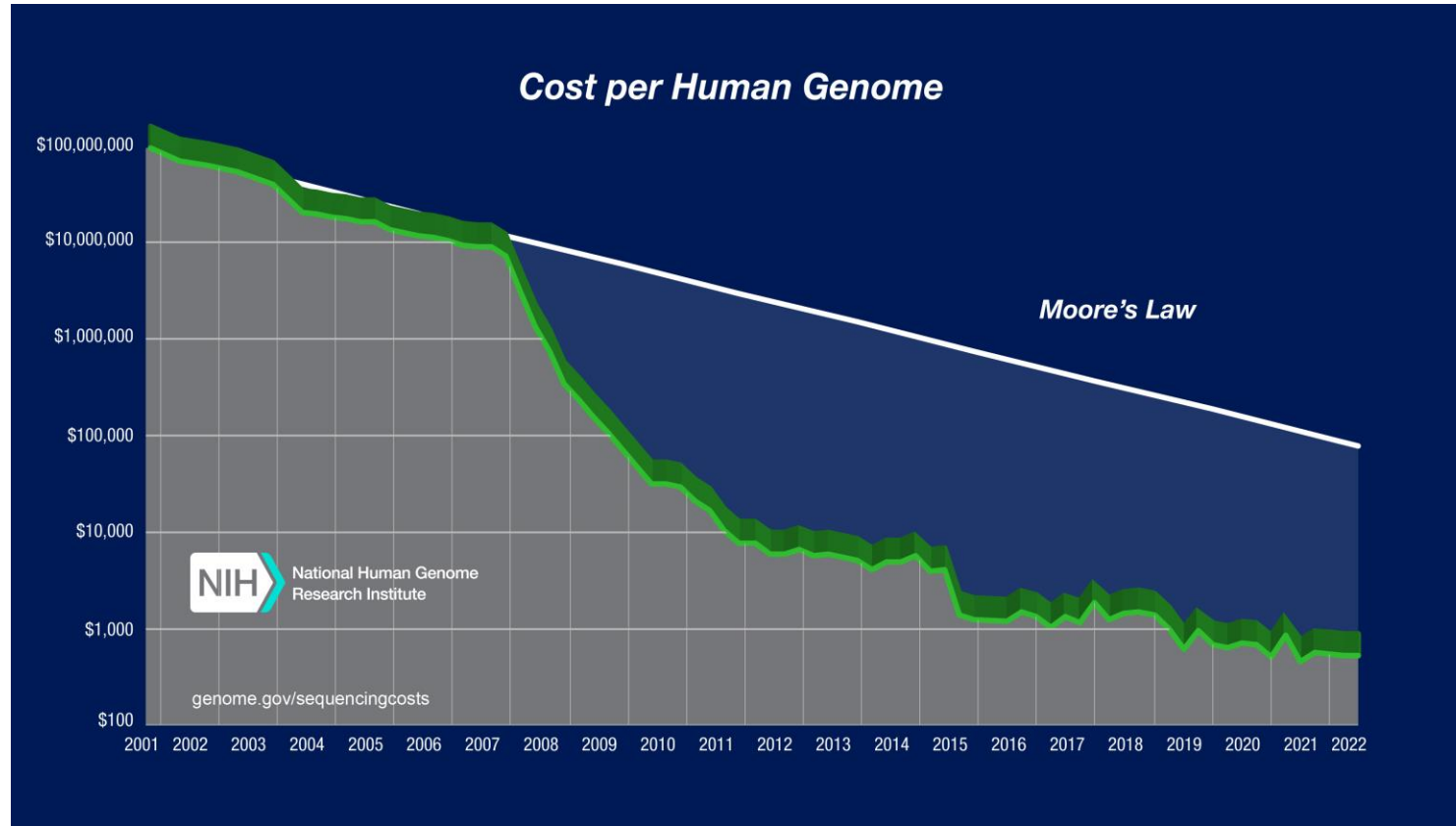


Session 3

Genomes, reads, assemblies,
and data repositories

Sequencing

Sequencing costs have fallen dramatically over the past couple of decades

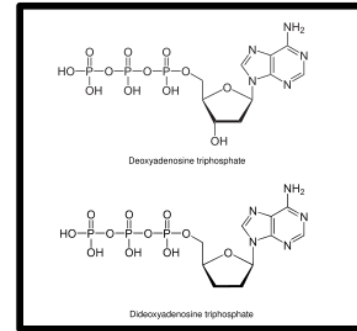
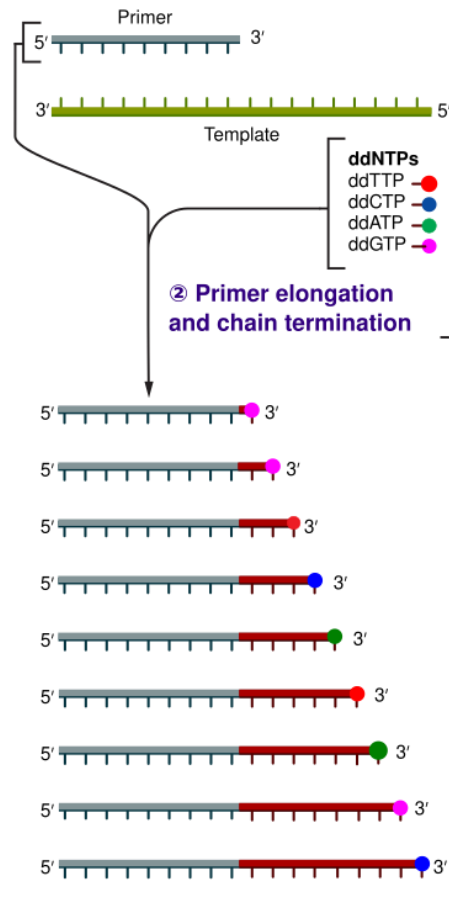


- The cost of sequencing varies by organism, lab, method, etc.
- The cost of a human genome is a common metric

