



BioEd Summit: Introduction to NCBI Resources – the what and why of NCBI

E. Sally Chang, PhD



National Library of Medicine
National Center for Biotechnology Information



NCBI created in 1988 by an Act of Congress.

- Create public databases and accept submissions of primary data
- Develop tools to analyze these data (*curate data to create a quality controlled, value-added reference datasets*)
- Conduct research in computational biology
- Disseminate biomedical information
- **Archive, Access and Analyze!**

- The birth of the NCBI -

Health Omnibus Programs Extension Act of 1988
Public Law 100-607 | Nov. 4, 1988

An Act
To amend the Public Health Service Act to establish certain health programs, to revise and extend certain health programs, and for other purposes.
Enacted by the Senate and House of Representatives of the United States of America in Congress assembled.

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) **SHORT TITLE.**—This Act may be cited as the “Health Omnibus Programs Extension Act of 1988”.

(b) **TABLE OF CONTENTS.**—

Title I—National Institute on Deafness and Other Communication Disorders and Health Research Extension Act of 1988

Title II—Programs with Respect to Acquired Immune Deficiency Syndrome

Title III—Preventive Health, Health Services, and Health Programs

Title IV—Organ Transplant Amendments of 1988

Title V—Food and Drug Administration

Title VI—Health Personnel Qualifications Act of 1988

Title VII—Nursing Charge Reduction and Education Extension Act of 1988

Title VIII—Services and Education of Progress of Health Care for the Handicapped

Title IX—Testing of Cervical Fluids

Subtitle B—Biotechnology Information

SEC. 303. ESTABLISHMENT OF NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION.

Part D of title IV (42 U.S.C. 286 et seq.) is amended by adding at the end the following new subject:

“Subpart B—National Center for Biotechnology Information

“PURPOSE, ORGANIZATION, FUNCTIONS, AND POWERS OF THE NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION

“SEC. 473. (a) In order to focus and expand the collection, storage, retrieval, and dissemination of the results of biotechnology research by information systems, and to support and enhance the development of new information technologies to aid in the understanding of the molecular processes that control health and disease, there is established the National Center for Biotechnology Information (hereinafter in this section referred to as the “Center”) in the National Library of Medicine.

(b) The Secretary, through the Center and subject to section 4604a, shall—


(1) design, develop, implement, and manage automated systems for the collection, storage, retrieval, analysis, and dissemination of knowledge concerning human molecular biology, biochemistry, and genetics;

(2) perform research into advanced methods of computer-based information processing capable of representing and analyzing the vast number of biologically important molecules and compounds;

(3) enable persons engaged in biotechnology research and medical care to use systems developed under paragraph (1) and methods described in paragraph (2); and

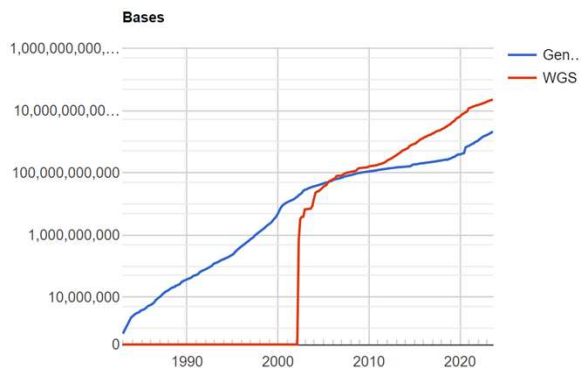
(4) coordinate, as much as is practicable, efforts to gather biotechnology information on an international basis.

(c) For the purpose of performing the duties specified in subsection (b), there are authorized to be appropriated \$1,000,000 for fiscal year 1989 and such sums as may be necessary for fiscal year 1990. Funds appropriated under this subsection shall remain available until expended.”.



Archive (and curate!)

- 35+ public databases
- Three main sources of data:
 - Direct submissions
 - Collaborations/agreements
 - Internal curation (Refseq)
- Ever-increasing data is exciting but poses challenges for storage, access and interpretation!

