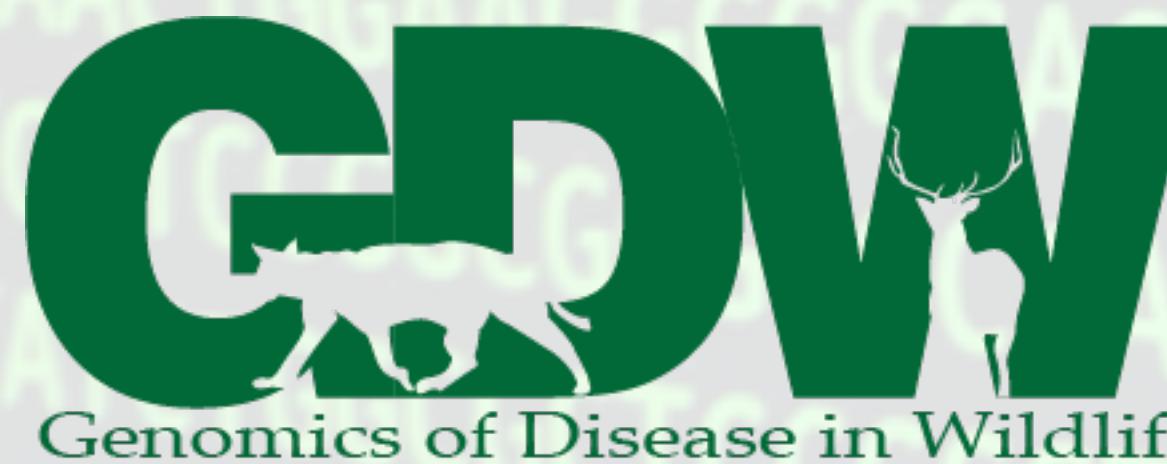


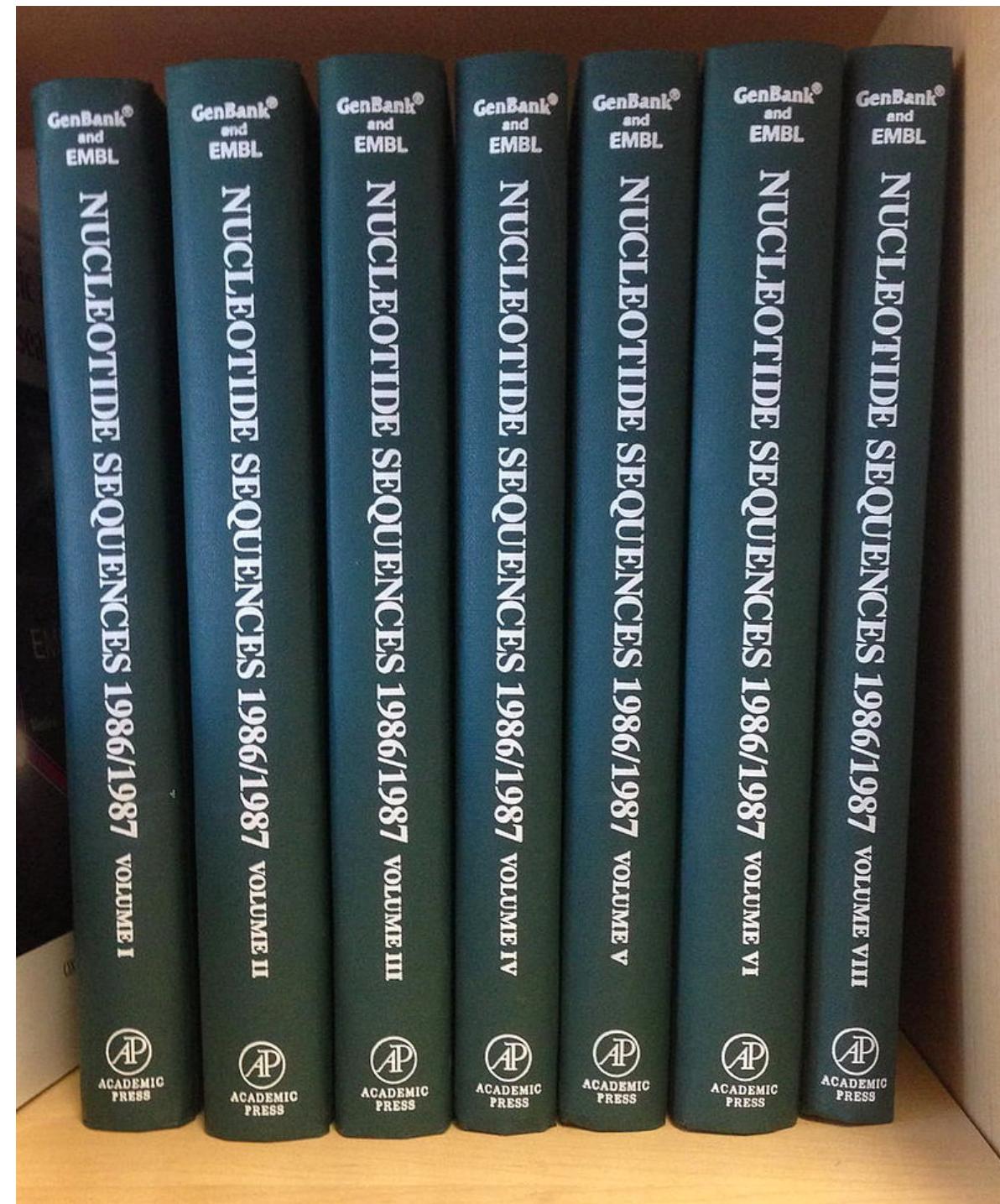
An overview of genomics databases and online resources: what they are and how to access them

