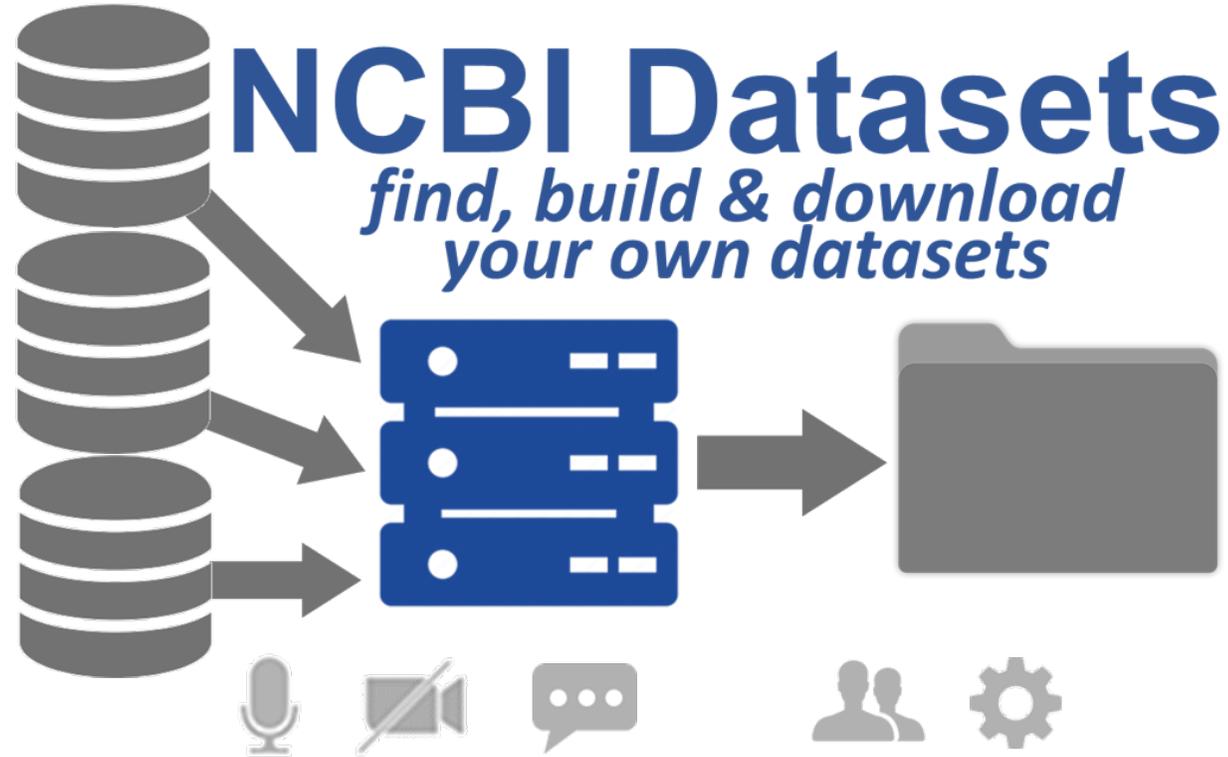


# ASHG CoLab Live!



# Learn how to retrieve custom gene and genome datasets using NCBI Datasets - a new & developing resource

Talk directly with us to let us know you how you use genomic datasets and help us learn what types and formats of data you'd love to be able to get and how!

Drs. O'Leary & Holmes are here to answer questions too!



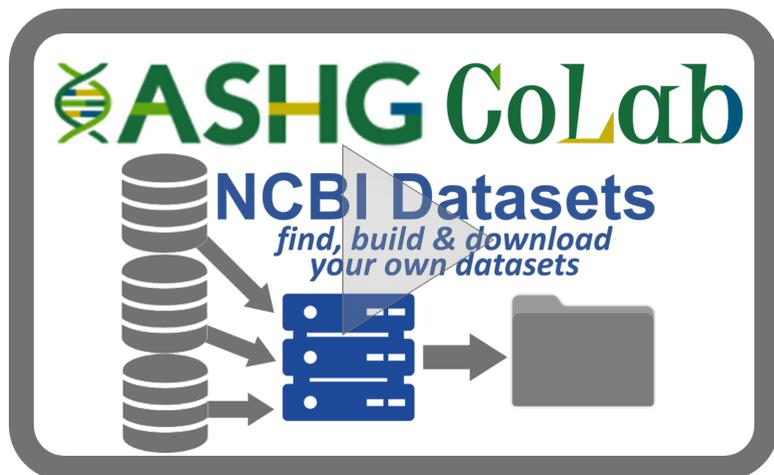
Nuala O'Leary, Ph.D.



Brad Holmes, Ph.D.



Rana Morris, Ph.D.

