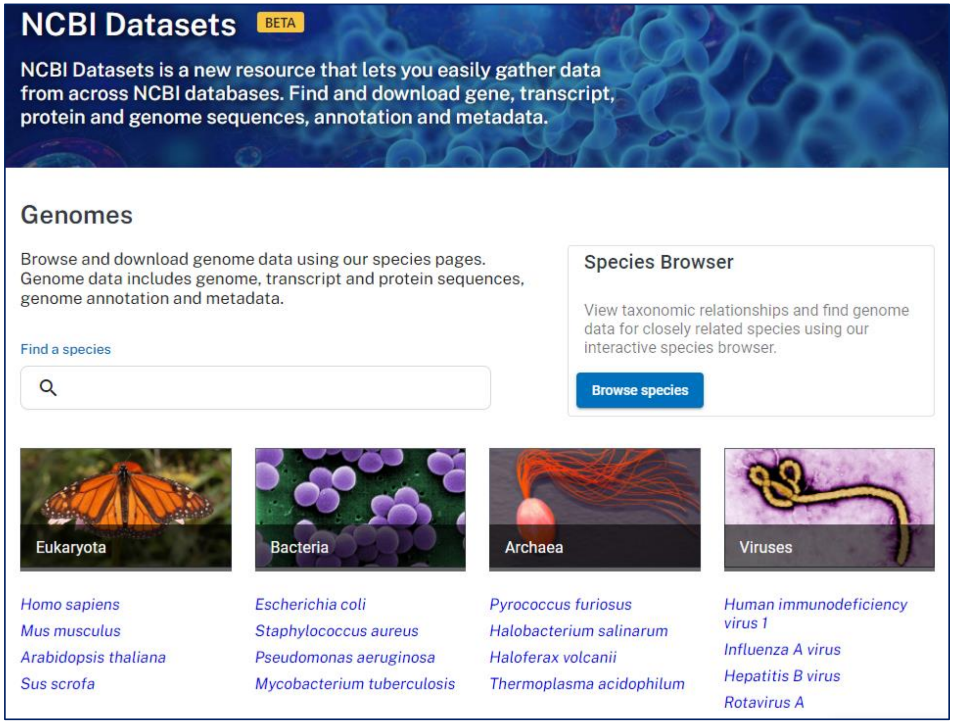
Welcome to the NCBI Datasets workshop!

Cold Spring Harbor Labs Genome Informatics 2021
November 2, 2021

START HERE
https://bit.ly/cshl2021

NIH National Library of Medicine
National Center for Biotechnology Information

What is NCBI Datasets?



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.

Species Browser

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

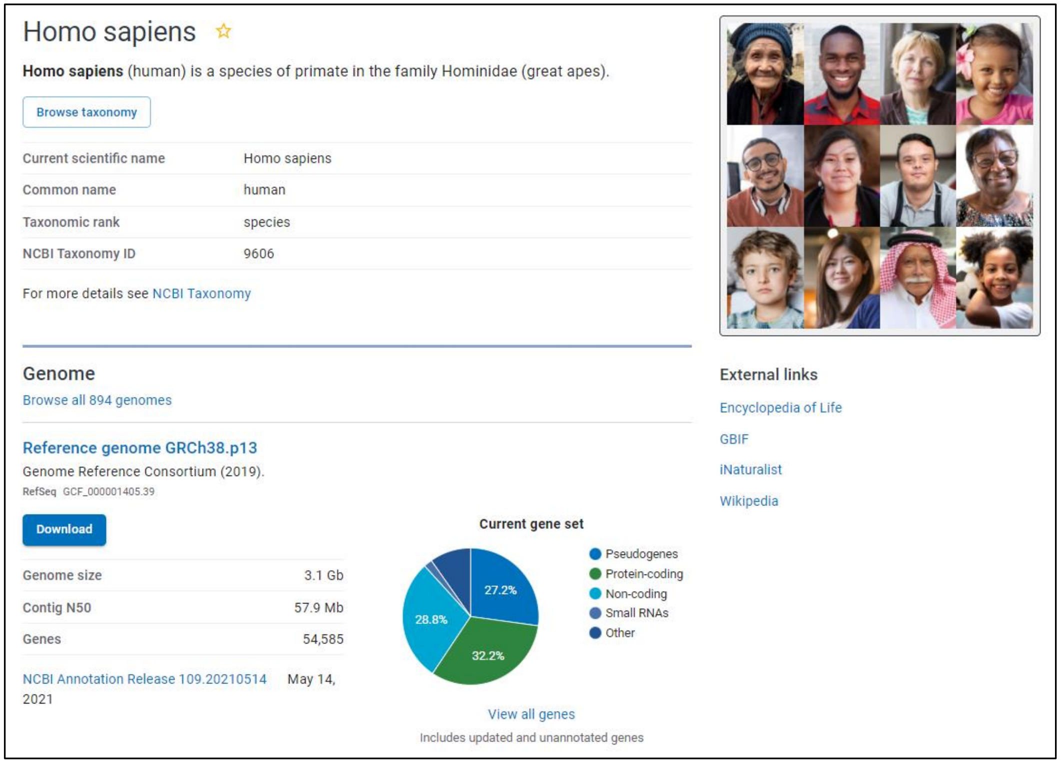
New resource that makes it easier for users to find and download NCBI sequence data

