

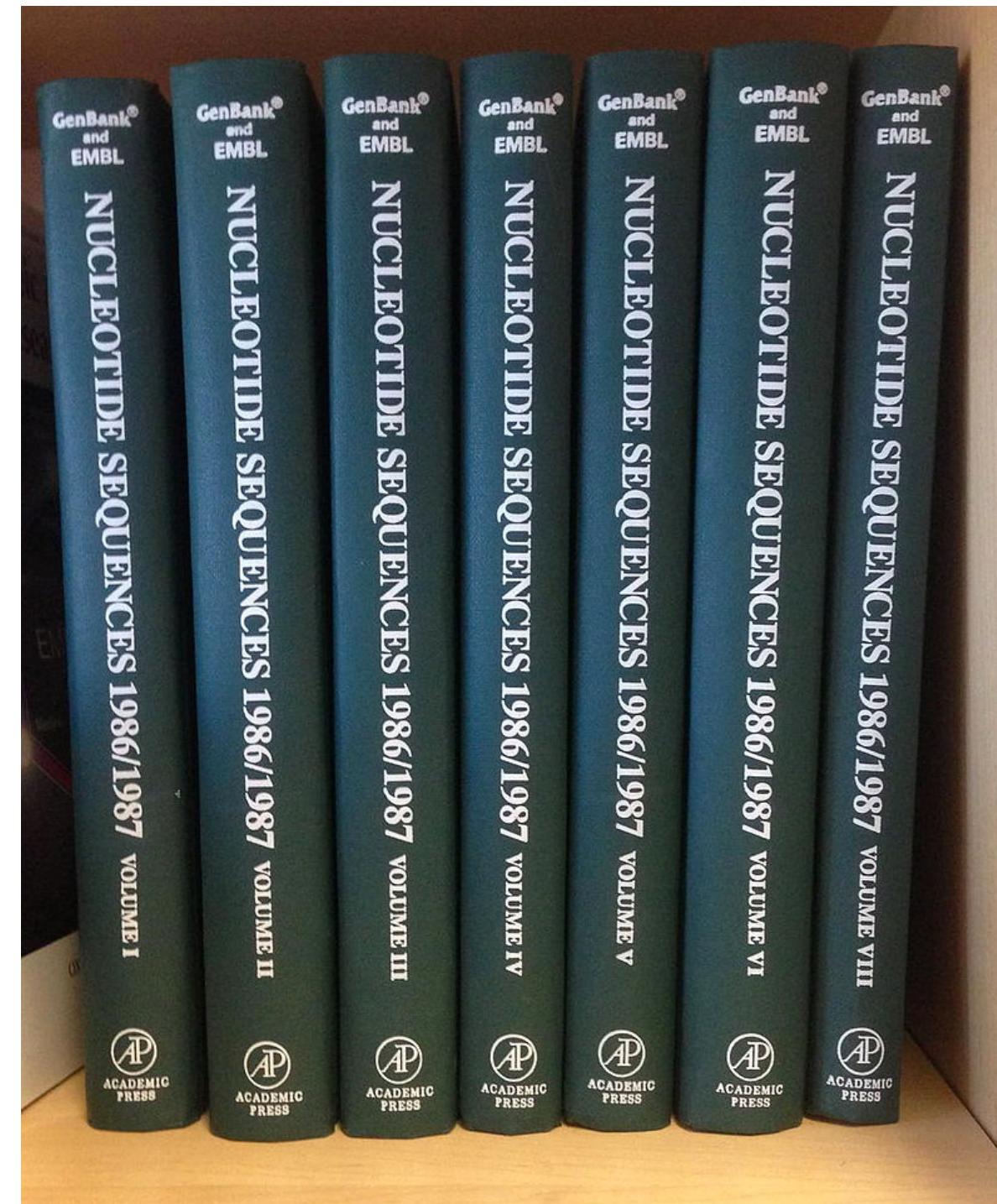
# An overview of bioinformatics databases and online resources: what they are and how to access them

Mark Stenglein, GDW



# GenBank was one of the earliest sequence databases.

GenBank circa 1987



~10,000 sequences

GenBank release 100 (1997)  
distributed by CDROM



Genbank today



>220,000,000 sequences

BOVCHYMOA NUCLEOTIDE SEQUENCES 1984										
SITES:	key	site	span	description	key	site	span	description		
refnumbr	21	1	numbered 1 in [1]		pept/pept	195	0	chymo propept end/ mature pept		
->pept	21	1	chymo prepropept cds start		start					
pept/pept	69	0	chymo prepropept end/ propept start pept<-	1166	1	chymo mature pept cds end				
ORIGIN:	20 bases upstream from codon 1									
SEQUENCE:	1275 bp	293 a	391 c	336 g	255 t					
1	cggcttgcacc	cgtatccaa	atggagggttc	tctgtttgtc	acttgtgtc	ttcgctctct	cccaggggcg	tgagatccac	aggatcccttc	tgttacaatgg
101	caatgttttg	aggaaaggcg	tgaaggagca	tgggtttgtc	gaggatcttc	tgcggaaaca	ggatgtatgg	tttagcaca	atgtttccgg	tttcggggag
201	gtggccggcgg	tgcctttgtac	caatctactg	gtatgtcgt	acttttggaa	gatttaccc	ggggatcccg	ccccgggggtt	caccgtgtcgt	tttgacactg
301	gtctttttgtg	tttcgttggta	ccctttatct	actgtcaagag	aatgtttttgc	aaaaaaccc	gggttttgcg	ccccggggaa	ttgtttccact	tccagaaacct
401	ggggcaagccc	ctgttttatac	actatggggac	aggccggatg	cgaggccatcc	tgggtttatg	caccgtttact	gtttttccaaa	ttttttccat	ccagcagaca
501	gttggccgttg	tgcacccatgg	ccccggggac	gttccatcac	atgtttttttt	cpacggggatc	cttggggatgg	ccatcccccc	gtttttccat	ggatgttttt
601	tacccatgttt	tgcacccatgg	atgttttttt	atgtttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
701	ggcccatccac	ccgtttttttt	atatggggcc	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
801	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
901	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
1001	actggccccc	acagaacatcg	tatgtatgt	tttgatctgt	ctggggatcc	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
1101	atccggatgt	attacatcg	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
1201	acacatcgat	acacatcgat	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt

~1,300,000 sequences

First release: 1982: 606 sequences

# Today, we'll focus mainly on NCBI databases and resources, and how to access them

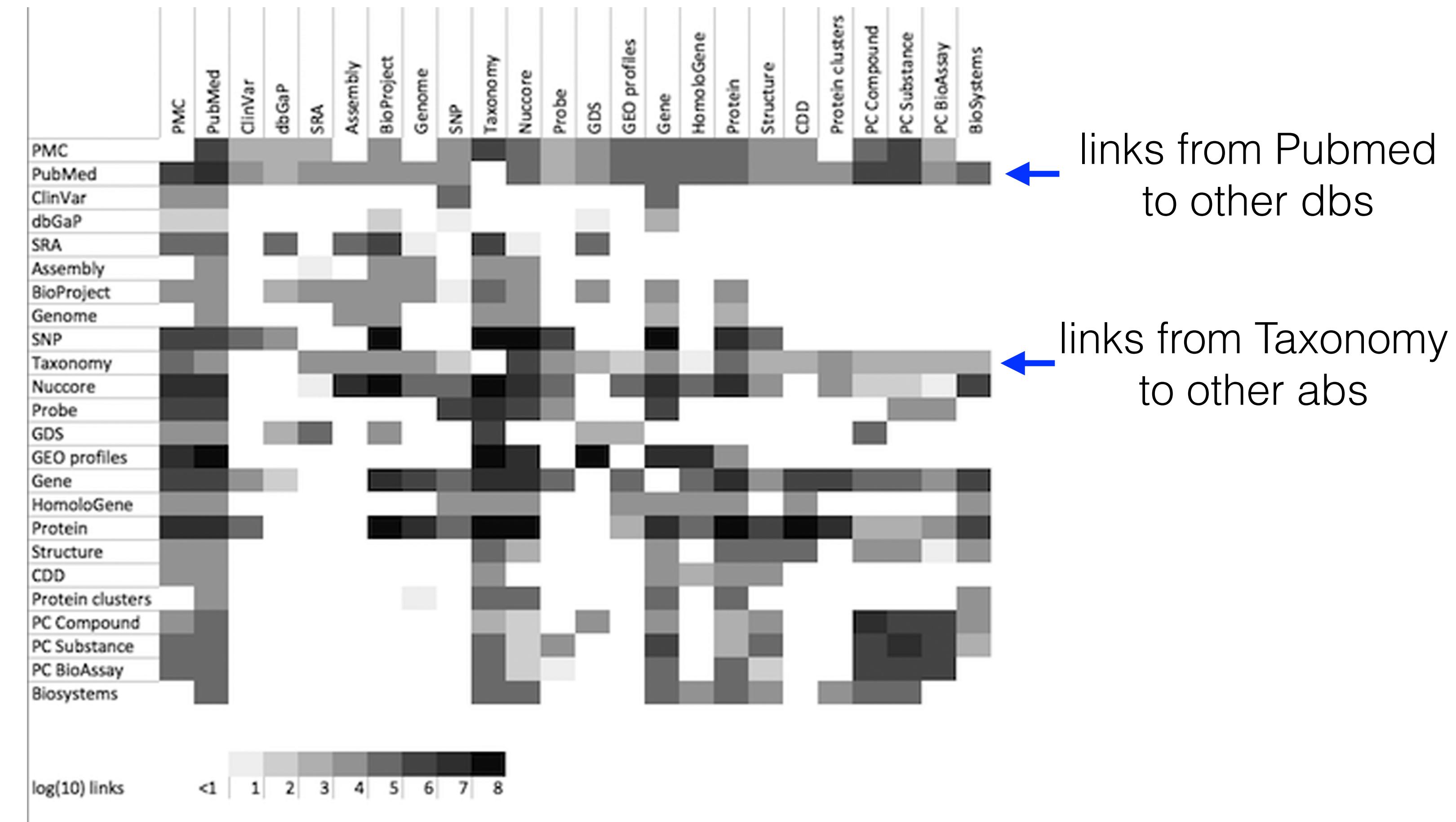
## Categories of NCBI databases

Category	Example NCBI db	Content
Literature	PubMed	Scientific and medical abstracts/ citations
Genomes	Assembly	Genome assembly information
Genes	Gene	Collected information about gene loci
Proteins	Protein	Protein sequences
Chemicals	PubChem Compound	Chemical information with structures, information and links
Health	dbGaP	Genotype/phenotype interaction studies

# One really useful feature of NCBI databases is that they link to each other

So, you can, for example:

- get all the nucleotide sequences associated with a taxon of interested
- get all the protein sequences predicted to be encoded by a genome
- get the SRA datasets associated with a particular paper in Pubmed



# The paper containing the original sequence description of SARS-CoV-2

## Article

# A new coronavirus associated with human respiratory disease in China

<https://doi.org/10.1038/s41586-020-2008-3>

Received: 7 January 2020

Accepted: 28 January 2020

Published online: 3 February 2020

Open access

 Check for updates

Fan Wu<sup>1,7</sup>, Su Zhao<sup>2,7</sup>, Bin Yu<sup>3,7</sup>, Yan-Mei Chen<sup>1,7</sup>, Wen Wang<sup>4,7</sup>, Zhi-Gang Song<sup>1,7</sup>, Yi Hu<sup>2,7</sup>, Zhao-Wu Tao<sup>2</sup>, Jun-Hua Tian<sup>3</sup>, Yuan-Yuan Pei<sup>1</sup>, Ming-Li Yuan<sup>2</sup>, Yu-Ling Zhang<sup>1</sup>, Fa-Hui Dai<sup>1</sup>, Yi Liu<sup>1</sup>, Qi-Min Wang<sup>1</sup>, Jiao-Jiao Zheng<sup>1</sup>, Lin Xu<sup>1</sup>, Edward C. Holmes<sup>1,5</sup> & Yong-Zhen Zhang<sup>1,4,6</sup>✉

Emerging infectious diseases, such as severe acute respiratory syndrome (SARS) and Zika virus disease, present a major threat to public health<sup>1–3</sup>. Despite intense research efforts, how, when and where new diseases appear are still a source of considerable uncertainty. A severe respiratory disease was recently reported in Wuhan, Hubei province, China. As of 25 January 2020, at least 1,975 cases had been reported since the first patient was hospitalized on 12 December 2019. Epidemiological investigations have suggested that the outbreak was associated with a seafood market in Wuhan.

Here we study a single patient who was a worker at the market and who was admitted to the Central Hospital of Wuhan on 26 December 2019 while experiencing a severe respiratory syndrome that included fever, dizziness and a cough. Metagenomic RNA sequencing<sup>4</sup> of a sample of bronchoalveolar lavage fluid from the patient identified a new RNA virus strain from the family *Coronaviridae*, which is designated here ‘WH-Human 1’ coronavirus (and has also been referred to as ‘2019-nCoV’).

Phylogenetic analysis of the complete viral genome (29,903 nucleotides) revealed that the virus was most closely related (89.1% nucleotide similarity) to a group of

# The pubmed record for that paper

← → ⌂ https://pubmed.ncbi.nlm.nih.gov/32015508/ ⌂ ⌂ ⌂

Most Visited blastn blastx blastp NCBI Tax Genbank SRA PubMed FoCo W ggplot theme Kuali CO-COVID\_stats Pre-K Saliva Nextflow :: Anacond...

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

Log in

**PubMed.gov**

Search

Advanced User Guide

Save Email Send to Display options

Case Reports > Nature. 2020 Mar;579(7798):265-269. doi: 10.1038/s41586-020-2008-3.  
Epub 2020 Feb 3.

A new coronavirus associated with human respiratory disease in China

Fan Wu # 1, Su Zhao # 2, Bin Yu # 3, Yan-Mei Chen # 1, Wen Wang # 4, Zhi-Gang Song # 1,  
Yi Hu # 2, Zhao-Wu Tao 2, Jun-Hua Tian 3, Yuan-Yuan Pei 1, Ming-Li Yuan 2, Yu-Ling Zhang 1,  
Fa-Hui Dai 1, Yi Liu 1, Qi-Min Wang 1, Jiao-Jiao Zheng 1, Lin Xu 1, Edward C Holmes 1 5,  
Yong-Zhen Zhang 6 7 8

Affiliations + expand

FULL TEXT LINKS

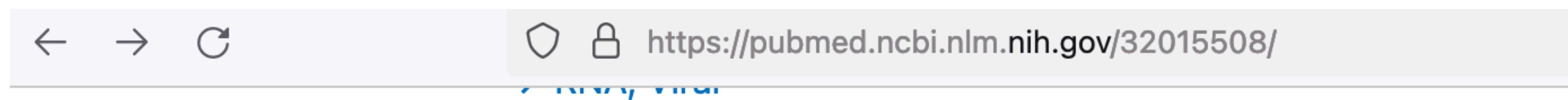
npg nature publishing group

PMC Full text

ACTIONS

Cite Favorites

At the bottom of the pubmed page: related information links



## Related information

[Assembly](#)

[Cited in Books](#)

[Domains](#)

[Gene](#)

[MedGen](#)

[Nucleotide](#) ←

[Nucleotide](#)

[Nucleotide \(Weighted\)](#)

[Protein](#)

[Protein \(RefSeq\)](#)

[Protein \(Weighted\)](#)

[Related Project](#)

[SRA](#)

[Taxonomy via GenBank](#)

Click this link to get to the actual virus genome sequence

# We've jumped to the NCBI nucleotide database (Genbank)

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Advanced Help

**COVID-19 Information**

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

GenBank Send to: ▾ Change region shown ▾

Customize view ▾

Analyze this sequence ▾

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

NCBI Virus ▾

Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

Related information ▾

Assembly

BioProject

Protein ←

PubMed

**Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome**

NCBI Reference Sequence: NC\_045512.2

[FASTA](#) [Graphics](#)

Go to: ▾

LOCUS NC\_045512 29903 bp ss-RNA linear VRL 18-JUL-2020

DEFINITION Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.