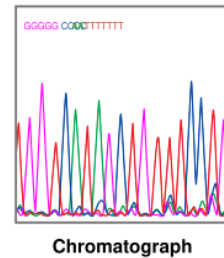
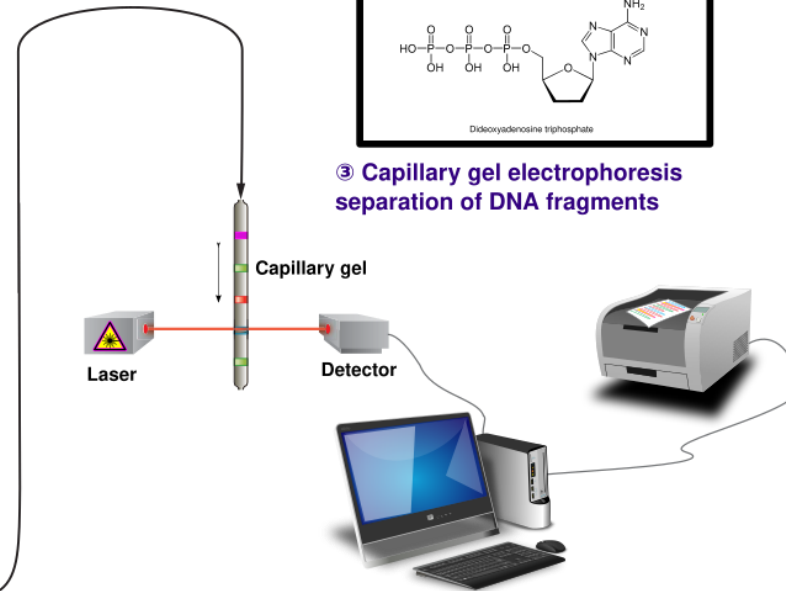
Sanger sequencing

① Reaction mixture

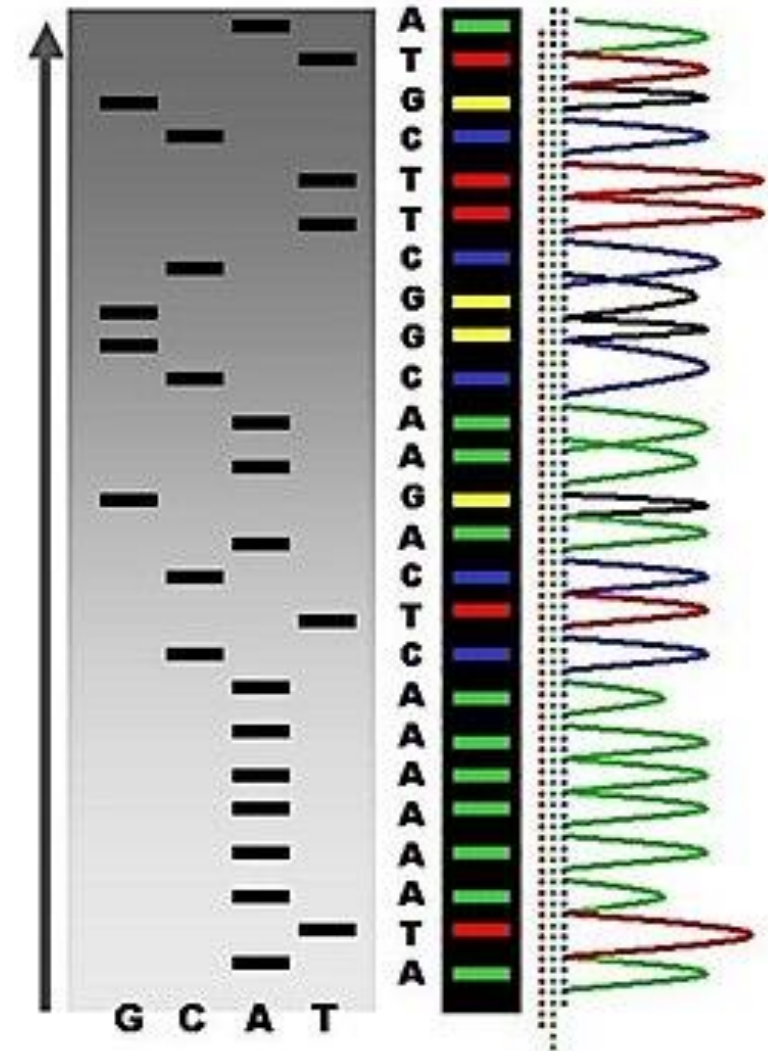
- ▶ Primer and DNA template
- ▶ DNA polymerase
- ▶ ddNTPs with flourochromes
- ▶ dNTPs (dATP, dCTP, dGTP, and dTTP)



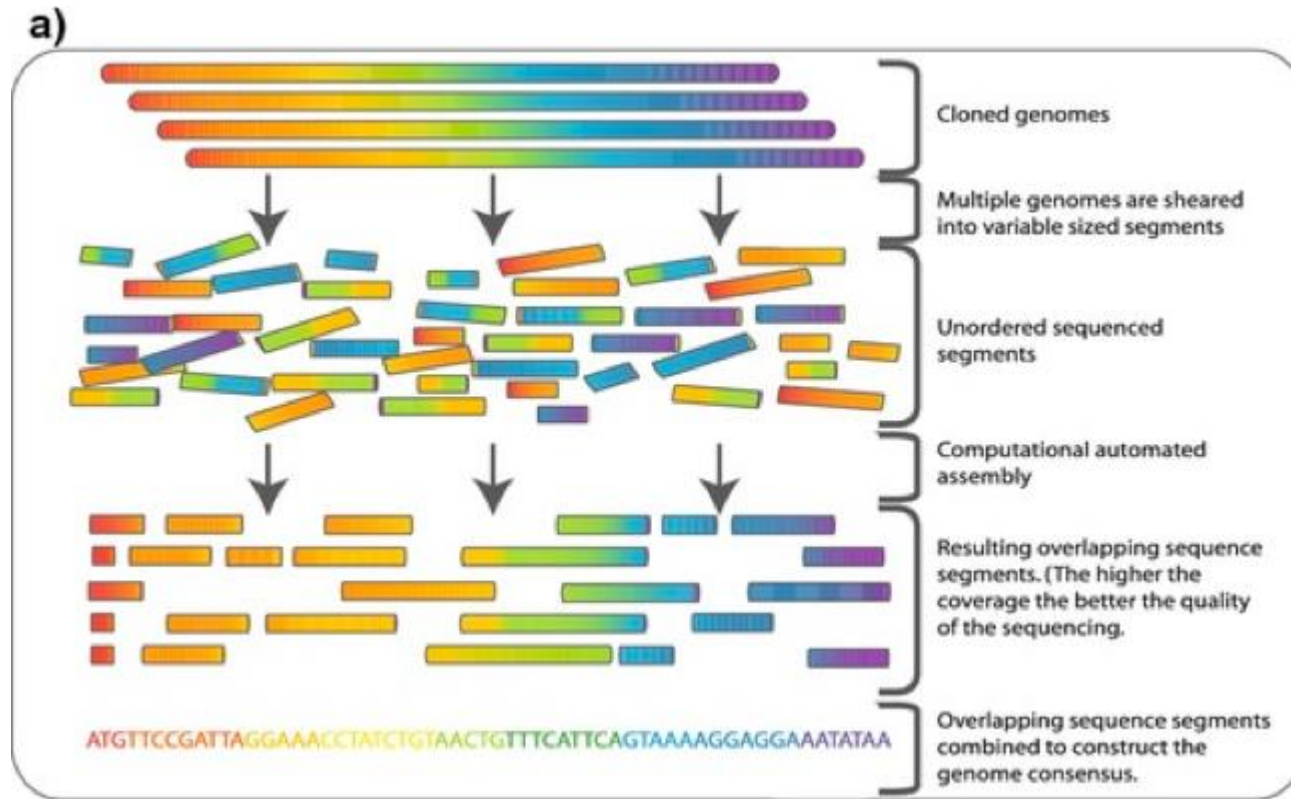
③ Capillary gel electrophoresis separation of DNA fragments



