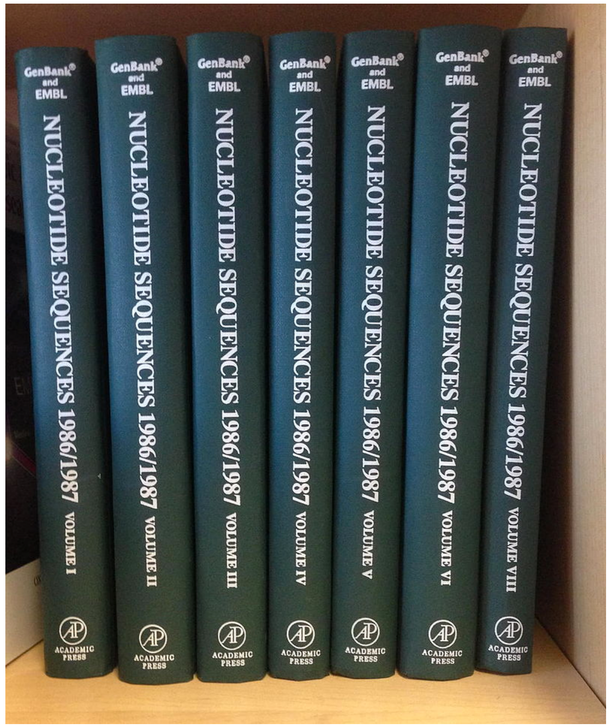
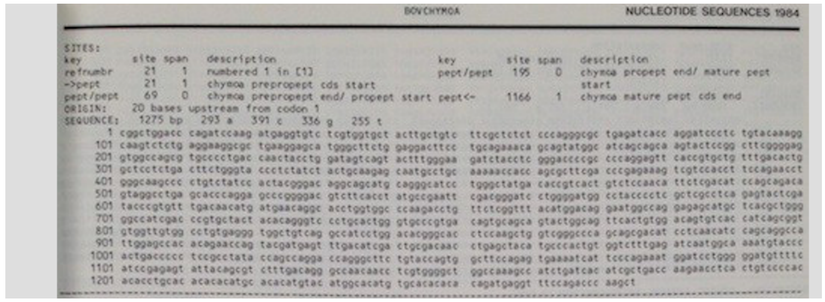
Sequence data formats and databases

GenBank was one of the earliest sequence databases.