ACCESSION NC\_045512

VERSION NC\_045512.2

DBLINK BioProject: [PRJNA485481](#)

KEYWORDS RefSeq.

SOURCE Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

ORGANISM [Severe acute respiratory syndrome coronavirus 2](#)

Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirinae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus.

REFERENCE 1 (bases 1 to 29903)

AUTHORS Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C. and Zhang, Y.Z.

TITLE A new coronavirus associated with human respiratory disease in China

JOURNAL [Nature](#) 570 (7762) 265-266 (2020)

Click this link to get to the protein sequences encoded by this genome

# Now we are in the NCBI protein database

Sequence length  
Custom range...

Molecular weight  
Custom range...

Release date  
Custom range...

Revision date  
Custom range...

[Clear all](#)

[Show additional filters](#)

[ORF7b \[Severe acute respiratory syndrome coronavirus 2\]](#)  
1. Accession: GI: 1820616061  
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[ORF1a polyprotein \[Severe acute respiratory syndrome coronavirus 2\]](#)  
2. Accession: GI: 1802476803  
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[ORF10 protein \[Severe acute respiratory syndrome coronavirus 2\]](#)  
3. 38 aa protein  
Accession: YP\_009725255.1 GI: 1798174256  
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[nucleocapsid phosphoprotein \[Severe acute respiratory syndrome coronavirus 2\]](#)  
4. 419 aa protein  
Accession: YP\_009724397.2 GI: 1798174255  
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[ORF8 protein \[Severe acute respiratory syndrome coronavirus 2\]](#)  
5. 121 aa protein  
Accession: YP\_009724396.1 GI: 1796318604  
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[ORF7a protein \[Severe acute respiratory syndrome coronavirus 2\]](#)  
6. 121 aa protein  
Accession: YP\_009724395.1 GI: 1796318603  
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

**Find related data**

Database: Select

[Find items](#)

**Recent activity**

[Turn Off](#) [Clear](#)

Protein Links for Nucleotide (Select 1798174254) (12) Protein

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, c Nucleotide

Protein Links for Nucleotide (Select 1798172431) (10) Protein

Nucleotide (Weighted) Links for PubMed (Select 32015508) (411230) Nucleotide

Nucleotide Links for PubMed (Select 32015508) (2) Nucleotide

[See more...](#)



You could click on these sequences one at a time to access them

← → ⌂ https://www.ncbi.nlm.nih.gov/protein/YP\_009724390.1

## surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]

NCBI Reference Sequence: YP\_009724390

[Identical Proteins](#) [FASTA](#) [Graphics](#)

---

Go to:

LOCUS	YP_009724390	1273 aa	linear	VRL 18-JUL-2020
DEFINITION	surface glycoprotein [Severe acute respiratory syndrome coronavirus 2].			
ACCESSION	YP_009724390			
VERSION	YP_009724390.1			
DBLINK	BioProject: <a href="#">PRJNA485481</a>			
DBSOURCE	REFSEQ: accession <a href="#">NC_045512.2</a>			
KEYWORDS	RefSeq.			
SOURCE	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)			
ORGANISM	<a href="#">Severe acute respiratory syndrome coronavirus 2</a> Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus.			
REFERENCE	1 (residues 1 to 1273)			
AUTHORS	Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C. and Zhang, Y.Z.			
TITLE	A new coronavirus associated with human respiratory disease in China			
JOURNAL	Nature 579 (7798), 265–269 (2020)			
PUBMED	<a href="#">32015508</a>			
REMARK	Erratum: Nature 2020 Apr; 580(7803):F7 PMID: 32296181			

# Or you can download them all at once, in various formats

Summary ▾ 20 per page ▾ Sort by Default order ▾

Send to: ▾ Filters: [Manage Filters](#)

**Items: 12**

✖ There were some problems retrieving the sequence. GI: 1820616061

✖ There were some problems retrieving the sequence. GI: 1802476803

[ORF7b \[Severe acute respiratory syndrome coronavirus 2\]](#)

1. Accession: GI: 1820616061

[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[ORF1a polyprotein \[Severe acute respiratory syndrome coronavirus 2\]](#)

2. Accession: GI: 1802476803

[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[ORF10 protein \[Severe acute respiratory syndrome coronavirus 2\]](#)

3. 38 aa protein

Accession: YP\_009725255.1 GI: 1798174256

[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[nucleocapsid phosphoprotein \[Severe acute respiratory syndrome coronavirus 2\]](#)

4. 419 aa protein

Accession: YP\_009724397.2 GI: 1798174255

[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

## Choose Destination

- File  Clipboard  
 Collections  Analysis Tool

Download 12 items.

## Format

- ✓ Summary
- GenPept
- GenPept (full)
- FASTA**
- ASN.1
- XML
- INSDSeq XML
- TinySeq XML
- Feature Table
- FASTA CDS
- Accession List
- GI List
- GFF3

sequences

with COBALT

ed Domains wit

ata

ct

## Recent activity

 Protein Links for Nucleotide 1798174254) (12)

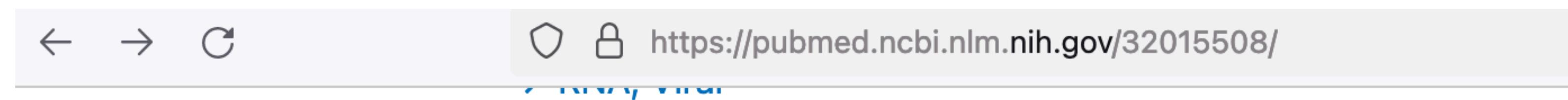
 Severe acute respiratory sy coronaviru 2 isolate Wuha

 Protein Links for Nucleotide 1798172431) (10)

 Nucleotide (Weighted) Link (Select 32015508) (411230

 Nucleotide Links for PubMed 32015508) (2)

At the bottom of the pubmed page: related information links



## Related information

[Assembly](#)  
[Cited in Books](#)  
[Domains](#)  
[Gene](#)  
[MedGen](#)  
[Nucleotide](#)  
[Nucleotide](#)  
[Nucleotide \(Weighted\)](#)  
[Protein](#)  
[Protein \(RefSeq\)](#)  
[Protein \(Weighted\)](#)  
[Related Project](#)  
[SRA](#) ←  
[Taxonomy via GenBank](#)

Click this link to get to the raw sequencing data from this paper



## COVID-19 Information



[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

[Full](#) [Send to:](#) 

### Related information

[BioProject](#)[BioSample](#)[PMC](#)[PubMed](#)[Taxonomy](#)

### Recent activity

[Turn Off](#) [Clear](#)

 SRA Links for PubMed (Select 32015508) (1) SRA

 Protein Links for Nucleotide (Select 1798174254) (12) Protein

 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, c Nucleotide

 Protein Links for Nucleotide (Select 1798172431) (10) Protein

 Nucleotide (Weighted) Links for PubMed (Select 32015508) (411230) Nucleotide

### Links from PubMed

#### [SRX7636886: Complete genome of a novel coronavirus associated with severe human respiratory disease in Wuhan, China](#)

1 ILLUMINA (Illumina MiniSeq) run: 28.3M spots, 8G bases, 2.6Gb downloads

**Design:** Total RNA was extracted from the BALF sample of a patient using the RNeasy Plus Universal Mini Kit (Qiagen) following the manufacturers instructions. An RNA library was then constructed using the SMARTer Stranded Total RNA-Seq Kit v2 (TaKaRa, Dalian, China). Ribosomal RNA (rRNA) depletion was performed during library construction following the manufacturers instructions. Paired-end (150 bp) sequencing of the RNA library was performed on the MiniSeq platform (Illumina).

