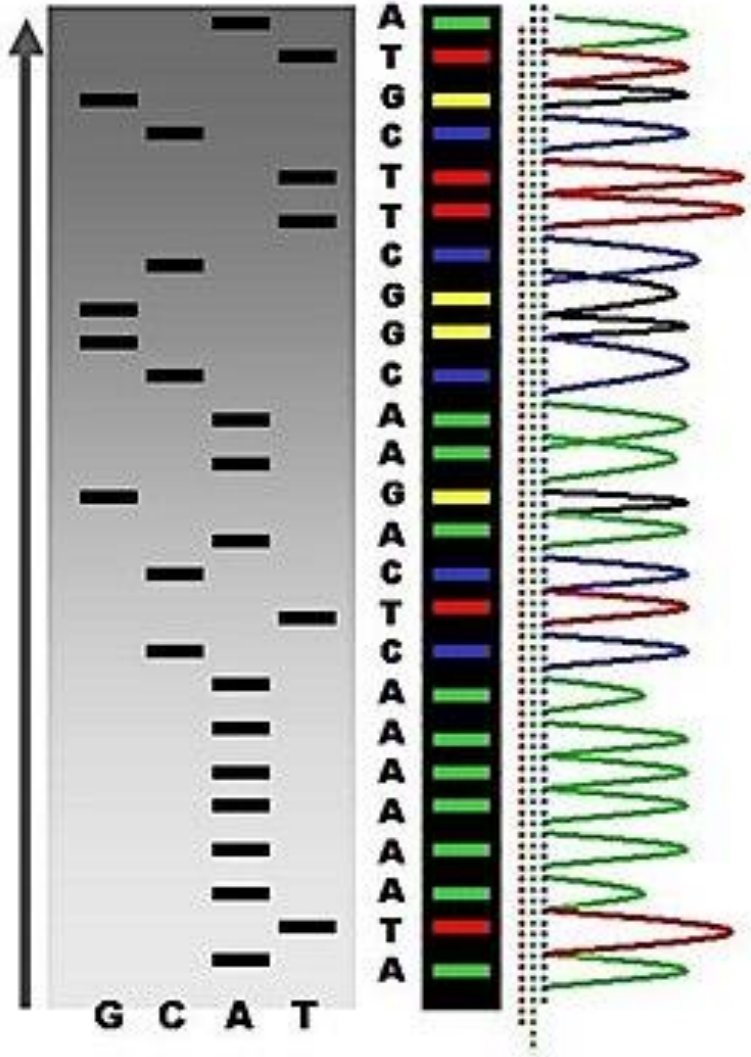


③ Capillary gel electrophoresis separation of DNA fragments

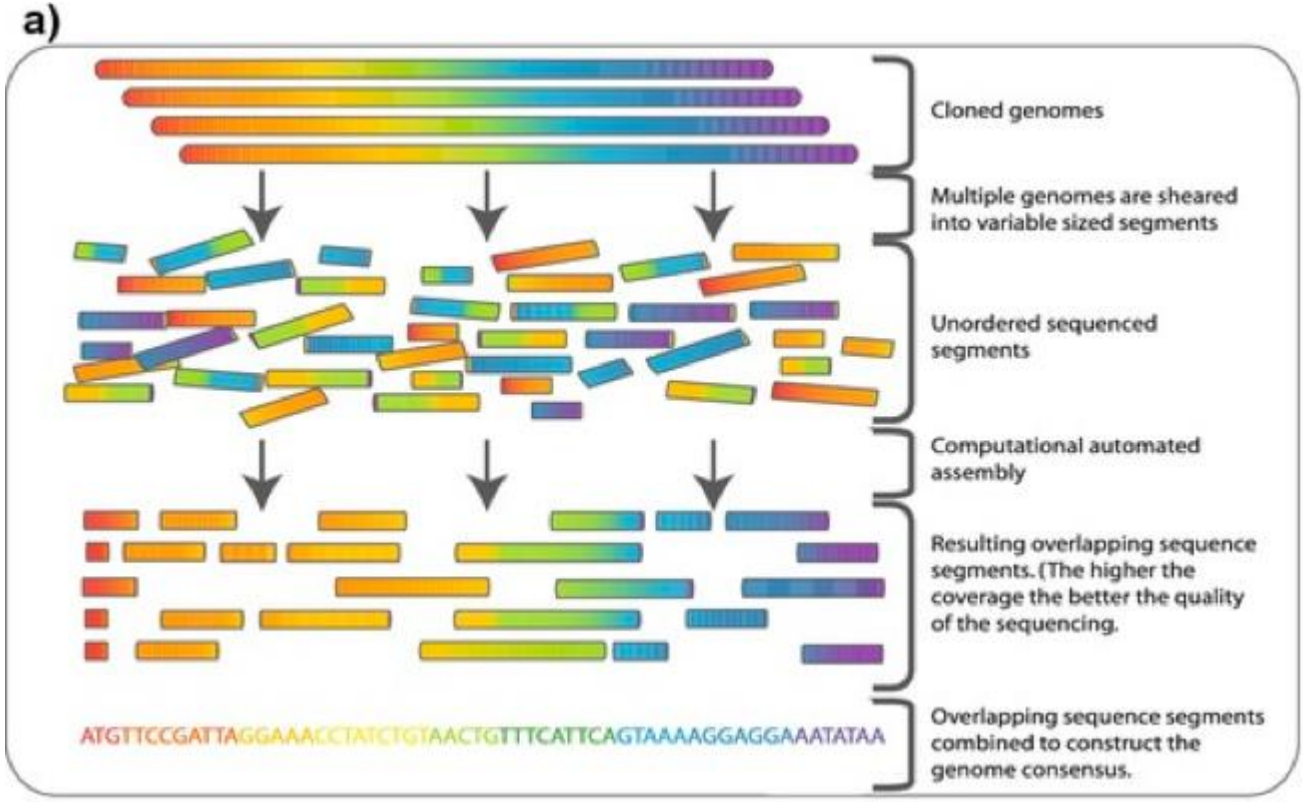




④ Laser detection of flourochromes and computational sequence analysis



Shotgun sequencing



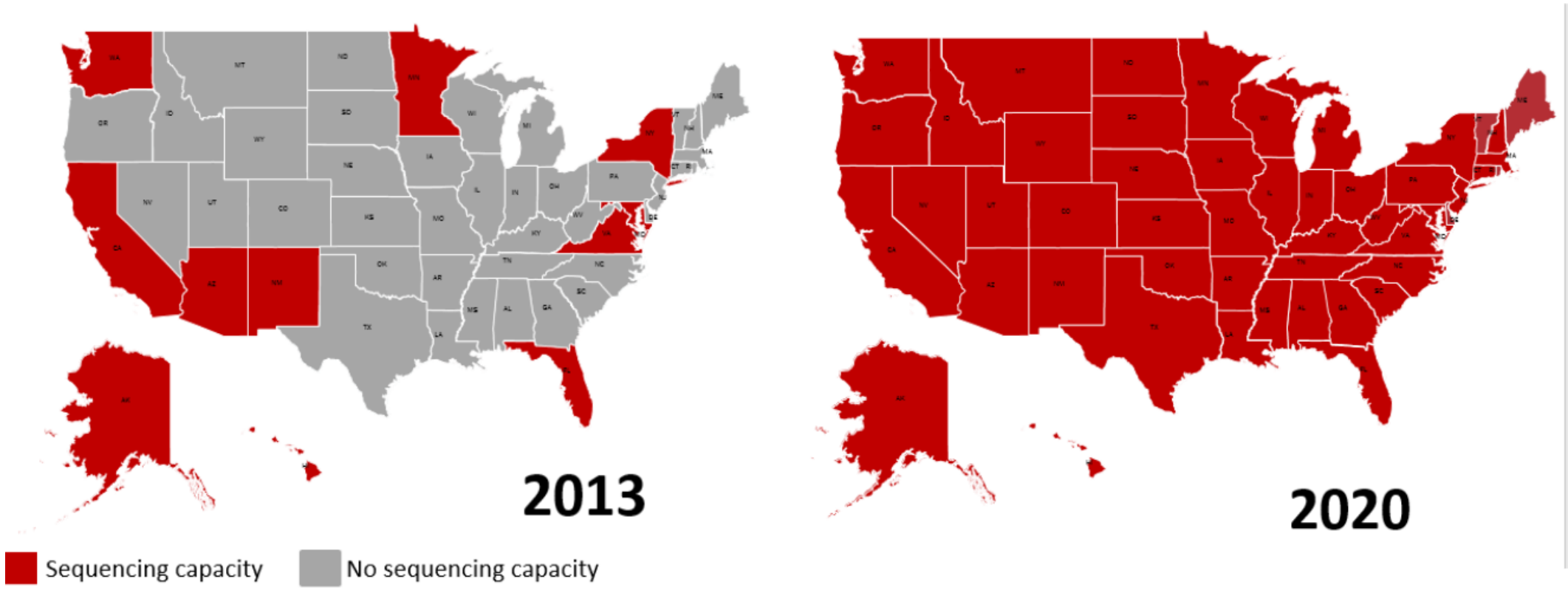
- Randomly chop up the genome, then perform Sanger sequencing
- Sanger sequencing can only be used for short sequences 100-1000bp long