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

View all genes

Includes updated and unannotated genes

External links

Encyclopedia of Life

GBIF

iNaturalist

Wikipedia

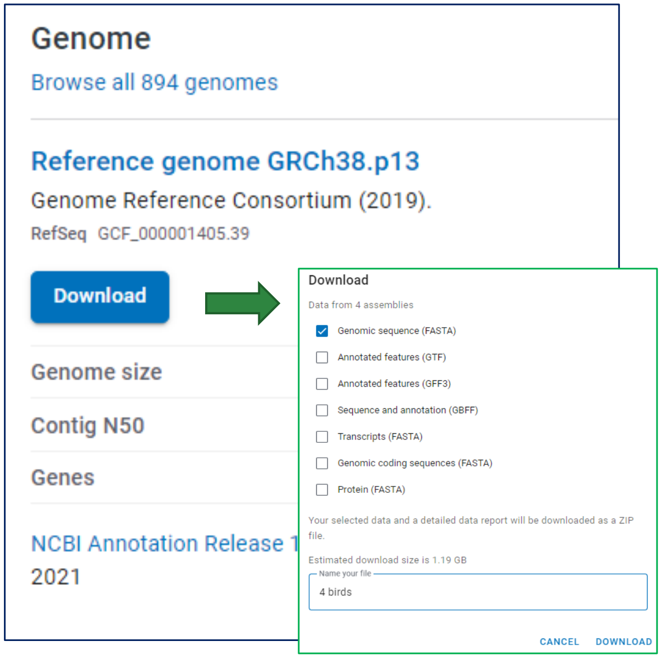
www.ncbi.nlm.nih.gov/data-hub/taxonomy/9606

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

Genome size

Contig N50

Genes

NCBI Annotation Release 1

2021

Download

Data from 4 assemblies

☑ Genomic sequence (FASTA)

□ Annotated features (GTF)

□ Annotated features (GFF3)

□ Sequence and annotation (GBFF)

□ Transcripts (FASTA)

□ Genomic coding sequences (FASTA)

□ Protein (FASTA)

Your selected data and a detailed data report will be downloaded as a ZIP file.