**Submitted by:** Shanghai Public Health Clinical Center & School of Public Health, Fudan University

**Study:** Complete genome of a novel coronavirus associated with severe human respiratory disease in Wuhan, China

[PRJNA603194](#) • [SRP245409](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

**Sample:**

[SAMN13922059](#) • [SRS6067521](#) • [All experiments](#) • [All runs](#)

**Organism:** [human lung metagenome](#)

**Library:**

**Name:** 1

**Instrument:** Illumina MiniSeq

**Strategy:** RNA-Seq

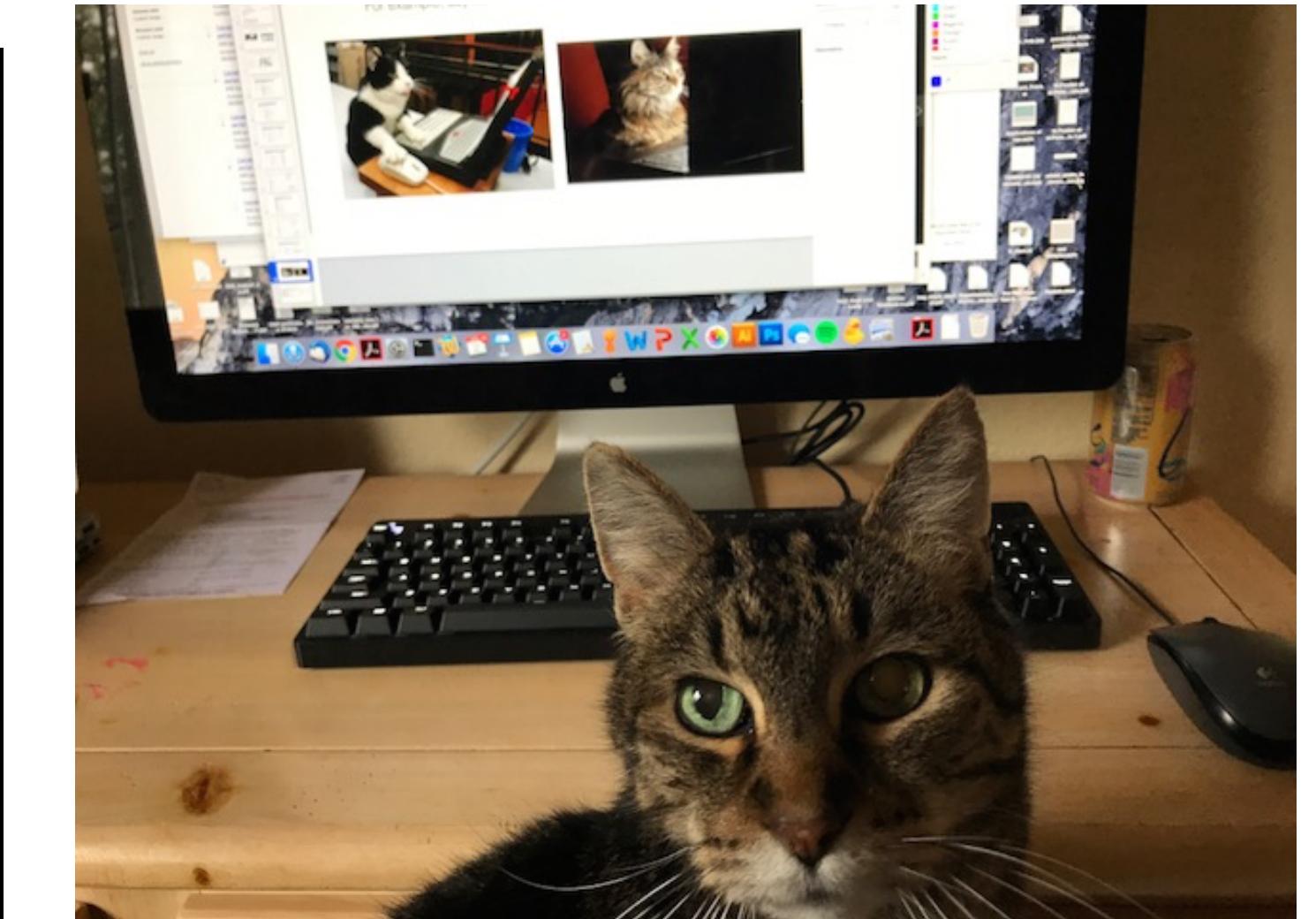
**Source:** METATRANSCRIPTOMIC

**Selection:** RANDOM

**Layout:** PAIRED

There are often many paths to the same data

For example, say we want to download the cat (*Felis catus*) genome



Kirby

# One of my favorite ways to access data in NCBI is via the Taxonomy database

The screenshot shows the NCBI Taxonomy database homepage. At the top, there is a navigation bar with links for 'Resources' and 'How To'. Below the navigation bar, there is a search bar with the word 'Taxonomy' and a dropdown menu. The main content area features a 'COVID-19 Information' banner with links to CDC, NIH, NCBI, HHS, and Spanish resources. To the left of the banner is a grid of butterfly images. The main title 'Taxonomy' is displayed prominently. A descriptive text block states: 'The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.' Below this, there are three columns of links: 'Using Taxonomy' (Quick Start Guide, FAQ, Handbook, Taxonomy FTP), 'Taxonomy Tools' (Browser, Common Tree, Statistics, Name/ID Status, Genetic Codes, Linking to Taxonomy, Extinct Organisms), and 'Other Resources' (GenBank, LinkOut, E-Utilities, Batch Entrez, INSDC).

← → ⌂ https://www.ncbi.nlm.nih.gov/taxonomy ⭐ 📄 ⌂

NCBI Resources How To Sign in to NCE

Taxonomy Taxonomy Search Limits Advanced Help

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

!

**Taxonomy**

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

**Using Taxonomy**

[Quick Start Guide](#)  
[FAQ](#)  
[Handbook](#)  
[Taxonomy FTP](#)

**Taxonomy Tools**

[Browser](#)  
[Common Tree](#)  
[Statistics](#)  
[Name/ID Status](#)  
[Genetic Codes](#)  
[Linking to Taxonomy](#)  
[Extinct Organisms](#)

**Other Resources**

[GenBank](#)  
[LinkOut](#)  
[E-Utilities](#)  
[Batch Entrez](#)  
[INSDC](#)

# One of my favorite ways to access data in NCBI is via the Taxonomy database

Screenshot of the NCBI Taxonomy Browser for Felis catus (Taxonomy ID: 9685). The page shows basic information about the species, its synonyms, and common names. To the right is a sidebar with links to various NCBI databases.

**Entrez records:**

Database name	Direct links
Nucleotide	<a href="#">92,472</a>
Protein	<a href="#">58,274</a>
Structure	<a href="#">21</a>
Genome	<a href="#">1</a>
Popset	<a href="#">207</a>
GEO Datasets	<a href="#">277</a>
PubMed Central	<a href="#">3,386</a>
Gene	<a href="#">46,051</a>
SRA Experiments	<a href="#">2,492</a>
Protein Clusters	<a href="#">12</a>
Identical Protein Groups	<a href="#">45,451</a>
Bio Project	<a href="#">110</a>
Bio Sample	<a href="#">1,649</a>
Bio Systems	<a href="#">495</a>
Assembly	<a href="#">8</a>
Probe	<a href="#">2,877</a>
PubChem BioAssay	<a href="#">1,118</a>
Taxonomy	<a href="#">1</a>

**Annotations:**

- A blue arrow points to the "Genome" link in the sidebar, labeled "genome".
- A blue arrow points to the "SRA Experiments" link in the sidebar, labeled "SRA datasets".