④ Laser detection of flouorochromes and computational sequence analysis



Shotgun sequencing



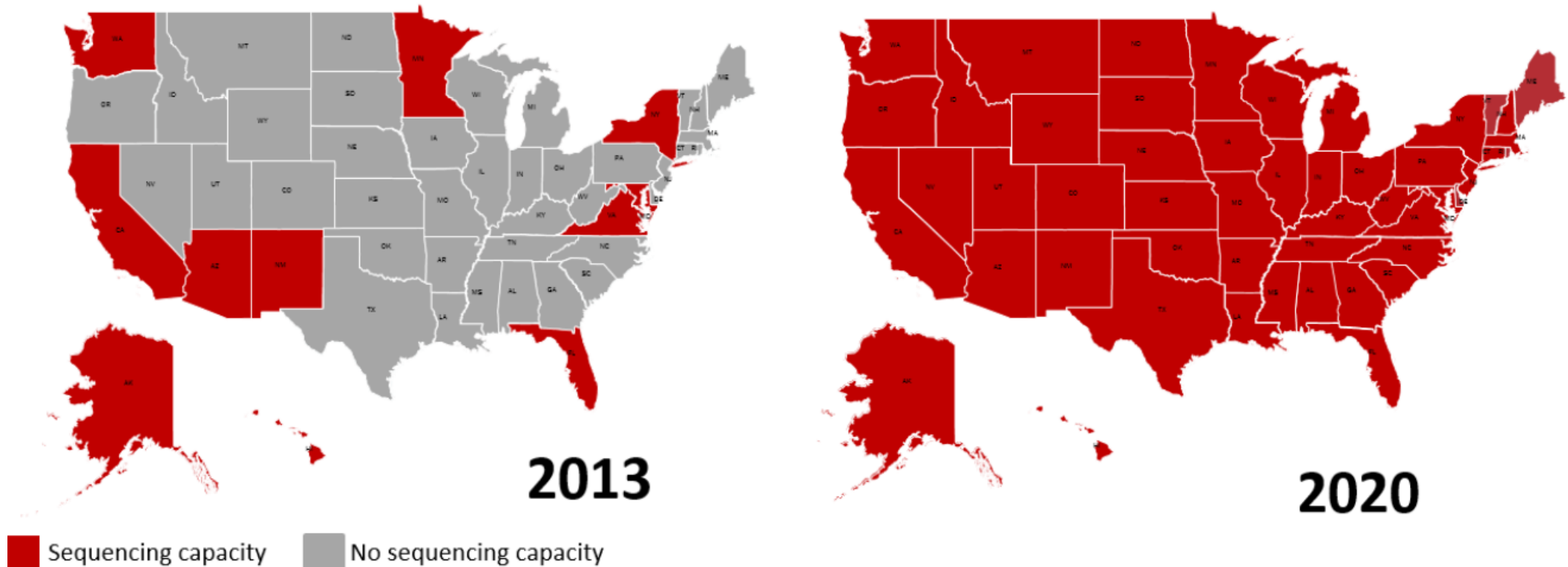
- Randomly chop up the genome, then perform Sanger sequencing
- Sanger sequencing can only be used for short sequences 100-1000bp long

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 States

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>

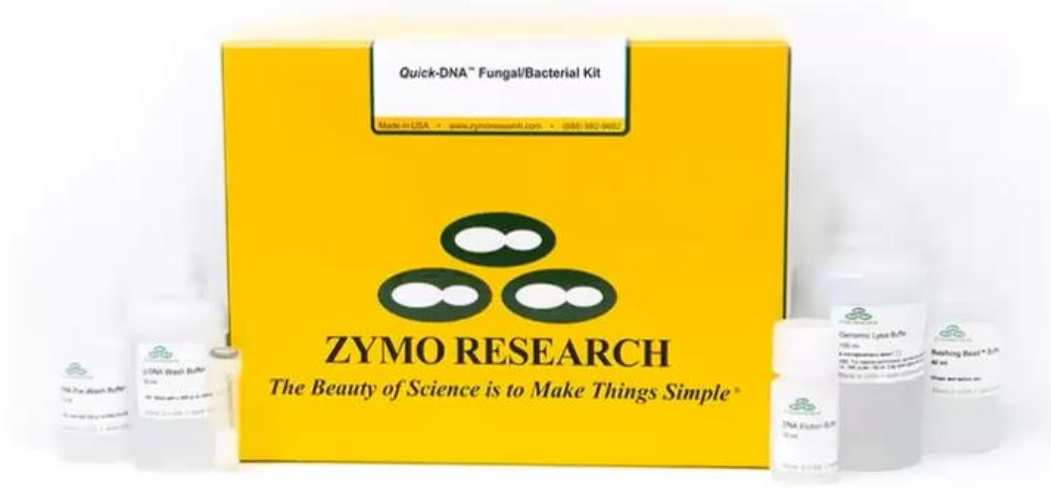
<https://www.cdc.gov/advanced-molecular-detection/php/training/module-3-6.html>

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)

Short-read sequencing by Illumina

DNA extraction



MiSeq sequencing

MiSeq[®] Reagent Kit v2 (500 cycle)



Library prep



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?autoplay=1>

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

Recall: the data we get from sequencing

```
@SRR123456.1 length=150  
AGCTTAGCTAGCTAGCTAGCTAGCTAGC  
+  
! '*((( (***) )%%%++) (%%%) .1****
```