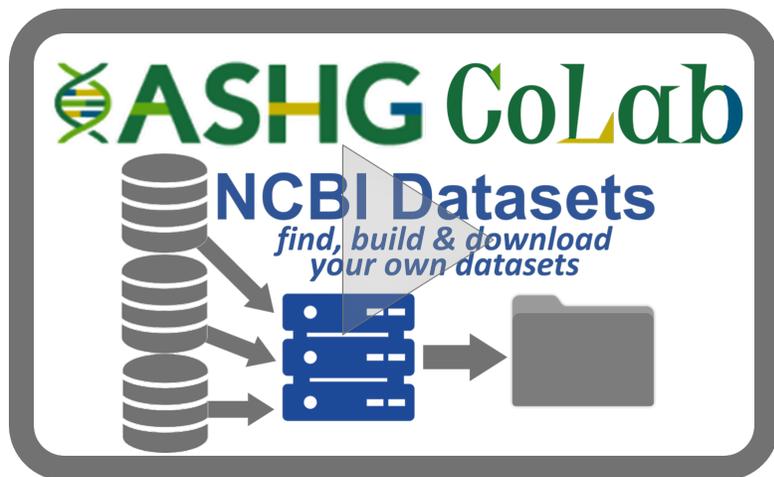


**ASHG CoLab Video Presented by:**  
*Nuala O'Leary, Ph.D. & Wayne Matten, Ph.D.*  
<https://youtu.be/79lakxFuNaQ>

# Learn how to retrieve custom gene and genome datasets using NCBI Datasets - a new & developing resource

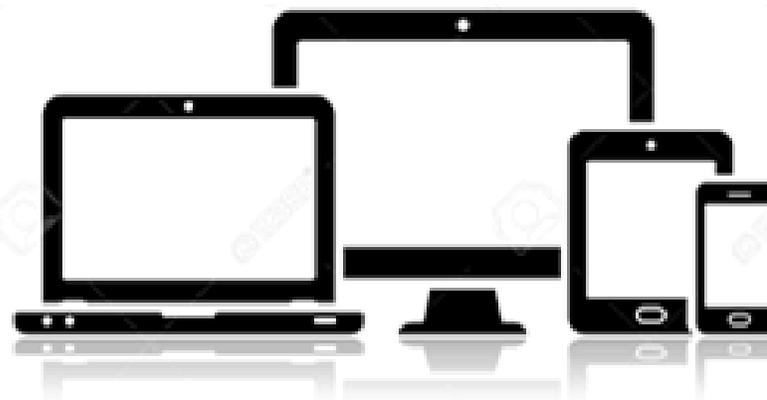
Talk directly with us to let us know you how you use genomic datasets and help us learn what types and formats of data you'd love to be able to get and how!

Drs. O'Leary & Holmes are here to answer questions too!



Welcome to the world of GoToWebinar....