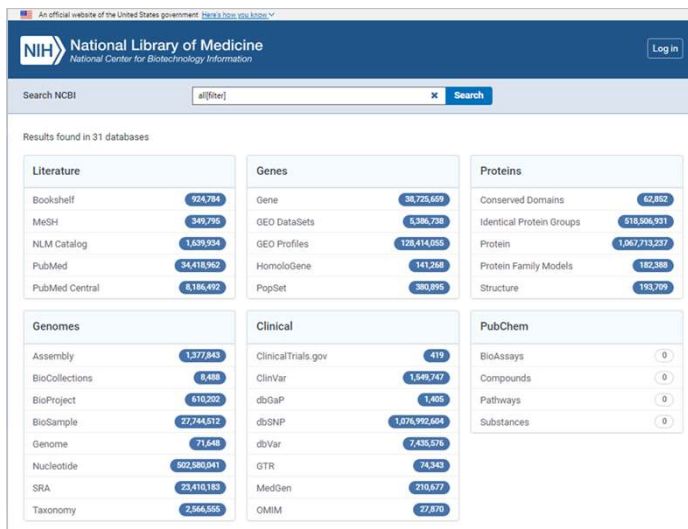


Growth of Genbank: The number of bases has doubles every 18 months!

How much data?

We have 40+ petabytes of data for people to access.

Largely connected by Entrez search system!



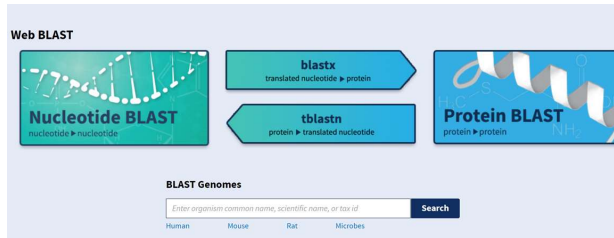
Access

Making sure users can find and obtain the data they actually need out of all those databases...

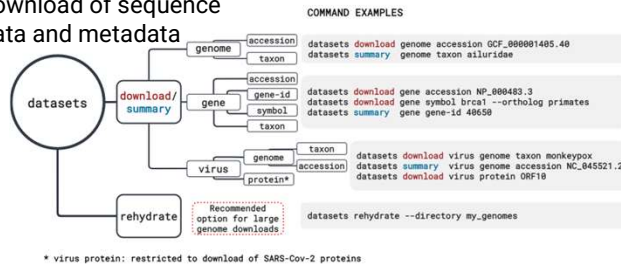
Searching of citations on PubMed



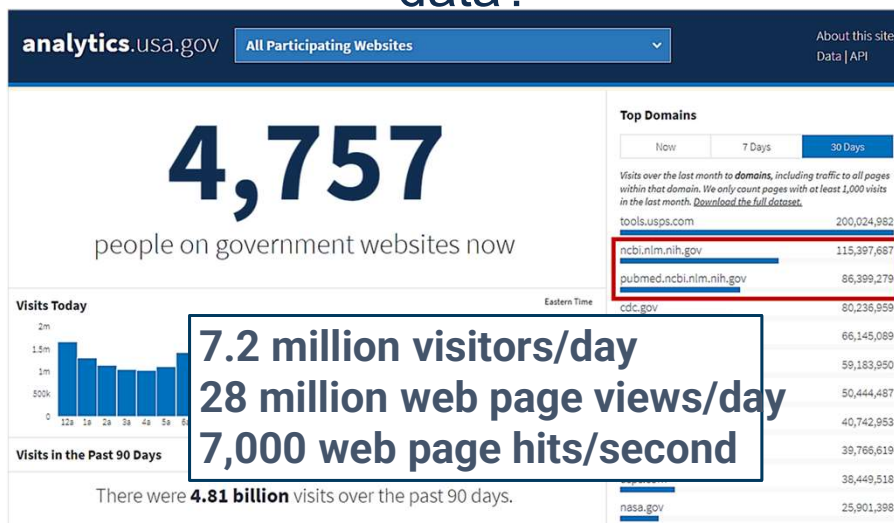
BLAST searches of sequence databases by sequence similarity



Datasets command line tool for bulk download of sequence data and metadata

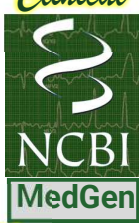


How many people visit our website to access data?




