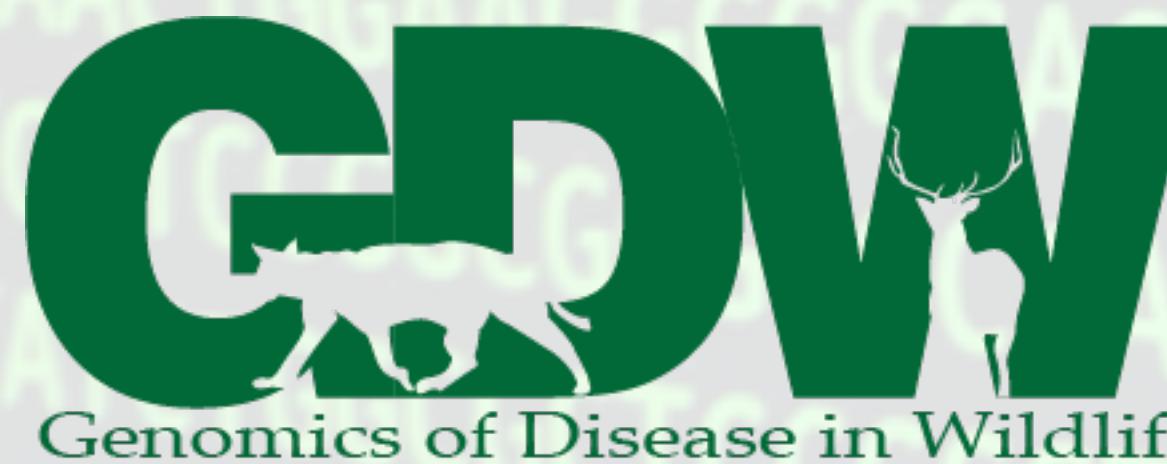


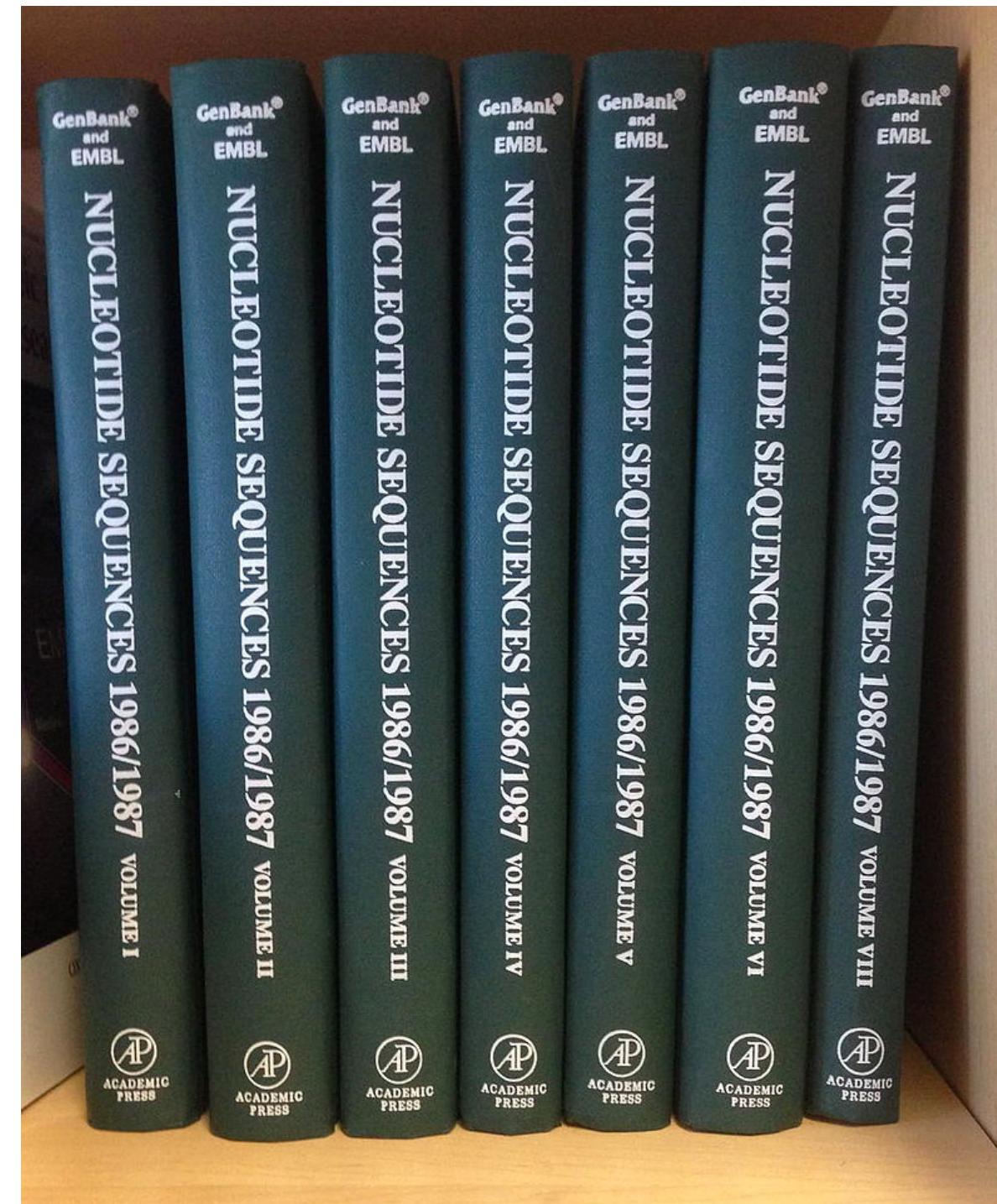
# An overview of genomics databases and online resources: what they are and how to access them

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



>237,000,000 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	cgcgttggc	cagatccaa	atggagggttc	tctgtttgtc	acttgtgtc	ttcgctctct	cccaggggcg	tgagatccac	aggatcccttc	tgttacaagg
101	caatgttttg	aggaaaggcg	tgaaggagca	tgggtttgtc	gaggatcttc	tgcggaaa	aaatggatgg	tttttttttt	tttttttttt	tttttttttt
201	gtggccageg	tgccttgc	caacttactt	gtatgtcgt	acttttggaa	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
301	gtcttttgg	ttttttgg	ccctttatct	actgtttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
401	ggggcaagcc	ctgtttttatc	actatggggat	aggccggat	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
501	gttggccgtt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
601	gttggccgtt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
701	gttggccgtt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
801	gttggccgtt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
901	tttggggccat	acagaacatcg	tacgtatgt	tttgcacatcg	tttgcacatcg	tttgcacatcg	tttgcacatcg	tttgcacatcg	tttgcacatcg	tttgcacatcg
1001	actggcccc	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
1101	atccggatgt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
1201	accccttgc	tttttttttt	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