*How does this thing work?*

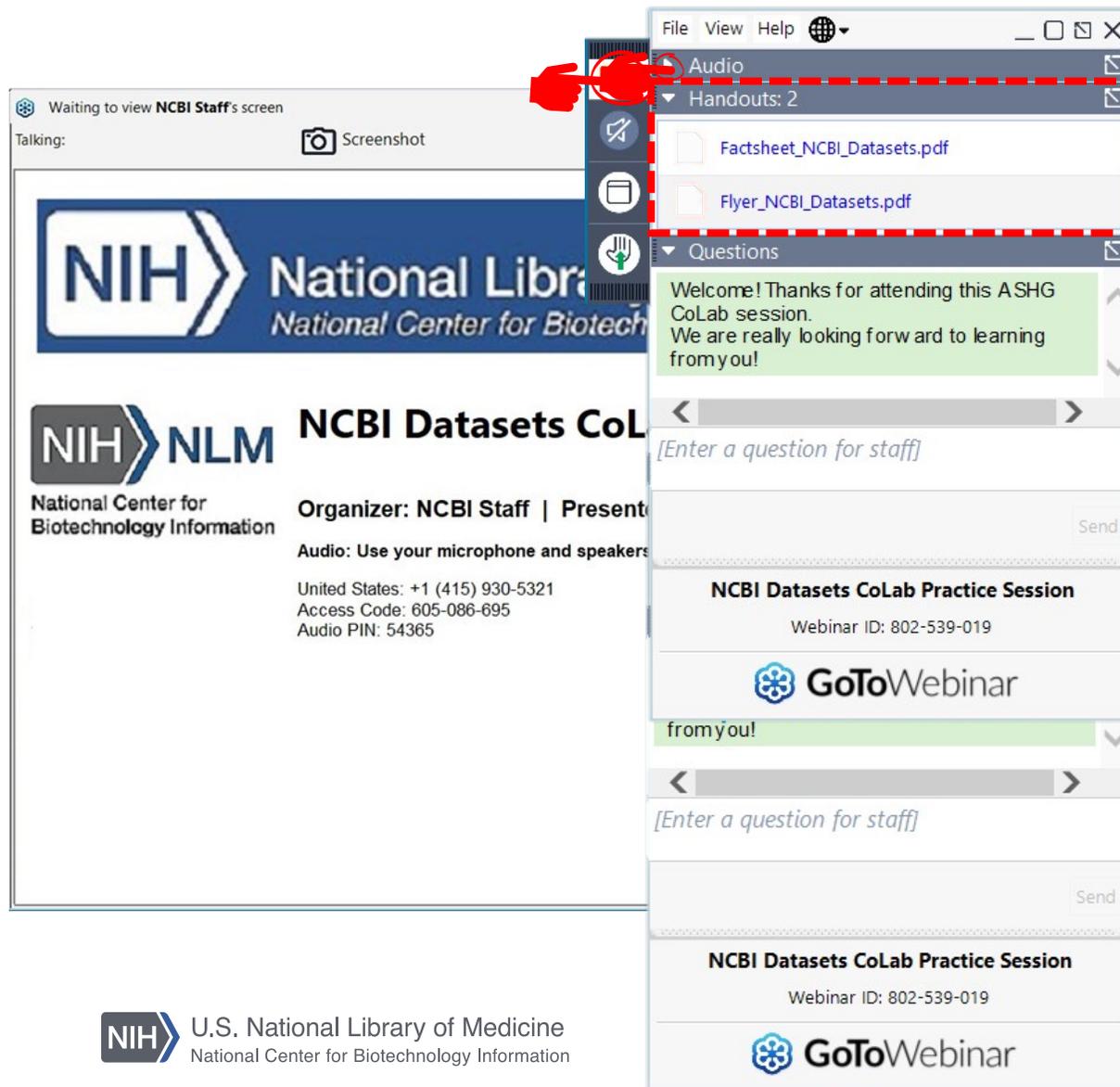


**ASHG CoLab Video Presented by:**

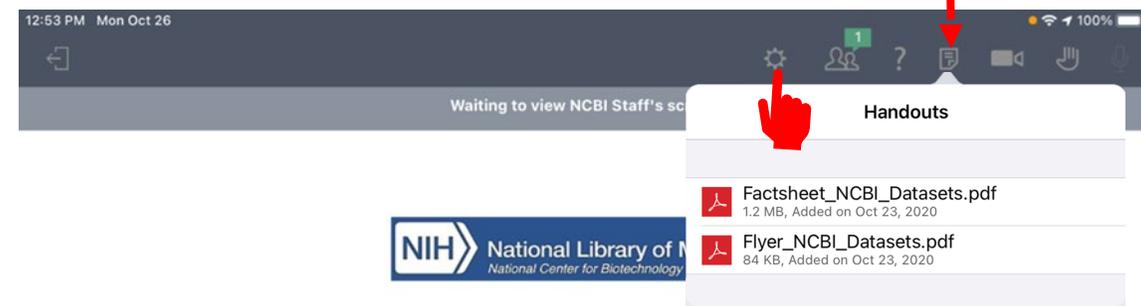
*Nuala O'Leary, Ph.D. & Wayne Matten, Ph.D.*

<https://youtu.be/79lakxFuNaQ>

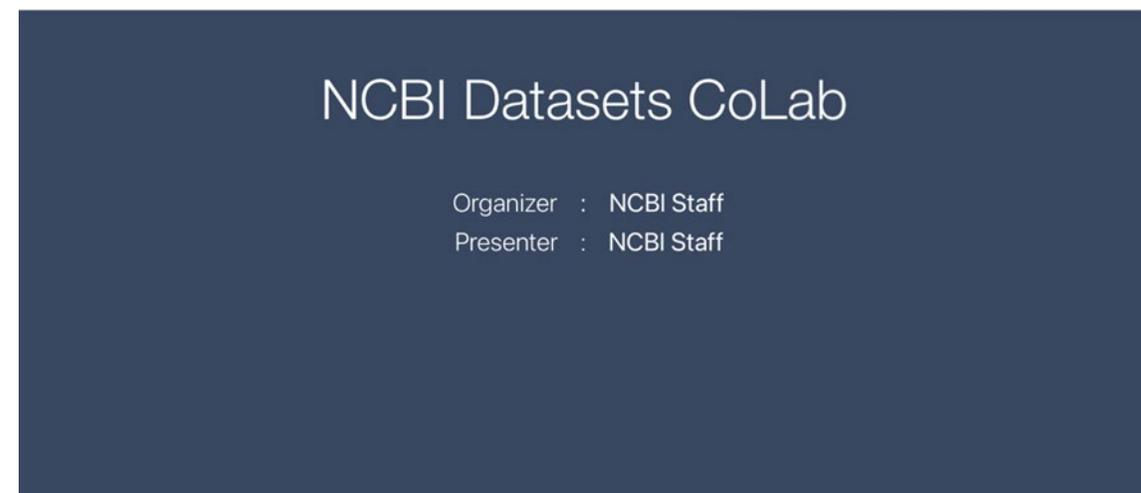
# WELCOME TO THE WORLD OF GOTOWEBINAR



The screenshot shows a mobile GoToWebinar interface. A file explorer overlay is open, displaying a menu with 'Audio', 'Handouts: 2', and 'Questions'. Under 'Handouts', two PDF files are listed: 'Factsheet\_NCBI\_Datasets.pdf' and 'Flyer\_NCBI\_Datasets.pdf'. A red dashed box highlights the 'Handouts' section, and a red arrow points from the 'Handouts' icon in the top right of the mobile app to this section. The background shows a waiting screen for 'NCBI Staff's screen' with a 'Screenshot' button. The main content area displays the NIH logo and 'National Library of Medicine National Center for Biotechnology Information'. Below this, it says 'NCBI Datasets CoLab' and 'Organizer: NCBI Staff | Presenter: NCBI Staff'. Contact information is provided: 'United States: +1 (415) 930-5321', 'Access Code: 805-086-695', and 'Audio PIN: 54365'. The GoToWebinar logo is also visible at the bottom.



This screenshot shows the top portion of the mobile GoToWebinar interface. The status bar at the top indicates the time is 12:53 PM on Monday, October 26, with 100% battery. The main header says 'Waiting to view NCBI Staff's screen'. A red arrow points to the 'Handouts' icon in the top right corner. A dropdown menu is open, showing the 'Handouts' section with two PDF files: 'Factsheet\_NCBI\_Datasets.pdf' (1.2 MB, Added on Oct 23, 2020) and 'Flyer\_NCBI\_Datasets.pdf' (84 KB, Added on Oct 23, 2020). A red hand icon points to the 'Handouts' header in the dropdown menu.