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	cggctgtacc	cagatccaa	atggagggttc	tctgtttgtc	acttgcgttc	ttcgctctct	cccaggggcg	tgagatccac	aggatcccttc	tgtacaaagg
101	caatgttttg	aggaaaggcg	tgaaggagca	tgggtttgtc	gaggatcttc	tgcggaaaca	ggatgtatgg	tttagcaca	atgatccgg	tttcggggag
201	gtggccageg	tgccttgc	caactacttg	gtatgtcgt	acttttggaa	gatttaccc	ggggcccccp	ccccgggggtt	caccgttgtc	tttgacactg
301	gtctcttgg	tttcgttgc	cccttatact	actgtcaagag	aaatggctgc	aaaaaaccc	ggcggttccp	ccccggggaa	tgcgttccact	tccagaaacct
401	ggggcaagccc	ctgtttatcc	actacggggc	aggccggatg	caaggccatcc	tgggtatgaa	caccgttact	gttcccaaaa	tttttgtcacat	ccagcagaca
501	gttggccgttgc	tgaccatccatg	ggccggggac	gttccgttccat	atggatggaa	cpacggggatc	ctggggatgg	ccatccccct	gttccgttca	gaggatctgg
601	tacccatgttt	tgaccaatcatg	atggatggc	atccgttggc	ccaaaggactt	ttctccgtt	atccatccat	gttccgttca	atccatccat	ttacgttgg
701	ggcccatccac	ccgttgtact	atacaggatc	cttgcacttg	gttccgttgc	ttctccgtt	atccatccat	gttccgttca	atccatccat	catcggccgt
801	gtggtttgg	ctctgtgggg	tggctgttc	gcatacttgc	atacggggac	ttccaaatgt	gttccgttca	atccatccat	gttccgttca	ttccaaatgt
901	ttggggccat	acagaatccatg	tacgtatgt	tttgatctgt	ctggccaaac	ttctccgtt	atccatccat	gttccgttca	atccatccat	ttccaaatgt
1001	actggccccc	tccgttgcata	ccaggccagg	ttgtgggttc	ttgtgggttc	ttctccgtt	atccatccat	gttccgttca	atccatccat	ttccaaatgt
1101	atccggatgt	attacatgc	ttttgtacgt	ttttgtacgt	ttgtgggttc	ttctccgtt	atccatccat	gttccgttca	atccatccat	ttccaaatgt
1201	acacatcgat	acacatcgatc	acacatcgatc	atggccatcg	ttgtgggttc	ttctccgtt	atccatccat	gttccgttca	atccatccat	ttccaaatgt

