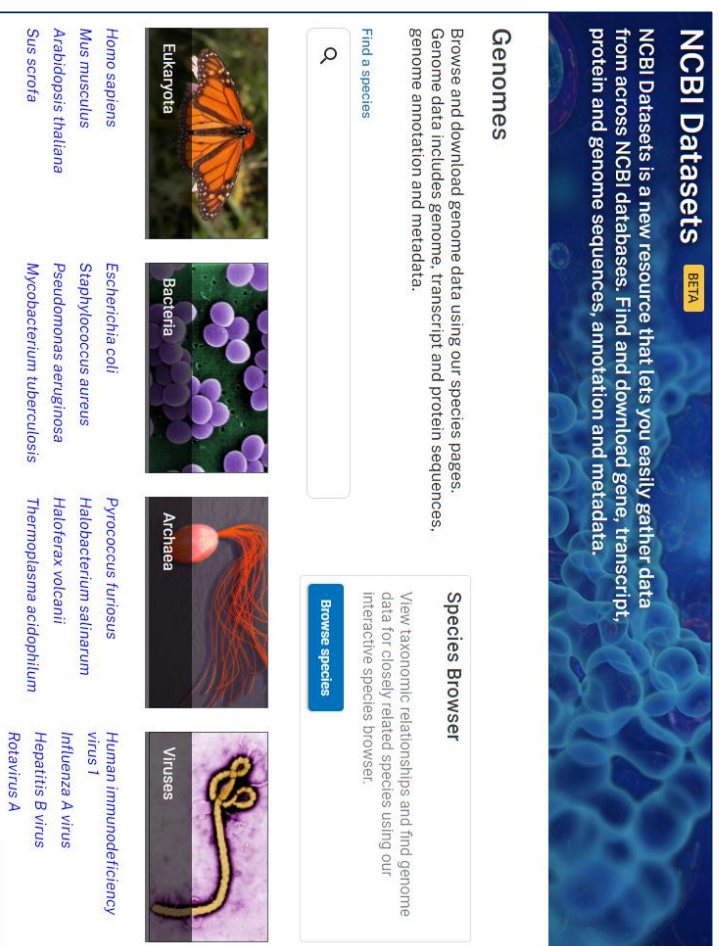


Welcome to the NCBI Datasets workshop!

Cold Spring Harbor Labs Genome Informatics 2021
November 2, 2021

START HERE
<https://bit.ly/cshl2021>

What is NCBI Datasets?



The screenshot shows the NCBI Datasets website. At the top, a banner reads "NCBI Datasets BETA" and "NCBI Datasets is a new resource that lets you easily gather data from across NCBI databases. Find and download gene, transcript, protein and genome sequences, annotation and metadata." Below this is a "Genomes" section with the text: "Browse and download genome data using our species pages. Genome data includes genome, transcript and protein sequences, genome annotation and metadata." To the right is a "Species Browser" section: "View taxonomic relationships and find genome data for closely related species using our interactive species browser." Below these are a search bar and a "Browse species" button. A grid of species categories follows, each with an image and a list of species names: Eukaryota (Monarch butterfly), Bacteria (purple bacteria), Archaea (thermophilic archaeon), and Viruses (various viruses). The species lists include Homo sapiens, Mus musculus, Arabidopsis thaliana, Sus scrofa, Escherichia coli, Staphylococcus aureus, Pseudomonas aeruginosa, Mycobacterium tuberculosis, Pyrococcus furiosus, Halobacterium salinarum, Haloterrax volcanii, Thermoplasma acidophilum, Human immunodeficiency virus 1, Influenza A virus, Hepatitis B virus, and Rotavirus A.

NCBI Datasets BETA

NCBI Datasets is a new resource that lets you easily gather data from across NCBI databases. Find and download gene, transcript, protein and genome sequences, annotation and metadata.

Genomes

Browse and download genome data using our species pages. Genome data includes genome, transcript and protein sequences, genome annotation and metadata.

Find a species

Species Browser

View taxonomic relationships and find genome data for closely related species using our interactive species browser.