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

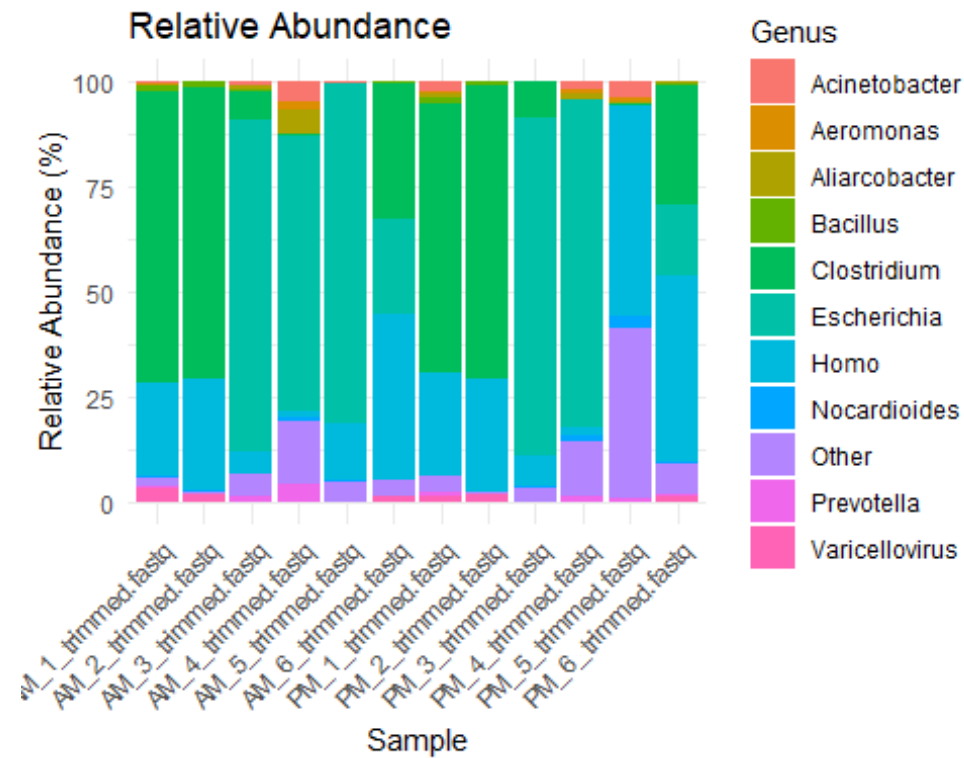
Read identifier
Nucleotide sequence
Quality scores (ASCII-encoded, reflects base call confidence)

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

	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

Assembly

What do we do with these short sequences?

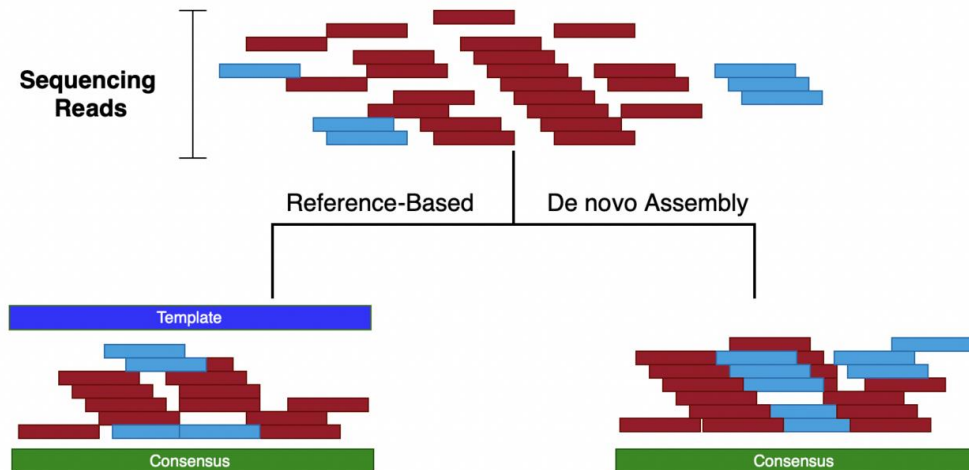


- Since we can't sequence the full genome end-to-end due to technology limitations, we end up with a collection of smaller sequences, called **reads**
- These reads can be used as data for analyses directly, or they can be assembled into **contigs** or **genomes**

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



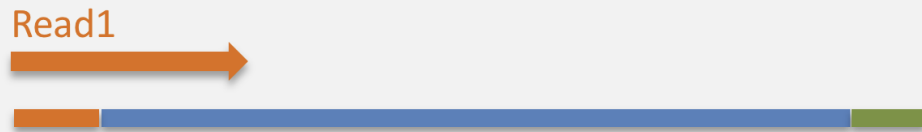
Contigs and scaffolds

- 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

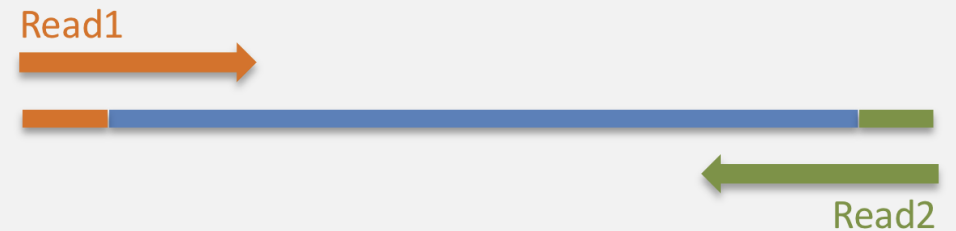
```
1: AGCTTAGCTAGGATTCCN-----NNNNNNNNNTGACTGAC
2: -----TGACTGAC
```

