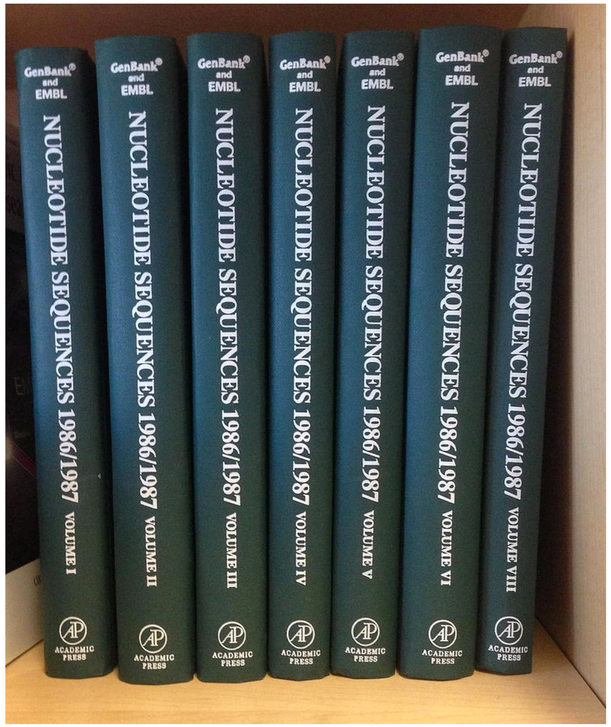
An overview of genomics databases and online resources: what they are and how to access them

Mark Stenglein, GDW

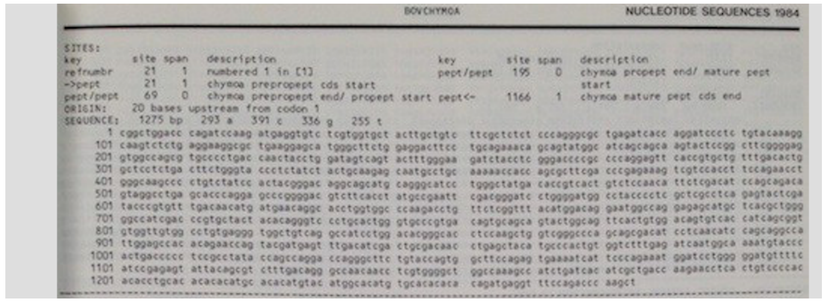
Genomics of Disease in Wildlife

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 mainly on NCBI databases and resources, and how to access them