This screenshot shows a slide titled 'NCBI Datasets CoLab'. The slide content includes: 'Organizer : NCBI Staff' and 'Presenter : NCBI Staff'. The slide has a dark blue background with white text.



# Datasets: A Resource for Genomic Data from NCBI

A portal with customizable tools to access genomic sequences and their related datasets  
<https://www.ncbi.nlm.nih.gov/datasets>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

## Introduction

Advances in sequencing technology has led to dramatic increase in available genomic sequence data for a large collection of organisms. This also poses a significant challenge on how to organize and present the available datasets, and how to make these datasets readily accessible. At NCBI, genomic assemblies are organized through versioned entries in the Assembly database [1], which provides a summary of the assembly and a link to the dataset stored in the NCBI FTP site. The Assembly database also provides a download tool to allow bulk download of the retrieved set.



NCBI Datasets is a new resource that lets you easily gather assembled genomic data and their related datasets from across NCBI databases. It provides a web search page to allow customization of datasets for browsing the downloading, an API service for integration with various third party tools or workflows, and a command line tool for bulk access. This handout will address the key features of this newly released portal. Currently, the access is limited to the eukaryotic entries. Prokaryotic and viral datasets are also available for downloading, but not for online browsing.

## Getting Started

The main entry point is through the web portal, shown to the right, which provide accesses to:

- An overview of this resource given at the top (A)
- Information on programmatic access by way of command-line tool or API through linked pages (B)
- List of genome assemblies available for major taxonomic groups and well studied species (C) linking to the web search page with results limited to that group or species
- A link to a new interface for gene-specific information (D) allowing the retrieval of information/datasets for a user-specified genes
- A link to all publicly available SARS-Cov-2 datasets (E) from NCBI, and
- A collection of FAQs (F) addressing common questions over this resources

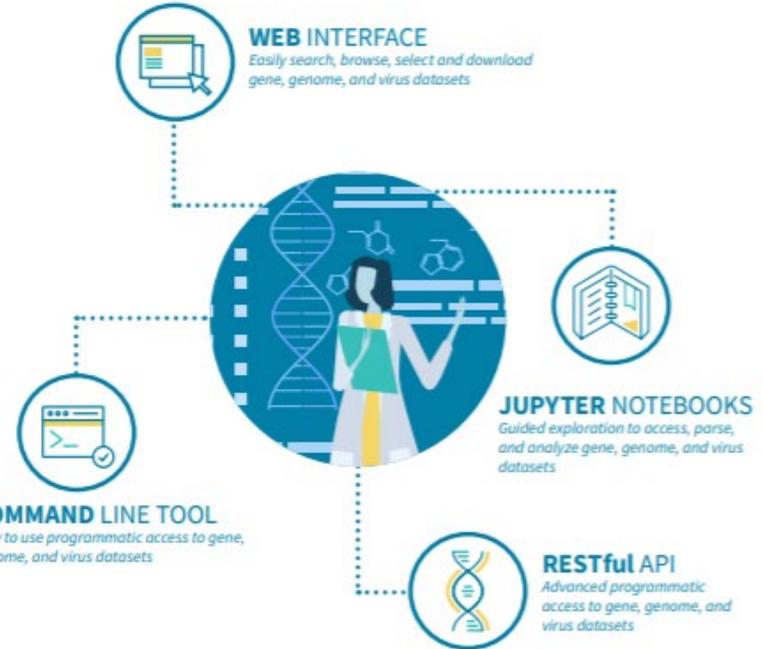
The screenshot shows the NCBI Datasets homepage. Callout A points to the 'Welcome to NCBI Datasets' section. Callout B points to the 'Programmatic access' section, which includes links for 'Command-line', 'GitHub', and 'Datasets API'. Callout C points to the 'Browsing genome datasets' section, which lists various organisms like Homo sapiens, Mus musculus, Arabidopsis thaliana, Rattus norvegicus, Drosophila melanogaster, and Bos taurus. Callout D points to the 'Data tables' section. Callout E points to the 'Coronavirus datasets' section. Callout F points to the 'FAQs' section.

NCBI Handout Series | NCBI Datasets | Last Updated on August 31, 2020

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

# NCBI Datasets

A new tool to build custom gene and genome datasets



## RESOURCES

Many file types to choose from!

- Genomic fasta
- Transcript fasta
- Protein fasta
- GFF
- GBFF
- GTF
- Metadata

Begin building your dataset



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

Follow us on Twitter @ncbi

Contact us at [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)

# WELCOME TO THE WORLD OF GOTOWEBINAR

Waiting to view NCBI Staff's screen

Talking: Screenshot

NIH National Library of Medicine National Center for Biotechnology Information

NIH NLM NCBI Datasets CoLab  
National Center for Biotechnology Information

Organizer: NCBI Staff | Presenter: NCBI Staff

Audio: Use your microphone and speakers

United States: +1 (415) 930-5321  
Access Code: 805-086-695  
Audio PIN: 54365

File View Help

Audio

Handouts: 2

Factsheet\_NCBI\_Datasets.pdf

Flyer\_NCBI\_Datasets.pdf

Questions

Welcome! Thanks for attending this ASHG CoLab session. We are really looking forward to learning from you!

Enter a question for staff

(Type a question here!)