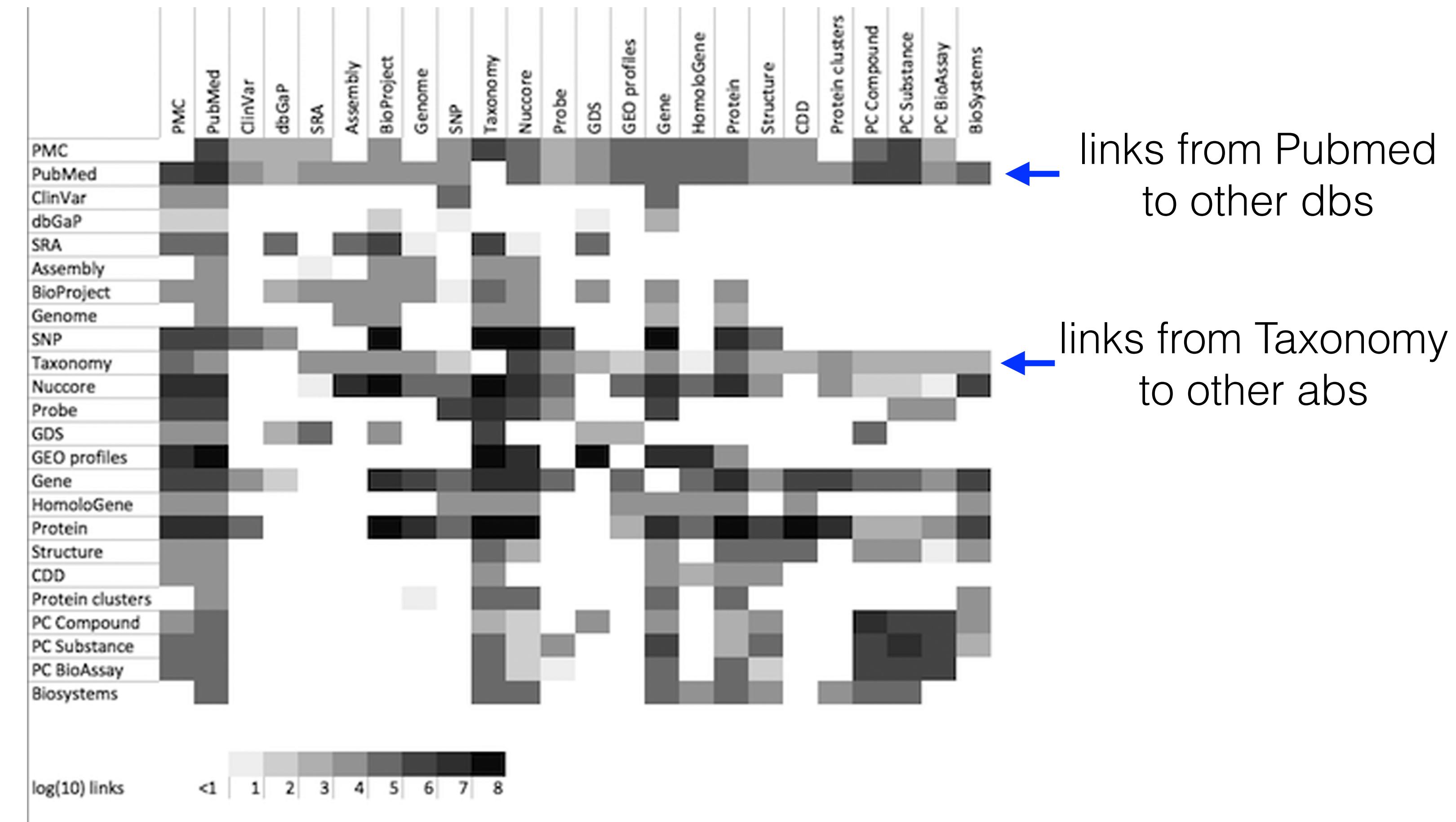
## Categories of NCBI databases

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/ 110%

**PubMed.gov** Search User Guide Advanced Clipboard (3)

Save Email Send to Display options

Case Reports > **Nature**. 2020 Mar;579(7798):265-269. doi: 10.1038/s41586-020-2008-3. FULL TEXT LINKS

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

PMID: 32015508 PMCID: PMC7094943 DOI: 10.1038/s41586-020-2008-3

Free PMC article

ACTIONS

Cite

Favorites

SHARE

# Exercise: navigate to the pubmed entry for the original SARS-CoV-2 paper

## PMID: 32015508

Screenshot of a web browser displaying the PubMed entry for the original SARS-CoV-2 paper (PMID: 32015508). The page includes the PubMed header, search bar, and various navigation links. The main content features the article title, authors, and publication details, along with links to full text, citation, and sharing options.

https://pubmed.ncbi.nlm.nih.gov/32015508/

**PubMed.gov** Search User Guide

Advanced Clipboard (3)

Save Email Send to Display options

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

Epub 2020 Feb 3.

**A new coronavirus associated with human respiratory disease in China**

**ACTIONS**

Cite Favorites

**SHARE**

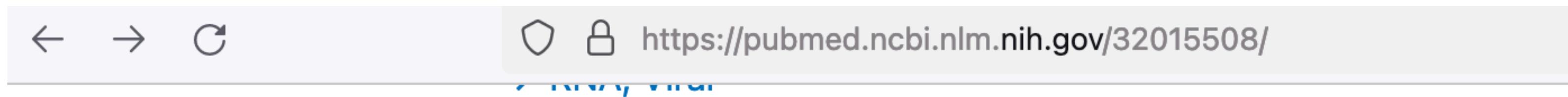
Fan Wu <sup># 1</sup>, Su Zhao <sup># 2</sup>, Bin Yu <sup># 3</sup>, Yan-Mei Chen <sup># 1</sup>, Wen Wang <sup># 4</sup>, Zhi-Gang Song <sup># 1</sup>,  
Yi Hu <sup># 2</sup>, Zhao-Wu Tao <sup>2</sup>, Jun-Hua Tian <sup>3</sup>, Yuan-Yuan Pei <sup>1</sup>, Ming-Li Yuan <sup>2</sup>, Yu-Ling Zhang <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>6 7 8</sup>

Affiliations + expand

PMID: 32015508 PMCID: PMC7094943 DOI: 10.1038/s41586-020-2008-3

Free PMC article

At the bottom of the pubmed page: related information - links other NCBI dbs



## 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)

Nucleotide      Nucleotide      Advanced

Species      Summary ▾      Sort by Default order ▾      Send to: ▾

Viruses (2)      Customize ...

Molecule types      Items: 2

genomic DNA/RNA (2)       [Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome](#)

Customize ...      1. 29,903 bp linear RNA

Accession: NC\_045512.2 GI: 1798174254      [Assembly](#) [BioProject](#) [Protein](#) [PubMed](#) [Taxonomy](#)

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

Source databases       [Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome](#)

INSDC (GenBank) (1)      2. 29,903 bp linear RNA

RefSeq (1)      Accession: MN908947.3 GI: 1798172431

Customize ...      [Assembly](#) [Protein](#) [PubMed](#) [Taxonomy](#)

Sequence Type      [GenBank](#) [FASTA](#) [Graphics](#)

Nucleotide (2)

Sequence length      Custom range...

Release date      Custom range...