Overview of some popular sequencing technologies

Helpful comparison: https://www.integra-biosciences.com/united-states/en/blog/article/sanger-sequencing-vs-ngs

Sequencing capacity in the United Sates

Seven years of building NGS capacity in state public health laboratories



www.aphl.org/aboutAPHL/publications/Documents/ID_NGSSurveyReport_52015.pdf

https://www.cdc.gov/advanced-molecular-detection/php/training/module-3-6.html

Which specimens to sequence?

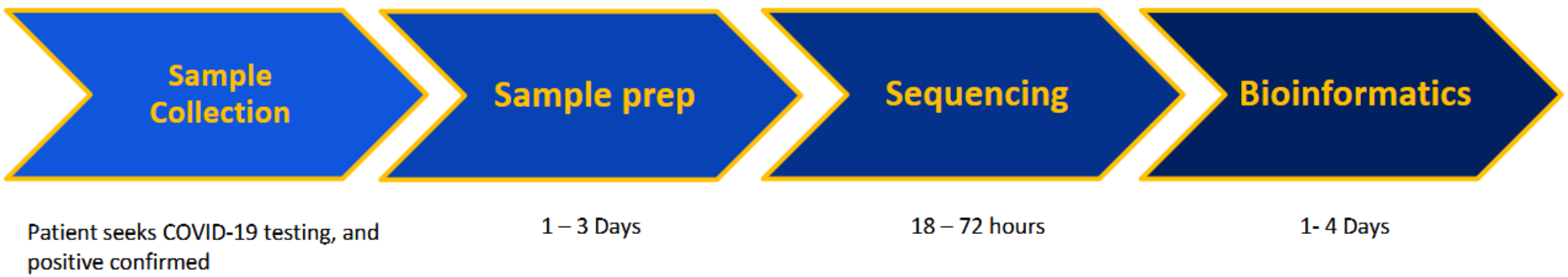
- Outbreak Investigations
  - High-risk groups, such as congregate living settings
    - eg, skilled nursing facilities, homeless shelters, correctional facilities
  - Super-spreader events
- Surveillance
  - Laboratory-based, for emerging strains and trends
    - eg, S-gene target failure, VOIs or VOCs
  - Epidemiologically defined, for cases of particular interest
    - eg, reinfection, vaccine breakthrough, travel exposure, severe COVID-19 in children

VOI – Variants of Interest, VOC – Variants of Concern : https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html

https://www.cdc.gov/advanced-molecular-detection/php/training/module-3-6.html

Technical considerations

- Genome Completeness
  - Is sequencing only the S-gene (spike protein) sufficient?
- Ct (cycle threshold) value
  - Ct value ⟺ genome sequence recovery
- Laboratory's sequencing capacity



https://www.aphl.org/programs/preparedness/Crisis-Management/Documents/APHL-COVID19-Ct-Values.pdf

https://www.cdc.gov/advanced-molecular-detection/php/training/module-3-6.html

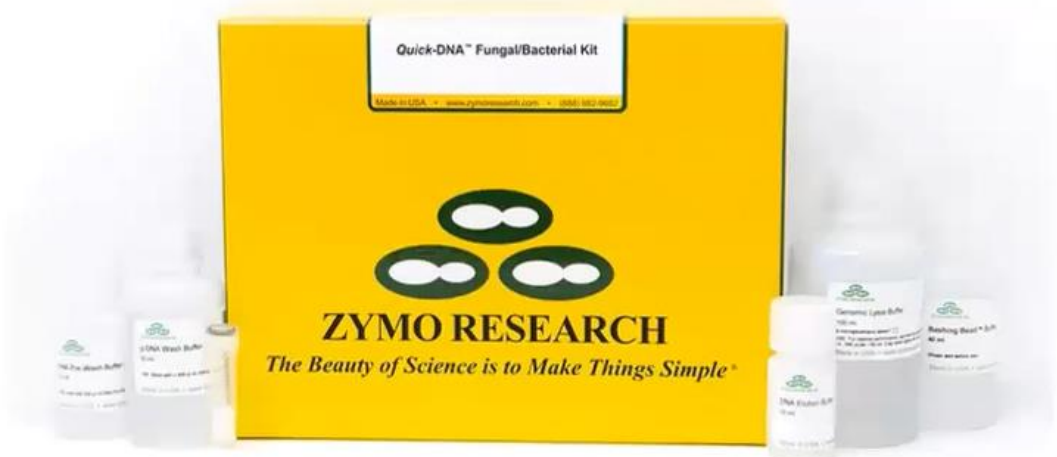
Sequencing capacity

- Some labs have no or minimal in-house wet-lab or dry-lab capacity
- Some labs have wet-lab and dry-lab capacity, however lack high throughput
- Some labs have high throughput wet-lab and dry lab capacities (e.g., sequencing instruments, access to computing environments, experience with genomic epi investigations)

https://www.cdc.gov/advanced-molecular-detection/php/training/module-3-6.html

Short-read sequencing by Illumina

DNA extraction



Library prep

MiSeq sequencing

MiSeq® Reagent Kit v2 (500 cycle)

Adapted from StaPH-B Monthly Webinar – August, 2020: https://www.youtube.com/embed/C3YCm3WOY0A?autoplay=

Long-read sequencing by PacBio

[Candida] auris
Representative genome: [Candida] auris (assembly Cand_auris_B11221_V1)
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against [Candida] auris **genome**, **transcript**, **protein**
All 59 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**
NEW Try **NCBI Datasets** - a new way to download genome sequence and annotation we're testing in NCBI Labs

Display Settings: Overview Send to:

Organism Overview ; Genome Assembly and Annotation report [59] ; Organelle Annotation Report [1] ID: 38761

[Candida] auris

An emerging multidrug-resistant (MDR) yeast

Lineage: Eukaryota[5270]; Fungi[2289]; Dikarya[2152]; Ascomycota[1687]; Saccharomycotina[463]; Saccharomycetes[463]; Saccharomycetales[463]; Metschnikowiaceae[65]; Clavispora[10]; Clavispora/Candida clade[8]; [Candida] auris[1]
This yeast, is causing invasive healthcare-associated infections with high mortality. Some strains have elevated minimum inhibitory concentrations (MICs) to the three major classes of antifungals, severely limiting treatment options.

Summary

Sequence data: genome assemblies: 59 (See Genome Assembly and Annotation report)
Statistics: median total length (Mb): 12.3603
median protein count: 5437
median GC%: 45.1249

Publications (limited to 20 most recent records)

1. Simultaneous Infection with *Enterobacteriaceae* and Pseudomonas aeruginosa Harboring Multiple Carbapenemases in a Returning Traveler Colonized with Candida auris. Khan A, et al. Antimicrob Agents Chemother 2020 Jan 27
2. Clade II Candida auris possess genomic structural variations related to an ancestral strain. Sekizuka T, et al. PLoS One 2019
3. Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species. Muñoz JF, et al. Nat Commun 2018 Dec 17

More...