Send

NCBI Datasets CoLab Practice Session

Webinar ID: 802-539-019

GoToWebinar

12:53 PM Mon Oct 26

Waiting to view NCBI Staff's screen

Questions

Q: Sent as a question.....  
A: answered question "Sent to All"

(Type a question here!)

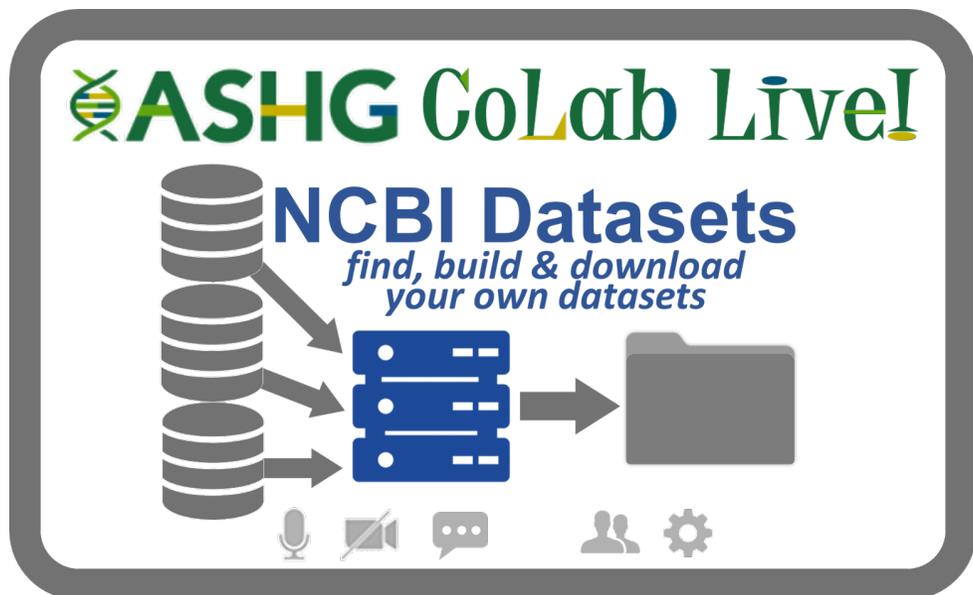
Send

NCBI Datasets CoLab

Organizer : NCBI Staff  
Presenter : NCBI Staff

The whole point of today is that we want to hear from and talk with you!

# Learn how to retrieve custom gene and genome datasets using NCBI Datasets - a new & developing resource



Nuala O'Leary, Ph.D.  
NCBI Datasets Team Lead



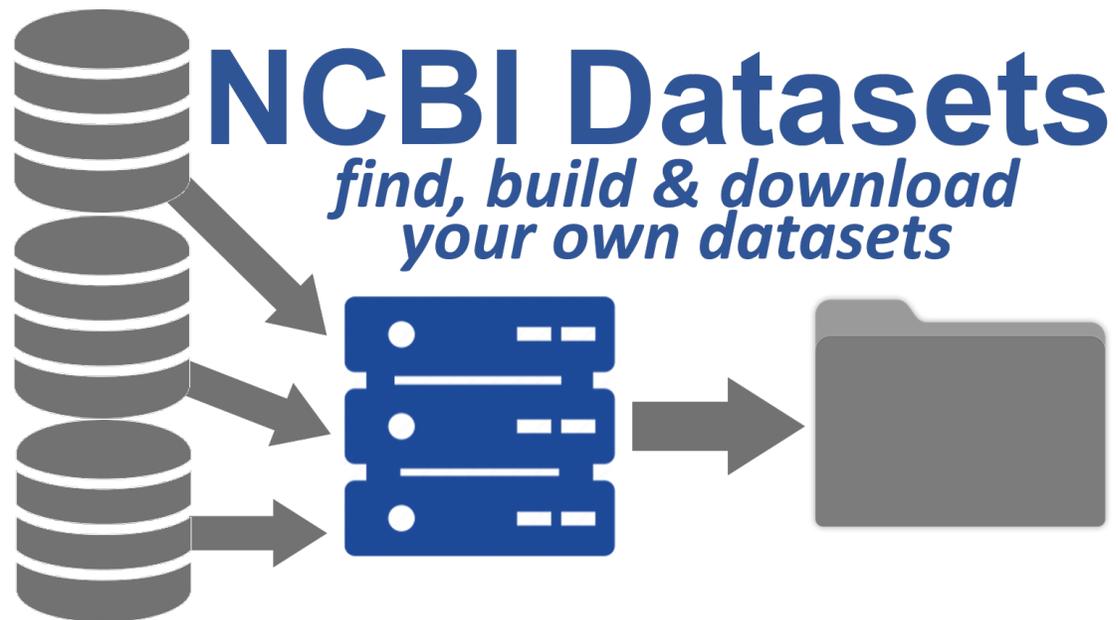
Brad Holmes, Ph.D.  
NCBI Datasets Team  
Software Developer

## Order of events:

- Video: "Introduction to NCBI Datasets"
- Quick view of NCBI Datasets by Dr. O'Leary
- A discussion with you!