This sequence is in the RefSeq database

RefSeq contains curated nucleotide and protein sequences

This is a subset of sequences in the nucleotide and protein databases

*<https://www.ncbi.nlm.nih.gov/refseq/>*

## **RefSeq: NCBI Reference Sequence Database**

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

RefSeq sequences have their own accessions,  
but are usually identical to a non-RefSeq sequence from which they derive

*[https://www.ncbi.nlm.nih.gov/nuccore/NC\\_045512.2](https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2)*

COMMENT     -----  
              REVIEWS [REFSEQ](#): This record has been curated by NCBI staff. The  
              reference sequence is identical to [MN908947](#).  
              On Jan 17, 2020 this sequence version replaced [NC\\_045512.1](#).  
              Annotation was added using homology to SARSr-CoV NC\_004718.3. ###

# Exercise: download the nucleotide sequence for the SARS-CoV-2 RefSeq sequence

First, download the sequence *without* annotation in FASTA format

Open this sequence in the BBEdit software on your laptops  
It will be in the Downloads folder, with a name like sequence.fasta

# FASTA: the most common format for sequence data

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, ←  
complete genome  
ATTAAAGGTTATACCTTCCCAGGTAAACAAACCAACTTCGATCTCTTAGATCTGTTCTAAA  
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAAC  
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTCGTCCGTG  
TTGCAGCCGATCATCAGCACATCTAGGTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTG  
CCTGGTTCAACGAGAAAACACACGTCCAACTCAGTTGCCTGTTTACAGGTTCGCGACGTGCTCGTAC  
GTGGCTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTAAAGATGGCACTTGTGG  
CTTAGTAGAAGTTGAAAAAGGCCTTGCCTCAACTGAACAGCCCTATGTGTTCATCAAACGTTGGAT  
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAAGTACGGTC  
GTAGTGGTGAGACACTGGTGTCCCTGTCATGTGGCGAAATACCAGTGGCTTACCGCAAGGTTCT  
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTGACTTA
```

- A “header line”.
- Only one line.
- Begins with “>”
- Name and other information about sequence.

# FASTA: the most common format for sequence data

>NC\_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,  
complete genome

```
ATTAAAGGTTATACCTTCCCAGGTAAACAAACCAACTTCGATCTCTTAGATCTGTTCTCTAAA  
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAAC  
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTCGTCCGTG  
TTGCAGCCGATCATCAGCACATCTAGGTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTG  
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTGCCTGTTTACAGGTTCGCGACGTGCTCGTAC  
GTGGCTTGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG  
CTTAGTAGAAGTTGAAAAAGGCCTTGCCTCAACTGAACAGCCCTATGTGTTCATCAAACGTTGGAT  
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAAGTACGGTC  
GTAGTGGTGAGACACTGGTGTCCCTGTCATGTGGCGAAATACCAGTGGCTTACCGCAAGGTTCT  
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTGACTTA
```



- The sequence
- Can be one line.
- Can be many lines.
- Nucleotide or amino acid sequence
- Uppercase or lower case OK

# FASTA files can contain more than one sequence

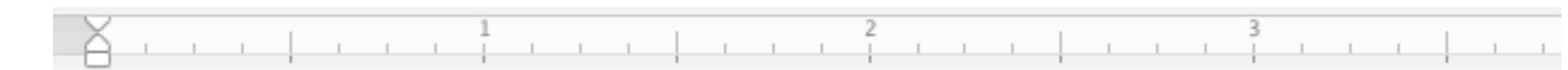
```
>YP_002124308.2 cytochrome c oxidase subunit III
MTHQSHAYHMKPSPWPLTGALSALLMTSGLAMWFHFHSTTLLMLGLLTNTLTMYQWWRDVTRESTYQGH
HTPPVQKGLRYGMVLFITSEVFFFAGFFWAFYHSSLAPTPQLGGHWPPGTITPLNPLEVPLLNTSVLLAS
GVSITWAHHSLMENNRRNQMIQALLITILLGLYFTLLQASEYFESPFTISDGIYGSTFFVATGFHGLHVII
GSTFLTICFIRQLMFHFTSKHHFGFEAAAWYWHFVDVVWLFLYVSIYWWS

>YP_002124314.2 cytochrome b
MTPMRKINPLMKLINHSFIDLPTPSNISAWwNFGSLLGACLILQITTGLFLAMHYSYPDASTAFSSIAHIT
RDVNYGWIIRYLHANGASMFFICLFLHIGRGLYYGSFLYSKTwNIGIILLLATMATAFMGYVLPWGQMSF
WGATVITNLLSAIPYIGTDLVQWIWGGYSVDSPTLTRFFTFFHFILEPFIIAALAALHLLFLHETGSNNPLG
ITSHSDKITFHPYYTIKDAGLFLLLSLMTLTLSPDLLGDPDNYTLANPLNTPHIKPEWYFLFAYTI
LRSVPNKLGGVLALLSILILAMIPILVSKQQSMMFRPLSQSLYWLLAADLLILT WIGGQPVSYPFIII
GQVASVLYFTTILILMPTISLIENKMLKWA

>YP_002124302.2 NADH dehydrogenase subunit 1
MANLLLLVVPILIAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPATSTITLYI
TAPTLALTIALLWTPLPMPNPLVNLNGLLFILATSSLAVYSILWSGWASNNSNYALIGALRAVAQTISY
EVTLAIILLSTLLMSGFNLSTLITAQEHLWLLLPSWPLAMMWFISTLAETNRTPFDLAEGESELVSGFN
IEYAAGPFALFFMAEYTNIIIMMNTTTTIFLGTTYNALSPELYTTYFVTKTLLTSFLWIRTAYPRFRY
DQLMHLLWKNFLPLTLALLMWYVSMPITISSIPPQT
```

# FASTA files are an example of “plain text format” files.