PacBio RSII

Adapted from StaPH-B Monthly Webinar – August, 2020: https://www.youtube.com/embed/C3YCm3WOY0A?autopl

Long-read sequencing by Oxford Nanopore Technologies

ASM827514v1

Organism name: [Candida] auris (budding yeasts)
Infraspecific name: Strain: B11245
BioSample: SAMN05379621
BioProject: PRJNA328792
Submitter: Centers for Disease Control and Prevention
Date: 2019/09/05
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_008275145.1 (latest)
RefSeq assembly accession: n/a
RefSeq assembly and GenBank assembly identical: n/a
Assembly method: Canu v. 1.5
Genome coverage: 88
Sequencing technology: Oxford_Nanopore

IDs: 4513771 [UID] 13609208 [GenBank]

History (Show revision history)

Global statistics

Total sequence length	12,431,639
Total ungapped length	12,431,639
Total number of chromosomes and plasmids	7

Adapted from StaPH-B Monthly Webinar – August, 2020: https://www.youtube.com/embed/C3YCm3WOY0A?autoplay=1

Recall: the data we get from sequencing

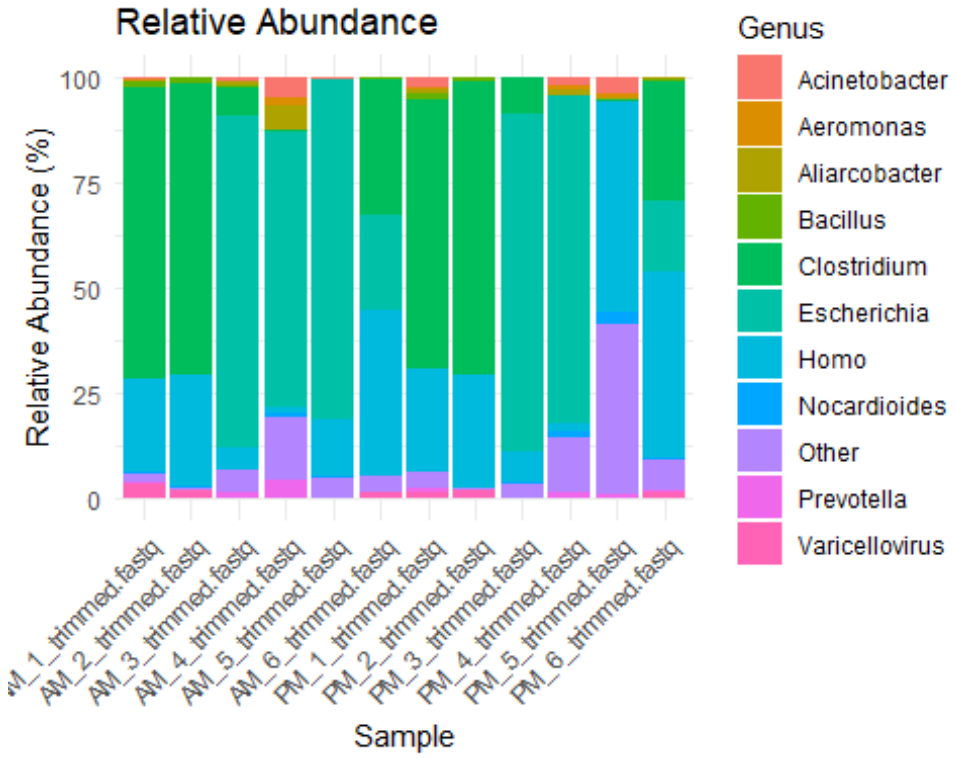
File name: sample1_R1.fastq.gz and sample1_R2.fastq.gz

	FASTQ	FASTA
Use case	Raw sequencing reads	Assembled sequences (contigs, scaffolds)
Contains	Nucleotide sequence and quality scores	Nucleotide sequence only
Format	4-line structure per read	2-line structure per sequence
Used for	Read-level QC, trimming, alignment	Assemblies, reference databases

- After sequencing, we get a FASTQ file
- These FASTQ files can be used as input for our analyses

Assembly

What do we do with these short sequences?

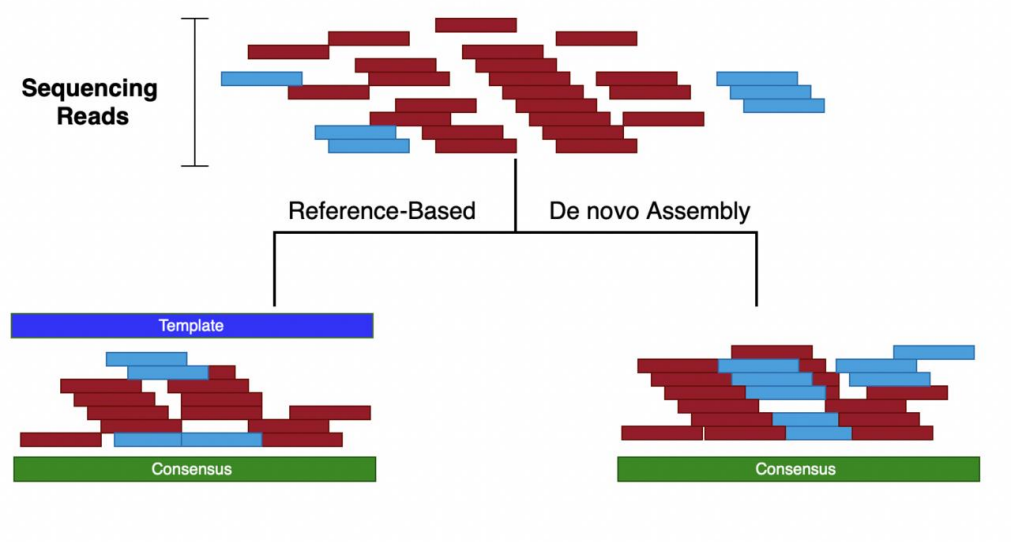


- Since we can't sequence the full genome end-to-end due to technology limitations, we end up with a collection of smaller sequences, called **reads**

- These reads can be used as data for analyses directly, or they can be assembled into **contigs** or **genomes**

Assembly

Strand	Sequence
Original	AGCATGCTGCAGTCATGCTTAGGCTA
First shotgun sequence	AGCATGCTGCAGTCATGCT------- -------------------TAGGCTA
Second shotgun sequence	AGCATG-------------------- ------CTGCAGTCATGCTTAGGCTA
Reconstruction	AGCATGCTGCAGTCATGCTTAGGCTA



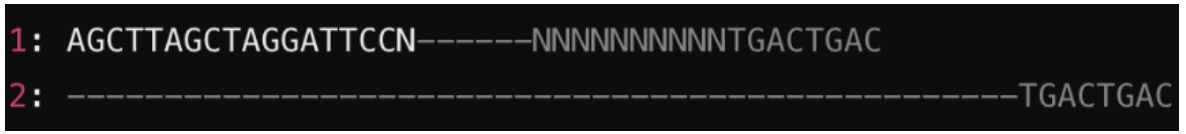
- Multiple short reads can be combined into a longer sequence
- By matching up sections of sequence common across reads, we can link them together
- This is done through genome assembly software like SPAdes