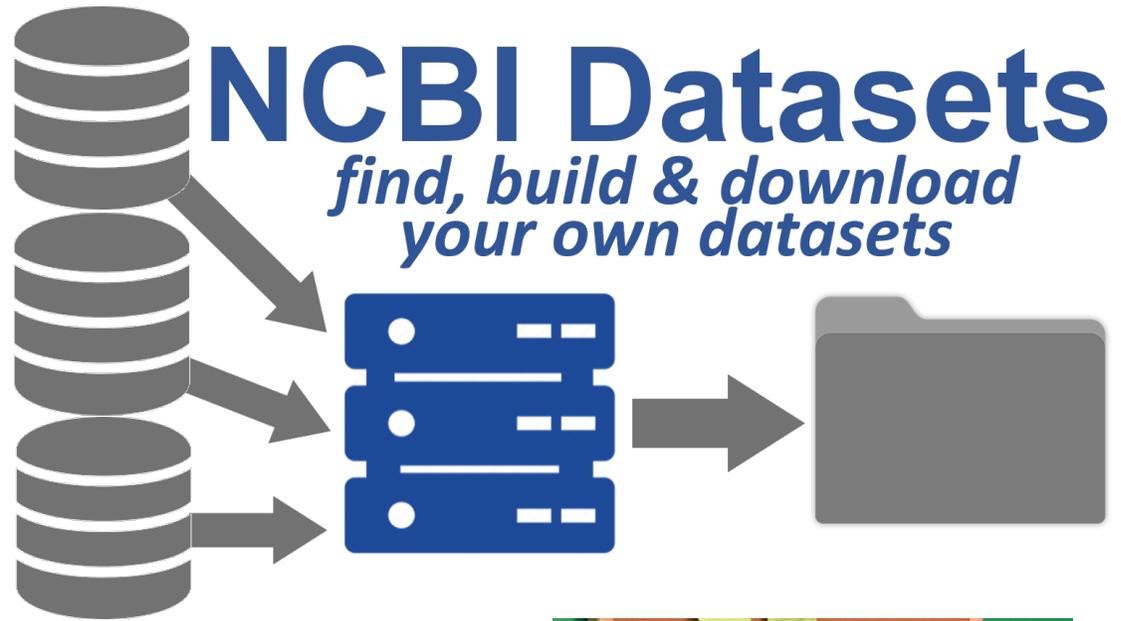


- Let us know what you prefer - *answer some polling questions*
- Talk with us - *unmute yourself*
- Chat with us - *type in the Questions section*



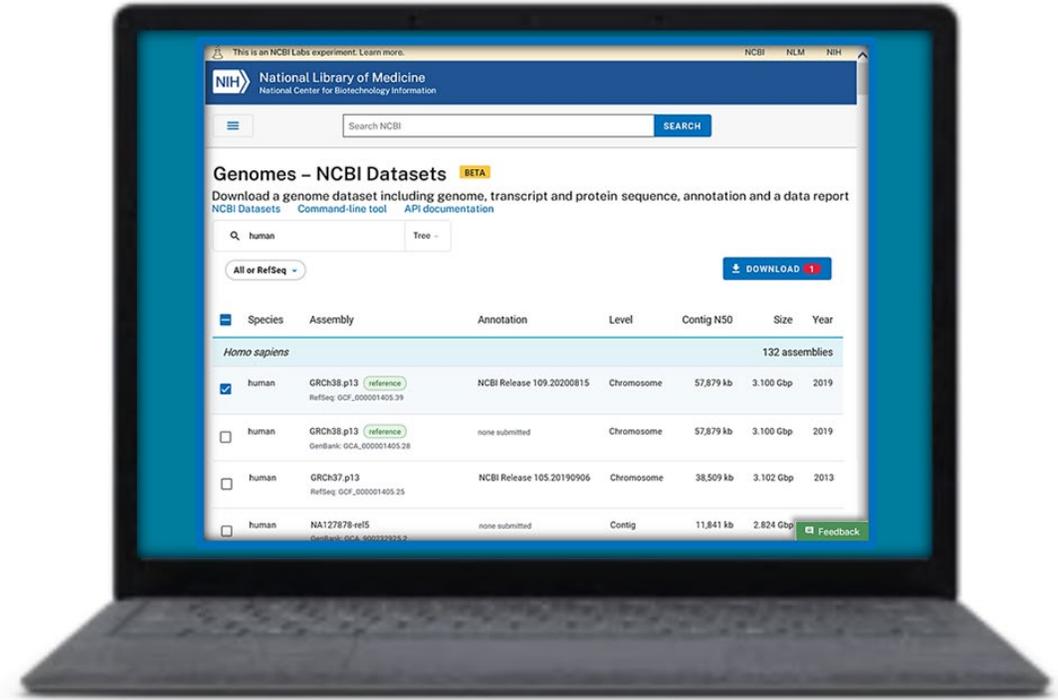
**13min ASHG CoLab Video:** <https://youtu.be/79lakxFuNaQ>