**Information on Felis catus:**

- Taxonomy ID: 9685 (for references in articles please use NCBI:txid9685)
- current name: **Felis catus** Linnaeus, 1758
  - homotypic synonym: **Felis silvestris catus**
  - includes: **Korat cats** L.
- Genbank common name: **domestic cat**
- NCBI BLAST name: **carnivores**
- Rank: **species**
- Genetic code: [Translation table 1 \(Standard\)](#)
- Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)
- Other names:
  - heterotypic synonym: **Felis domesticus**
  - common name(s): **cat, cats**
- Lineage (full):
  - [cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Craniata](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Dipnotetrapodomorpha](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Boreoeutheria](#); [Laurasiatheria](#); [Carnivora](#); [Feliformia](#); [Felidae](#); [Felinae](#); [Felis](#)

# Felis catus in the NCBI genome database

https://www.ncbi.nlm.nih.gov/genome/?term=txid9685[Organism:noexp]

Genome    Genome txid9685[Organism:noexp] Search Create alert Limits Advanced Help

**COVID-19 Information** X

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

**Felis catus (domestic cat)**  
Reference genome: [Felis catus \(assembly Felis\\_catus\\_9.0\)](#)  
Download sequences in FASTA format for [genome](#), [transcript](#), [protein](#)  
Download genome annotation in [GFF](#), [GenBank](#) or [tabular](#) format  
BLAST against Felis catus [genome](#), [transcript](#), [protein](#)

All 4 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: ID: 78

Organism Overview ; [Genome Assembly and Annotation report \[4\]](#) ; [Organelle Annotation Report \[1\]](#)

 **Felis catus (domestic cat)**  
domestic cat

Lineage: [Eukaryota](#)[7836]; [Metazoa](#)[3708]; [Chordata](#)[1775]; [Craniata](#)[1753]; [Vertebrata](#)[1753]; [Euteleostomi](#)[1737]; [Mammalia](#)[470]; [Eutheria](#)[445]; [Laurasiatheria](#)[255]; [Carnivora](#)[65]; [Feliformia](#)[24]; [Felidae](#)[16]; [Felinae](#)[11]; [Felis](#)[2]; [Felis catus](#)[1]

*Felis catus*, the domestic cat, provides several valuable models for infectious disease, including a model for human AIDS. With a large number of recognized breeds, the cat is also a valuable resource for studying phenotypic diversity and evolution. The cat genome will further facilitate research in human medicine as some rare diseases that occur [More...](#)

**Summary**

**Sequence data:** genome assemblies: 4; sequence reads: 2 (See [Genome Assembly and Annotation report](#))  
**Statistics:** median total length (Mb): 2507.5  
median protein count: 54726  
median GC%: 41.8903

NCBI Annotation Database: 101

**NCBI Resources**  
Genome Data Viewer

**Tools**  
BLAST Genome

**Related information**  
Assembly  
BioProject  
Gene  
Components  
Protein  
PubMed  
Taxonomy

**Search details**  
txid9685[Organism:noexp]

# There are actually 4 cat genome assemblies in NCBI

https://www.ncbi.nlm.nih.gov/genome/?term=txid9685[Organism:noexp]

Genome    Genome txid9685[Organism:noexp] Search Create alert Limits Advanced Help

**COVID-19 Information** X

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

**Felis catus (domestic cat)**  
Reference genome: [Felis catus \(assembly Felis\\_catus\\_9.0\)](#)  
Download sequences in FASTA format for [genome](#), [transcript](#), [protein](#)  
Download genome annotation in [GFF](#), [GenBank](#) or [tabular](#) format  
BLAST against Felis catus [genome](#), [transcript](#), [protein](#)

All 4 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: ID: 78

Organism Overview ; [Genome Assembly and Annotation report \[4\]](#) ; [Organelle Annotation Report \[1\]](#)

**Felis catus (domestic cat)**  
domestic cat

 Lineage: [Eukaryota](#)[7836]; [Metazoa](#)[3708]; [Chordata](#)[1775]; [Craniata](#)[1753]; [Vertebrata](#)[1753]; [Euteleostomi](#)[1737]; [Mammalia](#)[470]; [Eutheria](#)[445]; [Laurasiatheria](#)[255]; [Carnivora](#)[65]; [Feliformia](#)[24]; [Felidae](#)[16]; [Felinae](#)[11]; [Felis](#)[2]; [Felis catus](#)[1]

*Felis catus*, the domestic cat, provides several valuable models for infectious disease, including a model for human AIDS. With a large number of recognized breeds, the cat is also a valuable resource for studying phenotypic diversity and evolution. The cat genome will further facilitate research in human medicine as some rare diseases that occur [More...](#)

**Summary**

**Sequence data:** genome assemblies: 4; sequence reads: 2 (See [Genome Assembly and Annotation report](#))  
**Statistics:** median total length (Mb): 2507.5  
median protein count: 54726  
median GC%: 41.8903

NCBI Annotation Database: 101

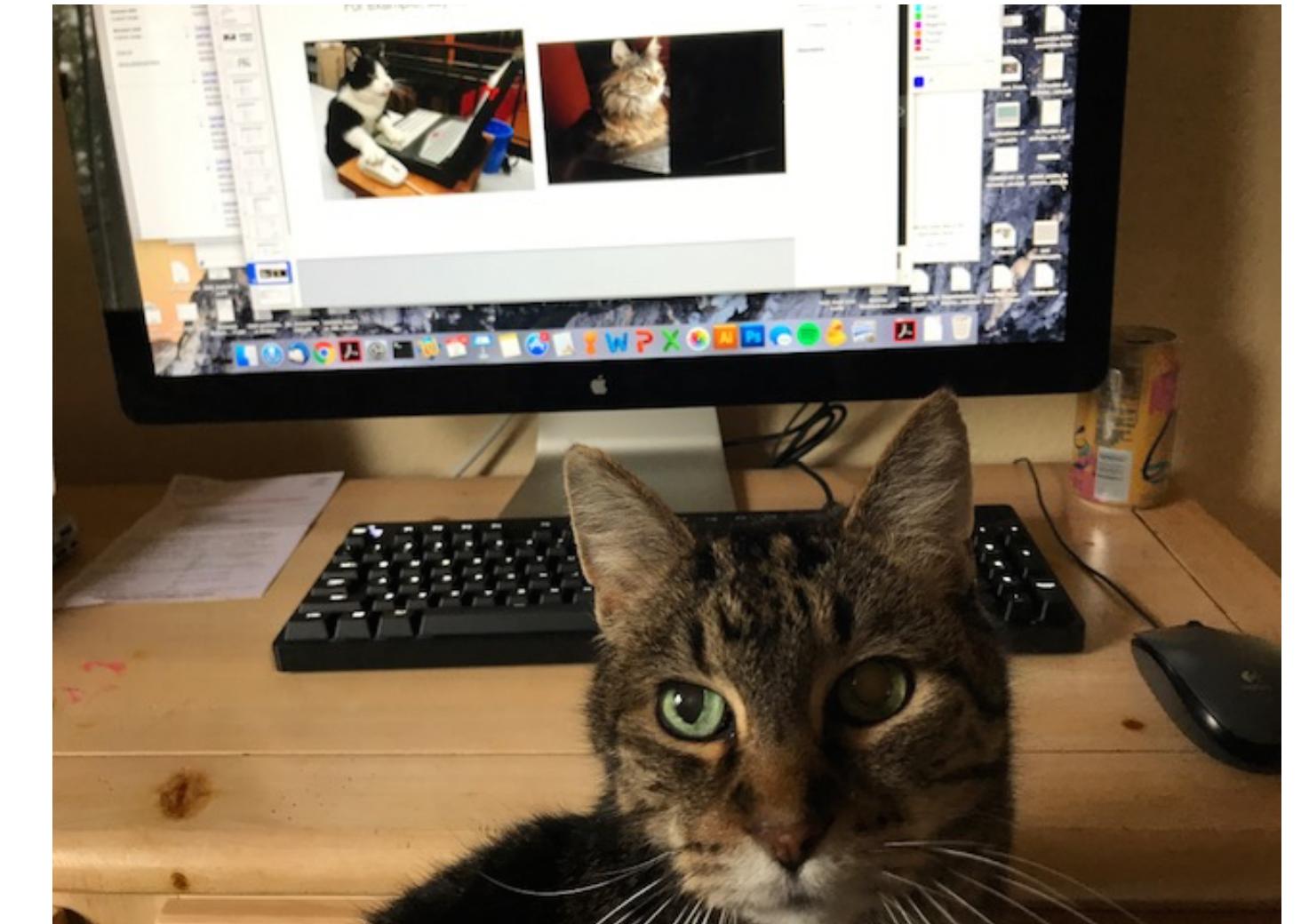
**NCBI Resources**  
Genome Data Viewer

**Tools**  
BLAST Genome

**Related information**  
Assembly  
BioProject  
Gene  
Components  
Protein  
PubMed  
Taxonomy

**Search details**  
txid9685 [Organism:noexp]

In reality, there are at least as many cat genomes as there are cats



Kirby

# You can go up the taxonomic tree in the Taxonomy db

## **Felis catus**

Taxonomy ID: 9685 (for references in articles please use NCBI:txid9685)

current name

***Felis catus*** Linnaeus, 1758

homotypic synonym: ***Felis silvestris catus***

includes: **Korat cats** L.