```
1 This is a plain text file.  
2  
3 There is no information about text formatting  
4 (font, bold, italics, etc.)  
5  
6 You can't do things like embed images in the file.  
7  
8 Most bioinformatics software uses plain text data.  
9  
10 Notepad++ is a plain text editor for Windows  
11  
12 BBEdit is a plain text editor for Mac OS.  
13  
14 This file only uses 350 bytes of storage on the disk.  
15
```



Microsoft word documents are not plain text

You can do **fancy** things with the text.

And embed images



File size: 12 kbytes (no image) / 340 kbytes (w/ image)



# FASTA format derives from a 1985 paper about sequence alignment software

## Rapid and Sensitive Protein Similarity Searches

David J. Lipman and William R. Pearson

Technical advances in molecular biology are providing scientists with the primary sequences of proteins implicated in such critical biological processes as differentiation and transformation. Frequently, the only information available about a potentially interesting protein is its amino acid sequence. This information has become more useful because of

product and platelet-derived growth factor (3). The similarity was so strong that it is likely that the chromosomal *sis* gene codes for a growth factor. This serendipitous finding has stimulated new interest in the role of growth factors in oncogenesis. (iii) The similarity of the T-cell receptor protein to immunoglobulin proteins (4).

Lipman and Pearson (1985) Science

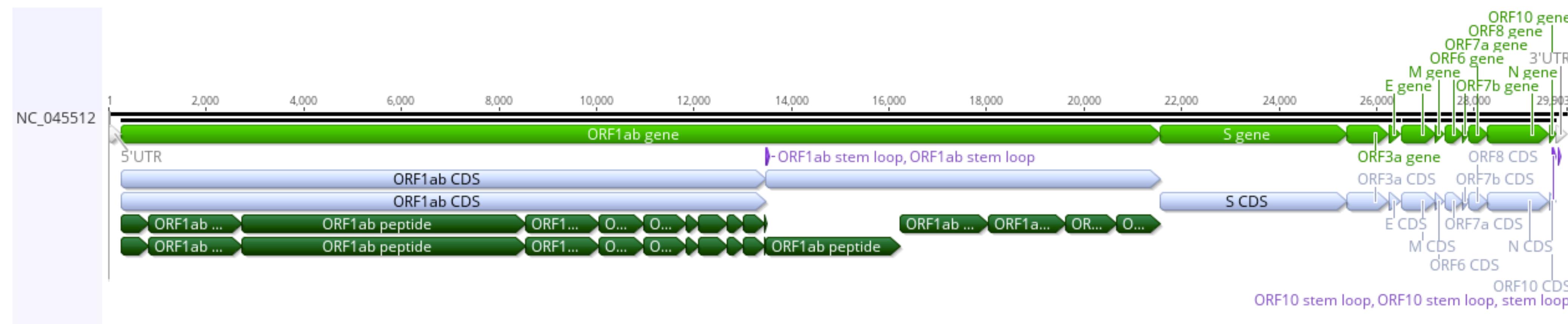
A 1985 Macintosh computer



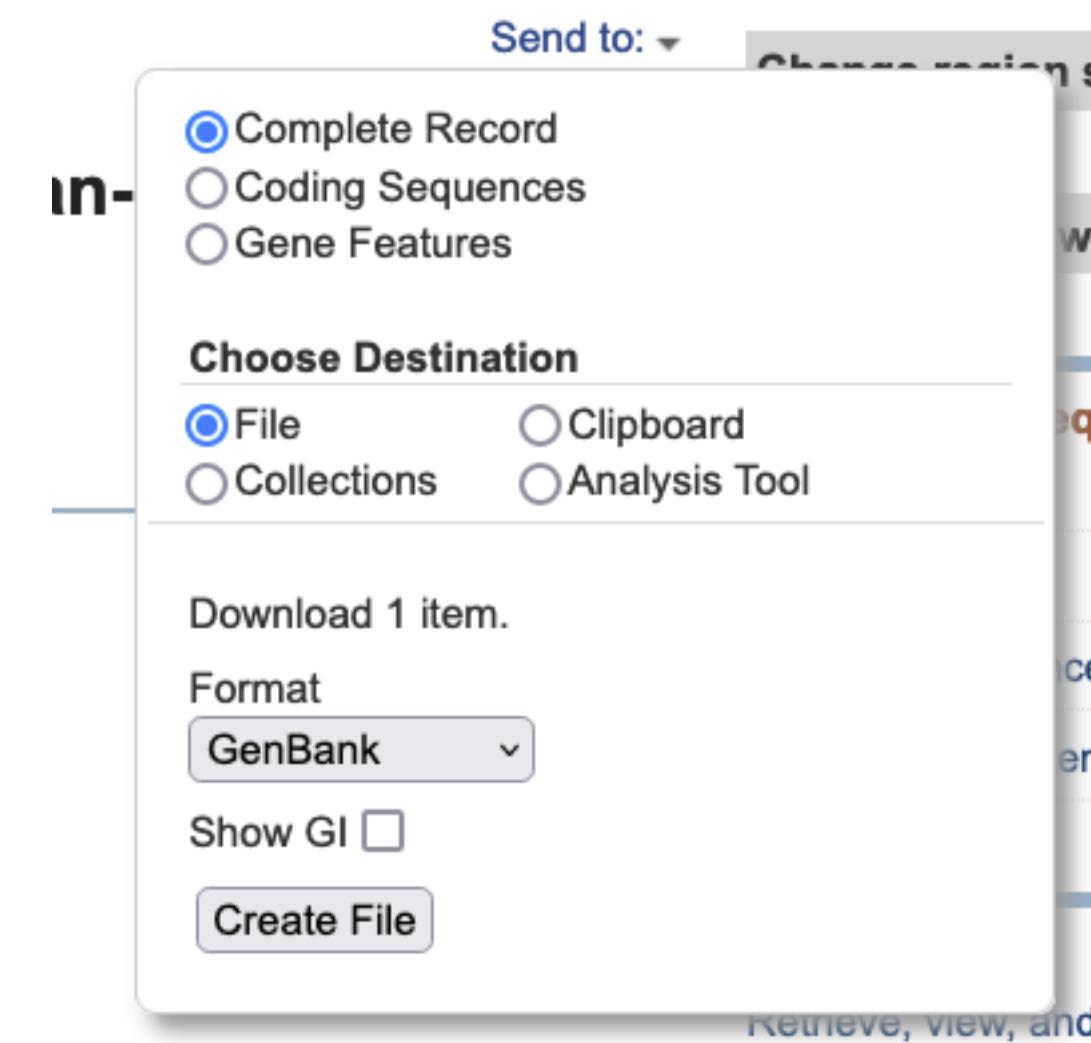
Storage: 400 kbyte floppy disks

Image credit: Sailko CC BY-SA 3.0 [Link](#)

Exercise: download the nucleotide sequence for the SARS-CoV-2 RefSeq sequence in a way that includes the annotation for this sequence



The Send To link in NCBI pages allows you to download one or more sequences in lots of different formats



# Genbank format includes annotation