~1,300,000 sequences

First release: 1982: 606 sequences

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

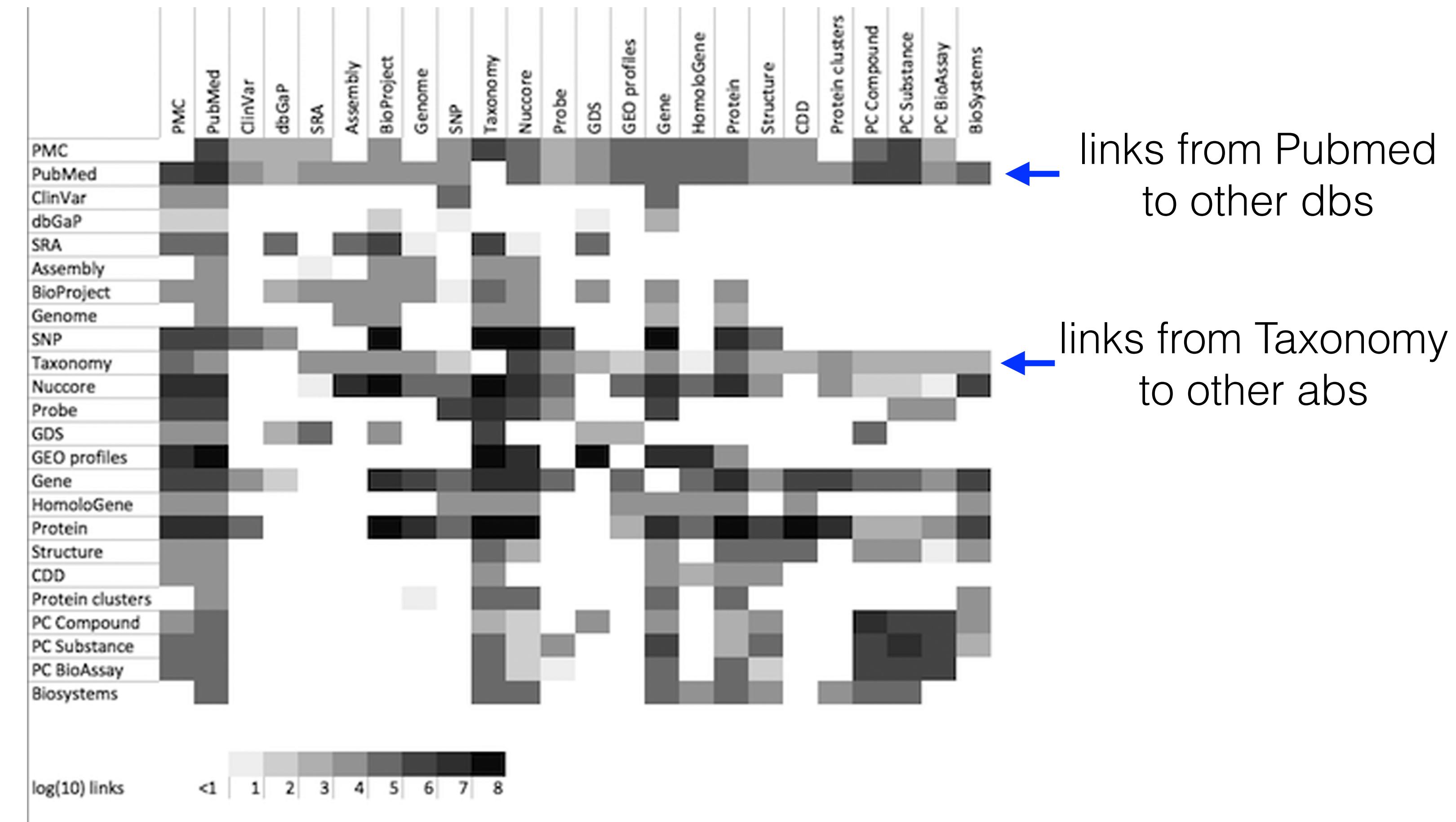
Categories of NCBI databases

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^{1,7}, Su Zhao^{2,7}, Bin Yu^{3,7}, Yan-Mei Chen^{1,7}, Wen Wang^{4,7}, Zhi-Gang Song^{1,7}, Yi Hu^{2,7}, Zhao-Wu Tao², Jun-Hua Tian³, Yuan-Yuan Pei¹, Ming-Li Yuan², Yu-Ling Zhang¹, Fa-Hui Dai¹, Yi Liu¹, Qi-Min Wang¹, Jiao-Jiao Zheng¹, Lin Xu¹, Edward C. Holmes^{1,5} & Yong-Zhen Zhang^{1,4,6}✉

Emerging infectious diseases, such as severe acute respiratory syndrome (SARS) and Zika virus disease, present a major threat to public health^{1–3}. 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 sequencing⁴ 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. npg nature publishing group

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

ACTIONS Cite Favorites

Affiliations + expand SHARE

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

Free PMC article

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

PMID: 32015508

The screenshot shows a web browser displaying the PubMed.gov website. The URL in the address bar is <https://pubmed.ncbi.nlm.nih.gov/32015508/>. The page header includes the PubMed logo, a search bar, and links for Advanced, Clipboard (3), User Guide, Save, Email, Send to, and Display options. The main content area displays the following information:

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

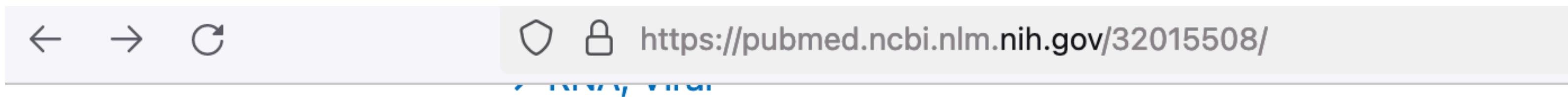
Fan Wu # 1, Su Zhao # 2, Bin Yu # 3, Yan-Mei Chen # 1, Wen Wang # 4, Zhi-Gang Song # 1, Yi Hu # 2, Zhao-Wu Tao 2, Jun-Hua Tian 3, Yuan-Yuan Pei 1, Ming-Li Yuan 2, Yu-Ling Zhang 1, Fa-Hui Dai 1, Yi Liu 1, Qi-Min Wang 1, Jiao-Jiao Zheng 1, Lin Xu 1, Edward C Holmes 1 5, Yong-Zhen Zhang 6 7 8

ACTIONS
“ Cite
☆ Favorites

Affiliations + expand
PMID: 32015508 PMCID: [PMC7094943](#) DOI: [10.1038/s41586-020-2008-3](https://doi.org/10.1038/s41586-020-2008-3)
Free PMC article

SHARE

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...