NCBI's Information Hubs

Clinical




NCBI
MedGen

Nucleotide Sequence




NCBI
Genome


NCBI Virus



Pathogens




Literature




NCBI
PubMed

BioProject
BioSample




Gene-based

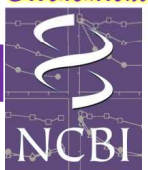


NCBI
Gene

Taxonomy




Biochemical




NCBI
PubChem
Compound


Protein



NCBI
Protein

Hubs of Data
Great places to start!




 National Library of Medicine
National Center for Biotechnology Information

If it gets overwhelming...we have resources!

NCBI Databases

What kinds of data does NCBI have?
How can I search for it?

Literature	Genomes
<p>The World's largest repository of medical and scientific literature, full-text articles, books and reports.</p> <p>PubMed Books and reports https://www.ncbi.nlm.nih.gov/books/</p> <p>PubMed Central Full-text articles https://pubmed.ncbi.nlm.nih.gov/</p> <p>PubMed Central Full-text articles https://pubmed.ncbi.nlm.nih.gov/</p>	<p>Genome sequence assemblies, large-scale functional genomic data, and diverse biological samples.</p> <p>Assembly Genome assembly information https://www.ncbi.nlm.nih.gov/genomes/</p> <p>BioCollections Nucleic acids and other biological collections https://www.ncbi.nlm.nih.gov/biocollect/</p> <p>BioProject Biological research projects with links to publications and data https://www.ncbi.nlm.nih.gov/bioproject/</p> <p>BioSample Biological sample metadata, description and links to data https://www.ncbi.nlm.nih.gov/biosample/</p> <p>Genome Genome sequencing projects with links to data https://www.ncbi.nlm.nih.gov/genome/</p> <p>Genomes Genome sequencing projects with links to data https://www.ncbi.nlm.nih.gov/genome/</p>
Clinical	Organisms
<p>Heritable DNA variations, associations with human pathologies, and clinical diagnostics and treatments.</p> <p>ClinicalTrials.gov Clinical trials and other clinical research conducted around the world https://clinicaltrials.gov/</p> <p>Genetics Human relations of clinical significance https://www.ncbi.nlm.nih.gov/genetics/</p> <p>OMIM Online Mendelian Inheritance in Man https://www.ncbi.nlm.nih.gov/omim/</p> <p>OMIM Online Mendelian Inheritance in Man https://www.ncbi.nlm.nih.gov/omim/</p> <p>OMIM Online Mendelian Inheritance in Man https://www.ncbi.nlm.nih.gov/omim/</p>	<p>Species classification and nomenclature https://www.ncbi.nlm.nih.gov/taxonomy/</p> <p>Organisms Species classification and nomenclature https://www.ncbi.nlm.nih.gov/taxonomy/</p>
Gene-related	Genes
<p>Gene Gene sequences and annotations used as references for the study of structure, expression, and evolution.</p> <p>Gene Gene sequences and annotations used as references for the study of structure, expression, and evolution.</p> <p>Gene Gene sequences and annotations used as references for the study of structure, expression, and evolution.</p>	<p>Gene sequences and annotations used as references for the study of structure, expression, and evolution.</p> <p>Gene Gene sequences and annotations used as references for the study of structure, expression, and evolution.</p> <p>Gene Gene sequences and annotations used as references for the study of structure, expression, and evolution.</p>

[Link](#)


NCBI Data Access

How to find NCBI Data for your work

4 WAYS TO ACCESS NCBI DATA

- 1 Web
- 2 FTP
- 3 APIs/Command-line Tools
- 4 Cloud Computing Buckets

Web

1. Textbox searching with text/terms

Resources have "send to" links and download buttons to facilitate downloading data.