```
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; Cornidovirineae; 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 579 (7798), 265–269 (2020)
PUBMED   32015508
REMARK    Erratum: [Nature. 2020 Apr;580(7803):E7. PMID: 32296181]
```

Note that Genbank format is still a plain text format!

Exercise: according to the annotation, what is the collection date and host of this first SARS-CoV-2 sequence?

# The nucleotide database links to other databases.

Nucleotide      Nucleotide      Advanced      Search      Help

GenBank ▾      Send to: ▾      Change region shown

## 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; Cornidovirineae; 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 579 (7798), 265–269 (2020)

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

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

# Exercise: download all the protein sequences encoded by this SARS-CoV-2 genome

Nucleotide      Nucleotide  Advanced Help

GenBank ▾ Send to: ▾ Change region shown

## 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; Cornidovirineae; 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 579 (7798), 265–269 (2020)

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

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

# You can download the protein sequences all at once, in various formats

Protein

Species Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾ Filters: [Manage Filters](#)

Viruses (12)  File  Clipboard

Customize ...  Collections  Analysis Tool

Source databases RefSeq (12) [GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

RefSeq (12) [BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

Customize ...

Sequence length Custom range... [Download 12 items.](#)

Molecular weight Custom range... [Format](#)

Release date Custom range... [✓ Summary](#)

Revision date Custom range... [GenPept](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[Clear all](#) [Show additional filters](#)

**Items: 12**

[ORF7b \[Severe acute respiratory syndrome coronavirus 2\]](#)  
1. 43 aa protein  
Accession: YP\_009725318.1 GI: 1820616061  
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[ORF1a polyprotein \[Severe acute respiratory syndrome coronavirus 2\]](#)  
2. 4405 aa protein  
Accession: YP\_009725295.1 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](#)

**Choose Destination**

File  Clipboard

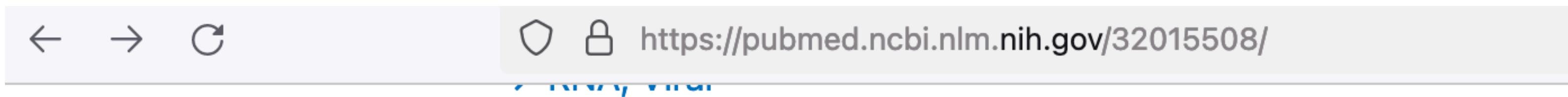
Collections  Analysis Tool

Recent activity

Protein Links for Nucleotide (1798174254) (12)

Severe acute respiratory syndrome coronavirus 2 isolate

The paper also links to the SRA database: a database of NGS datasets



## 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

# The NGS dataset that led to the initial identification of SARS-CoV-2

**SRA**    Help

**Full**

**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](#)

**Related information**

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

**Recent activity**

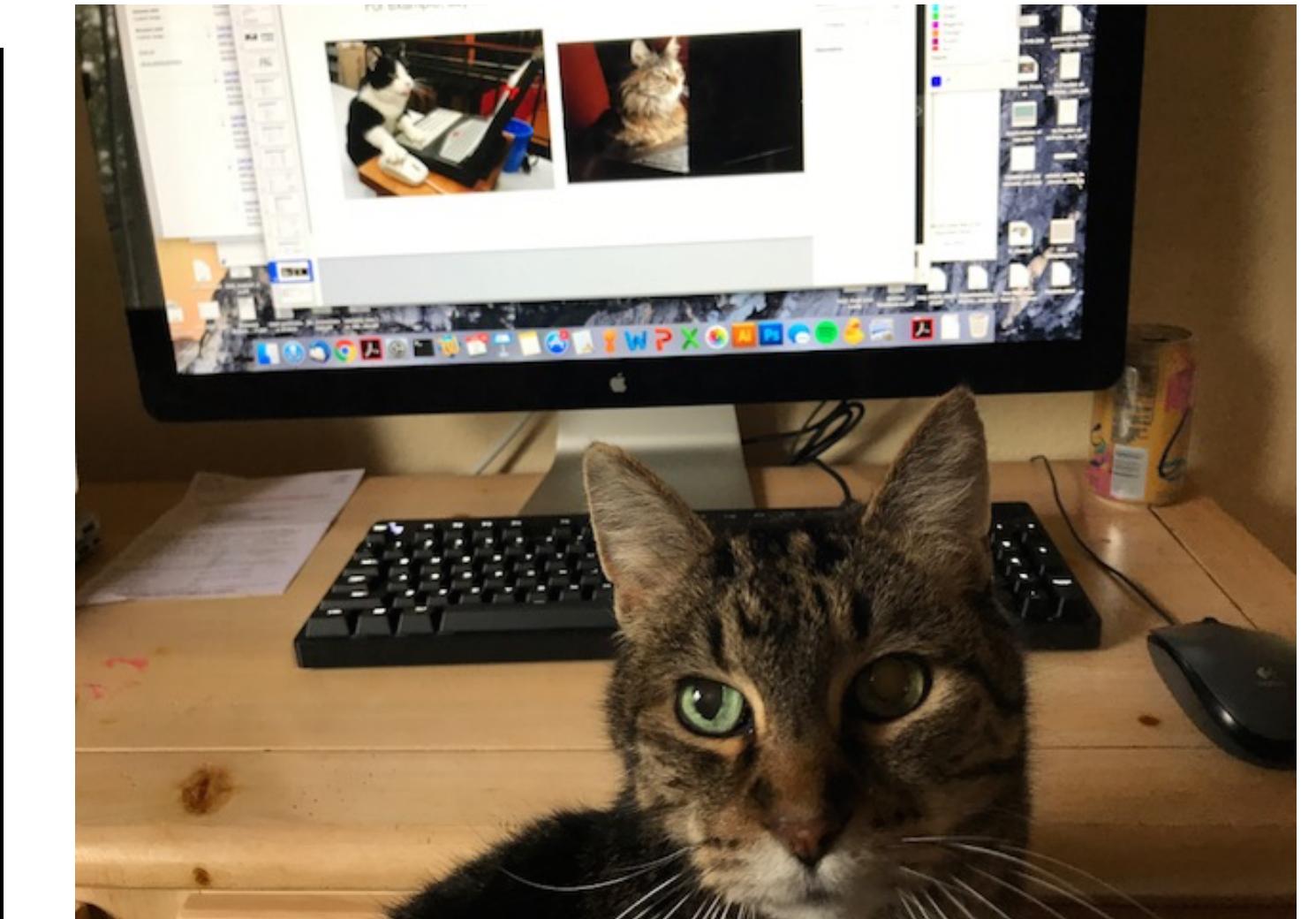
[Turn Off](#) [Clear](#)

[SRA Links for PubMed \(Select 32015508\)](#)  
(1) [SRA](#)

[Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, c Nucleotide](#)

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

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

Genome

Genome txid9685[Organism:noexp] X Search

Create alert Limits Advanced Help

**! Important Update**

By the end of June 2023, Genome record pages will be redirected to new NCBI Datasets [Taxonomy pages](#). [Learn more](#).

**Felis catus (domestic cat)**

**Reference genome:** [Felis catus \(assembly F.catus\\_Fca126\\_mat1.0\)](#)

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 5 genomes for species:**  
Browse the [list](#)  
Download sequence and annotation from [RefSeq](#) or [GenBank](#)

Display Settings: Overview Send to: ID: 78