Reference-based assembly

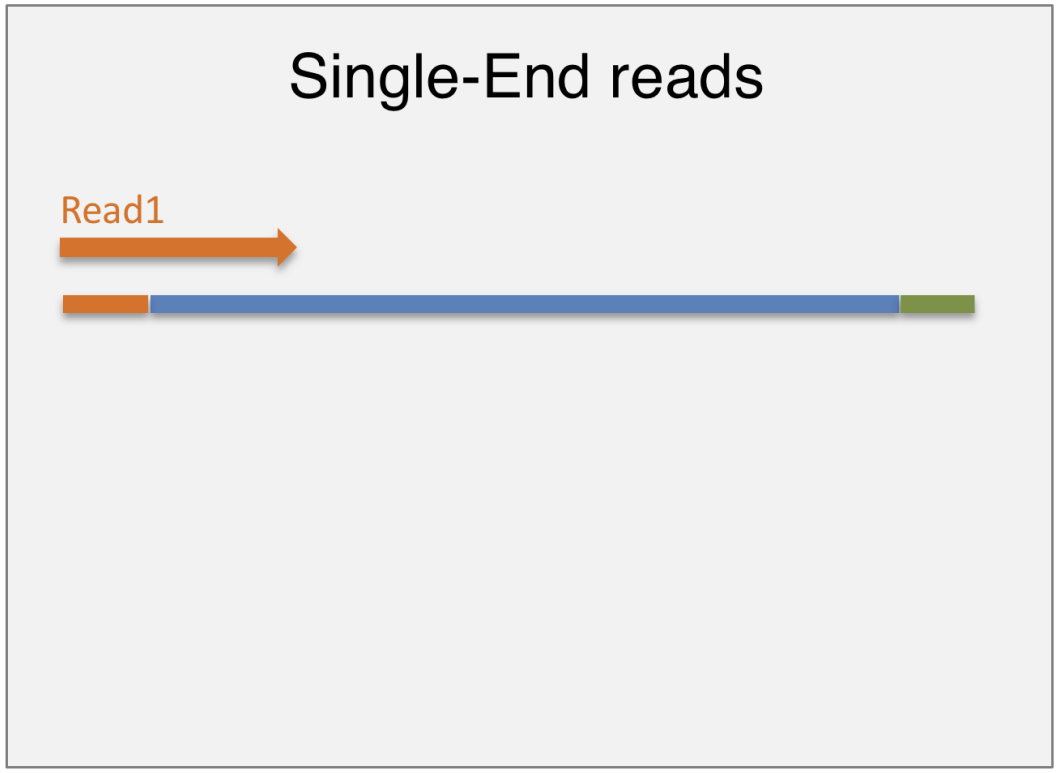
De novo assembly

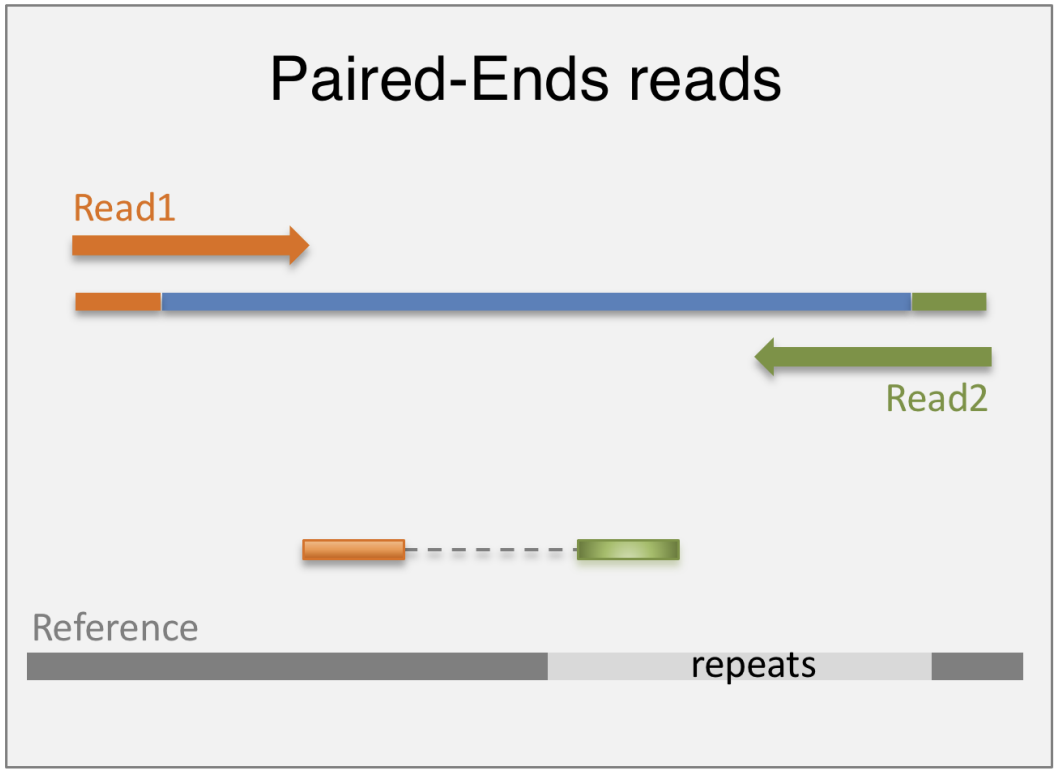
Contigs and scaffolds

- To make sense of our data, we stitch together overlapping reads into contiguous sequences without gaps, called **contigs**
  - Contigs may represent genes or partial genomes
  - The longer the contig, the better the assembly
- **Scaffolds** are a collection of contigs ordered and oriented using paired-end reads or long-read sequences with gaps between them
  - Helps reconstruct larger genome structures
  - Connected by estimated gaps
  - More complete, but often missing sequences



Single vs. Paired-end reads





Single vs. Paired-end reads

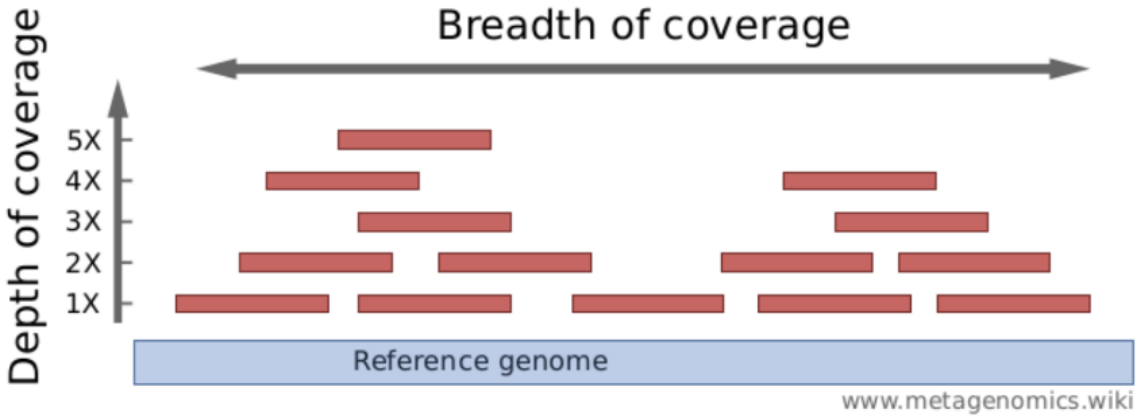
Single-end reads

- Each fragment is sequenced from one end
- Benefits: low cost and more computationally efficient
- Downsides: more difficult to assemble and lower accuracy, particularly for more complex samples like wastewater

Paired-end reads

- Each fragment is sequenced from both ends, generating 2 reads per fragment
- Benefits: better assembly and more accurate assignment of taxonomy
- Downsides: more expensive and more time consuming

How do we know if our data is any good?



Calculating optimal sample size:

$$Sequencing\ Coverage = \frac{Total\ \#\ bases\ mapped\ to\ a\ region}{Length\ of\ region}$$

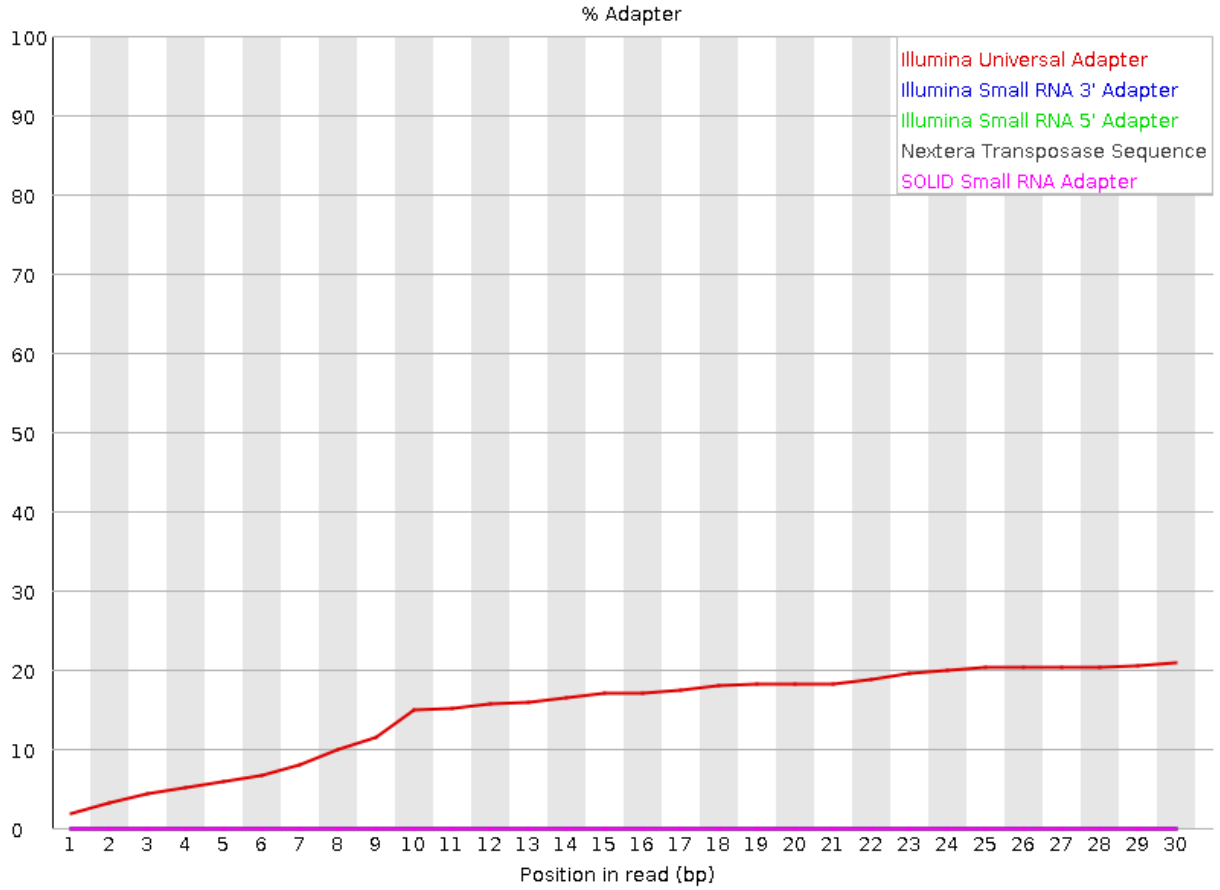
e.g. if you sequence 30 billion bases (Gb) of data for a 3 Gb hum