All Database Search page	https://www.ncbi.nlm.nih.gov/ncbi/
Molecular Database Searching (Entrez) Help/Doc	https://www.ncbi.nlm.nih.gov/books/NBK98977/
PubMed User Guide	https://pubmed.ncbi.nlm.nih.gov/about/

2. BLAST Searching with sequences


BLAST Homepage	https://blast.ncbi.nlm.nih.gov/Blast.cgi
BLAST Help page	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=HELP&BLAST=blast2.cgi


3. WAST Searching with 3D biomolecular structures

WAST Homepage	https://www.ncbi.nlm.nih.gov/structure/wast/wast.html
WAST Help page	https://www.ncbi.nlm.nih.gov/structure/wast/wast.html

4. PubChem Searching with chemical information

PubChem Homepage	https://pubchem.ncbi.nlm.nih.gov/
PubChem Help page	https://pubchemdocs.ncbi.nlm.nih.gov/about

[Link](#)


 National Library of Medicine
National Center for Biotechnology Information

Introducing a few hubs:

- Gene
- Bioproject
- Taxonomy
- PubChem
- MedGen



NCBI Gene Record as a hub for exploring a gene in the news: GLP1R

Examine Sequences in an Interactive Browser



Download Sequence Data and Pre-computed Orthologs

Retrieve FASTA format for: GLP1R
 Download protein: GLP1R

Download protein: GLP1R

Protein structure: GLP1R

View genomic regions, transcripts, products and variants, and go directly to their records



The Human GLP1R Gene page

GLP1R glucagon like peptide 1 receptor [Homo sapiens (human)] [Download Defaults](#)

Gene ID: 2749, updated on 10-MAR-2024

Summary

Official Symbol: [GLP1R](#) (provided by [HGNC](#))

Official Full Name: [glucagon like peptide 1 receptor](#) (provided by [HGNC](#))

Primary Source: [HGNC: 29593](#)

RefSeq status: [REVIEWED](#)

Gene type: [protein coding](#)

Organism: [Homo sapiens](#)

Lineage: [Eukaryota](#); [Mammalia](#); [Chordata](#); [Chorata](#); [Vertebrata](#); [Euteleostomi](#); [Mammalia](#); [Eutheria](#); [Primates](#); [Haplorhina](#); [Catarrhini](#); [Hominoidea](#); [Hominidae](#); [Homo](#)

Also known as: [GLP-1](#); [GLP-1R](#)

Summary

This gene encodes a G transmembrane protein that functions as a receptor (GLP-1) for the hormone which stimulates glucose induced insulin secretion. The protein binds to and activates the G protein-coupled receptor (GPCR) signaling pathway. The protein is highly conserved and is an important target for the treatment of type 2 diabetes and cancer. After results in multiple transcript variants (provided by RefSeq, Apr 2018)

Based expression in near (chromat 1 q), brain (cortex 1 q) and 10 other tissues (General gene information)

Expression

tissues: [all](#)

Orthologs

By the way (Access [RSC](#))

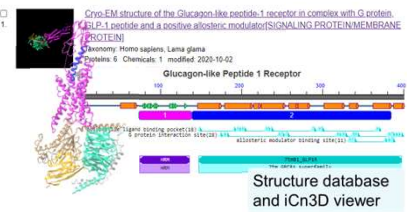
by the new (Access [RSC](#))

General protein information

NCBI Reference Sequences (RefSeq)

Related sequences

Explore and Visualize Protein Structures



Access Experimental Data



Gene Expression Omnibus (GEO)

PubChem BioAssay

Find an example Gene Record using Entrez

Search the **Gene database** for a **gene record** that:

Is related to diabetes, has associated gene expression profiles and clinical variation data from ClinVar and a MedGen record

Specify that "diabetes" should be treated as a disease/phenotype field

Filter for records with GEO gene expression data

Filter for genes with a ClinVar record

Filter for genes with a MedGen record

Search results
Items: 1 to 20 of 189

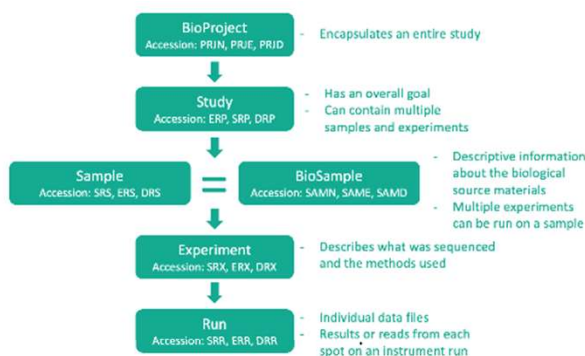
Name/Gene ID	Description	Location
<input type="checkbox"/> VEGFA ID: 7422	vascular endothelial growth factor A [<i>Homo sapiens</i> (human)]	Chromosome 6, NC_000006.12 (43770211..43786487)
<input type="checkbox"/> IL6 ID: 2550	Interleukin 6 [<i>Homo sapiens</i> (human)]	Chromosome 7, NC_000007.14 (22727200..22731998)

<https://www.ncbi.nlm.nih.gov/gene/advanced>

NIH National Library of Medicine
National Center for Biotechnology Information

NCBI BioProject as Hub

- Collection of data related to a single research effort
- A BioProject record provides users a single place to find links to the diverse data types
- Can search BioProjects or Browse by Attributes



tuberculosis Search

Please note: Searches on this page are limited to the fields available in this table. For more information, see Help.

Filters Download

#	Accession	Project Title	Organism	Organism Groups	Strain	Data Type	Has Data	Has Pub	Registration date
1	PRJNA1139960	Central Asian Outbreak (CAO) isolates from tuberculosis patients in Kemerovo	Multiple			Raw sequence reads	Yes	No	2024-07-25
2	PRJDB17033	Transcriptional Regulators SPI10 and SPI140 Modulate Inflammatory Response Genes in Mycobacterium tuberculosis-Infected Human Macrophages	<i>Homo sapiens</i>	Eukaryota; Animals; Mammals		Transcriptome or Gene expression	Yes	No	2024-07-22

<https://www.ncbi.nlm.nih.gov/bioproject/browse>

NIH National Library of Medicine
National Center for Biotechnology Information

NCBI BioProject as a data hub

Description

Mycobacterium tuberculosis Accession: PRJDB15809 ID: 1019517
 Heterogeneity in transmission dynamics among lineage 2 Mycobacterium tuberculosis strains in Kobe, Japan revealed by population-based whole-genome sequencing analysis

Attributes

Accession	PRJDB15809
Data Type	Genome sequencing
Scope	Monoisolate
Organism	Mycobacterium tuberculosis [Taxonomy ID: 1773] Bacteria; Actinomycetota; Actinomycetes; Mycobacteriales; Mycobacteriaceae; Mycobacterium; Mycobacterium tuberculosis complex; Mycobacterium tuberculosis
Grants	<ul style="list-style-type: none"> 21K10433, Japan Society of Promotion of Science 22fk0108607s0302, Japan Agency for Medical Research and Development
Submission	Registration date: 20-Sep-2023 Infectious Diseases, Kobe Institute of Health

See Genome Information for Mycobacterium tuberculosis

NAVIGATE ACROSS
 3158 additional projects are related by organism

Links to related data

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	550
OTHER DATASETS	
BioSample	550