Single vs. Paired-end reads

Single-End reads



Paired-Ends reads



Reference

repeats



Single vs. Paired-end reads

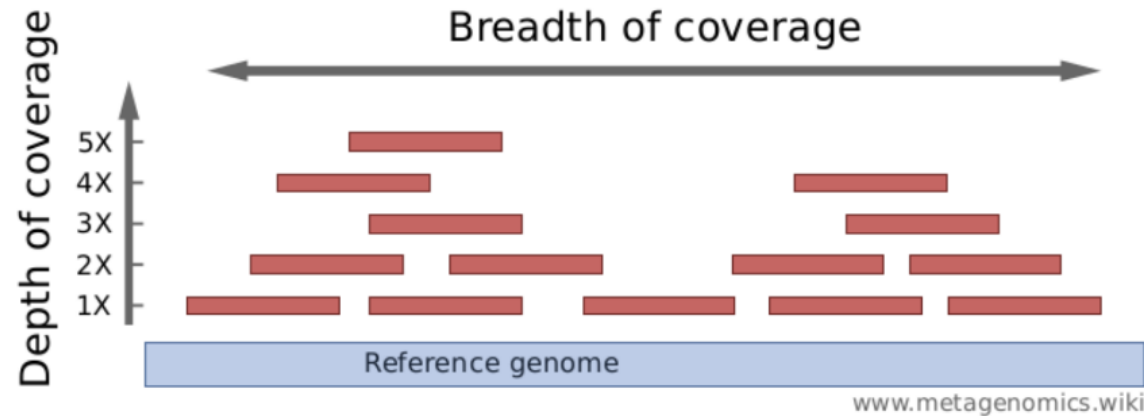
Single-end reads

- 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:

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

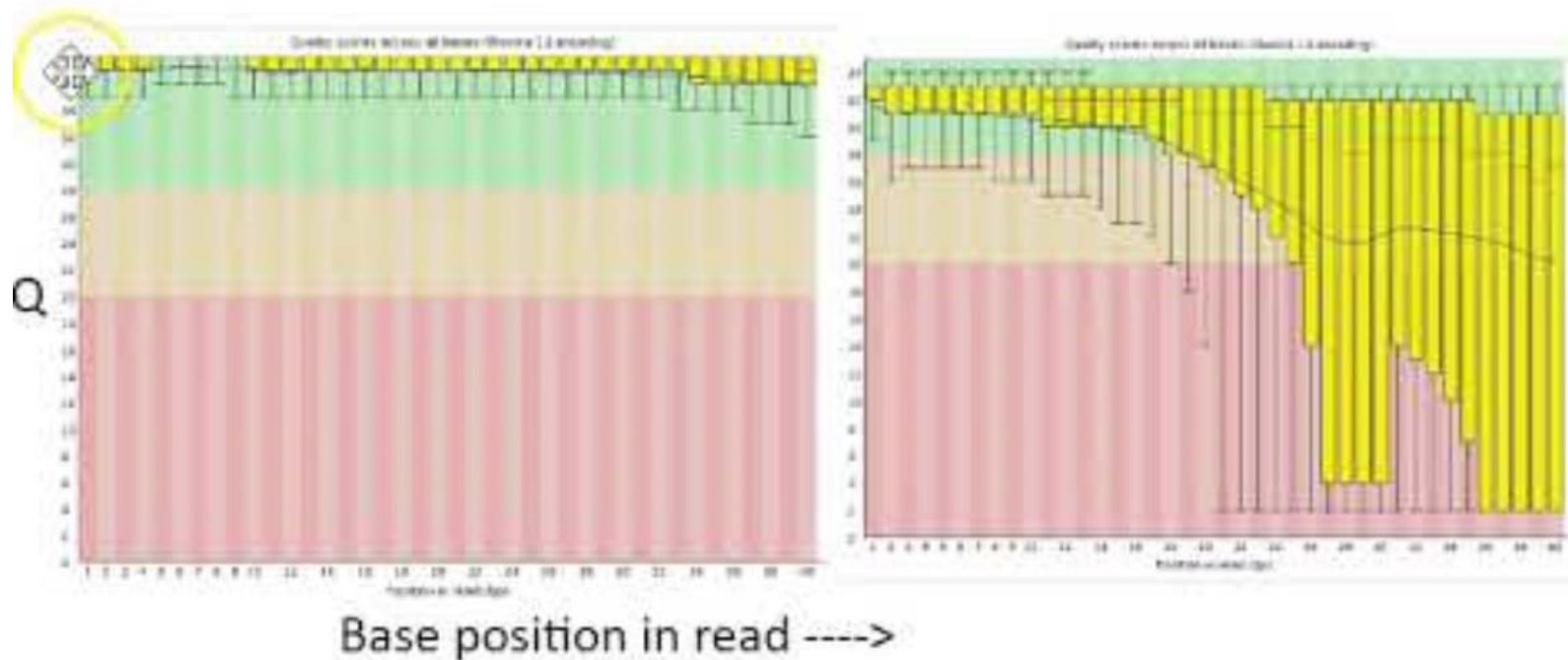
e.g. if you sequence 30 billion bases (Gb)
of data for a 3 Gb hum

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

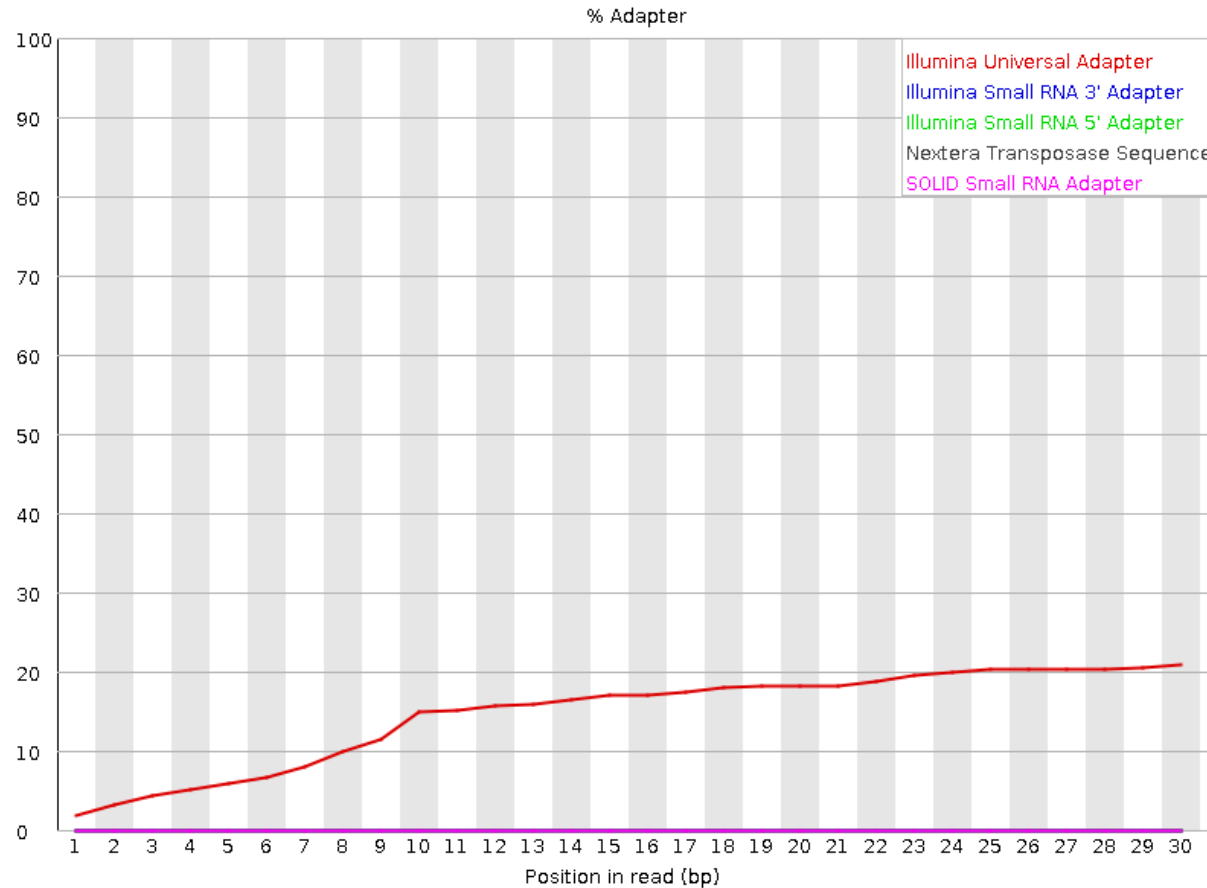
- Two key metrics: coverage and depth
- Coverage is the percentage of the genome sequenced
- Depth is the number of times we have sequenced a specific location of the genome (usually averaged over the entire genome)