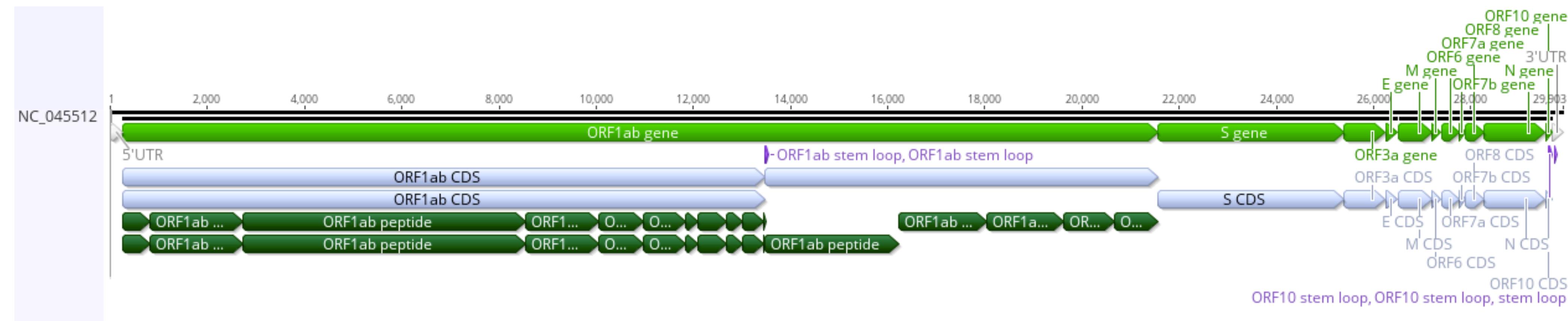
Release date Custom range...

This sequence is in the RefSeq database

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

You've practiced downloading sequences FASTA format, which does not include annotation