Parameter	Value
Data volume, Gbases	283
Data volume, Tbytes	0.16

Links to underlying data and samples



<https://ncbi.nlm.nih.gov/bioproject/1019517>

NCBI Taxonomy: Home for Organism-Based data

From the [organism page](#):
 Learn more about the taxonomy & link directly to organism-relevant data

Search term: mycobacterium_tuberculosis[orgn]
Search term: txid1773[orgn]

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	1,462,342	1,233,404
Protein	20,178,795	12,910,399
Structure	3,295	1,446
Genome	6	1
Popset	458	431
Conserved Domains	5	5
GEO Datasets	7,780	4,039
PubMed Central	86,309	81,628
Gene	127,633	1
SRA Experiments	194,537	170,989
GEO Profiles	22,220	16,198
Protein Clusters	2,380	2,880
Identical Protein Groups	796,017	736,741
BioProject	4,217	1,423
BioSample	206,240	184,108
Assembly	7,604	5,568
PubChem BioAssay	10,952	7,539
Taxonomy	2,915	1



New Taxonomy-Based Genome Browsing

Selected taxa

Haemaphysalis longicornis (longhorned tick)

Taxonomic name	Genomes
▼ <i>Eukaryota</i> (eukaryotes)	34,077
▼ <i>Metazoa</i> (animals)	12,395
▼ <i>Arthropoda</i> (arthropods)	4,797
▼ <i>Arachnida</i> (arachnids)	150
▼ <i>Ixodida</i> (ticks)	44
▼ <i>Ixodidae</i> (hardbacked ticks)	30
▼ <i>Haemaphysalis</i>	4
<i>Haemaphysalis longicornis</i> (longhorned tick)	4

Taxonomy landing page (points to *Haemaphysalis*)

Genome table (points to the 4 genomes for *Haemaphysalis longicornis*)

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



NCBI Microbiology Data

NCBI Virus

Quick Access to SARS-CoV-2 Data!

- Novel Severe acute respiratory syndrome coronavirus 2 RefSeq genomes, nucleotide, and protein sequences.
- View our new SARS-CoV-2 interactive dashboard.
- How to submit SARS-CoV-2 sequences.
- Visit our new SARS-CoV-2 Variants Overview New

NCBI Virus is a 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 sequence

Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.

Search by virus

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

<https://www.ncbi.nlm.nih.gov/labs/virus>



NCBI Pathogen Detection Project

Species	New Isolates	Total Isolates
Salmonella enterica	337	647,477
E.coli and Shigella	34	400,070
Campylobacter jejuni	151	124,460
Listeria monocytogenes	6	68,173

[See more organisms...](#)

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

PubChem as a Hub

- Small molecules
- Larger molecules
 - Nucleotides
 - Carbohydrates
 - Lipids
 - Peptides
 - Chemically-modified macromolecules
- World's largest collection of freely accessible chemical information



Compound Page Result – links to lots of data!

CONTENTS	
Title and Summary	
1 Structures	▼
2 Names and Identifiers	▼
3 Chemical and Physical Properties	▼
4 Spectral Information	▼
5 Related Records	▼
6 Chemical Vendors	
7 Drug and Medication Information	▼
8 Pharmacology and Biochemistry	▼
9 Use and Manufacturing	▼
10 Identification	▼
11 Safety and Hazards	▼
12 Toxicity	▼
13 Associated Disorders and Diseases	
14 Literature	▼
15 Patents	▼

MedGen: Hub for Clinical Info

Achondroplasia (ACH)

MedGen UID: 1289 • Concept ID: C0001080 • Congenital Abnormality

Synonyms: ACH; Achondroplastic dwarfism

SNOMED CT: Achondroplasia (86268005); Chondrodystrophia fetalis (86268005); Achondroplastic dwarf (86268005); Osteosclerosis congenita (86268005); Congenital osteosclerosis (86268005); Achondroplastic dwarfism (86268005)

Modes of inheritance: Autosomal dominant inheritance (Orphanet)

Gene (location): FGFR3 (4p16.3)