Categories of NCBI databases

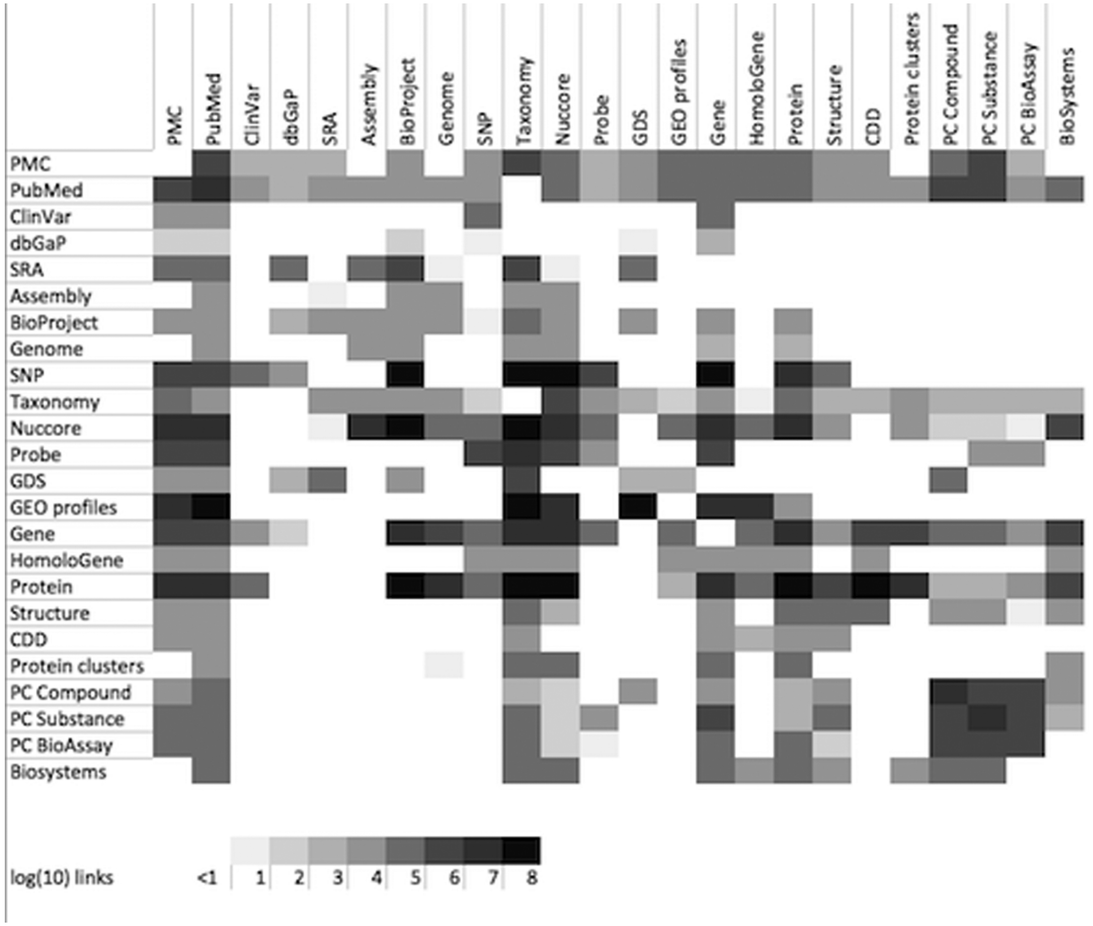
Category	Example NCBI db	Content
Literature	PubMed	Scientific and medical abstracts/ citations
Genomes	Assembly	Genome assembly information
Genes	Gene	Collected information about gene loci
Proteins	Protein	Protein sequences
Chemicals	PubChem Compound	Chemical information with structures, information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