FastQC for assessing sequencing read quality

Good vs. Bad. Box Whisker Plots of Per Base Sequence 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
 - Illuminaclic: 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!

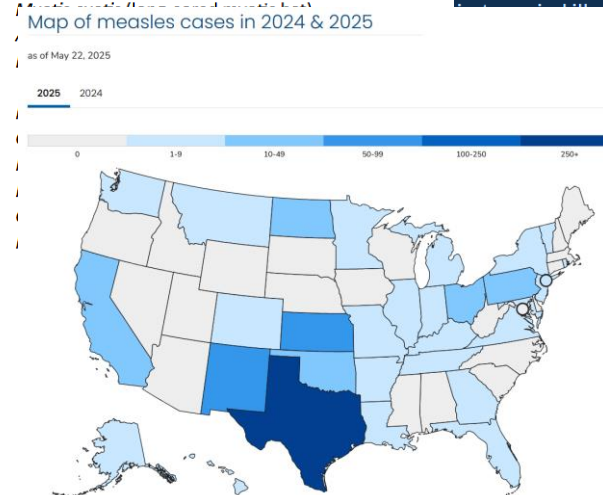
Data repositories

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
<i>Order Chiroptera</i> (unspecified)	11,883	490	4.1%
<i>Eptesicus fuscus</i> (big brown bat)	8,955	351	3.9%
<i>Tadarida brasiliensis</i> (Mexican free-tailed bat)	734	189	25.7%
<i>Nycticeius humeralis</i> (evening bat)	427	16	3.7%
<i>Lasiurus borealis</i> (Eastern red bat)	335	43	12.8%
<i>Myotis lucifugus</i> (little brown bat)	237	8	3.4%
<i>Myotis</i> spp. (not further differentiated)	132	15	11.4%
<i>Lasionycteris noctivagans</i> (silver-haired bat)	128	8	6.3%
<i>Parastrellus hesperus</i> (canyon bat)	122	44	36.1%
<i>Myotis californicus</i> (California myotis bat)			
<i>Lasiurus seminolus</i> (seminole bat)			
<i>Lasiurus cinereus</i> (hoary bat)			
<i>Myotis yumanensis</i> (Yuma myotis bat)			
<i>Lasiurus ega</i> (southern yellow bat)			
<i>Myotis austroriparius</i> (southeastern myotis bat)			



Georgia DPH COVID-19 Status Report

To ensure the most updated version of this webpage, please refresh your browser.

h's COVID-19 Status Report will be updated for the last time on Wednesday, M
-19 remains an important public health threat, but it is no longer the emergen
including influenza and RSV. DPH will continue to post a
[Respiratory Diseases](#). Additionally, tools to learn more abou
[Respiratory Viruses](#) and [CDC COVID Data Tracker](#).

via Department of Public Health (DPH) as of 3/27/2024, 2:5

latest: What's new with this report?

[de: Understand the data \(PDF\)](#)

[vload the data \(CSV\)](#)