Genbank common name: **domestic cat**

NCBI BLAST name: **carnivores**

Rank: **species**

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)

Other names:

heterotypic synonym

***Felis domesticus***

common name(s)

**cat, cats**

### [Lineage](#)(full )

[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Craniata](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Dipnotetrapodomorpha](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Boreoeutheria](#); [Laurasiatheria](#); [Carnivora](#); [Feliformia](#); [Felidae](#); [Felinae](#); [Felis](#)



**Felidae**

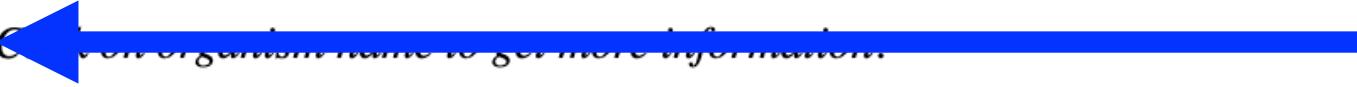
# You can go up the taxonomic tree in the Taxonomy db

Search for  as complete name  lock

Display 3 levels using filter: none

Nucleotide    Protein    Structure    Genome    Popset    SNP  
 Gene    HomoloGene    SRA Experiments    LinkOut    BLAST    GEO Profiles  
 Bio Project    Bio Sample    Bio Systems    Assembly    dbVar    Genetic Testing Registry  
 PubChem BioAssay    Conserved Domains    GEO Datasets    PubMed Central  
 Protein Clusters    Identical Protein Groups    SPARCLE  
 Host    Viral Host    Probe

[Lineage](#) (full): cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Carnivora; Feliformia

- [Felidae](#) (cat family) 38 38 genomes for species in Felidae
  - [Acinonychinae](#) 1
    - [Acinonyx](#) 1
      - [Acinonyx jubatus](#) (cheetah) 1
  - [Felinae](#) 30
    - [Caracal](#) 1
      - [Caracal caracal](#) 1
    - [Catopuma](#) 2
      - [Catopuma badia](#) (bay cat) 1
      - [Catopuma temminckii](#) (Asiatic golden cat) 1
    - [Felinae intergeneric hybrids](#) 1
      - [Felis catus x Leopardis geoffroyi](#) 1
      - [Felis catus x Prionailurus bengalensis](#)
      - [Leptailurus serval x Caracal caracal](#)
    - [Felis](#) 5
      - [Felis catus](#) (domestic cat) 1
      - [Felis chaus](#) (jungle cat) 1
      - [Felis chaus x Felis catus](#)
      - [Felis margarita](#) (sand cat) 1
      - [Felis nigripes](#) (black-footed cat) 1
    - [Felis silvestris](#) (wild cat) 1
    - [unclassified Felis](#)
    - [environmental samples](#)
  - [Leopardus](#) 7

# Felidae genomes

← → ⌂ https://www.ncbi.nlm.nih.gov/genome/?term=txid9681[Organism:exp] ⌂ Sign in to NCBI

NCBI Resources How To

Genome Genome txid9681[Organism:exp] Search Create alert Limits Advanced Help

**COVID-19 Information** X

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

See also 22 organelle- and plasmid-only records matching your search

Display Settings: Summary, 20 per page Send to: Database: Select

**Search results**

Items: 16

[Panthera tigris](#)  
1. tiger  
Kingdom: Eukaryota; Subgroup: Mammals  
Sequence data: genome assemblies:3  
Haploid chromosomes: 19; Organelles: 1  
Date: 2013/09/05  
ID: 10802

[Panthera leo](#)  
2. [Panthera leo overview](#)  
Kingdom: Eukaryota; Subgroup: Mammals  
Sequence data: genome assemblies:3  
Haploid chromosomes: 19  
Date: 2019/10/01  
ID: 13342

**Filters:** [Manage Filters](#)

**Find related data**

**Search details**

txid9681[Organism:exp]

**Recent activity**

Turn Off Clear

txid9681[Organism:exp] (16) Genome

felis catus (1) Taxonomy

You can download genome sequences from the genome database

← → C https://www.ncbi.nlm.nih.gov/genome/10802 ⌂ ⌃ ⌁ ⌂ Sign in to NCBI

NCBI Resources How To

Genome **Genome** Search Limits Advanced Help

**COVID-19 Information** X

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

**Panthera tigris (tiger)**  
Representative genome: [Panthera tigris altaica \(assembly PanTig1.0\)](#)  
Download sequences in FASTA format for [genome](#), [transcript](#), [protein](#) ← **Download links**  
Download genome annotation in [GFF](#), [GenBank](#) or [tabular](#) format  
BLAST against Panthera tigris [genome](#), [transcript](#), [protein](#)

All 3 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: ID: 10802

[Organism Overview](#) ; [Genome Assembly and Annotation report \[3\]](#) ; [Organelle Annotation Report \[2\]](#)

**Panthera tigris (tiger)**  
tiger 

Lineage: [Eukaryota\[7836\]](#); [Metazoa\[3708\]](#); [Chordata\[1775\]](#); [Craniata\[1753\]](#); [Vertebrata\[1753\]](#); [Euteleostomi\[1737\]](#); [Mammalia\[470\]](#); [Eutheria\[445\]](#); [Laurasiatheria\[255\]](#); [Carnivora\[65\]](#); [Feliformia\[24\]](#); [Felidae\[16\]](#); [Pantherinae\[4\]](#); [Panthera\[4\]](#); [Panthera tigris\[1\]](#)

Tiger is the largest wild cat, and along with other members of the genus *Panthera*, the most endangered species in the world. There are five extant subspecies, namely the Bengal tiger, South China or Amoy tiger, Indochinese tiger, Sumatran tiger and the Siberian tiger. Three subspecies, the Bali tiger, Javan tiger and the Caspian tiger, are extinct.

**Summary**

Sequence data: genome assemblies: 3; sequence reads: 3 (See [Genome Assembly and Annotation report](#))  
Statistics: median total length (Mb): 2424.64

NCBI Resources

Genome Data Viewer

Tools

BLAST Genome

Related information

Assembly

BioProject

Gene

Components

Protein

PubMed

Taxonomy

Recent activity

Turn Off Clear

# Non-NCBI databases include GISAID: a repository for virus sequences

© 2008 - 2021 | Terms of Use | Privacy Notice | Contact  
You are logged in as **Mark Stenglein** - [logout](#)

Registered Users EpiFlu™ **EpiCoV™** EpiRSV™ My profile

[!\[\]\(b7d27235b6dd1a5ec2f50548477b4f0a\_img.jpg\) EpiCoV™](#) [!\[\]\(df066fe55682b2c9295d48d23cc06f50\_img.jpg\) Search](#) [!\[\]\(16c8dda5b2c8d903678ae1392fab538f\_img.jpg\) Downloads](#) [!\[\]\(9a613ab50383b8e14cacbed708ace0b5\_img.jpg\) Upload](#)

## Search

Accession ID  Virus name   complete  high coverage  
Location  Host  low coverage excl  w/Patient status  
Collection  to  Submission  to   collection date compl  
Clade all Lineage  Substitutions  Variants  [Reset](#) [Fulltext ▲](#)