**1min Short Video:** [https://www.youtube.com/watch?v=2\\_57xOi-aSg](https://www.youtube.com/watch?v=2_57xOi-aSg)

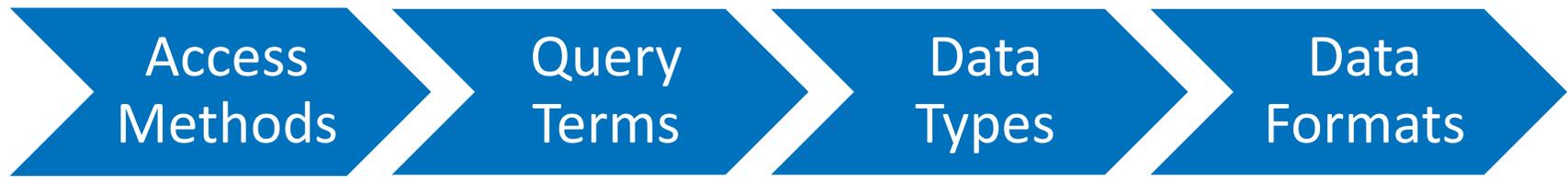


# NCBI Datasets

*find, build & download your own datasets*



Nuala O'Leary, Ph.D.  
NCBI Datasets Team Lead



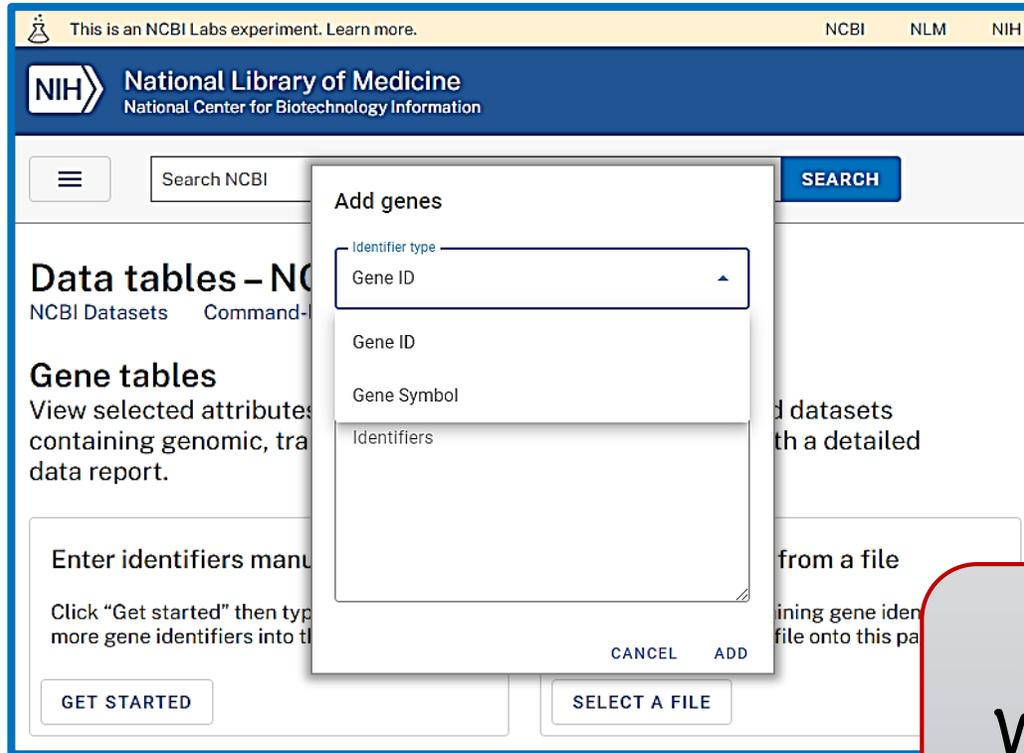


Access  
Methods

Query  
Terms

Data  
Types

Data  
Formats



## SEARCH QUERY TERMS

**Genomes:** NCBI Genomic Assembly Accession  
NCBI TaxID  
“taxonomic term”

**Genes:** NCBI GeneID  
NCBI Gene Symbol  
“taxonomic term”

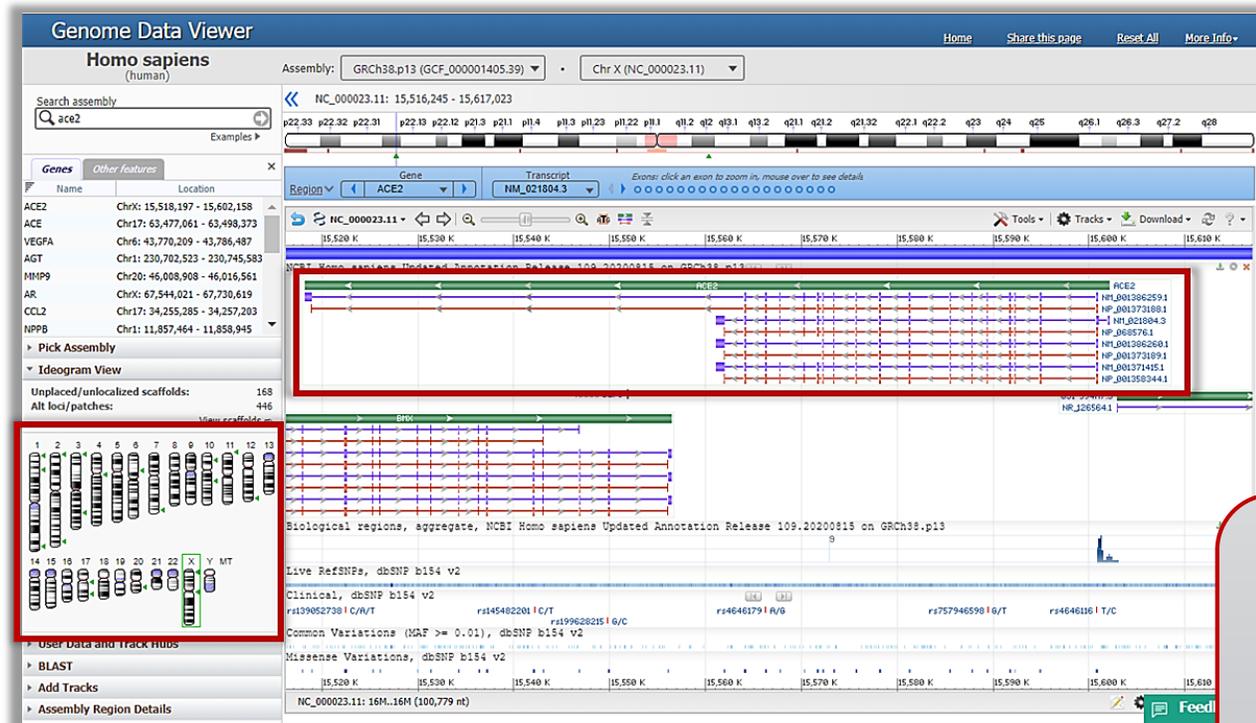
**POLL QUESTION**  
What do you use to  
query for data?