GenBank circa 1987



~10,000 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   chymoa propept end/ mature pept
->pept        21    1   chymoa prepropept cds start                                     start
pept/pept     69    0   chymoa prepropept end/ propept start pept<-      1166    1   chymoa mature pept cds end
ORIGIN:    20 bases upstream from codon 1
SEQUENCE:  1275 bp   293 a   391 c   336 g   255 t
       1 cggctggacc cagatccaag atgaggtgtc tcgtggtgct acttgctgtc  ttcgctctct cccagggcgc tgagatcacc aggatccctc tgtacaaagg
     101 caagtctctg aggaaggcgc tgaaggagca tgggcttctg gaggacttcc  tgcagaaaca gcagtatggc atcagcagca agtactccgg cttcgggggag
     201 gtggccagcg tgcccctgac caactacctg gatagtcagt actttgggaa  gatctacctc gggaccccgc cccaggagtt caccgtgctg tttgacactg
     301 gctcctctga cttctgggta ccctctatct actgcaagag caatgcctgc  aaaaaccacc agcgcttcga cccgagaaag tcgtccacct tccagaacct
     401 gggcaagccc ctgtctatcc actacgggac aggcagcatg cagggcatcc  tgggctatga caccgtcact gtctccaaca ttctcgacat ccagcagaca
     501 gtaggcctga gcacccagga gcccgggggc gtcttcacct atgccgaatt  cgacgggatc ctggggatgg cctaccccta gctcgcctca gagtactcga
     601 taccccgtgtt tgacaacatg atgaacaggc acctggtggc ccaagacctg  ttctcggttt acatggacag gaatggccag gagagcatgc tcacgctggg
     701 ggccatcgac ccgtgctact acacagggtc cctgcactgg gtgcccgtga  cagtgcagca gtactggcag ttcactgtgg acagtgtcac catcagcggt
     801 gtggttgtgg cctgtgaggg tggctgtcag gccatcctgg acacgggcac  ctccaagctg gtcgggccca gcagcgacat cctcaacatc cagcaggcca
     901 ttggagccac acagaaccag tacgatgagt ttgacatcga ctgcgacaac  ctgagctaca tgcccactgt ggtctttgag atcaatggca aaatgtaccc
    1001 actgaccccc tccgcctata ccagccagga ccagggcttc tgtaccagtg  gcttccagag tgaaaatcat tcccagaaat ggatcctggg ggatgttttc
    1101 atccgagagt attacagcgt cttttgacag gccaacaacc tcgtgggggt  ggccaaggcc atctgatcac atcgctgacc aagaacctca ctgtccccac
    1201 acacctgcac acacacatgc acacatgtac atggcacatg tgcacacaca  cagatgaggt ttccagaccc aagct
```

GenBank release 100 (1997)
distributed by CDROM

~1,300,000 sequences

Genbank today

>237,000,000 sequences

First release: 1982: 606 sequences

Today, we'll focus on NCBI databases

National Library of Medicine
National Center for Biotechnology Information

Category	Example NCBI db	Content
Literature	PubMed	Papers
Nucleotide	Nucleotide	Nucleotide sequences
Genome	Assembly	Genome assemblies
Taxonomy	Taxonomy	Information about species and higher order taxa
Proteins	Protein	Protein sequences
Raw Data	SRA (Sequence Read Archive)	NGS datasets

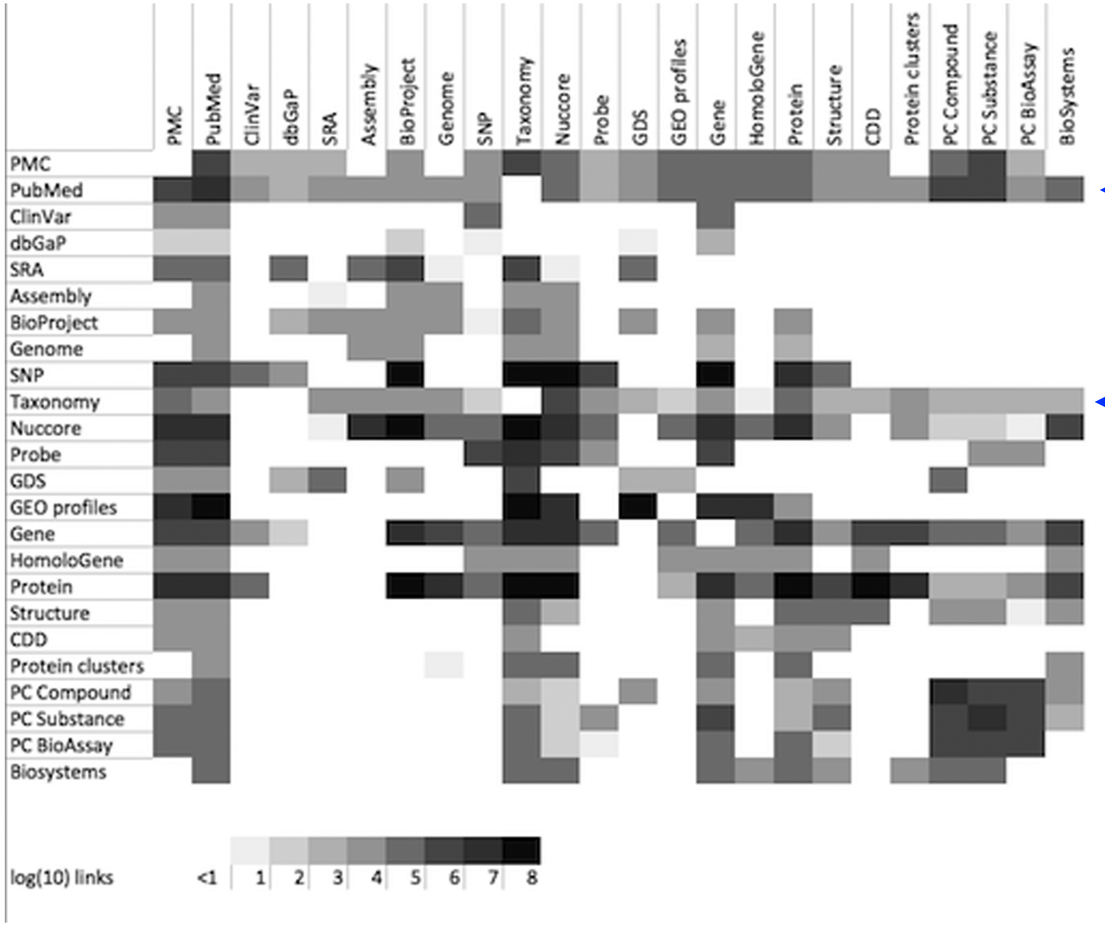
Plus many more…

https://academic.oup.com/nar/issue/45/D1

One useful feature of NCBI databases is that they connect to each other

So, you can, for example:

- get all the sequences associated with a species or genus
- get all the protein sequences encoded by a genome
- get the NGS datasets associated with a paper



links from Pubmed to other dbs

links from Taxonomy to other dbs

Nucleic Acids Res (2017) 45 (D1): D12-D17

The 2020 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

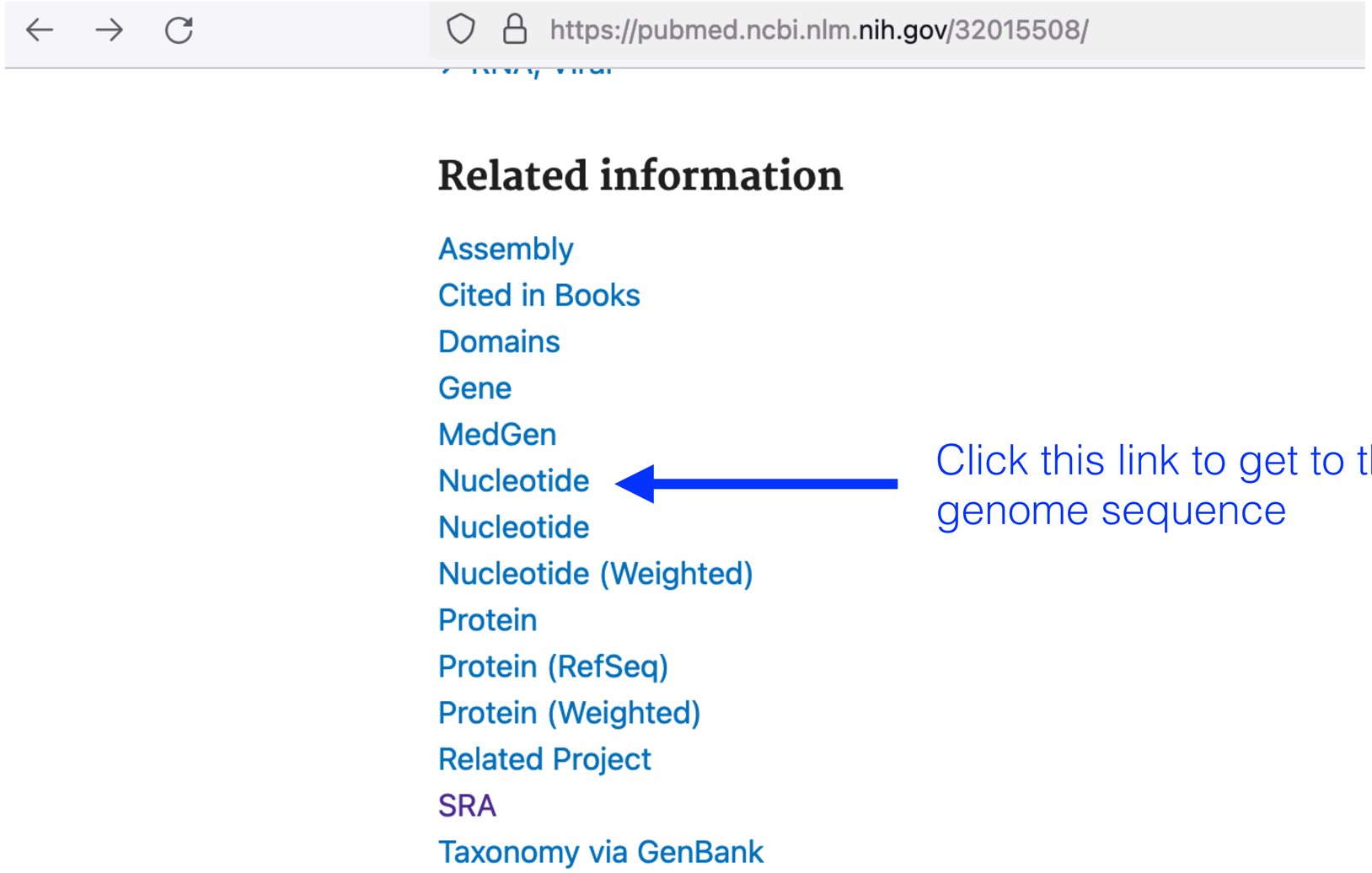
Fan Wu[1,7], Su Zhao[2,7], Bin Yu[3,7], Yan-Mei Chen[1,7], Wen Wang[4,7], Zhi-Gang Song[1,7], Yi Hu[2,7], Zhao-Wu Tao[2], Jun-Hua Tian[3], Yuan-Yuan Pei[1], Ming-Li Yuan[2], Yu-Ling Zhang[1], Fa-Hui Dai[1], Yi Liu[1], Qi-Min Wang[1], Jiao-Jiao Zheng[1], Lin Xu[1], Edward C. Holmes[1,5] & Yong-Zhen Zhang[1,4,6]✉

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 sequencing[4] 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

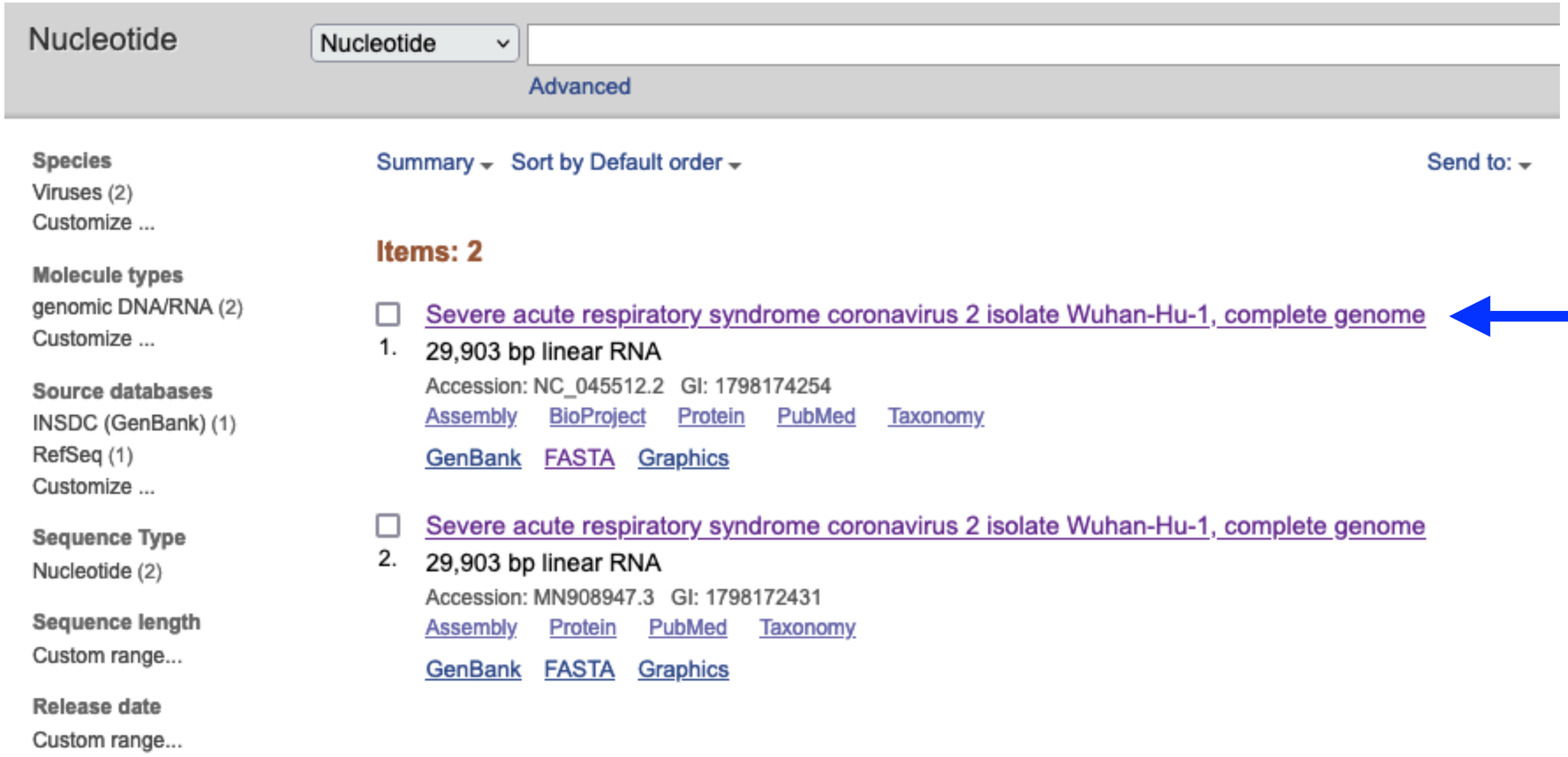
The pubmed record for that paper

Exercise: find the pubmed entry for the original SARS-CoV-2 paper PMID: 32015508

At the bottom of the pubmed page: related information links



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

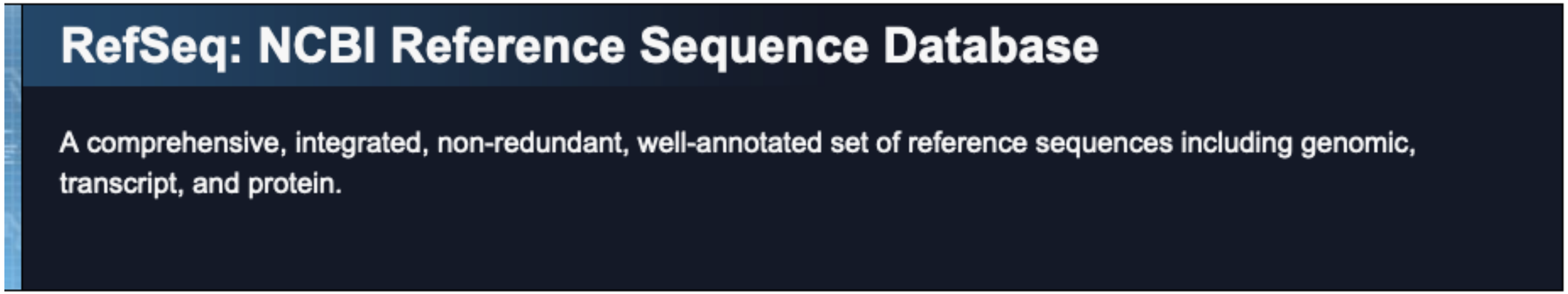


This sequence is in the RefSeq database

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 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