$$Sequencing\ Depth = \frac{(Average\ genome\ size * desired\ coverage)}{Average\ microbial\ abundance}$$

- Two key metrics: coverage and depth
- Coverage is the percentage of the genome sequenced
- Depth is the number of times we have sequenced a specific location of the genome (usually averaged over the entire genome)

FastQC for assessing sequencing read quality

How do we know we need to trim our sequences?



- The plots generated with FastQC offer a clue!
- The “Adapter Content” plot will show several lines. If any of these raise significantly above 0, it indicates trimming is needed.

Trimming and removing adapter contamination

- Read trimming is important to improve read quality
- Trimmomatic, while customizable, provides an easy-to-use first pass at read trimming
  - Illuminaclip: specifies the path to adapters, PCR sequences, etc. for trimming. These are included with the tool, but experts may have custom files for this
  - MINLEN: the minimum length of reads which we'll keep, anything trimmed to shorter is removed

Assembly QC with QUAST

- Before moving forward with our assemblies, we need to know if they’re any good!
- QUAST generates several useful metrics for assemblies including:
  - Number of contigs greater than 0 bp: ideally, your assembly should have as few pieces as possible
  - Total assembly length: ideally, this should be close to the expected genome size for your target organism
  - N50 of the assembly, the average length of largest contigs accounting for 50% of the genome size: higher numbers are better!

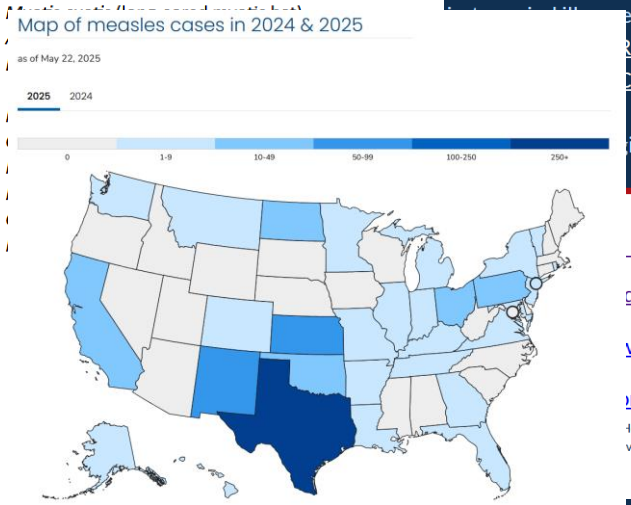
Data repositories

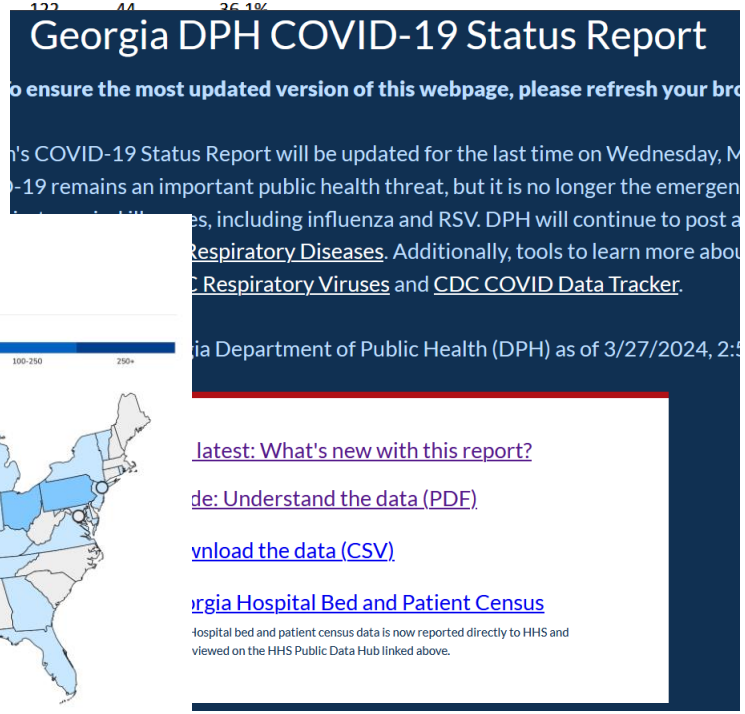
De-identified data is often publicly available to download

- CDC and health department websites
  - Dashboards may have “Download data” link under figure
- US Census for population estimates
- USDA for livestock case counts
- Supplementary materials in journal articles
- Data requests for additional data (if available)

Supplementary Table S1—Species of bats submitted for rabies testing in the United States during 2022.

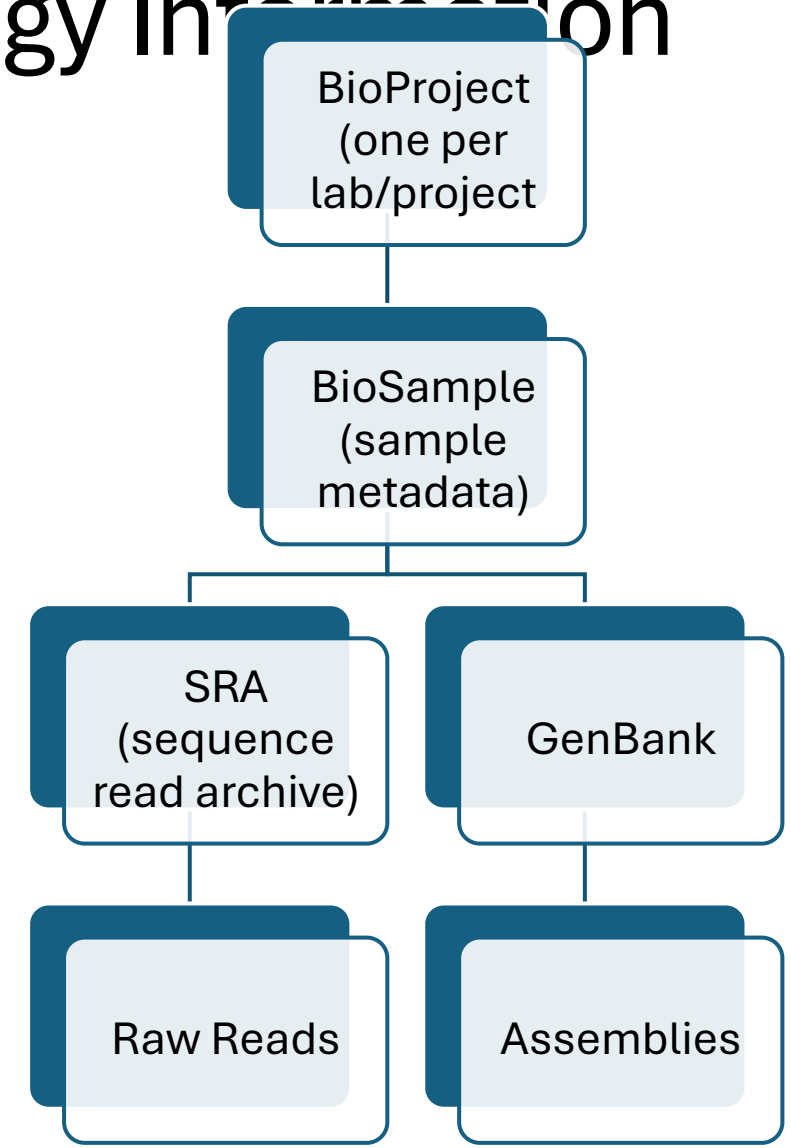
Species (common name)	No. of tested	No. of positive	Percentage of positive
Order Chiroptera (unspecified)	11,883	490	4.1%
Eptesicus fuscus (big brown bat)	8,955	351	3.9%
Tadarida brasiliensis (Mexican free-tailed bat)	734	189	25.7%
Nycticeius humeralis (evening bat)	427	16	3.7%
Lasiurus borealis (Eastern red bat)	335	43	12.8%
Myotis lucifugus (little brown bat)	237	8	3.4%
Myotis spp. (not further differentiated)	132	15	11.4%
Lasionycteris noctivagans (silver-haired bat)	128	8	6.3%
Parastrellus hesperus (canyon bat)	122	44	36.1%
Myotis californicus (California myotis bat)			
Lasiurus seminolus (seminole bat)			
Lasiurus cinereus (hoary bat)			
Myotis yumanensis (Yuma myotis bat)			
Lasiurus ega (southern yellow bat)			
Myotis austroriparius (southeastern myotis bat)			