Access  
Methods

Query  
Terms

Data  
Types

Data  
Formats



**GENE PRODUCTS  
& GENOMES**  
*sequences & metadata*

**POLL QUESTION**  
What type of data  
would you like to get?

Access  
Methods

Query  
Terms

Data  
Types

Data  
Formats

## Metadata **FILE FORMATS**

Genome dataset report  
& Gene data table:

TSV

JSONL

Genome annotations:

GFF3

GTF

\*GBFF (w/annotations)

## Sequence **FILE FORMATS**

FASTA

\*GBFF (w/annotations)

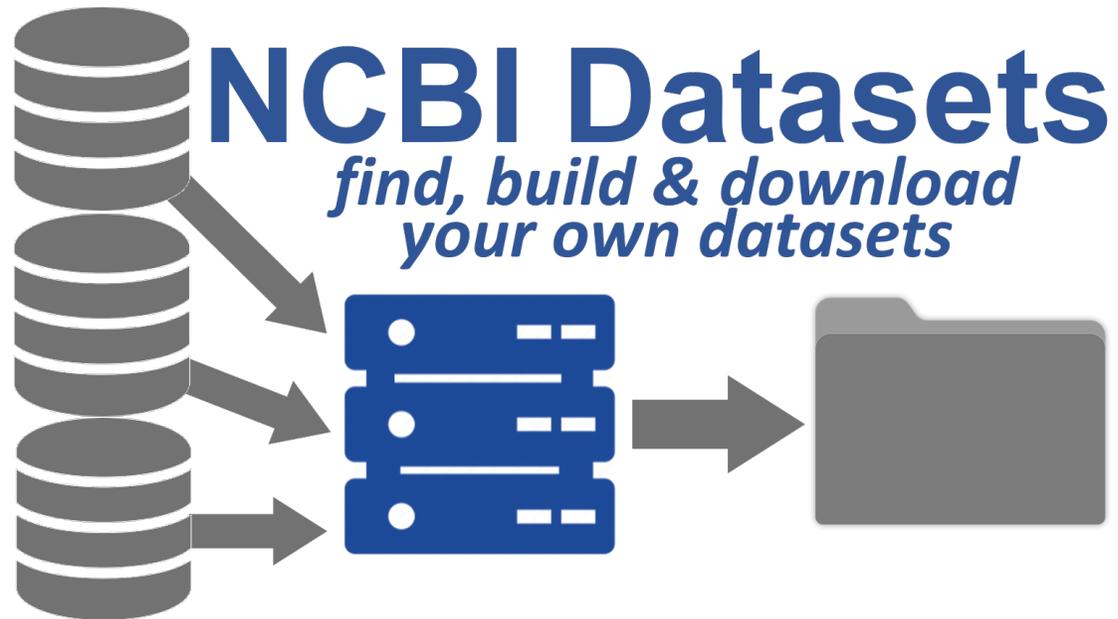
The screenshot shows the NCBI Genomes download interface. A 'Download' dialog box is open, listing the following options:

- Genomic sequence (FASTA)
- Annotated features (GTF)
- Annotated features (GFF3)
- Sequence and annotation (GBFF)
- Transcripts (FASTA)
- Protein (FASTA)

The dialog also indicates that the data will be downloaded as a ZIP file and provides an estimated download size of 976.56 KB. A 'Name your file' input field is present at the bottom of the dialog. The background shows the 'Genomes - NCBI Data' page with a search for 'human' and a table of assemblies.

Contig	Size	Year	
132 assemblies			
57,879 kb	3.100 Gbp	2019	
57,879 kb	3.100 Gbp	2019	
Chromosome	38,509 kb	3.102 Gbp	2013

**POLL QUESTION**  
What file formats  
(sequence & metadata)  
do you want?



**POLL QUESTION**  
Would you be willing to talk with us or beta test it?

The laptop screen displays the NCBI Datasets website. The page title is "Genomes - NCBI Database". Below the title, there is a search bar with "human" entered. A table lists genome datasets for "Homo sapiens".

Species	Assembly	Genome	Size	Year
human	GRCh38 p13	NCBI Release 109.20200815	3,100 Gbp	2019
human	GRCh38 p13	Chromosome	57,879 kb	2019
human	GRCh37 p13	NCBI Release 105.20190906	3,102 Gbp	2013
human	NA127878-re15	Contig	11,841 kb	2,824 Gbp

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

To learn more keep an eye on the NCBI Insights Blog & social media streams:



<https://ncbiinsights.nlm.nih.gov>