[Organism Overview](#) ; [Genome Assembly and Annotation report \[5\]](#) ; [Organelle Annotation Report \[1\]](#)



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

Lineage: [Eukaryota\[11633\]](#); [Metazoa\[5514\]](#); [Chordata\[2675\]](#); [Craniata\[2650\]](#); [Vertebrata\[2650\]](#); [Euteleostomi\[2628\]](#); [Mammalia\[720\]](#); [Eutheria\[546\]](#); [Laurasiatheria\[303\]](#); [Carnivora\[84\]](#); [Feliformia\[30\]](#); [Felidae\[22\]](#); [Felinae\[14\]](#); [Felis\[3\]](#); [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...](#)

**NCBI Resources**

[Genome Data Viewer](#)

**Tools**

[BLAST Genome](#)

**Related information**

[Assembly](#)  
[BioProject](#)  
[Gene](#)  
[Components](#)  
[Protein](#)  
[PubMed](#)  
[Taxonomy](#)

# There are actually 5 genome assemblies for *Felis catus*

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

**Important Update**  
By the end of June 2023, Genome record pages will be redirected to new NCBI Datasets [Taxonomy pages](#). [Learn more](#).

**Felis catus (domestic cat)**  
**Reference genome:** [Felis catus \(assembly F.catus\\_Fca126\\_mat1.0\)](#)  
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 5 genomes for species:** 

Browse the [list](#)  
Download sequence and annotation from [RefSeq](#) or [GenBank](#)

Display Settings: Overview      Send to: ID: 78

[Organism Overview](#) ; [Genome Assembly and Annotation report \[5\]](#) ; [Organelle Annotation Report \[1\]](#)

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

Lineage: [Eukaryota\[11633\]](#); [Metazoa\[5514\]](#); [Chordata\[2675\]](#); [Craniata\[2650\]](#); [Vertebrata\[2650\]](#); [Euteleostomi\[2628\]](#); [Mammalia\[720\]](#); [Eutheria\[546\]](#); [Laurasiatheria\[303\]](#); [Carnivora\[84\]](#); [Feliformia\[30\]](#); [Felidae\[22\]](#); [Felinae\[14\]](#); [Felis\[3\]](#); [Felis catus\[1\]](#)

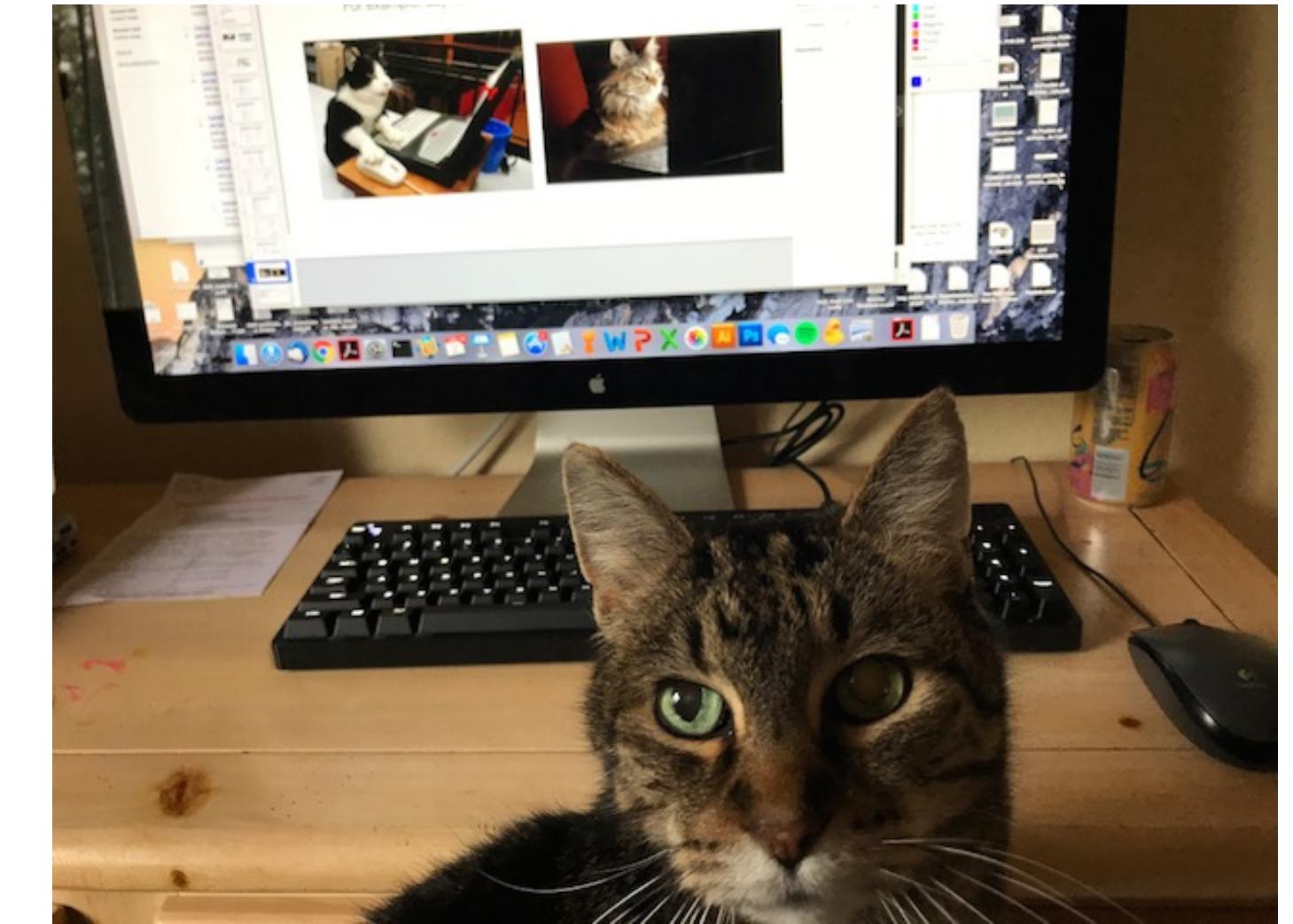
*Felis catus*, the domestic cat, provides several valuable models for infectious disease, including a model for human AIDS. With a large number of recognized breeds, the cat is also a valuable resource for studying phenotypic diversity and evolution. The cat genome will further facilitate research in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

**NCBI Resources**  
[Genome Data Viewer](#)

**Tools**  
[BLAST Genome](#)

**Related information**  
[Assembly](#)  
[BioProject](#)  
[Gene](#)  
[Components](#)  
[Protein](#)  
[PubMed](#)  
[Taxonomy](#)

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**

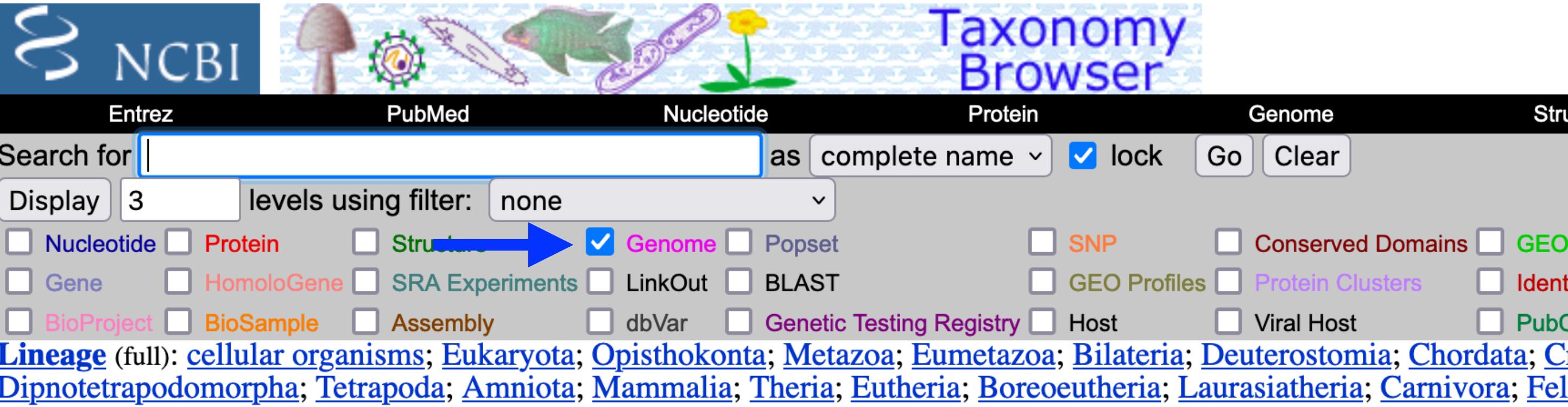
### [Lineage](#)(full )

[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

The screenshot shows the NCBI Taxonomy Browser interface. At the top, there are links for Entrez, PubMed, Nucleotide, Protein, Genome, and Structure. Below the links is a search bar with the placeholder "Search for". To the right of the search bar are buttons for "as complete name" (with a dropdown arrow), "lock" (checked), "Go", and "Clear". Under the search bar, there is a "Display" section with a dropdown menu set to "3 levels using filter: none". A blue arrow points from this dropdown to the "Genome" checkbox, which is checked. Other filter options include "Nucleotide", "Protein", "Gene", "HomoloGene", "SRA Experiments", "LinkOut", "BLAST", "BioProject", "BioSample", "Assembly", "dbVar", "Genetic Testing Registry", "Host", "Viral Host", "SNP", "Conserved Domains", "GEO Profiles", "Protein Clusters", "Identifiers", and "PubChem". Below the filters, the "Lineage (full)" section lists the taxonomic hierarchy: cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Carnivora; Feliformia; Felidae; Felinae; Felinae intergeneric hybrids; Felis; Felis catus; Felis chaus; Felis chaus x Felis catus; Felis margarita; Caracal; Catopuma; Catopuma badia; Catopuma temminckii; Acinonychiae; Acinonyx; Acinonyx jubatus; and Caracal caracal.