National Center for Biotechnology Information (NCBI)

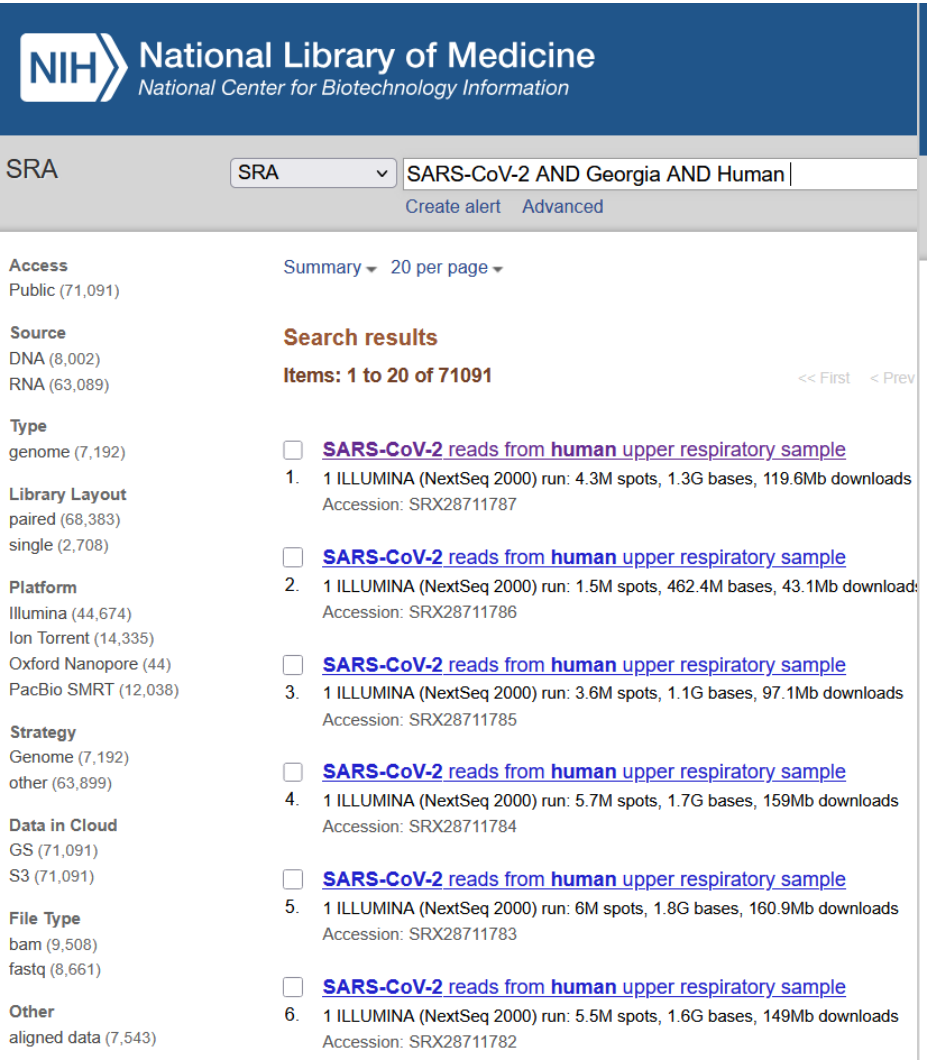
- Part of the International Nucleotide Sequence Database Collaboration (INSDC)
- Includes reads (SRA) and assemblies (GenBank)
- NCBI datasets is a new feature for easier searching: https://www.ncbi.nlm.nih.gov/datasets/genome/