```
COMMENT     REVIEWED 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

Let’s assume that you don’t have Geneious, so do this from your browser.

First, download the sequence *without* annotation

FASTA: the most common format for sequence data

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
```

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
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
```

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
MTHQSHAYHMVKPSPWPLTGALSALLMTSGLAMWFHFHSTTLLMLGLLTNTLTMYQWWRDVTRESTYQGH
HTPPVQKGLRYGMVLFITSEVFFFAGFFWAFYHSSLAPTPQLGGHWPPTGITPLNPLEVPLLNTSVLLAS
GVSITWAHHSLMENNRNQMIQALLITILLGLYFTLLQASEYFESPFTISDGIYGSTFFVATGFHGLHVII
GSTFLTICFIRQLMFHFTSKHHFGFEAAAWYWHFVDVVWLFLYVSIYWWGS

>YP_002124314.2 cytochrome b
MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDASTAFSSIAHIT
RDVNYGWIIRYLHANGASMFFICLFLHIGRGLYYGSFLYSKTWNIGIILLLATMATAFMGYVLPWGQMSF
WGATVITNLLSAIPYIGTDLVQWIWGGYSVDSPTLTRFFTFHFILPFIIAALAALHLLFLHETGSNNPLG
ITSHSDKITFHPYYTIKDALGLFLFLLSLMTLTLLSPDLLGDPDNYTLANPLNTPPHIKPEWYFLFAYTI
LRSVPNKLGGVLALLLSILILAMIPILHVSKQQSMMFRPLSQSLYWLLAADLLILTWIGGQPVSYPFIII
GQVASVLYFTTILILMPTISLIENKMLKWA

>YP_002124302.2 NADH dehydrogenase subunit 1
MANLLLLVVPILIAMAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPATSTITLYI
TAPTLALTIALLLWTPLPMPNPLVNLNLGLLFILATSSLAVYSILWSGWASNSNYALIGALRAVAQTISY
EVTLAIILLSTLLMSGSFNLSTLITAQEHLWLLLPSWPLAMMWFISTLAETNRTPFDLAEGESELVSGFN
IEYAAGPFALFFMAEYTNIIMMNTLTTTIFLGTTYNALSPELYTTYFVTKTLLLTSLFLWIRTAYPRFRY
DQLMHLLWKNFLPLTLALLMWYVSMPITISSIPPQT
```

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

```
This is a plain text file.

There is no information about text formatting
(font, bold, italics, etc.)

You can't do things like embed images in the file.

Most bioinformatics software uses plain text data.

Notepad++ is a plain text editor for Windows

BBEdit is a plain text editor for MacOS.

This file only uses 350 bytes of storage on the disk.

```

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*).

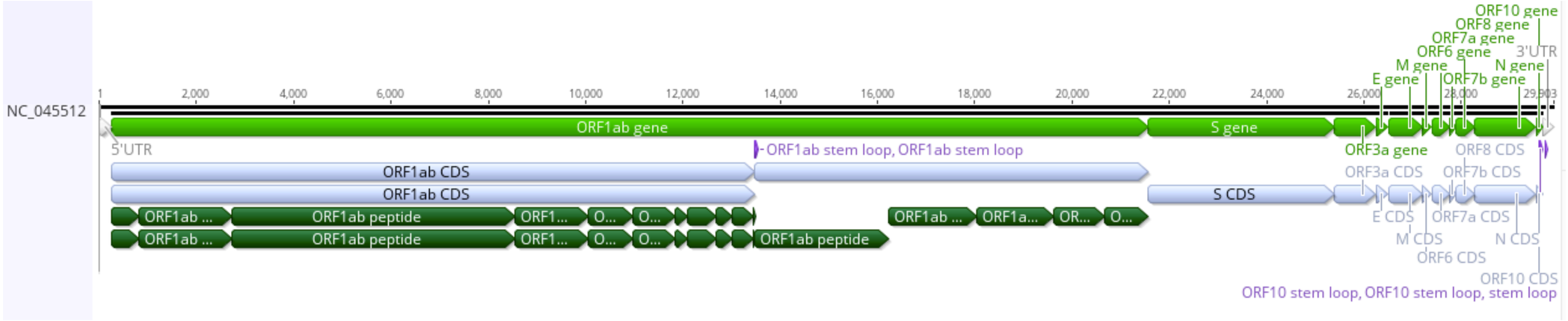
Lipman and Pearson (1985) Science

A 1985 Macintosh computer

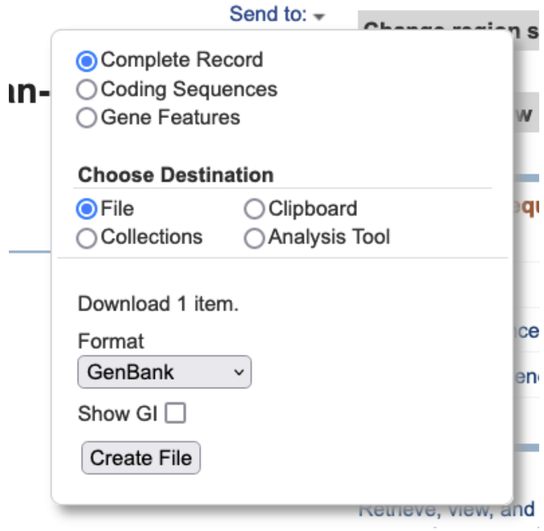
Storage: 400 kbyte floppy disks

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



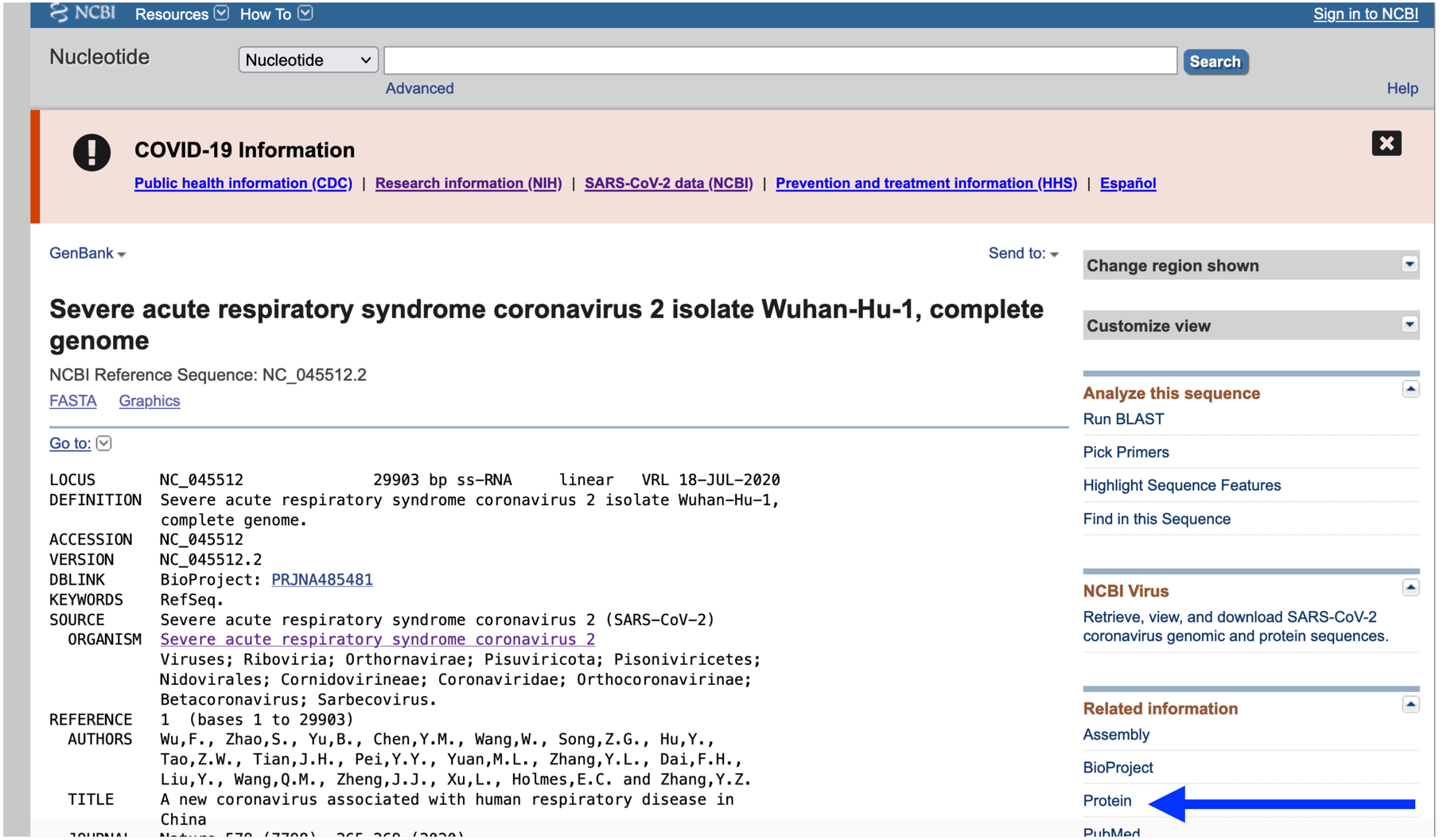
Genbank format includes annotation