Browse species

Eukaryota

Homo sapiens
Mus musculus
Arabidopsis thaliana
Sus scrofa

Bacteria

Escherichia coli
Staphylococcus aureus
Pseudomonas aeruginosa
Mycobacterium tuberculosis

Archaea

Pyrococcus furiosus
Halobacterium salinarum
Haloterrax volcanii
Thermoplasma acidophilum

Viruses

Human immunodeficiency virus 1
Influenza A virus
Hepatitis B virus
Rotavirus A

New resource that makes it easier for users to find and download NCBI sequence data

We're achieving this by creating better web and programmatic interfaces for gathering NCBI sequence data

What data can you get from Datasets?

Datasets data packages include sequence, annotation and metadata for

- Genomes
- Genes
- Orthologs
- SARS-CoV-2 genomes

Homo sapiens

Homo sapiens (human) is a species of primate in the family Hominidae (great apes).

[Browse taxonomy](#)

Current scientific name	Homo sapiens
Common name	human
Taxonomic rank	species
NCBI Taxonomy ID	9606

For more details see NCBI Taxonomy

Genome

Browse all 894 genomes

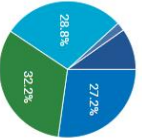
Reference genome GRCh38.p13
Genome Reference Consortium (2019).
RefSeq: GCF_000001405.39

[Download](#)

Genome size	3.1 Gb
Contig N50	57.9 Mb
Genes	54,585

NCBI Annotation Release 109.20210514 May 14, 2021

Current gene set

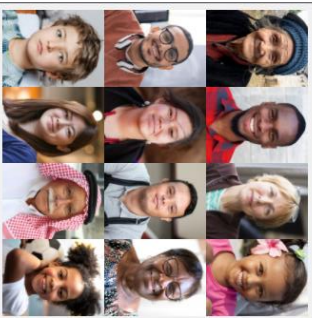


Includes updated and unannotated genes

[View all genes](#)

External links

[Encyclopedia of Life](#)
[GBIF](#)
[iNaturalist](#)
[Wikipedia](#)



How can you access the data packages?

Web
interface

Command
line tools

Client
libraries

Genome

Browse all 894 genomes

Reference genome GRCh38.p13

Genome Reference Consortium (2019).

RefSeq GCF_000001405.39

Download

➔

Download

Data from 4 assemblies

☒ Genomic sequence (FASTA)

☐ Annotated features (GTF)

☐ Annotated features (GFF3)

☐ Sequence and annotation (GFF)

☐ Transcripts (FASTA)

☐ Genomic coding sequences (FASTA)

☐ Protein (FASTA)

File

Your selected data and a detailed data report will be downloaded as a zip

Estimated download size is 1.19 GB

View full file

4 bids

CANCEL DOWNLOAD

Genome size

Contig N50

Genes

NCBI Annotation Release 107

2021

> datasets download genome taxon human
Examples
dataset's download genome accession
GCF_000001405.39
dataset's download genome taxon "bos taurus"
dataset's download genome taxon human
dataset's download genome accession
PRJNA289059

OpenAPI 3.0 REST API

NCBI Datasets REST API

The [OpenAPI 3.0 Specification](#) is an open-source format for APIs. An OpenAPI 3.0 specification serves as the core definition interfaces, and is the single access mechanism used by all interfaces, and is the single access mechanism used by all interfaces.

The [NCBI Datasets OpenAPI 3.0 spec](#) is available in YAML, a variety of open-source tools and other software framework interacting with the REST API in a way that is idiomatic for the environment.

NCBI Datasets API v1

NCBI Datasets is a resource that lets you easily

The Datasets API is still in alpha, and we're updating it often

Explore examples in our documentation

- Quick starts and How-to guides
- Command line reference
- Data schemas
- Data package descriptions
- OpenAPI docs

www.ncbi.nlm.nih.gov/datasets/docs/

Documentation	
Quickstart guides	Documentation / Quickstart guides
Command line tools	<h2>Quickstart Guides</h2> <div>Quickstart: command line tools Install and use the NCBI Datasets command line tools</div> <div>Quickstart: SARS-CoV-2 genomes Use the NCBI Datasets SARS-CoV-2 genome page</div> <div>Quickstart: SARS-CoV-2 proteins Use the NCBI Datasets SARS-CoV-2 protein page</div> <div>Quickstart: genomes Use the NCBI Datasets genome page</div> <div>Quickstart: genes Use the NCBI Datasets gene page</div> <div>Quickstart: orthologs Use the NCBI datasets ortholog page</div>
SARS-CoV-2 genomes	
Genomes	
Genes	
Orthologs	
Command line	
How-to guides	
Data packages	
Programming languages	
Reference	