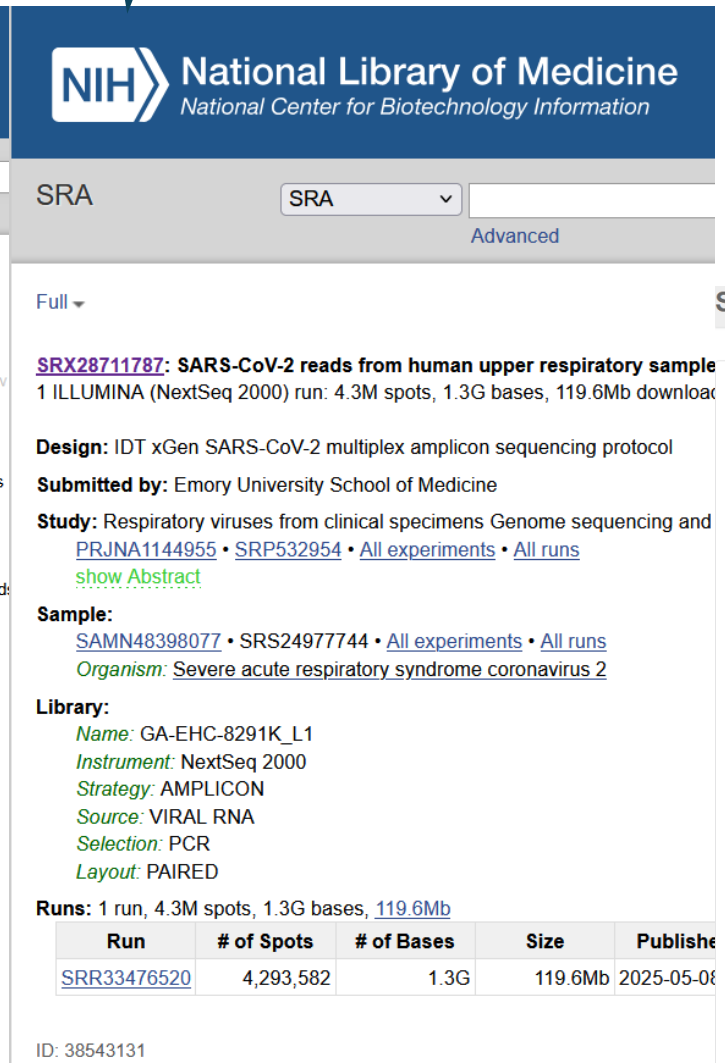


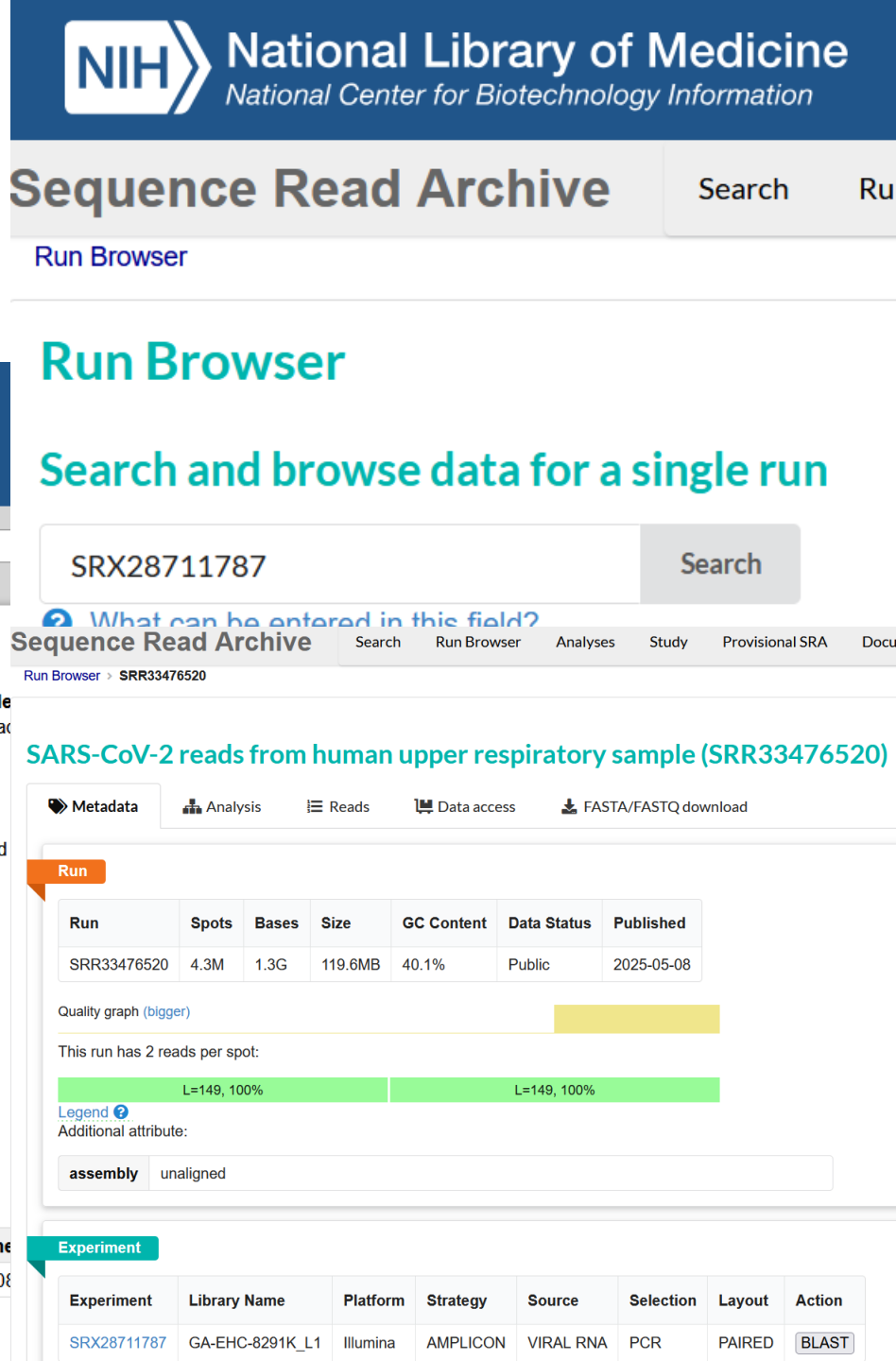
NIH National Library of Medicine
National Center for Biotechnology Information

NCBI SRA

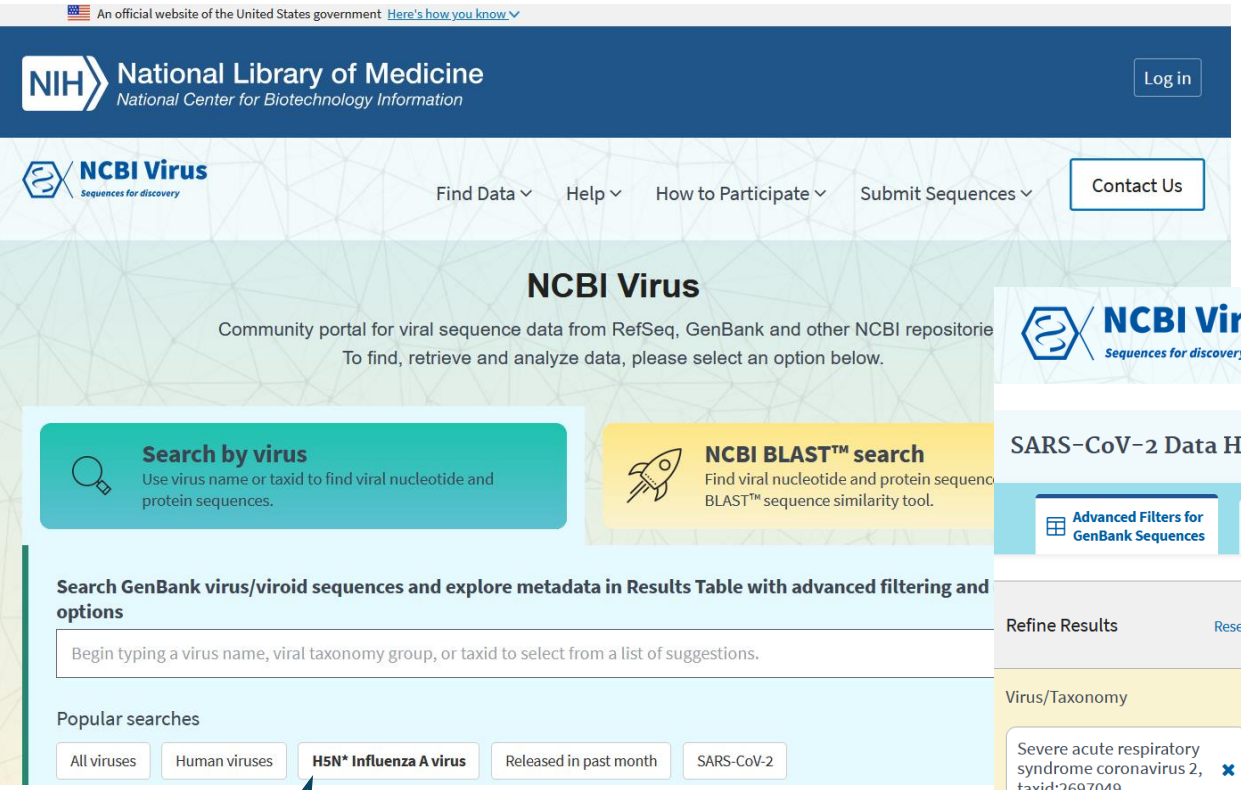
These are big files...you can only download them one at a time via the web interface