```
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!

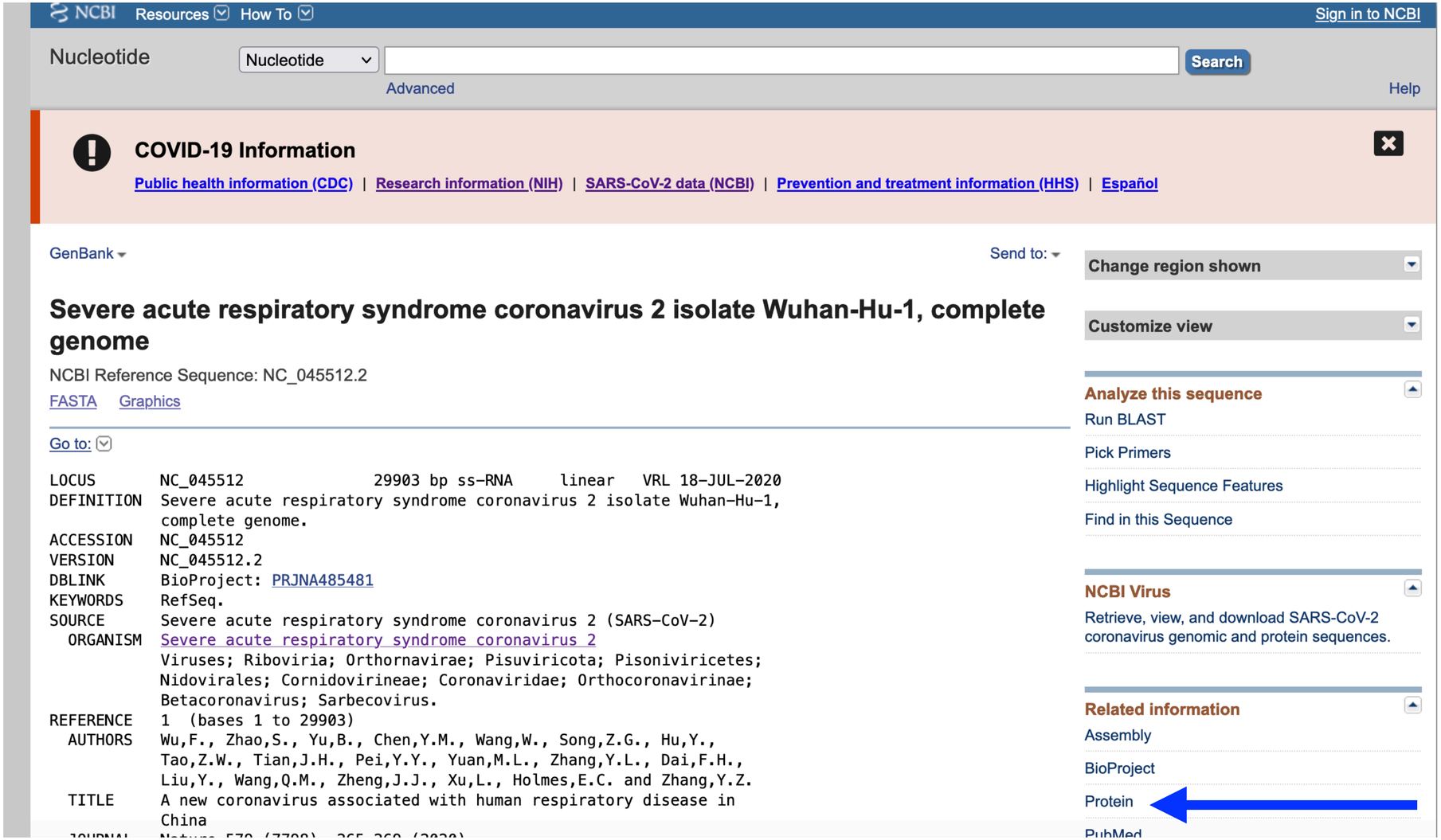
Exercise: according to the annotation, what is the collection data and host of this first SARS-CoV-2 sequence?

The nucleotide database links to other databases.



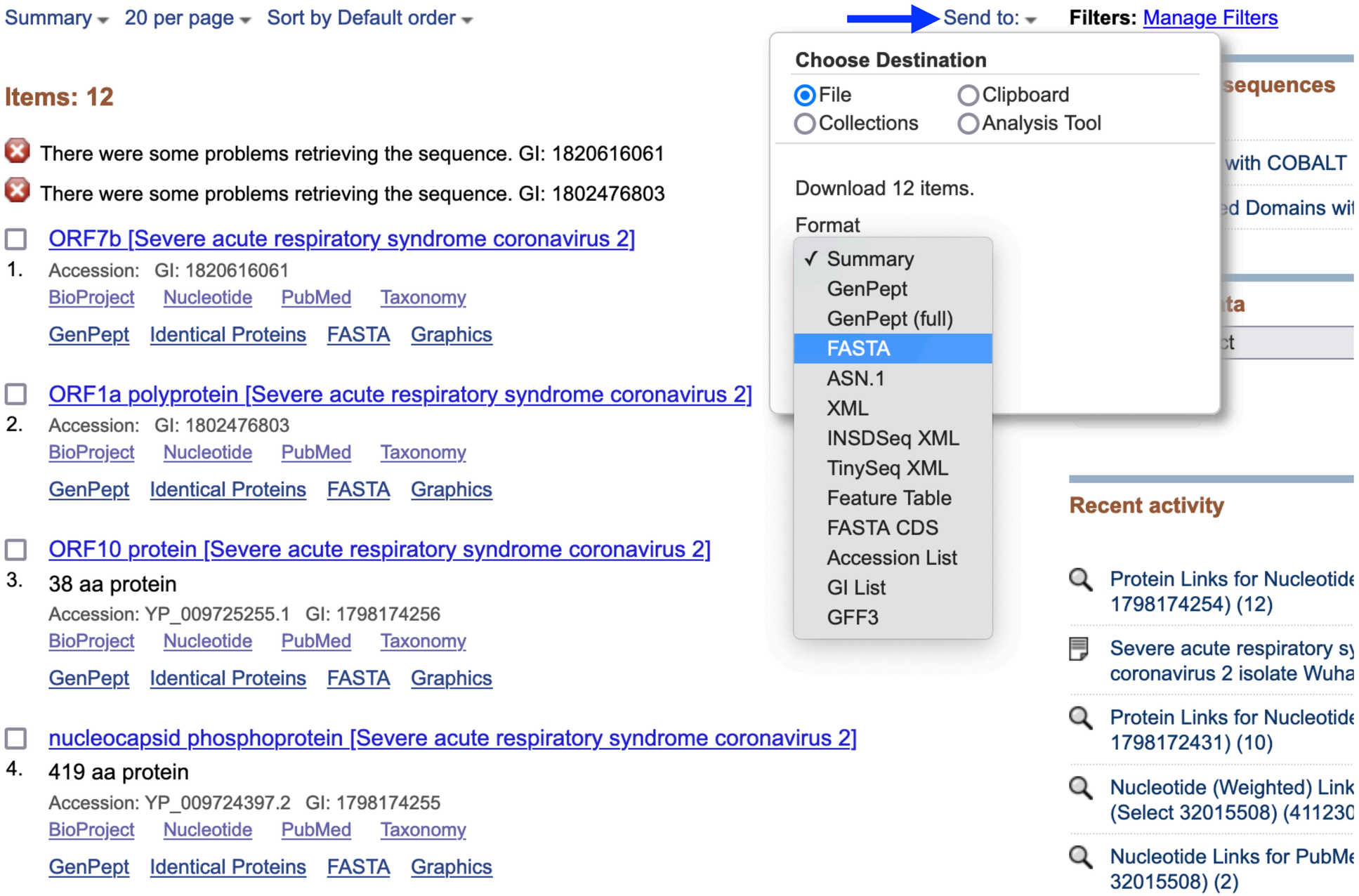
Click this link to get to the protein sequences encoded by this genome