- **Felidae** (cat family) 22 Click on organism name to get more information. ← 22 genomes for species in Felidae
  - **Acinonychiae** 1
    - **Acinonyx** 1
      - **Acinonyx jubatus** (cheetah) 1
  - **Felinae** 14
    - **Caracal** 1
      - **Caracal caracal** 1
    - **Catopuma**
      - **Catopuma badia** (bay cat)
      - **Catopuma temminckii** (Asiatic golden cat)
    - **Felinae intergeneric hybrids**
      - **Felis catus x Leopardus geoffroyi**
      - **Felis catus x Prionailurus bengalensis**
      - **Leptailurus serval x Caracal caracal**
    - **Felis** 3
      - **Felis catus** (domestic cat) 1
      - **Felis chaus** (jungle cat) 1
      - **Felis chaus x Felis catus**
      - **Felis margarita** (sand cat)

# 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

Genome      Genome txid9694[Organism:exp] | Search      Create alert   Limits   Advanced      Help

**Important Update**  
By the end of June 2023, Genome record pages will be redirected to new NCBI Datasets [Taxonomy pages](#). [Learn more](#).

**Panthera tigris (tiger)**  
**Representative genome:** [Panthera tigris \(assembly P.tigris\\_Pti1\\_mat1.1\)](#)  
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 6 genomes for species:**  
[Browse the list](#)  
Download sequence and annotation from [RefSeq](#) or [GenBank](#)

Display Settings: Overview      Send to: ID: 10802

[Organism Overview](#) ; [Genome Assembly and Annotation report \[6\]](#) ; [Organelle Annotation Report \[2\]](#)  
 **Panthera tigris (tiger)**  
tiger  
Lineage: [Eukaryota](#)[11633]; [Metazoa](#)[5514]; [Chordata](#)[2675]; [Craniata](#)[2650]; [Vertebrata](#)[2650]; [Euteleostomi](#)[2628]; [Mammalia](#)[720]; [Eutheria](#)[546]; [Laurasiatheria](#)[303]; [Carnivora](#)[84]; [Feliformia](#)[30]; [Felidae](#)[22]; [Pantherinae](#)[7]; [Panthera](#)[5]; [Panthera tigris](#)[1]

**NCBI Resources**  
[Genome Data Viewer](#)

**Tools**  
[BLAST Genome](#)

**Related information**  
[Assembly](#)  
[BioProject](#)  
[Gene](#)  
[Components](#)

Links to genome, transcriptome, proteome, annotation

These links work from the command line using tools like curl or wget

# You can download data from the command line

This is often useful when you're working on a server.

The screenshot shows the NCBI genome browser interface for the Felis catus genome. At the top, there is a search bar with the query "felis catus[orgn]". Below the search bar, there are links for "Create alert", "Limits", and "Advanced". A large blue arrow points from the text "FTP links" to the "Reference genome" section, which contains links for FASTA, GFF, GenBank, and tabular formats. There is also a link for BLAST against the Felis catus genome. Below this, there is a section for "All 2 genomes for species" with a link to "Browse the list" and "Download sequence and annotation from RefSeq or GenBank". At the bottom, there is an "Organism Overview" section with a thumbnail image of a cat, the species name "Felis catus (domestic cat)", its common name "domestic cat", and its lineage information: Eukaryota[2334]; Metazoa[779]; Chordata[332]; Craniata[324]; Vertebrata[324]; Euteleostomi[319]; Mammalia[136]; Eutheria[131]; Laurasiatheria[61]; Carnivora[13]; Feliformia[4]; Felidae[4]; Felinae[1]; Felis[1]; Felis catus[1]. A brief description follows, mentioning the cat's value as a model for infectious diseases and its role in research.

The screenshot shows a Mac OS X Terminal window with the command "curl -O ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/181/335/GCF\_000181335.2\_Felis\_catus\_8.0/GCF\_000181335.2\_Felis\_catus\_8.0\_genomic.fna.gz" being run. The output shows the progress of the file download, including the total size (786M), received bytes (5834k), transferred bytes (226k), and download speed (309k).