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

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



links from Pubmed to other dbs

links from Taxonomy to other abs

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

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

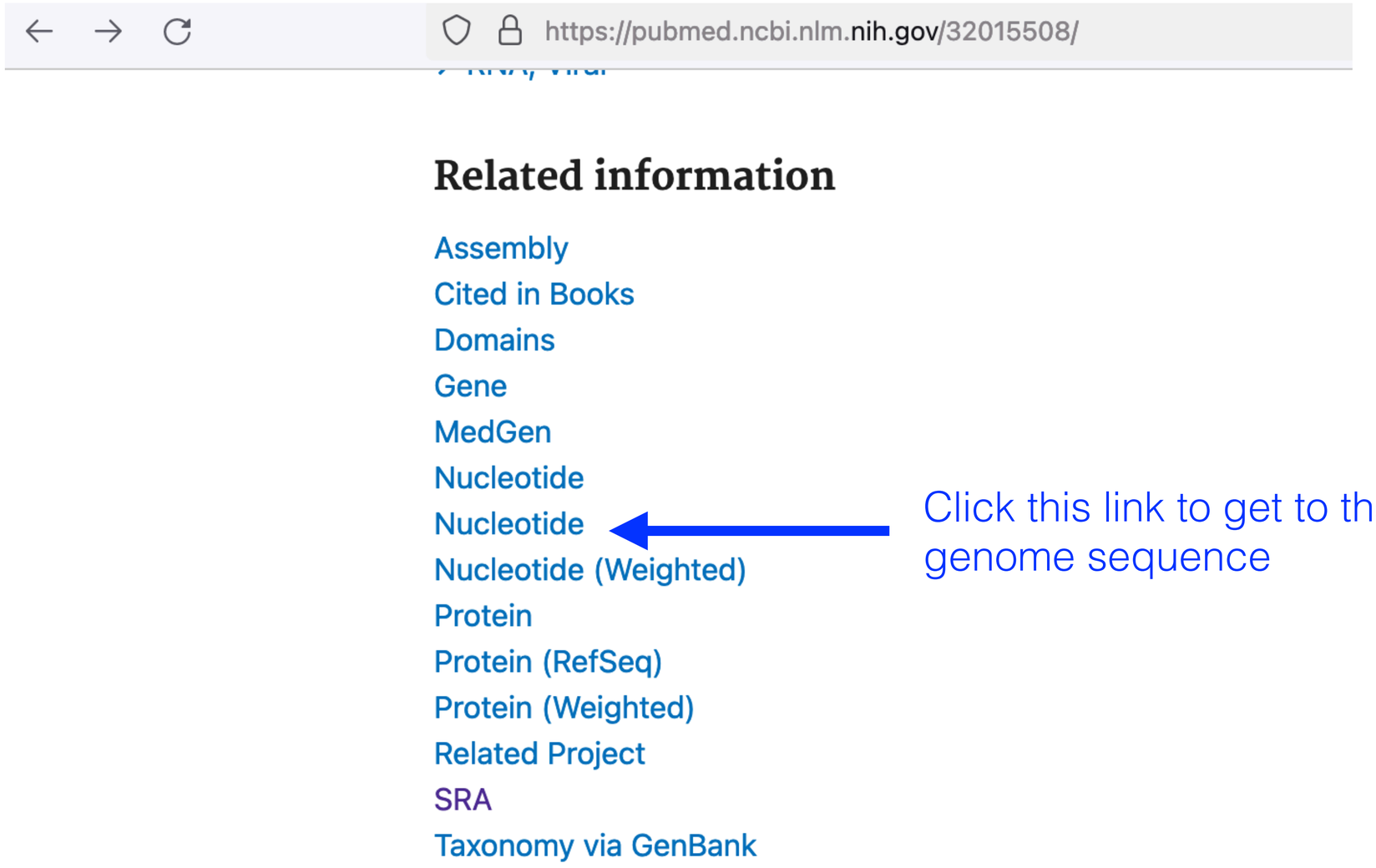
Fan Wu[1,7], Su Zhao[2,7], Bin Yu[3,7], Yan-Mei Chen[1,7], Wen Wang[4,7], Zhi-Gang Song[1,7], Yi Hu[2,7], Zhao-Wu Tao[2], Jun-Hua Tian[3], Yuan-Yuan Pei[1], Ming-Li Yuan[2], Yu-Ling Zhang[1], Fa-Hui Dai[1], Yi Liu[1], Qi-Min Wang[1], Jiao-Jiao Zheng[1], Lin Xu[1], Edward C. Holmes[1,5] & Yong-Zhen Zhang[1,4,6]✉