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

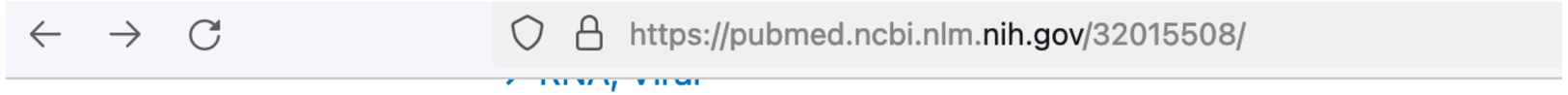


Click this link to get to the protein sequences encoded by this genome

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



At the bottom of the pubmed page: related information links

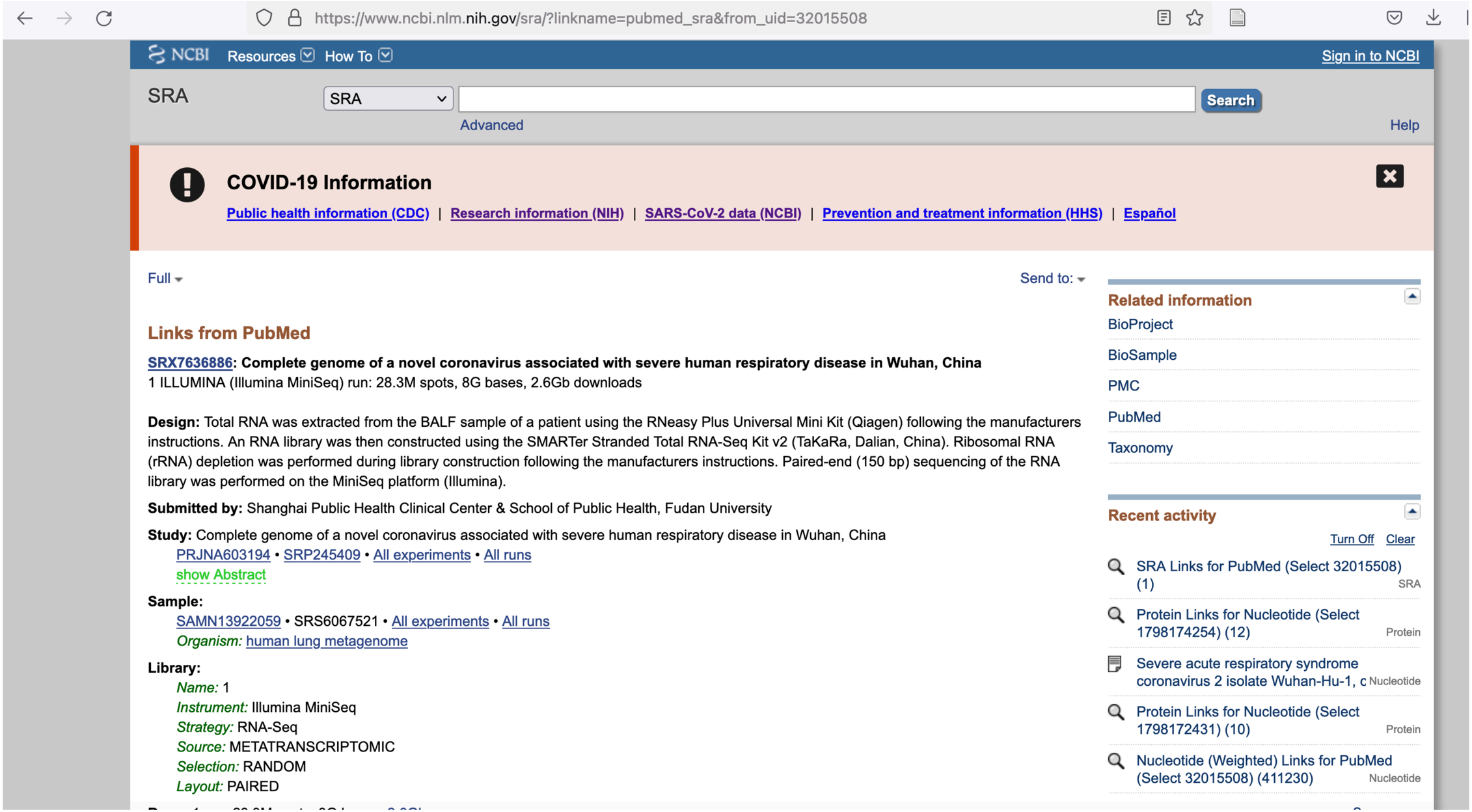


Related information

Assembly
Cited in Books
Domains
Gene
MedGen
Nucleotide
Nucleotide
Nucleotide (Weighted)
Protein
Protein (RefSeq)
Protein (Weighted)
Related Project
SRA
Taxonomy via GenBank

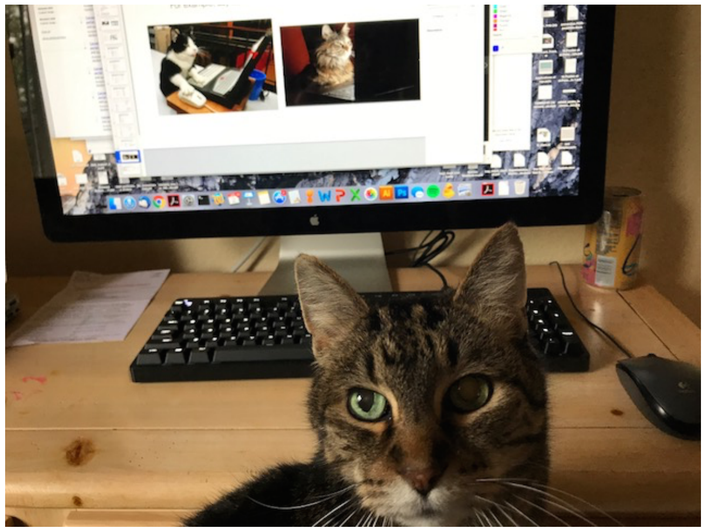
Click this link to get to the NGS data from this paper

The NCBI SRA database contains NGS datasets



A great way to find sequence data for an organism you are interested in is via the NCBI Taxonomy database

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



Kirby

The Taxonomy database

https://www.ncbi.nlm.nih.gov/taxonomy

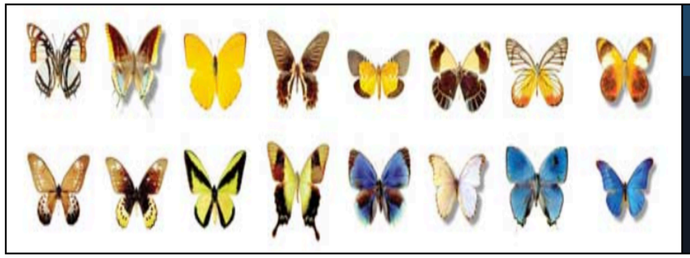
NCBI Resources How To Sign in to NCB

Taxonomy Taxonomy Search

Limits Advanced Hel

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

The Taxonomy page for *Felis catus*

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 [complete name] [lock] Go Clear

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

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

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

Download genome, transcriptome, proteome, annotation

Display Settings: Overview | Send to:

Organism Overview ; Genome Assembly and Annotation report [4] ; Organelle Annotation Report [1] ID: 78

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 Release: 104

NCBI Resources
Genome Data Viewer

Tools
BLAST Genome

Related information
Assembly
BioProject
Gene
Components
Protein
PubMed
Taxonomy

Search details
txid9685[Organism:noexp]

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

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 ← Download annotation
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:

Organism Overview ; Genome Assembly and Annotation report [4] ; Organelle Annotation Report [1] ID: 78