<input type="checkbox"/>	Virus name	Passage de	Accession ID	Collection da	Submission D	<a href="#">i</a>	Length	Host	Location	Originating
<input type="checkbox"/>	hCoV-19/Japan/YCH0433/2021	Original	EPI_ISL_3183964	2021-07-27	2021-08-02	<a href="#">i</a>	29,825	Human	Asia / Japan / Ya	Department
<input type="checkbox"/>	hCoV-19/Japan/YCH0432/2021	Original	EPI_ISL_3183963	2021-07-27	2021-08-02	<a href="#">i</a>	29,823	Human	Asia / Japan / Ya	Department
<input type="checkbox"/>	hCoV-19/Japan/YCH0431/2021	Original	EPI_ISL_3183962	2021-07-27	2021-08-02	<a href="#">i</a>	29,823	Human	Asia / Japan / Ya	Department
<input type="checkbox"/>	hCoV-19/Japan/YCH0430/2021	Original	EPI_ISL_3183961	2021-07-27	2021-08-02	<a href="#">i</a>	29,822	Human	Asia / Japan / Ya	Department
<input type="checkbox"/>	hCoV-19/Japan/YCH0429/2021	Original	EPI_ISL_3183960	2021-07-28	2021-08-02	<a href="#">i</a>	29,830	Human	Asia / Japan / Ya	Department
<input type="checkbox"/>	hCoV-19/Japan/YCH0428/2021	Original	EPI_ISL_3183959	2021-07-28	2021-08-02	<a href="#">i</a>	29,822	Human	Asia / Japan / Ya	Department
<input type="checkbox"/>	hCoV-19/Japan/YCH0427/2021	Original	EPI_ISL_3183958	2021-07-27	2021-08-02	<a href="#">i</a>	29,838	Human	Asia / Japan / Ya	Genome Ar
<input type="checkbox"/>	hCoV-19/Japan/YCH0426/2021	Original	EPI_ISL_3183957	2021-07-27	2021-08-02	<a href="#">i</a>	29,822	Human	Asia / Japan / Ya	Genome Ar
<input type="checkbox"/>	hCoV-19/Japan/YCH0425/2021	Original	EPI_ISL_3183956	2021-07-27	2021-08-02	<a href="#">i</a>	29,816	Human	Asia / Japan / Ya	Genome Ar
<input type="checkbox"/>	hCoV-19/Japan/YCH0424/2021	Original	EPI_ISL_3183955	2021-07-27	2021-08-02	<a href="#">i</a>	29,835	Human	Asia / Japan / Ya	Genome Ar

Total: 2,567,276 viruses

<< < 1 2 3 4 5 > >>

Select  Analysis  Download

2.5 million SARS-CoV-2 sequences!!

# GISAID has some nice features, but is limited to a few pathogens

The screenshot shows the GISAID search interface. At the top, there are tabs for Registered Users, EpiFlu™ (selected), EpiCoV™, EpiRSV™, and My profile. Below the tabs are navigation icons: Search, Back to results, Worksets, Upload, Batch Upload, Settings, and Analysis. Status counts are displayed: Count 133 viruses, GISAID published 189,014 viruses (879,573 sequences), Total count 342,557 viruses (1,480,042 sequences). A section titled 'Basic filters' includes a dropdown for Predefined search, radio buttons for Search in Released files or Worksets (Released files is selected), and a search patterns input field. The main search area displays filters for Type (A, B, C), H (1-10), N (1-10), Lineage (empty), Host (-all-, Human, Animal, Avian, Chicken, Curlew, Duck, Eagle, Falcon, Goose), and Location (empty). The 'Avian' host and 'Asia' location filters are highlighted with a blue dashed border.

Type	H	N	Lineage	Host	Location
A	1	1		-all-	Bahrain
B	2	2		Human	Bangladesh
C	3	3		Animal	Bhutan
	4	4		Avian	Antarctica
	5	5		Chicken	Asia
	6	6		Curlew	British Indian Ocean Territory
	7	7		Duck	Brunei
	8	8		Eagle	Cambodia
	9	9		Falcon	China
	10	10		Goose	Christmas Island

Download all the IAV H5N1 sequences from birds in Hong Kong  
(133 viruses)

# GISAID requires approval to access data and has restrictive terms of use

## GISAID EPIFLU™ DATABASE ACCESS AGREEMENT

Effective: March 16, 2011

**WHEREAS** Freunde von GISAID e.V. ("GISAID") maintains a global database for influenza gene sequences along with associated data, including virological, clinical, epidemiological and demographic information (if available) for all influenza viruses, including but not limited to H5N1 sequences, (the "GISAID EpiFlu™ Database") for the purpose of facilitating the sharing, research and investigation of such sequences and associated data.

**NOW, therefore,** this Database Access Agreement (the "Agreement") is entered into by and between the undersigned ("You") and GISAID.

- Access to the GISAID EpiFlu™ Database, Data.** Access to, and use of, the GISAID EpiFlu™ Database and Data, as defined herein, is governed by this Agreement. By accessing or otherwise using the GISAID EpiFlu™ Database, whether as a provider or user of Data, You accept and agree to be bound by the terms of this Agreement. For purposes of this Agreement, the term "**Data**" means any and all (i) sequence data and other associated data and information contained in the GISAID EpiFlu™ Database pertaining to influenza viruses, (ii) any annotations, corrections, updates, modifications, improvements, derivatives or other enhancements to any such data contained in the GISAID EpiFlu™ Database, and (iii) any safety information relevant to use of the data or to regulatory approval of vaccines or other therapies that embody or utilize the data contained in the GISAID EpiFlu™ Database.
- License Terms.** You are hereby granted a non-exclusive, worldwide, royalty-free, non-transferable and revocable license to access and use the GISAID EpiFlu™ Database and Data solely in accordance with this Agreement in all its terms. Without limiting the foregoing, your access to and use of the GISAID