Now download this SARS-CoV-2 sequence *with* annotation in Genbank format



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

FASTA and Genbank files are both examples 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 MacOS.  
13  
14 This file only uses 350 bytes of storage on the disk.  
15
```

Most bioinformatics data files are plain text



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)

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?

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

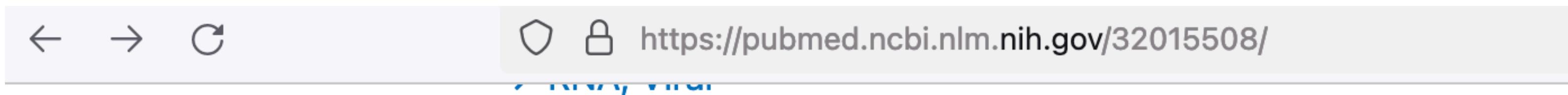
Collections Analysis Tool

[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 SRA Search Help

Full Send to: Related information

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](#)

BioProject
BioSample
PMC
PubMed
Taxonomy

Recent activity Turn Off Clear

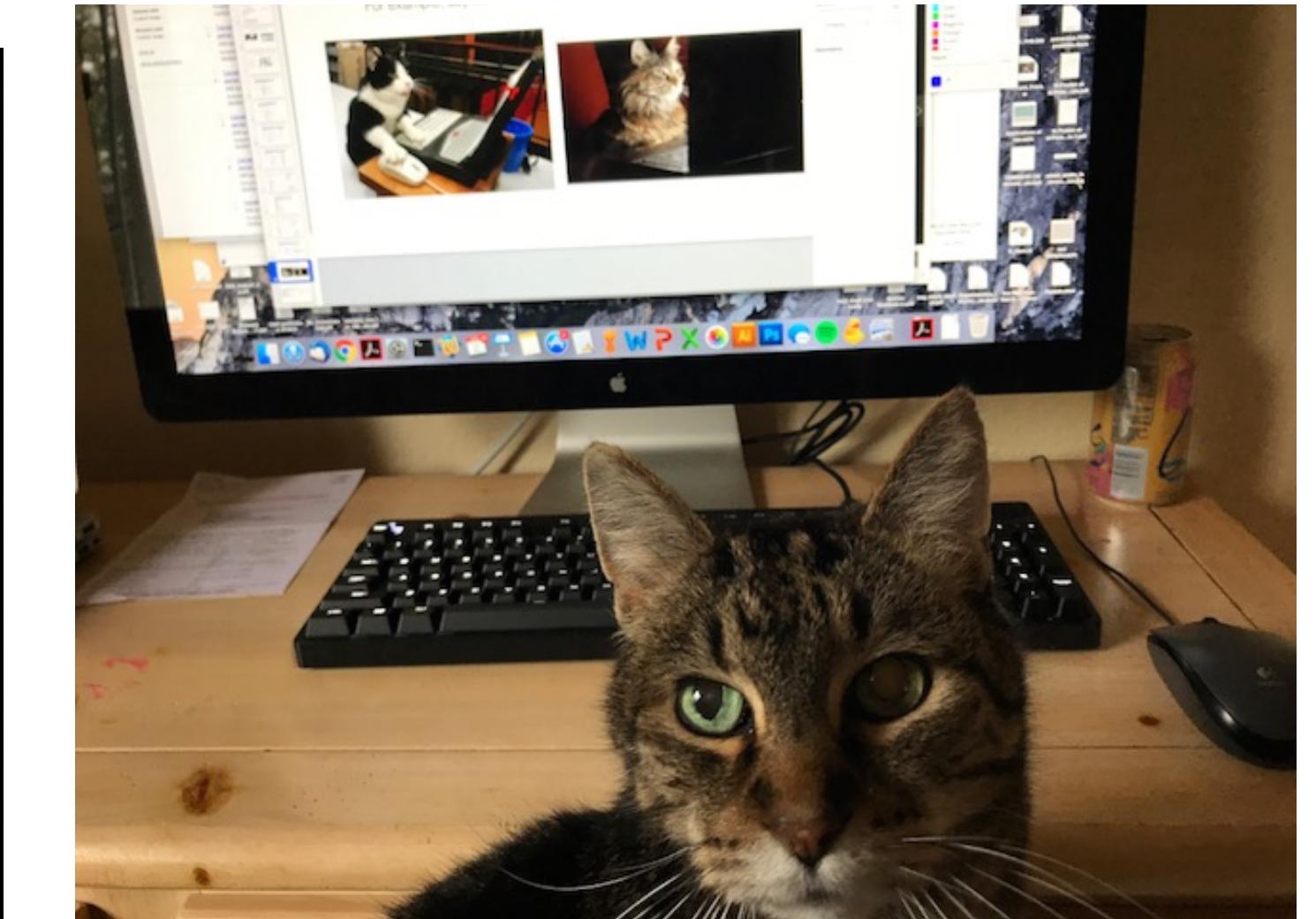
SRA Links for PubMed (Select 32015508) (1) SRA

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

We will download NGS datasets from the SRA database later

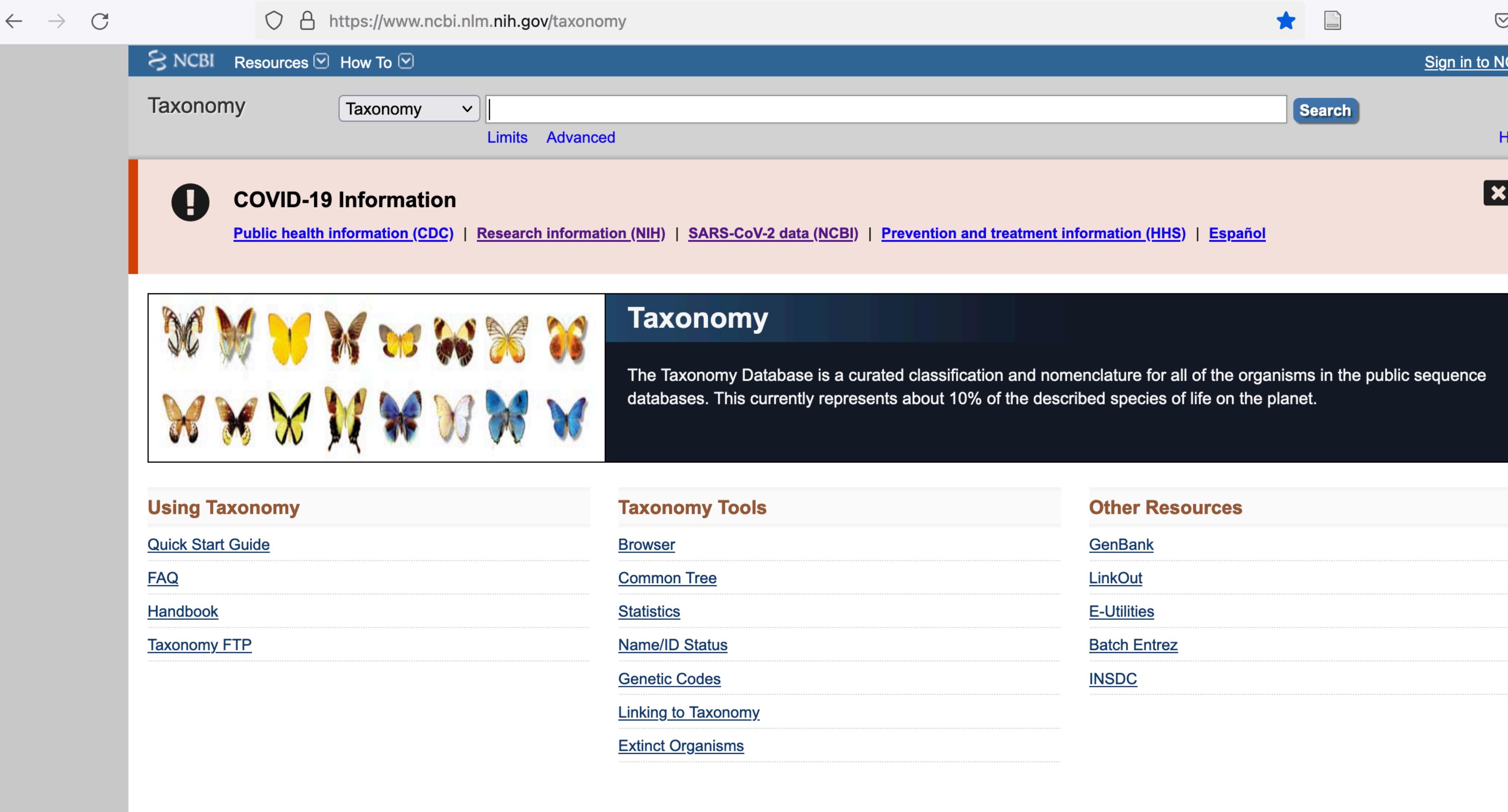
There are often many paths to the same data