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 Release: 104

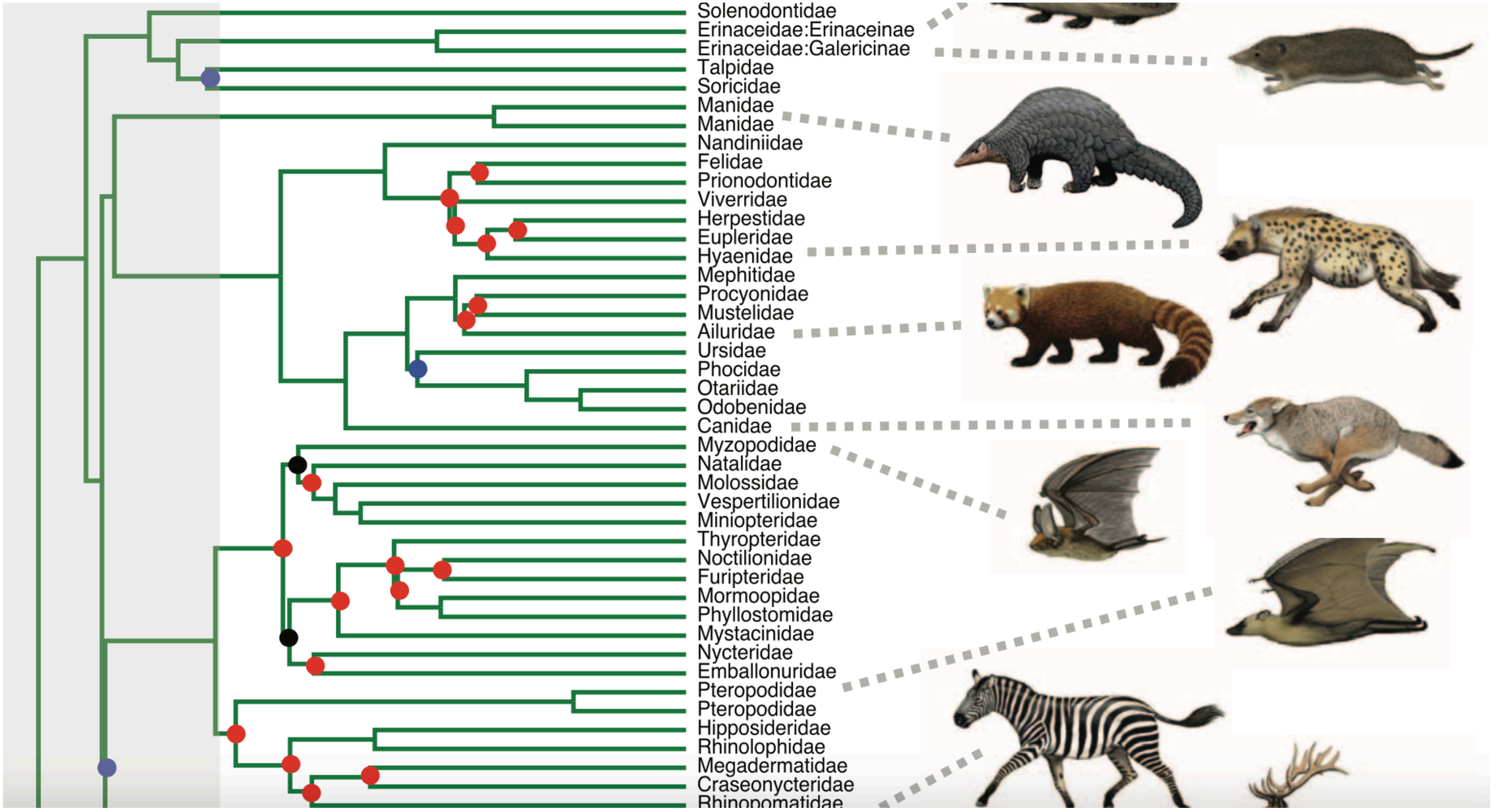
NCBI Resources
Genome Data Viewer

Tools
BLAST Genome

Related information
Assembly
BioProject
Gene
Components
Protein
PubMed
Taxonomy

Search details
txid9685[Organism:noexp]

The taxonomy database includes the taxonomic lineage of organisms



Meredith (2011) Science

You can go to any point in the taxonomic tree of life 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*)

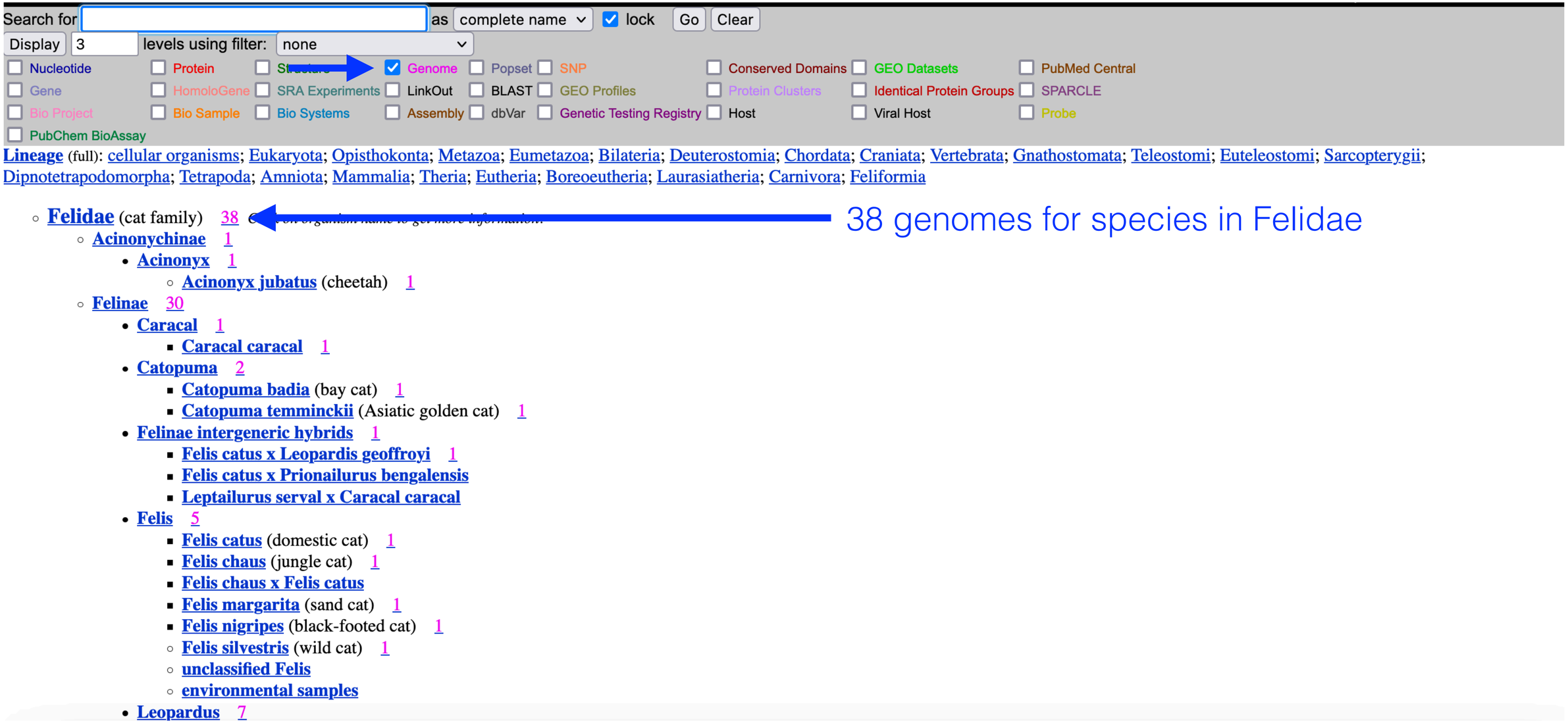
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Carnivora; Feliformia;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