```
mdstengl@Marks-MBP:~$ curl -O ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/181/335/GCF_000181335.2_Felis_catus_8.0/GCF_000181335.2_Felis_catus_8.0_genomic.fna.gz
% Total    % Received % Xferd  Average Speed   Time     Time     Current
          Dload  Upload Total   Spent   Left  Speed
0  786M      0  5834k    0      0  226k  0:59:09  0:00:25  0:58:44  309k
```

curl is a file transfer utility built into Linux, MacOS

similar utilities exist for Windows

# You can download sequences from Geneious too, which can be convenient

The screenshot shows the Geneious Prime software interface. The left sidebar displays a navigation tree with categories like Local, Shared Databases, NCBI, and Nucleotide (1). The main search bar at the top contains the identifier "NC\_045512". Below the search bar, a table lists the results of the search, showing one entry: "NC\_045512" with the description "Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome". The main workspace is a "Sequence View" showing the genome sequence NC\_045512. The sequence is represented by a horizontal line with various gene annotations. Annotations include the 5' UTR, ORF1ab gene, S gene, M gene, E gene, N gene, and 3' UTR. Within the ORF1ab gene, there are multiple ORF1ab peptide and ORF1a annotations. The "Live Annotate & Predict" panel on the right allows users to add annotations from MDS oligos with a similarity of 100% and find ORFs with a minimum size of 250. The bottom status bar indicates "919 / 14,380 MB Memory" and provides instructions for zooming.

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

Registered Users   EpiFlu™   **EpiCoV™**   EpiRSV™   EpiPox™   My Profile

**EpiCoV™** | **Search** | **Downloads** | **Upload**

**Search ▼** **Reset filters**

EPI\_ISL ID  Virus name  EPI\_SET ID   Complete ?  
Location  Host   High coverage ?  
Collection  to  Submission  to   Low coverage excluded ?  
Clade  Lineage  Variant   With patient status ?  
AA Substitutions ? Nucl Mutations ?   Collection date complete ?  
 Under investigation

*Text Search*

<input type="checkbox"/>	Virus name	Passage de	Accession ID	Collection da	Submission D		Length	Host	Location	Originating
<input type="checkbox"/>	hCoV-19/Spain/MD-HULP-014591595/2023	Original	EPI_ISL_17761364	2023-05-24	2023-06-03		29,773	Human	Europe / Spain /	Servicio Mi
<input type="checkbox"/>	hCoV-19/Spain/MD-HULP-062138879/2023	Original	EPI_ISL_17761363	2023-05-22	2023-06-03		29,470	Human	Europe / Spain /	Servicio Mi
<input type="checkbox"/>	hCoV-19/Spain/MD-HULP-062147224/2023	Original	EPI_ISL_17761362	2023-05-23	2023-06-03		29,749	Human	Europe / Spain /	Servicio Mi
<input type="checkbox"/>	hCoV-19/Spain/MD-HULP-062153481/2023	Original	EPI_ISL_17761361	2023-05-22	2023-06-03		29,774	Human	Europe / Spain /	Servicio Mi
<input type="checkbox"/>	hCoV-19/Spain/MD-HULP-062157816/2023	Original	EPI_ISL_17761360	2023-05-22	2023-06-03		29,774	Human	Europe / Spain /	Servicio Mi

Total: 15,646,070 viruses

<< < 1 2 3 4 5 > >>

**EPI\_SET** **Select** **Analysis** **Download**

15 million SARS-CoV-2 sequences!! (~7M in Genbank)

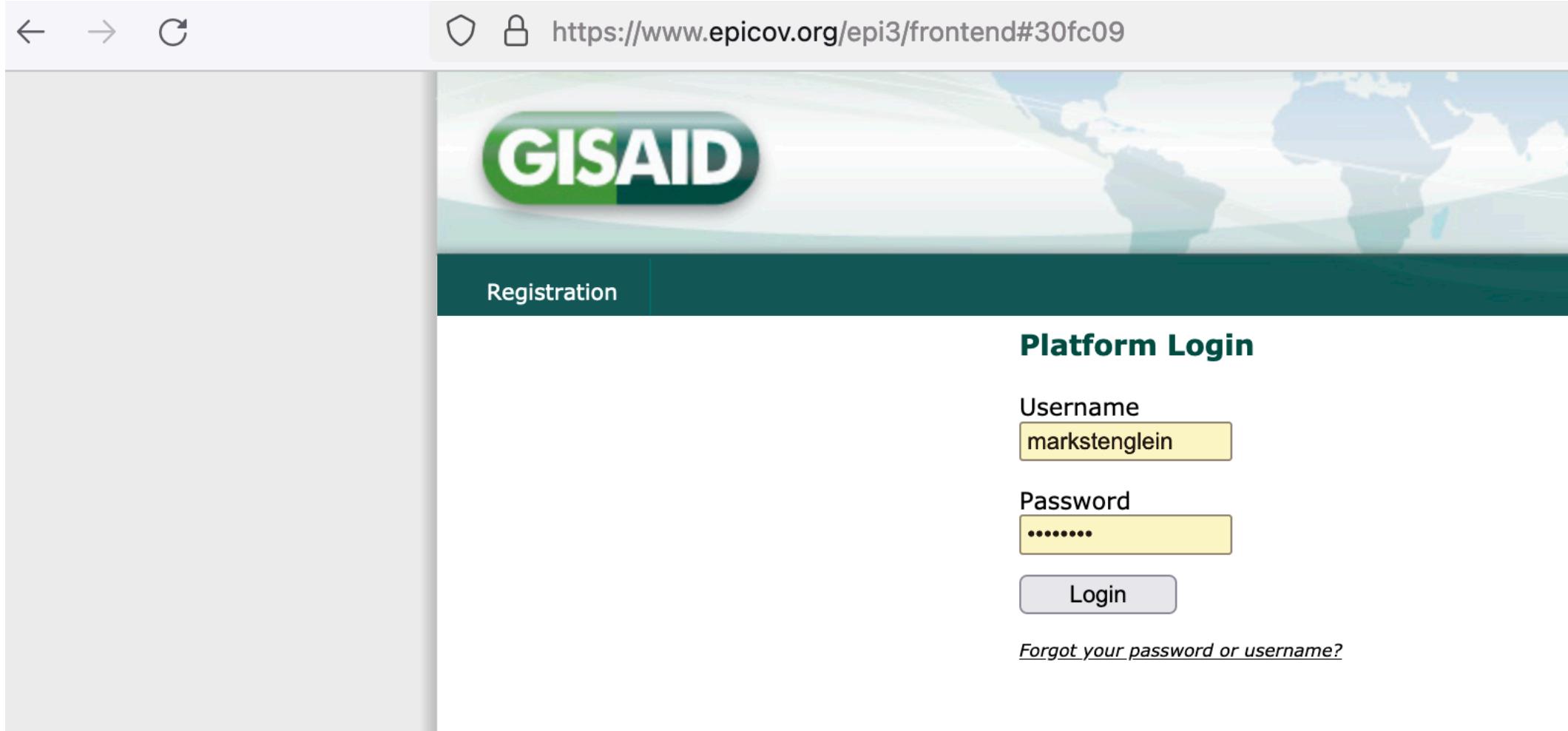
# GISAID has some nice features, but is limited to a few pathogens

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 'Basic filters' section includes a Predefined search dropdown (Select ...), a 'Search in' radio button group (Released files selected, Worksets), and a 'Search patterns' input field. The main search area displays a grid of 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 (-all-, Bahrain, Bangladesh, Bhutan, British Indian Ocean Territory, Brunei, Cambodia, China, Christmas Island, Georgia, Hong Kong (SAR)). 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	Europe
	7	7		Duck	North America
	8	8		Eagle	Oceania
	9	9		Falcon	South America
	10	10		Goose	

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". Below the title, it says "Assemblathon 2 assemblies." and "Dataset type: Genomic". It also indicates "Data released on June 24, 2013". A large block of text lists numerous authors' names in green: 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>

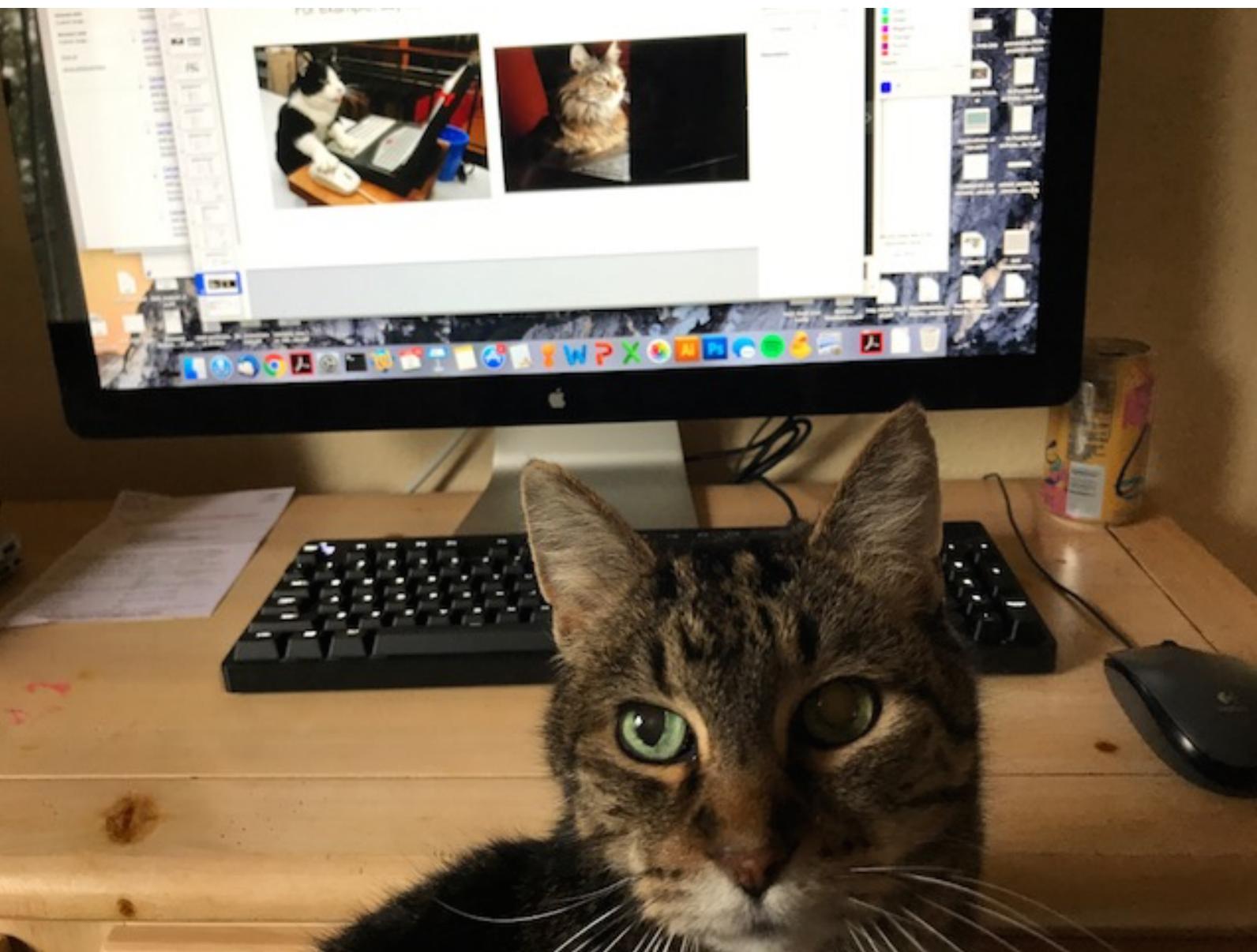
DOI 10.5524/100060

Table Settings

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