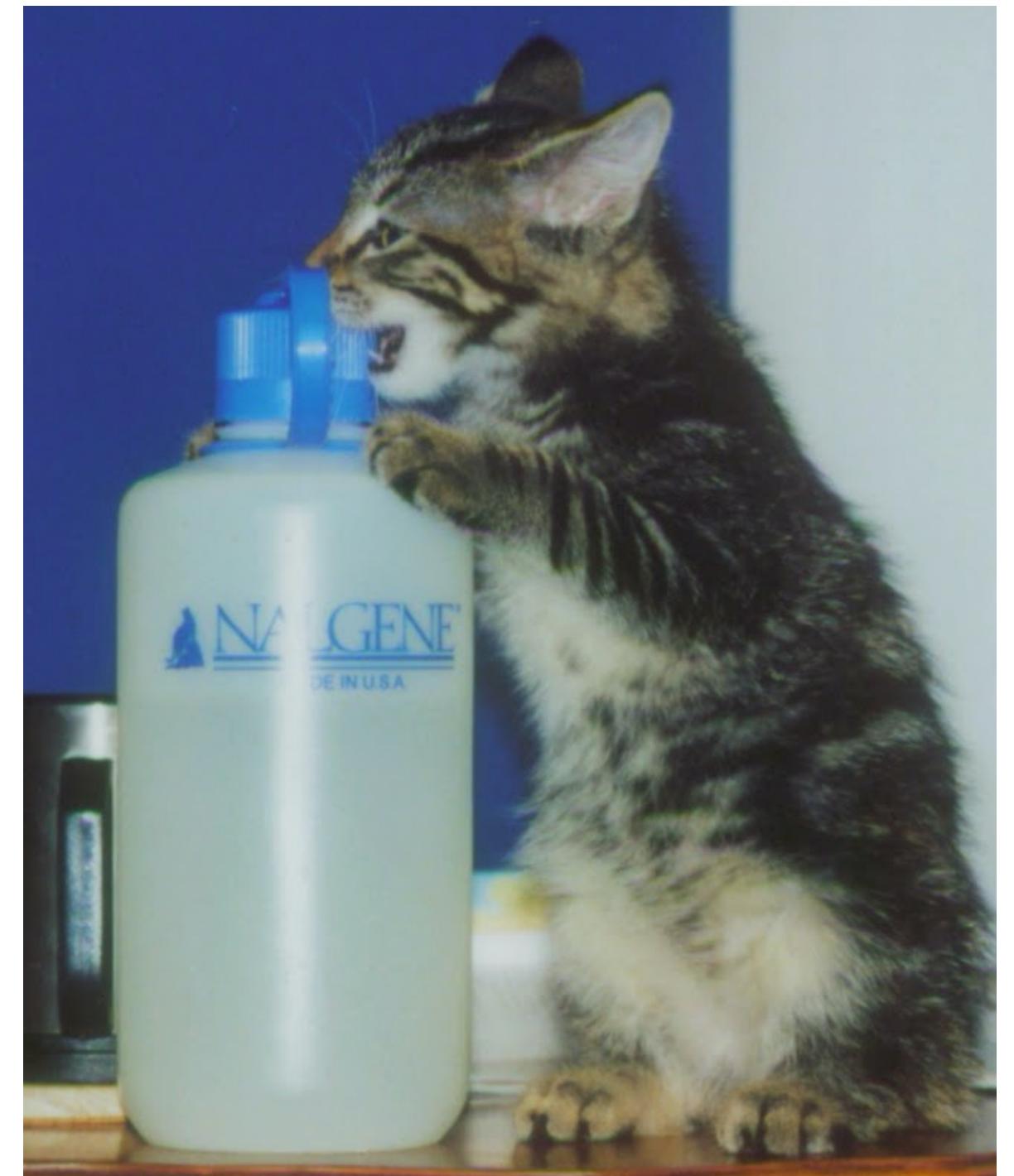
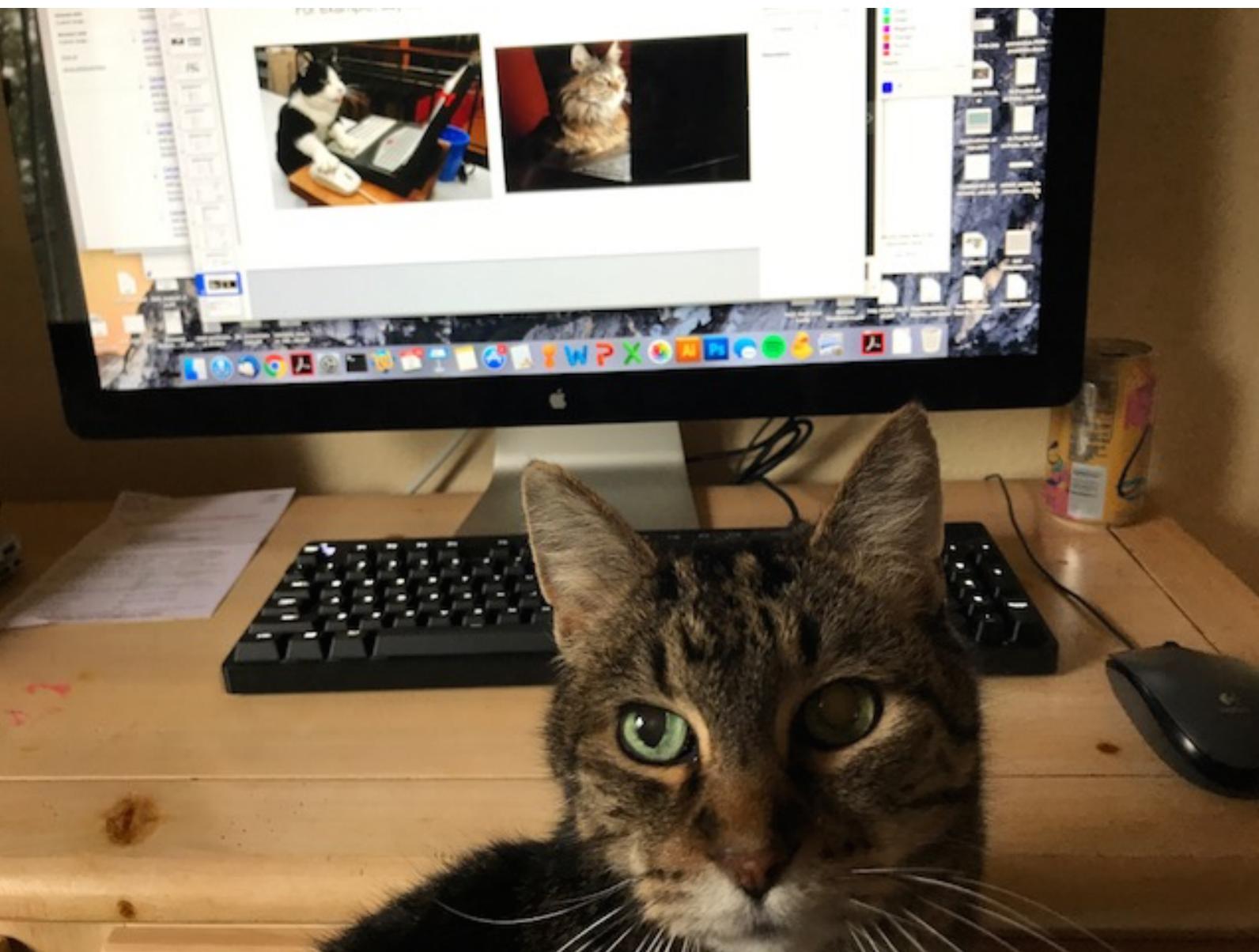
# Some data lives in non-standard locations

The screenshot shows a web browser window with the URL [gigadb.org/dataset/100060](http://gigadb.org/dataset/100060). The page title is "ASSEMBLATHON 2". It features a green sidebar on the left with the text "Assemblathon 2 assemblies." and "Dataset type: Genomic Data released on June 24, 2013". The main content area contains a large block of author names and a citation: "Bradnam KR; Fass JN; Alexandrov A; Baranay P; Bechner M; Birol I; Boisvert S; Chapman JA; Chapuis G; Chikhi R; Chitsaz H; Chou W; Corbeil J; Del Fabbro C; Docking TRR; Durbin R; Earl D; Emrich S; Fedotov P; Fonseca NA; Ganapathy G; Gibbs RA; Gnerre S; Godzarisidis ♀; Goldstein S; Haimel M; Hall G; Haussler D; Hiatt JB; Ho I; Howard JT; Hunt M; Jackman SD; Jaffe DB; Jarvis ED; Jiang H; Kazakov S; Kersey PJ; Kitzman JO; Knight JR; Koren S; Lam T; Lavenier D; Laviolette F; Li Y; Li Z; Liu B; Liu Y; Luo R; MacCallum I; MacManes MD; Maillet N; Melnikov S; Naquin D; Ning Z; Otto TD; Paten B; Paulo OS; Phillippy AM; Pina-Martins F; Place M; Przybylski D; Qin X; Qu C; Ribeiro FJ; Richards S; Rokhsar DS; Ruby JG; Scalabrin S; Schatz MC; Schwartz DC; Sergushichev A; Sharpe T; Shaw TI; Shendure J; Shi Y; Simpson JT; Song H; Tsarev F; Vezzi F; Vicedomini R; Vieira BM; Wang J; Worley KC; Yin S; Yiu S; Yuan J; Zhang G; Zhang H; Zhou S; Korf IF (2013): Assemblathon 2 assemblies. GigaScience Database. <http://dx.doi.org/10.5524/100060>". Below this is a DOI button labeled "DOI 10.5524/100060". To the right is a "Table Settings" button. At the bottom is a table with three rows of sample data.

Sample ID	Taxonomic ID	Common Name	Genbank Name	Scientific Name	Sample Attributes
ERS218597	499168	Boa constrictor constrictor		Boa constrictor constrictor	
ERS222880	13146	Melopsittacus undulatus	budgerigar	Melopsittacus undulatus	Cell type:blood Sex:male [PATO:0000384] Common name:budgerigar
SRS140425	106582	Maylandia zebra	zebra mbuna	Maylandia zebra	Sex:male [PATO:0000384] Tissue:muscle and heart Common name:zebra mbuna fish ... +

Displaying 1-3 of 3 Sample(s).

# Questions?



Kirby in 2000, wondering where his GenBank CDROMs are