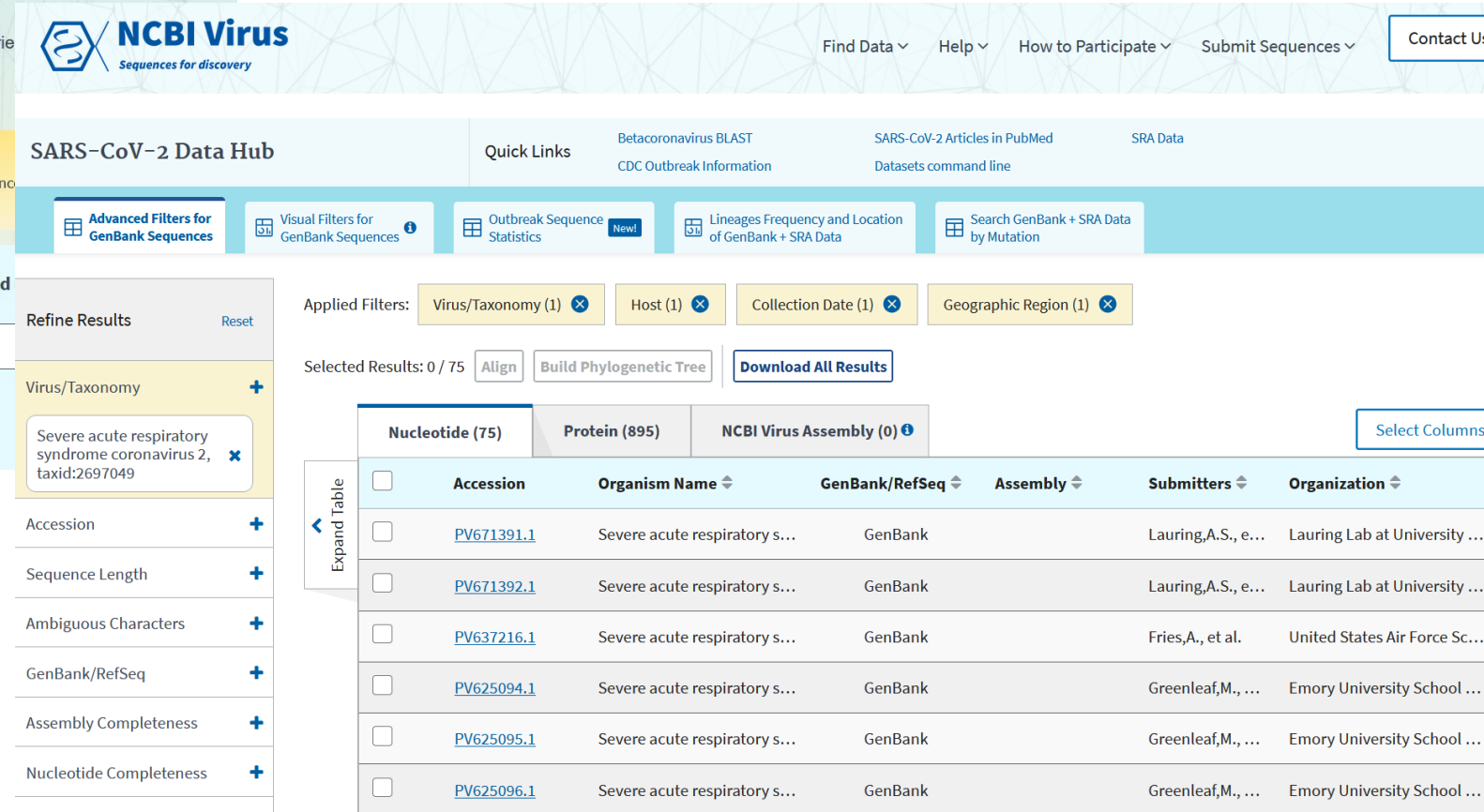






NCBI GenBank





You can use the new feature, NCBI Datasets, to download any pathogen

BACTERIAL AND VIRAL BIOINFORMATICS RESOURCE CENTER

Welcome to the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), a comprehensive resource for bacterial and viral infectious disease research. BV-BRC provides integrated data, advanced bioinformatics tools, and workflows to support the scientific community in understanding and combating infectious diseases. Explore our **Quick Start Guide** to learn more about the platform's features and capabilities. For questions or assistance, please don't hesitate to **contact us**.

SEARCH All Data Types *Find a gene, genome, microarray, etc* All terms

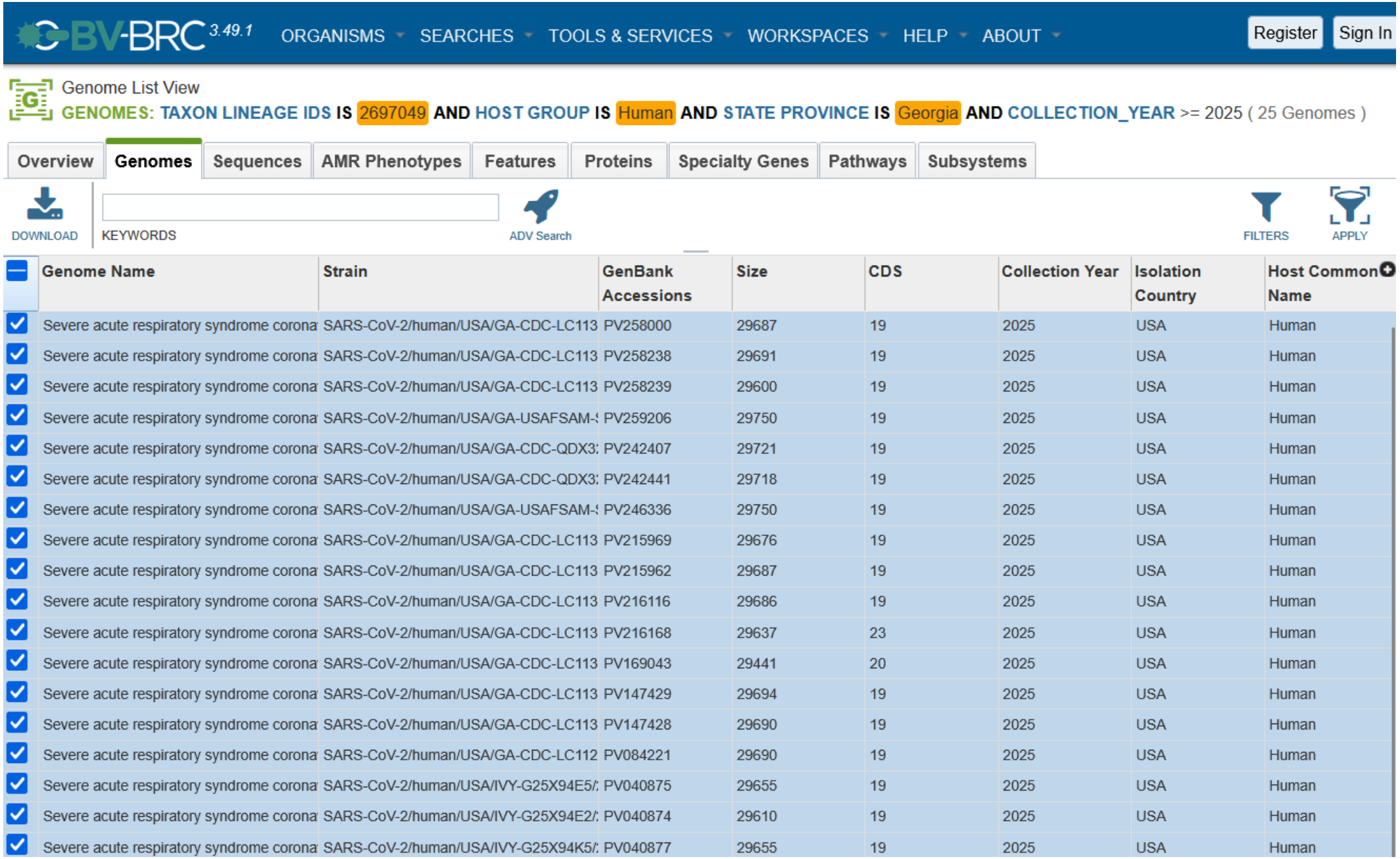
BROWSE BACTERIA ARCHAEA VIRUSES EUKARYOTIC HOSTS

ANALYZE DATA IN BV-BRC

Upload and analyze your data in the private workspace. **Register** or **Login** to get started.

SEARCH	ANALYZE	MANAGE DATA	BATCH ACCESS
Taxa	Assembly	Access Private Workspace	Command-Line Interface
Genomes	Annotation	Save Search Results	Data API
Proteins	BLAST	Upload Data	FTP
Specialty Genes	MSA	Access Analysis Jobs	
All Searches...	All Tools & Services...	Share and Publish	

QUICK START TUTORIALS REFERENCE GUIDES



Genome Name	Strain	GenBank Accessions	Size	CDS	Collection Year	Isolation Country	Host Common Name
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV258000	29687	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV258238	29691	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV258239	29600	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-USAFSAM-S	PV259206	29750	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-QDX3:	PV242407	29721	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-QDX3:	PV242441	29718	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-USAFSAM-S	PV246336	29750	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV215969	29676	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV215962	29687	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV216116	29686	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV216168	29637	23	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV169043	29441	20	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV147429	29694	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC113	PV147428	29690	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/GA-CDC-LC112	PV084221	29690	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/IVY-G25X94E5/	PV040875	29655	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/IVY-G25X94E2/	PV040874	29610	19	2025	USA	Human
Severe acute respiratory syndrome corona	SARS-CoV-2/human/USA/IVY-G25X94K5/	PV040877	29655	19	2025	USA	Human

GISAID has several databases, with different features for each

- EpiFlu
- EpiCoV
- EpiRSV
- EpiPox