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

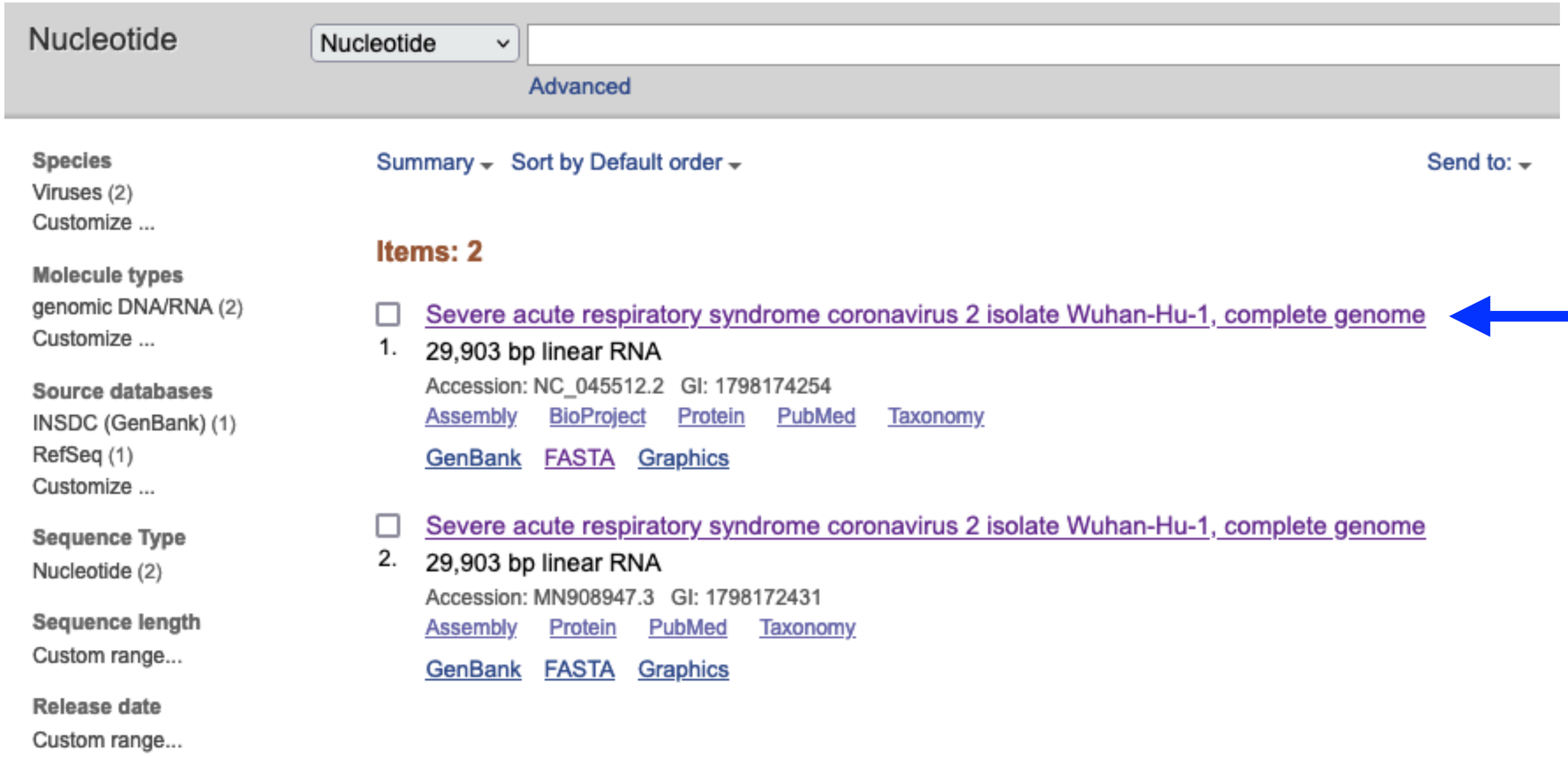
The pubmed record for that paper

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

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



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

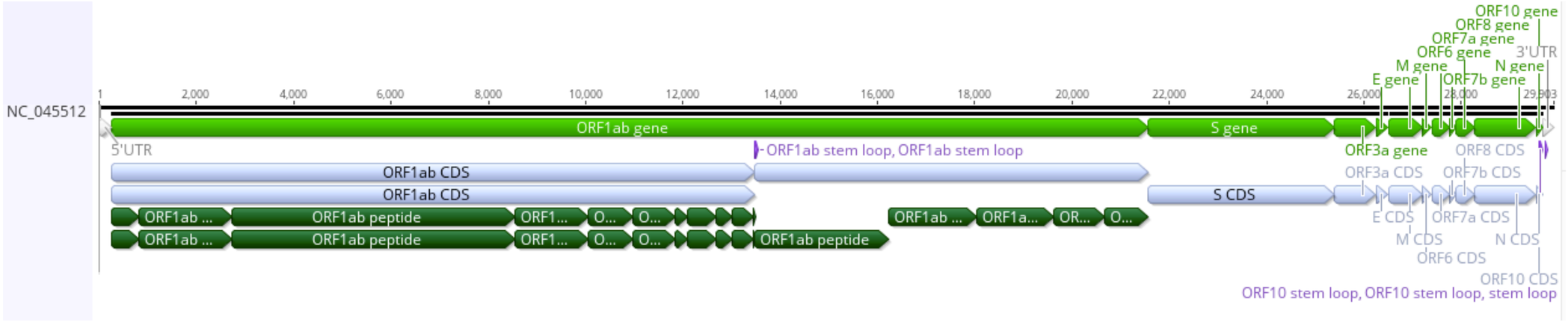


This sequence is in the RefSeq database

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.

```
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.