Eukaryotes

Mammals

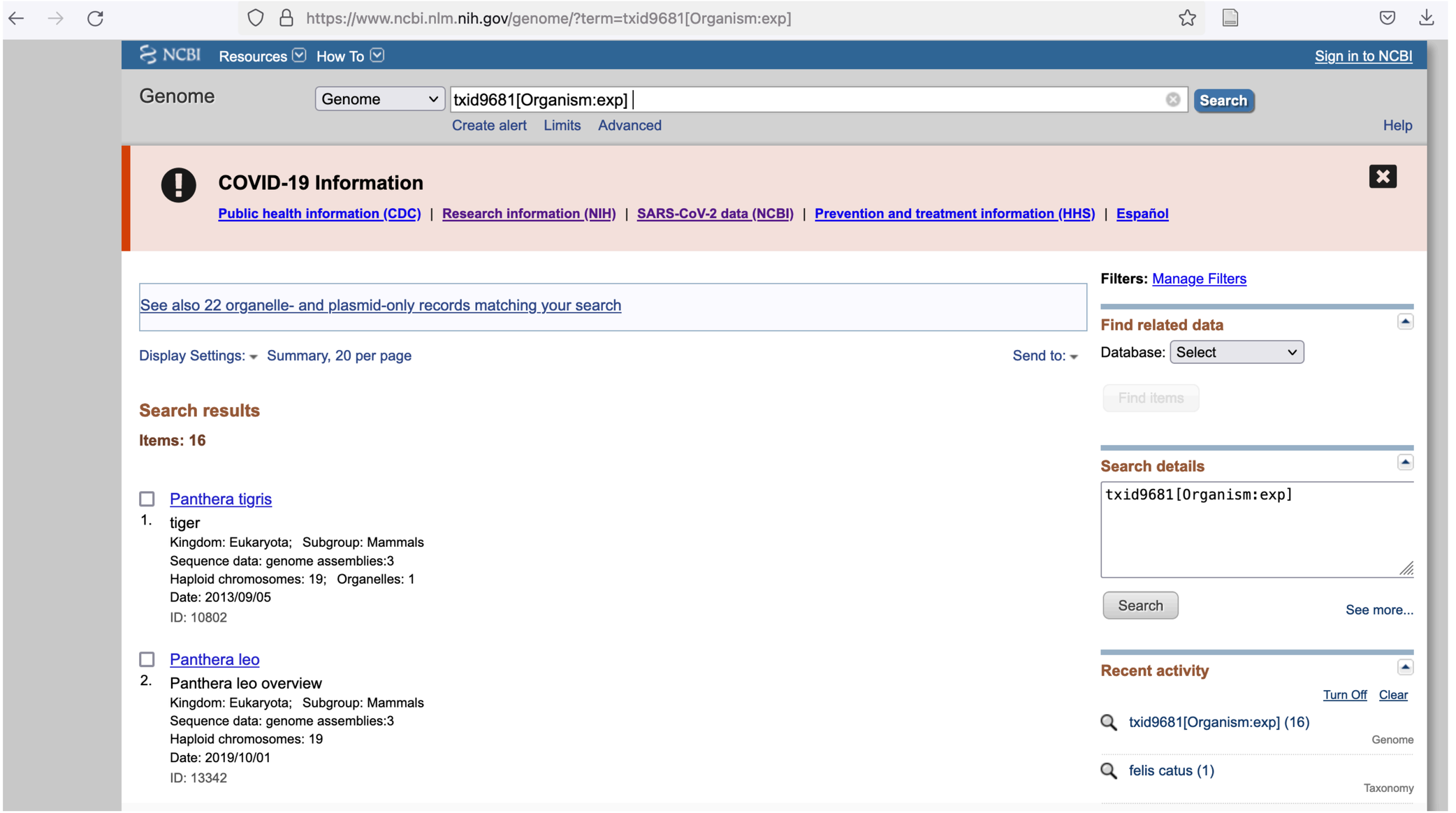
Felidae

All the available genomes for species in the Felidae family



38 genomes for species in Felidae

Felidae genomes

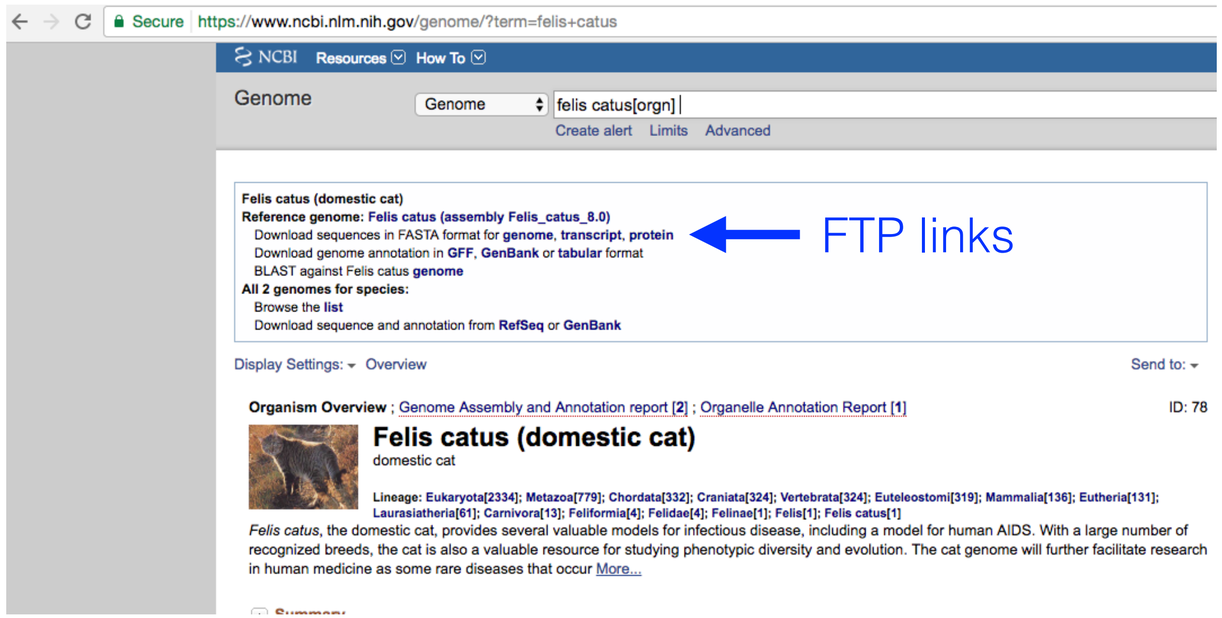


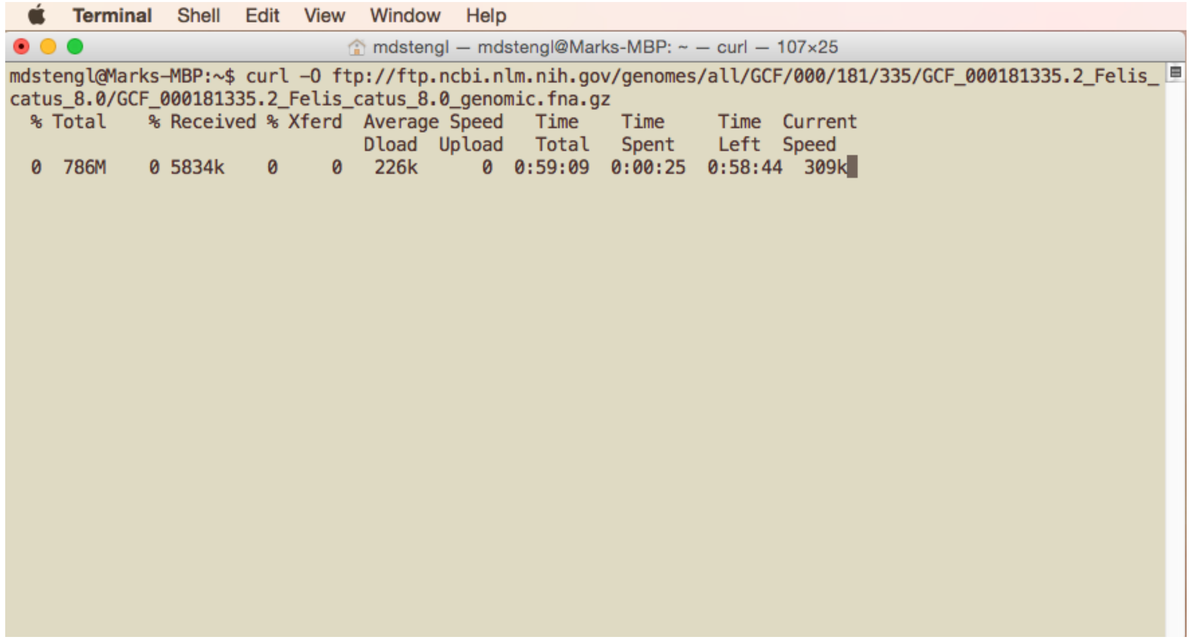
Exercise: NCBI Scavenger Hunt!