Monarch Initiative: MONDO:0007037

OMIM®: 100800

Orphanet: ORPHA15

GTR MeSH Orphanet

C Clinical test, **R** Research test, **O** OMIM, **G** GeneReviews, **V** ClinVar



<input type="checkbox"/> NM_000142.5(FGFR3):c.89G>A(p.Arg30His)	FGFR3 (R30H)	Single nucleotide variant (missense variant +1 more)
<input type="checkbox"/> NM_000142.5(FGFR3):c.797T>C(p.Val266Ala)	FGFR3 (V266A)	Single nucleotide variant (missense variant +1 more)

National Library of Medicine
National Center for Biotechnology Information

<https://www.ncbi.nlm.nih.gov/medgen/1289>

Table of contents

- Disease characteristics
- Additional descriptions
- Clinical features
- Term Hierarchy
- Professional guidelines
- Recent clinical studies
- Recent systematic reviews

Those are just a few examples!

National Library of Medicine
National Center for Biotechnology Information

Our role! Workshops & Codeathons

RECRUITMENT, SELECTION & PARTICIPATION

18,233 people *applied* from 138 countries
 7,369 applicants were *accepted* from 120 countries
 3,989 people *participated* from 99 countries



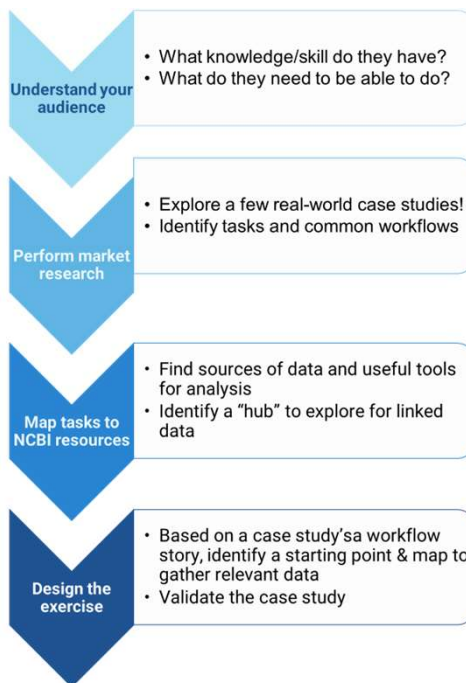
SELECTED EDUCATIONAL EVENTS

WORKSHOPS: Clinical Genetics • Pathogen Genome Analysis • Primer Design
 Human Genetic Variation Analysis • Data Management • Metagenomic Analysis Model
 Organism Genomics • Chemical & Drug Discovery • Literature Mining
 Sequence & Structure Visualization • Bioinformatics/Data Science • Federal Granting
 and more...

CODEATHONS: Petabyte-Scale Sequence Search • Population Genomics VCF Analysis
 Genomic Variant Analysis & Visualization • PubMed Search Results
 Representation to name just a few...

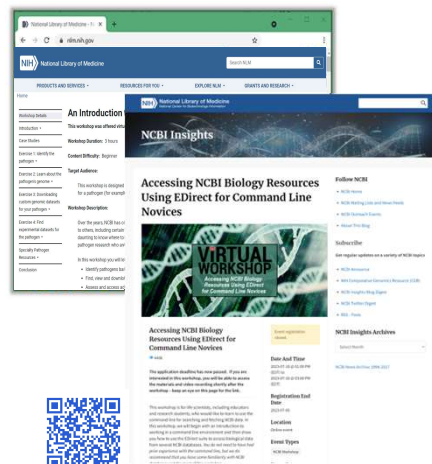
<https://ncbiinsights.ncbi.nlm.nih.gov/ncbi-outreach-events/>

A teaser on our curriculum development process – learn a lot more tomorrow!



To find out more:

- Check out our **Outreach page** for upcoming events: Our past events provide a link to all workshop materials including video recordings!
- Check out the fliers we shared
- **Resource specialist meeting this afternoon!**



[Link](#)