```

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)

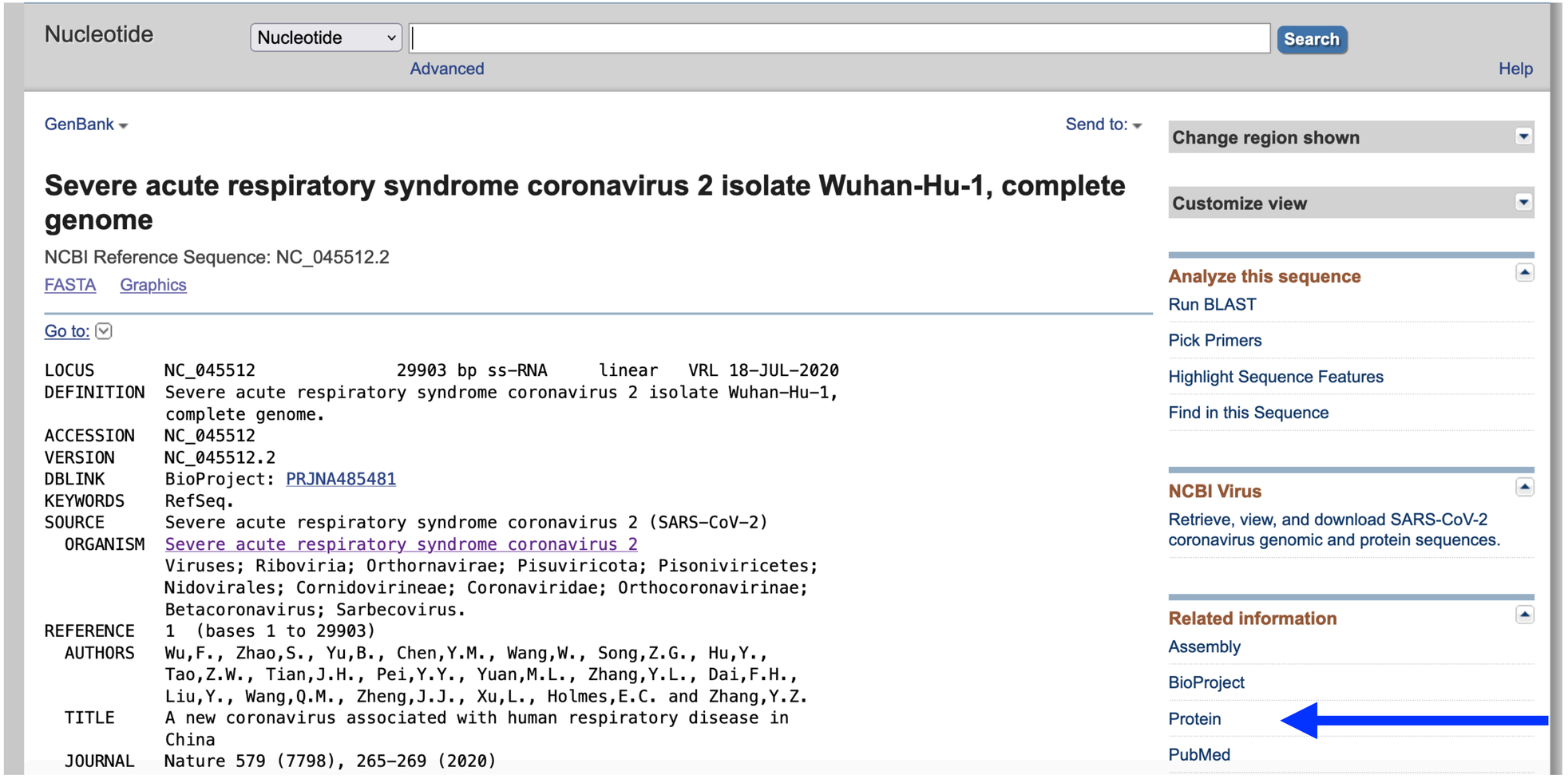
Genbank format includes annotation

```
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!