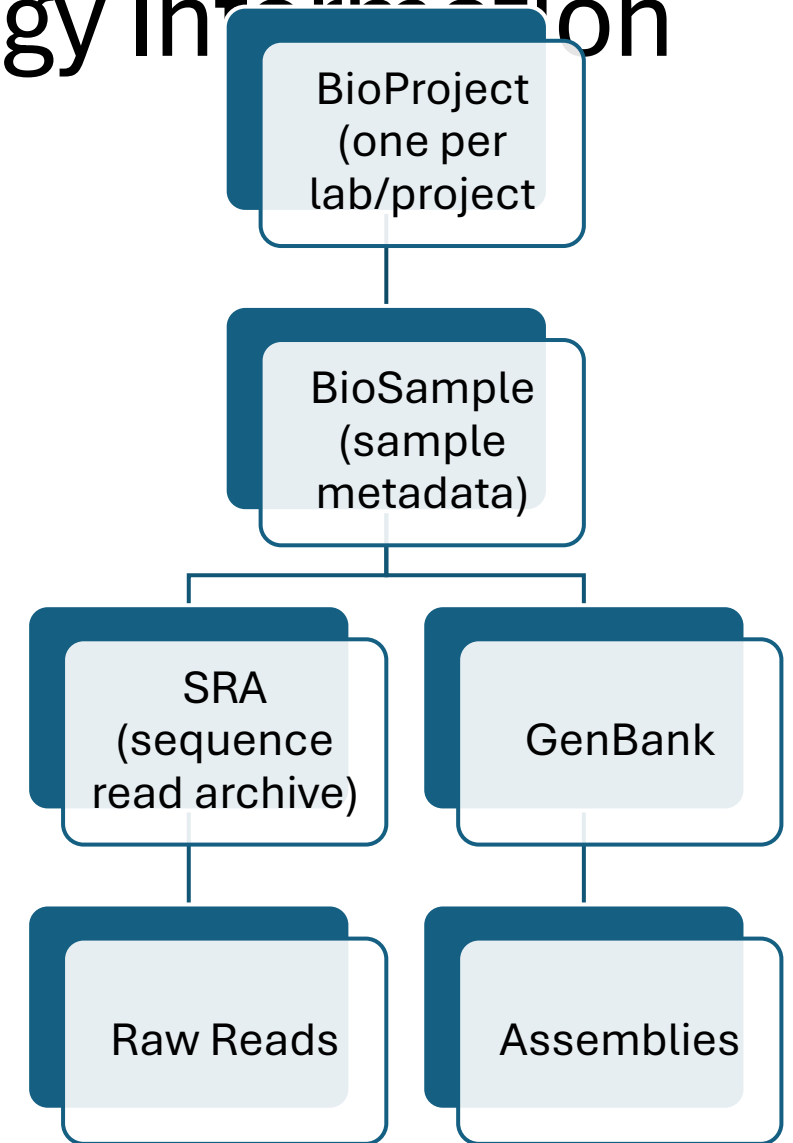
Georgia Hospital Bed and Patient Census

Hospital bed and patient census data is now reported directly to HHS and viewed on the HHS Public Data Hub linked above.

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/>



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

 **National Library of Medicine**
National Center for Biotechnology Information

SRA

SRA

SARS-CoV-2 AND Georgia AND Human

Create alert Advanced

Access
Public (71,091)

Source
DNA (8,002)
RNA (63,089)

Type
genome (7,192)

Library Layout
paired (68,383)
single (2,708)

Platform
Illumina (44,674)
Ion Torrent (14,335)
Oxford Nanopore (44)
PacBio SMRT (12,038)

Strategy
Genome (7,192)
other (63,899)

Data in Cloud
GS (71,091)
S3 (71,091)

File Type
bam (9,508)
fastq (8,661)

Other
aligned data (7,543)

Summary 20 per page

Search results

Items: 1 to 20 of 71091

☐ [SARS-CoV-2 reads from human upper respiratory sample](#)

1. 1 ILLUMINA (NextSeq 2000) run: 4.3M spots, 1.3G bases, 119.6Mb downloads
Accession: SRX28711787

☐ [SARS-CoV-2 reads from human upper respiratory sample](#)

2. 1 ILLUMINA (NextSeq 2000) run: 1.5M spots, 462.4M bases, 43.1Mb downloads
Accession: SRX28711786

☐ [SARS-CoV-2 reads from human upper respiratory sample](#)

3. 1 ILLUMINA (NextSeq 2000) run: 3.6M spots, 1.1G bases, 97.1Mb downloads
Accession: SRX28711785

☐ [SARS-CoV-2 reads from human upper respiratory sample](#)

4. 1 ILLUMINA (NextSeq 2000) run: 5.7M spots, 1.7G bases, 159Mb downloads
Accession: SRX28711784

☐ [SARS-CoV-2 reads from human upper respiratory sample](#)

5. 1 ILLUMINA (NextSeq 2000) run: 6M spots, 1.8G bases, 160.9Mb downloads
Accession: SRX28711783

☐ [SARS-CoV-2 reads from human upper respiratory sample](#)

6. 1 ILLUMINA (NextSeq 2000) run: 5.5M spots, 1.6G bases, 149Mb downloads
Accession: SRX28711782

 **National Library of Medicine**
National Center for Biotechnology Information

SRA

SRA

Advanced

Full

SRX28711787: SARS-CoV-2 reads from human upper respiratory sample
1 ILLUMINA (NextSeq 2000) run: 4.3M spots, 1.3G bases, 119.6Mb downloads

Design: IDT xGen SARS-CoV-2 multiplex amplicon sequencing protocol

Submitted by: Emory University School of Medicine

Study: Respiratory viruses from clinical specimens Genome sequencing and
[PRJNA1144955](#) • [SRP532954](#) • [All experiments](#) • [All runs](#)
[show Abstract](#)

Sample:
[SAMN48398077](#) • [SRS24977744](#) • [All experiments](#) • [All runs](#)
Organism: [Severe acute respiratory syndrome coronavirus 2](#)

Library:
Name: GA-EHC-8291K_L1
Instrument: NextSeq 2000
Strategy: AMPLICON
Source: VIRAL RNA
Selection: PCR
Layout: PAIRED

Runs: 1 run, 4.3M spots, 1.3G bases, [119.6Mb](#)

Run	# of Spots	# of Bases	Size	Published
SRR33476520	4,293,582	1.3G	119.6Mb	2025-05-08

ID: 38543131

 **National Library of Medicine**
National Center for Biotechnology Information

Sequence Read Archive

Search Run

Run Browser

Run Browser

Search and browse data for a single run

SRX28711787

Search

What can be entered in this field?

Sequence Read Archive

Search Run Browser Analyses Study Provisional SRA Docu

Run Browser > SRR33476520

SARS-CoV-2 reads from human upper respiratory sample (SRR33476520)

Metadata Analysis Reads Data access FASTA/FASTQ download

Run

Run	Spots	Bases	Size	GC Content	Data Status	Published
SRR33476520	4.3M	1.3G	119.6MB	40.1%	Public	2025-05-08

Quality graph (bigger)

This run has 2 reads per spot:

L=149, 100%

L=149, 100%

Legend Additional attribute:

assembly unaligned

Experiment

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX28711787	GA-EHC-8291K_L1	Illumina	AMPLICON	VIRAL RNA	PCR	PAIRED	BLAST

NCBI GenBank

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

NCBI Virus
Sequences for discovery

Find Data ▾ Help ▾ How to Participate ▾ Submit Sequences ▾ [Contact Us](#)

NCBI Virus

Community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories. To find, retrieve and analyze data, please select an option below.



Search by virus

Use virus name or taxid to find viral nucleotide and protein sequences.



NCBI BLAST™ search

Find viral nucleotide and protein sequences using NCBI BLAST™ sequence similarity tool.

Search GenBank virus/viroid sequences and explore metadata in Results Table with advanced filtering and options

Begin typing a virus name, viral taxonomy group, or taxid to select from a list of suggestions.