Image credit: Harvard Law Record CC BY 2.0 Link

You can download data from the command line

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

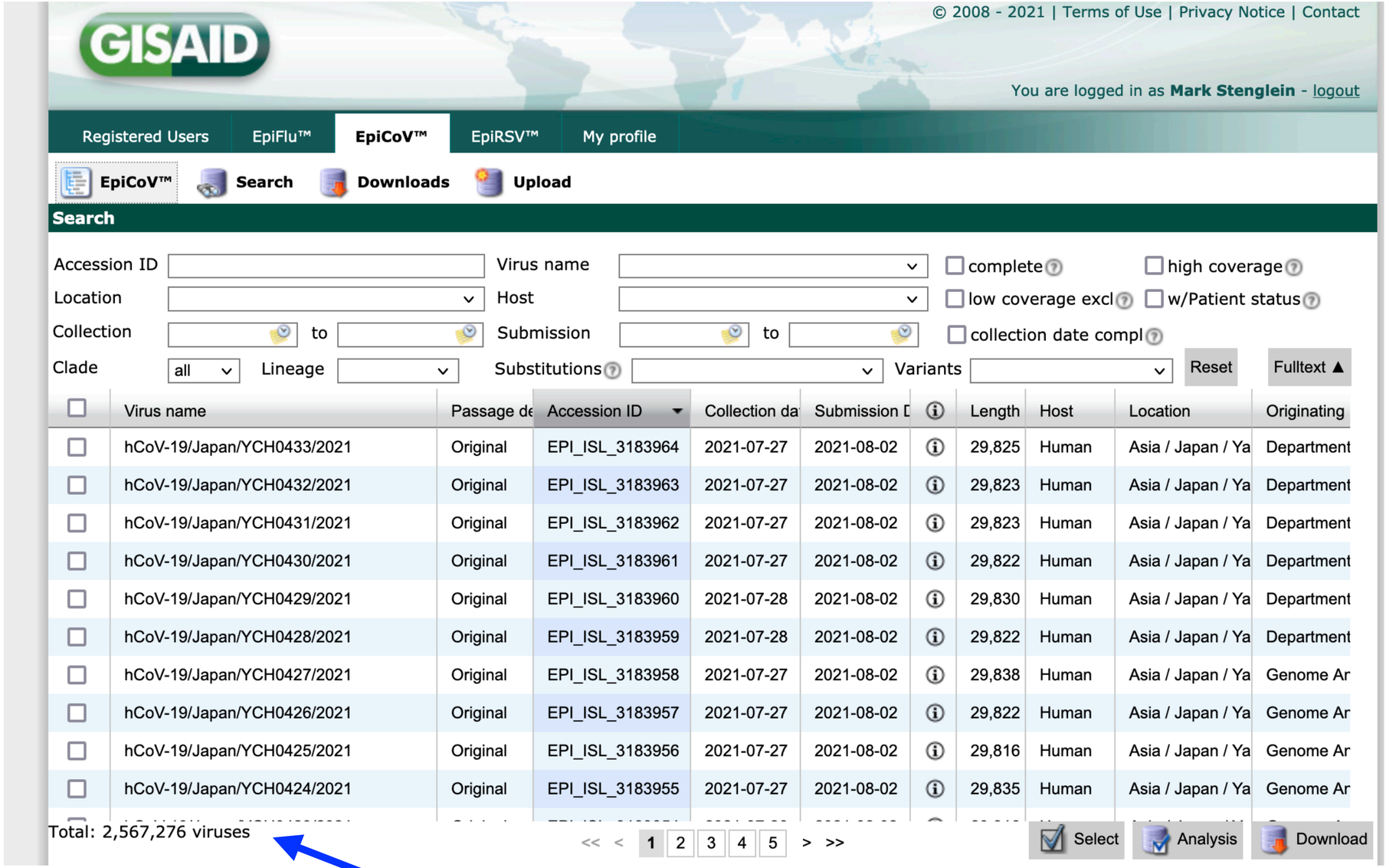




curl is a file transfer utility built into Linux, MacOS

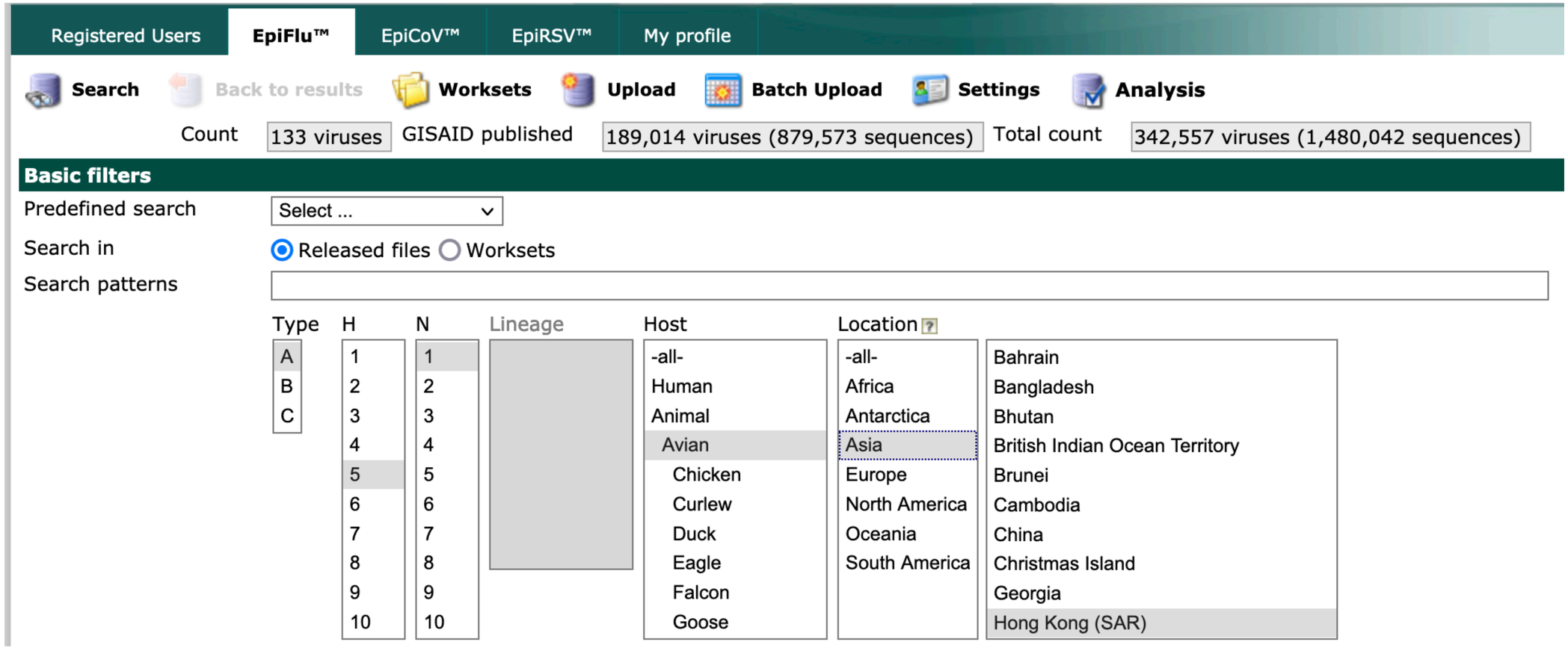
similar utilities exist for Windows

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



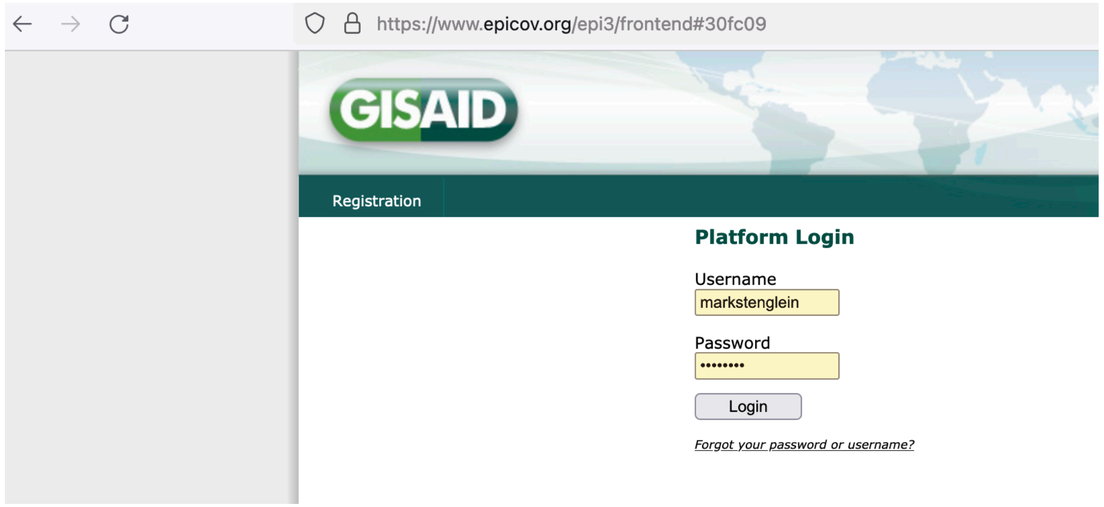
12.9 million SARS-CoV-2 sequences!! (~6M in Genbank)

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



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



Anonymous GISAID User
@GISAID_anon

Tweeting what scientists are afraid to post publicly for fear of retribution from GISAID
DMs are open and will be treated confidentially

Joined August 2022

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.

1. **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.
2. **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

gigadb.org/dataset/100060

ASSEMBLATHON 2

Assemblathon 2 assemblies.

Dataset type: Genomic
Data released on June 24, 2013

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; Godzaridis ; 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).