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's a navigation bar with links for 'NCBI Resources' and 'How To'. A search bar is present, along with a 'Taxonomy' dropdown menu and 'Search' button. A 'COVID-19 Information' banner is displayed, containing links to CDC, NIH, NCBI, HHS, and Spanish resources. The main content area features a title 'Taxonomy' and a subtext explaining the database as a curated classification and nomenclature for all organisms. Below this is a grid of butterfly images. The page is divided into sections: '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).

I ❤️
NCBI
Taxonomy

A database representation of the tree of life

The NCBI Taxonomy page for the domestic cat

← → C https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9685 ☆ ☰

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy BioCollections

Search for as lock

Display 3 levels using filter: none

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](#)

Entrez records

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

genome ←

SRA datasets ←

Felis catus in the NCBI genome database

Genome Genome 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: Send to:

[Organism Overview](#) ; [Genome Assembly and Annotation report \[5\]](#) ; [Organelle Annotation Report \[1\]](#) ID: 78

 **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](#)

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

NCBI Taxonomy Browser

Search for: as: complete name lock

Display: 3 levels using filter: none

Nucleotide Protein Structure Genome Popset SNP Conserved Domains GEO D...
 Gene HomoloGene SRA Experiments LinkOut BLAST GEO Profiles Protein Clusters Identic...
 BioProject BioSample Assembly dbVar Genetic Testing Registry Host Viral Host PubCh...

Lineage (full): cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Cr...
Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Carnivora; Felidae; Felid...

- **Felidae** (cat family) 22 Click on organism name to get more information. ← 22 genomes for species in Felidae
- **Acinonychinae** 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 Find items

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 executed. The output shows the progress of the file download, including the total size (786M), received bytes (5834k), transferred files (0), average speed (226k), and download/upload times.

% Total	% Received	% Xferd	Average Speed	Time	Time	Time	Current
Dload	Upload	Total	Spent	Left	Speed		
0	786M	0	5834k	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, using a standard genetic code and ATG start codons.