Popular searches

All viruses

Human viruses

H5N1* Influenza A virus

Released in past month

SARS-CoV-2

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

NCBI Virus
Sequences for discovery

Find Data ▾ Help ▾ How to Participate ▾ Submit Sequences ▾ [Contact Us](#)

SARS-CoV-2 Data Hub

Quick Links

[Betacoronavirus BLAST](#)
[CDC Outbreak Information](#)

[SARS-CoV-2 Articles in PubMed](#)
[Datasets command line](#)

[SRA Data](#)

[Advanced Filters for GenBank Sequences](#)

[Visual Filters for GenBank Sequences](#)

[Outbreak Sequence Statistics](#)

[Lineages Frequency and Location of GenBank + SRA Data](#)

[Search GenBank + SRA Data by Mutation](#)

Refine Results

[Reset](#)

Virus/Taxonomy

Severe acute respiratory syndrome coronavirus 2, taxid:2697049

Accession

Sequence Length

Ambiguous Characters

GenBank/RefSeq

Assembly Completeness

Nucleotide Completeness

Applied Filters:

Virus/Taxonomy (1)

Host (1)

Collection Date (1)

Geographic Region (1)

Selected Results: 0 / 75

[Align](#)

[Build Phylogenetic Tree](#)

[Download All Results](#)

Nucleotide (75)

Protein (895)

NCBI Virus Assembly (0)

[Select Columns](#)

Expand Table

<input type="checkbox"/>	Accession	Organism Name	GenBank/RefSeq	Assembly	Submitters	Organization
<input type="checkbox"/>	PV671391.1	Severe acute respiratory s...	GenBank		Lauring,A.S., e...	Lauring Lab at University ...
<input type="checkbox"/>	PV671392.1	Severe acute respiratory s...	GenBank		Lauring,A.S., e...	Lauring Lab at University ...
<input type="checkbox"/>	PV637216.1	Severe acute respiratory s...	GenBank		Fries,A., et al.	United States Air Force Sc...
<input type="checkbox"/>	PV625094.1	Severe acute respiratory s...	GenBank		Greenleaf,M., ...	Emory University School ...
<input type="checkbox"/>	PV625095.1	Severe acute respiratory s...	GenBank		Greenleaf,M., ...	Emory University School ...
<input type="checkbox"/>	PV625096.1	Severe acute respiratory s...	GenBank		Greenleaf,M., ...	Emory University School ...

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

Q ⓘ 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

Taxa

Genomes

Proteins

Specialty Genes

[All Searches...](#)

ANALYZE

Assembly

Annotation

BLAST

MSA

[All Tools & Services...](#)

MANAGE DATA

[Access Private Workspace](#)

[Save Search Results](#)

[Upload Data](#)

[Access Analysis Jobs](#)

[Share and Publish](#)

BATCH ACCESS

[Command-Line Interface](#)

[Data API](#)

[FTP](#)

QUICK START

TUTORIALS

REFERENCE GUIDES

Genome List View

GENOMES: TAXON LINEAGE IDS IS 2697049 AND HOST GROUP IS Human AND STATE PROVINCE IS Georgia AND COLLECTION_YEAR >= 2025 (25 Genomes)

[Overview](#)
[Genomes](#)
[Sequences](#)
[AMR Phenotypes](#)
[Features](#)
[Proteins](#)
[Specialty Genes](#)
[Pathways](#)
[Subsystems](#)



DOWNLOAD

KEYWORDS



ADV Search



FILTERS



APPLY

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

Registered Users EpiFlu™ **EpiCoV™** EpiRSV™ EpiPox™ EpiArbo™ My Profile

EpiCoV™ | [Search](#) | [Downloads](#) | [Upload](#)

Search ▼ Reset filters

EPI_ISL ID

Virus name

EPI_SET ID

☐ Complete [?]

Location

Host

☐ High coverage [?]

Collection to

Submission to

☐ Low coverage excluded [?]

Clade

Lineage

Variant

☐ With patient status [?]

AA Substitutions [?]

Nucl Mutations [?]

☐ Collection date complete [?]

☐ Under investigation

<input checked="" type="checkbox"/>	VIRUS NAME	PASSAGE D	ACCESSION ID	COLLECTION	SUBMISSION	<input type="checkbox"/>	LENGTH	HOST	LOCATION	ORIGINATING
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-UM-G25Y95F5/2025	Original	EPI_ISL_19858287	2025-04-10	2025-05-01	<input type="checkbox"/>	29,655	Human	North America / USA / Georgia	University of Georgia
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-CDC-LC1136800/2025	Original	EPI_ISL_19843005	2025-04-08	2025-04-22	<input type="checkbox"/>	29,688	Human	North America / USA / Georgia	Laboratory
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-UM-G25Y94Z2/2025	Original	EPI_ISL_19858165	2025-03-30	2025-05-01	<input type="checkbox"/>	29,655	Human	North America / USA / Georgia	University of Georgia
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-UM-G25Y94Y0/2025	Original	EPI_ISL_19858174	2025-03-30	2025-05-01	<input type="checkbox"/>	29,661	Human	North America / USA / Georgia	University of Georgia
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-UM-G25Y94X7/2025	Original	EPI_ISL_19858416	2025-03-30	2025-05-01	<input type="checkbox"/>	29,658	Human	North America / USA / Georgia	University of Georgia
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-UM-G25Y94Y3/2025	Original	EPI_ISL_19858180	2025-03-30	2025-05-01	<input type="checkbox"/>	29,655	Human	North America / USA / Georgia	University of Georgia
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-UM-G25Y94X4/2025	Original	EPI_ISL_19858169	2025-03-29	2025-05-01	<input type="checkbox"/>	29,655	Human	North America / USA / Georgia	University of Georgia
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-CDC-LC1136379/2025	Original	EPI_ISL_19821928	2025-03-27	2025-04-08	<input type="checkbox"/>	29,691	Human	North America / USA / Georgia	Laboratory
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-USAFSAM-S23564/2025	Original	EPI_ISL_19867073	2025-03-26	2025-05-09	<input type="checkbox"/>	29,750	Human	North America / USA / Georgia	US Air Force
<input checked="" type="checkbox"/>	hCoV-19/USA/GA-CDC-LC1136226/2025	Original	EPI_ISL_19821802	2025-03-25	2025-04-08	<input type="checkbox"/>	29,681	Human	North America / USA / Georgia	Laboratory

Total: 73 viruses

[Charts](#) [EPI_SET](#) [Select](#) [Analysis](#) [Download](#)

Important note: In the [GISAID EpiFlu™ Database Access Agreement](#), you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the [GISAID EpiFlu™ Database Access Agreement](#) in respect of such data in the same manner as if they were data relating to influenza viruses.