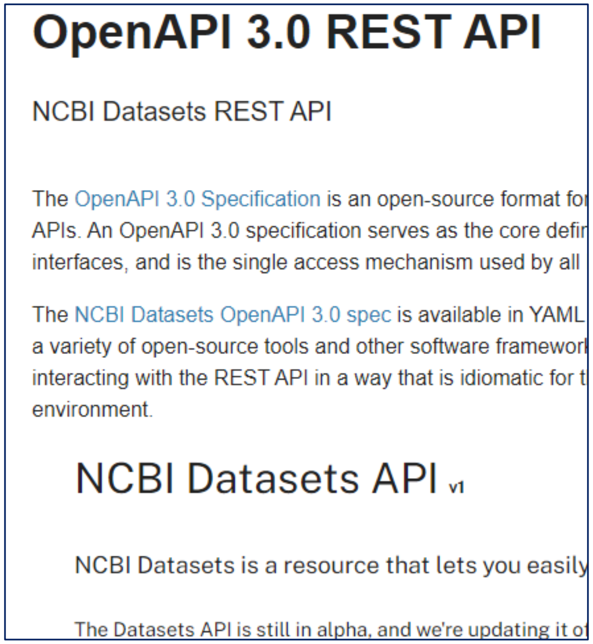
Estimated download size is 1.19 GB

Name your file: 4 birds

CANCEL DOWNLOAD

```
> datasets download genome taxon human

Examples
datasets download genome accession
GCF_000001405.39
datasets  download genome taxon "bos taurus"
datasets download genome taxon human
datasets download genome accession
PRJNA289059
```



OpenAPI 3.0 REST API

NCBI Datasets REST API

The OpenAPI 3.0 Specification is an open-source format fo
APIs. An OpenAPI 3.0 specification serves as the core defi
interfaces, and is the single access mechanism used by all

The NCBI Datasets OpenAPI 3.0 spec is available in YAML
a variety of open-source tools and other software framewor
interacting with the REST API in a way that is idiomatic for t
environment.

NCBI Datasets API v1

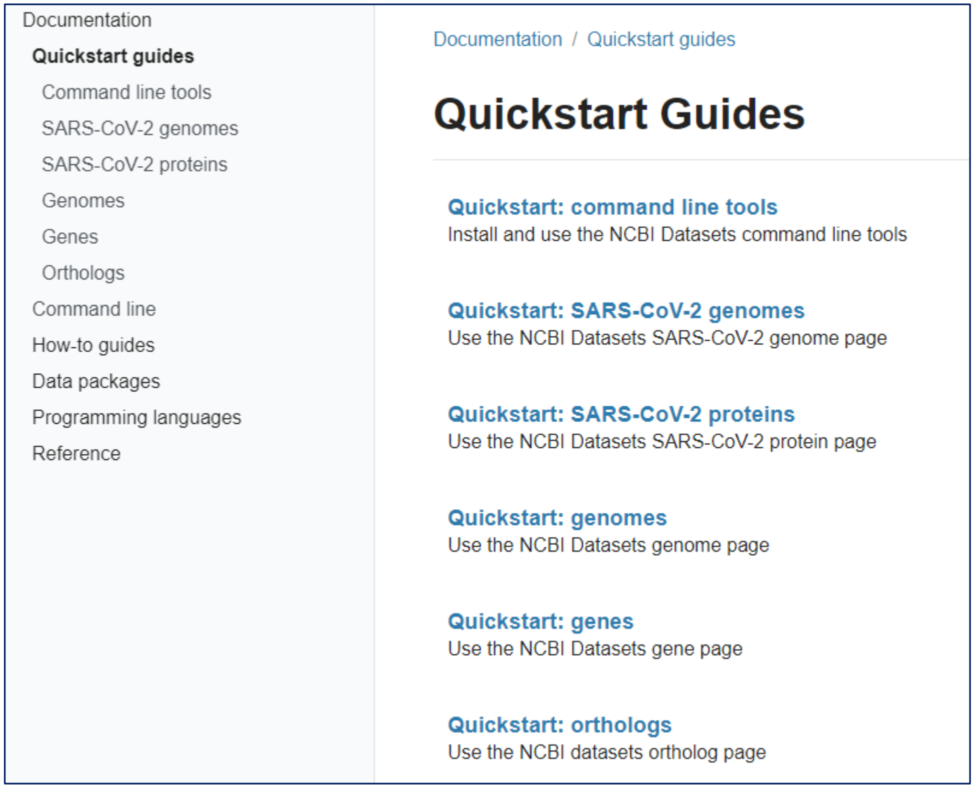
NCBI Datasets is a resource that lets you easily

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

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/



Quickstart Guides

Quickstart: command line tools
Install and use the NCBI Datasets command line tools

Quickstart: SARS-CoV-2 genomes
Use the NCBI Datasets SARS-CoV-2 genome page

Quickstart: SARS-CoV-2 proteins
Use the NCBI Datasets SARS-CoV-2 protein page

Quickstart: genomes
Use the NCBI Datasets genome page

Quickstart: genes
Use the NCBI Datasets gene page

Quickstart: orthologs
Use the NCBI datasets ortholog page