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)

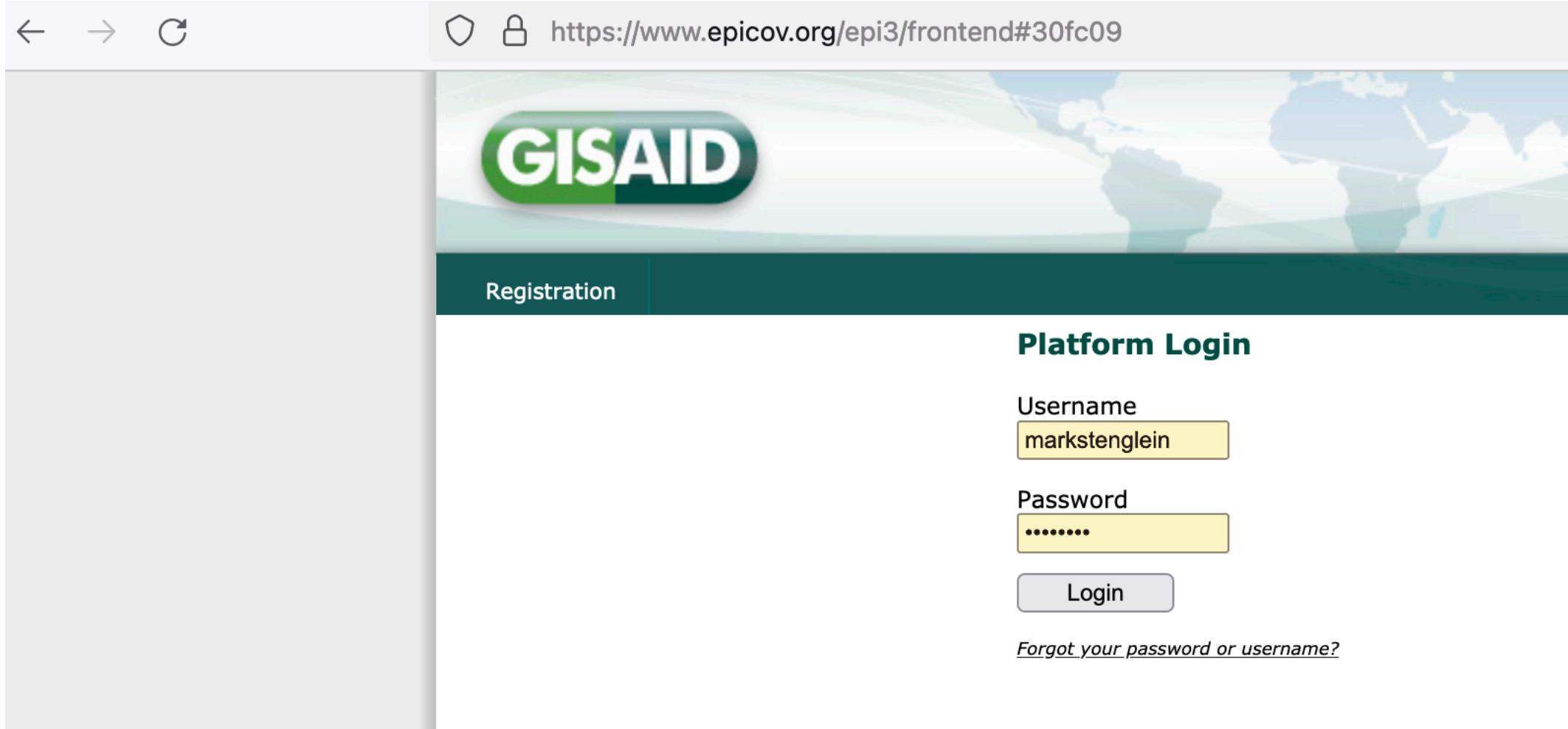
GISAID has some nice features, but is limited to a few pathogens

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 several navigation icons: Search, Back to results, Worksets, Upload, Batch Upload, Settings, and Analysis. Below these are status counts: 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, a radio button for Search in Released files (selected over 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 (empty). The location dropdown is expanded, showing options like Bahrain, Bangladesh, Bhutan, British Indian Ocean Territory, Brunei, Cambodia, China, Christmas Island, Georgia, and Hong Kong (SAR). The "Asia" option is selected.

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

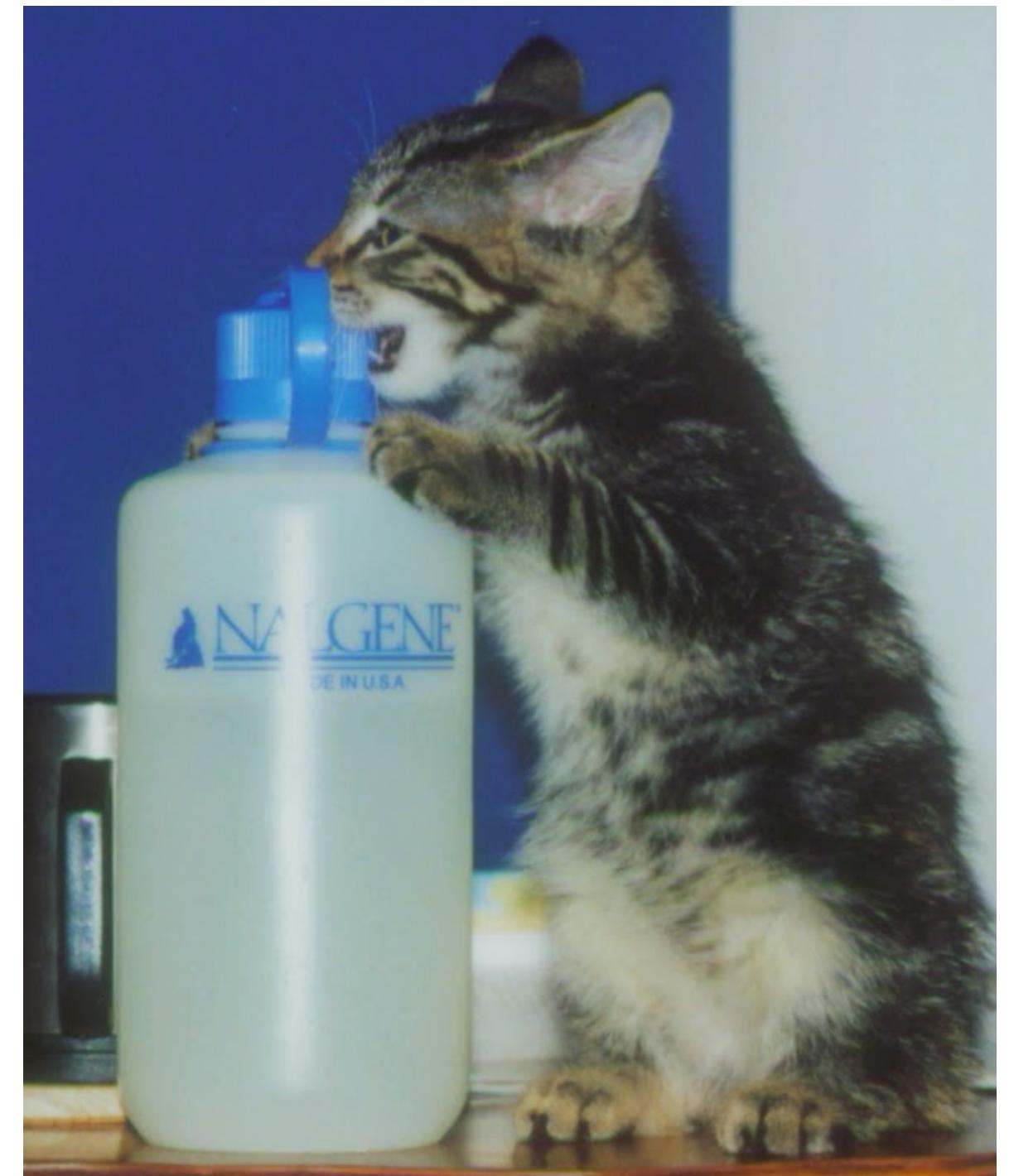
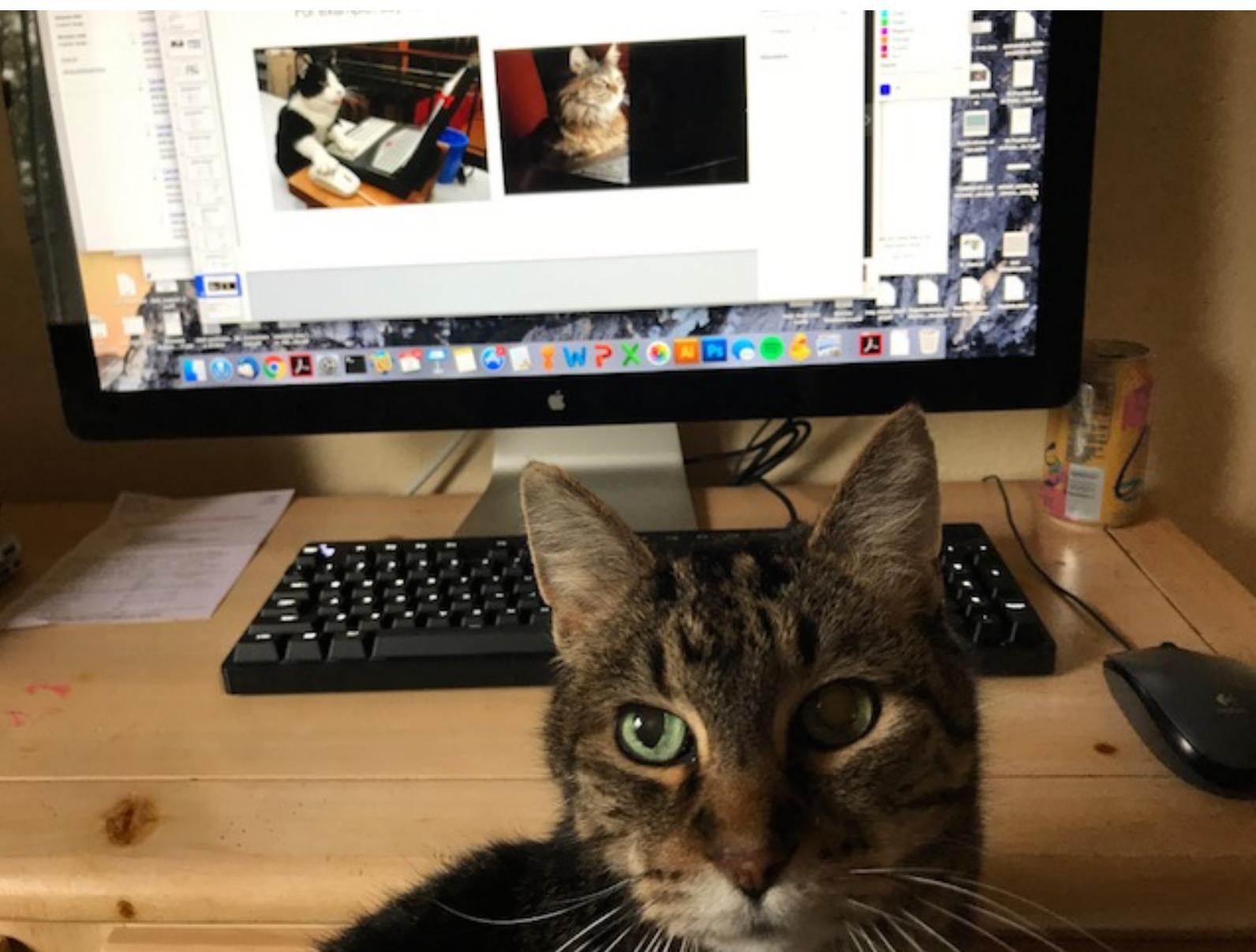
Some data lives in non-standard locations

The screenshot shows a web browser window with the URL gigadb.org/dataset/100060. 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:

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 ... +

At the very bottom, it says "Displaying 1-3 of 3 Sample(s)."

Questions?



Kirby in 2000, wondering where his GenBank CDROMs are