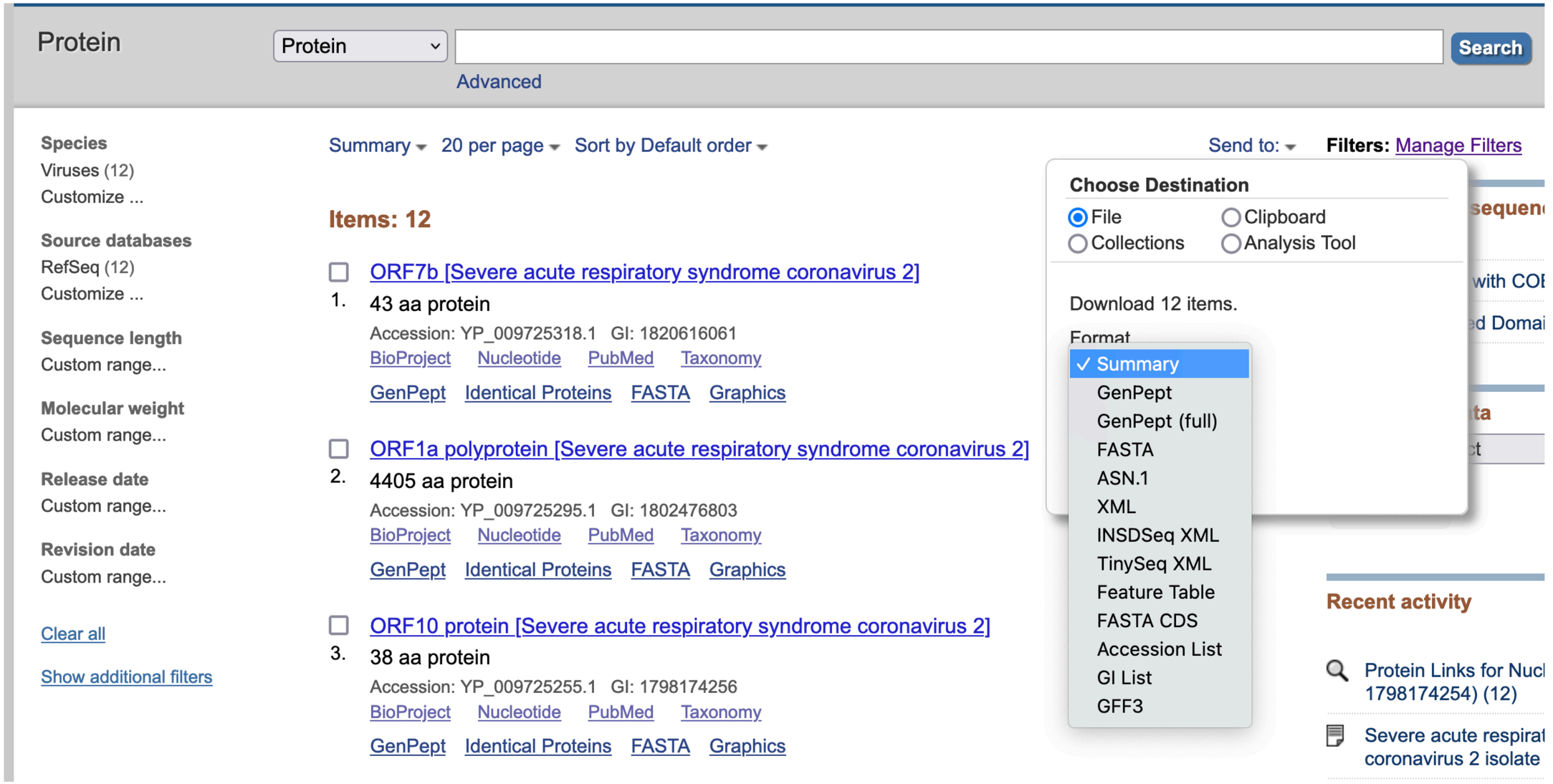
Exercise: according to the annotation, what is the collection date and host of this first SARS-CoV-2 sequence?

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

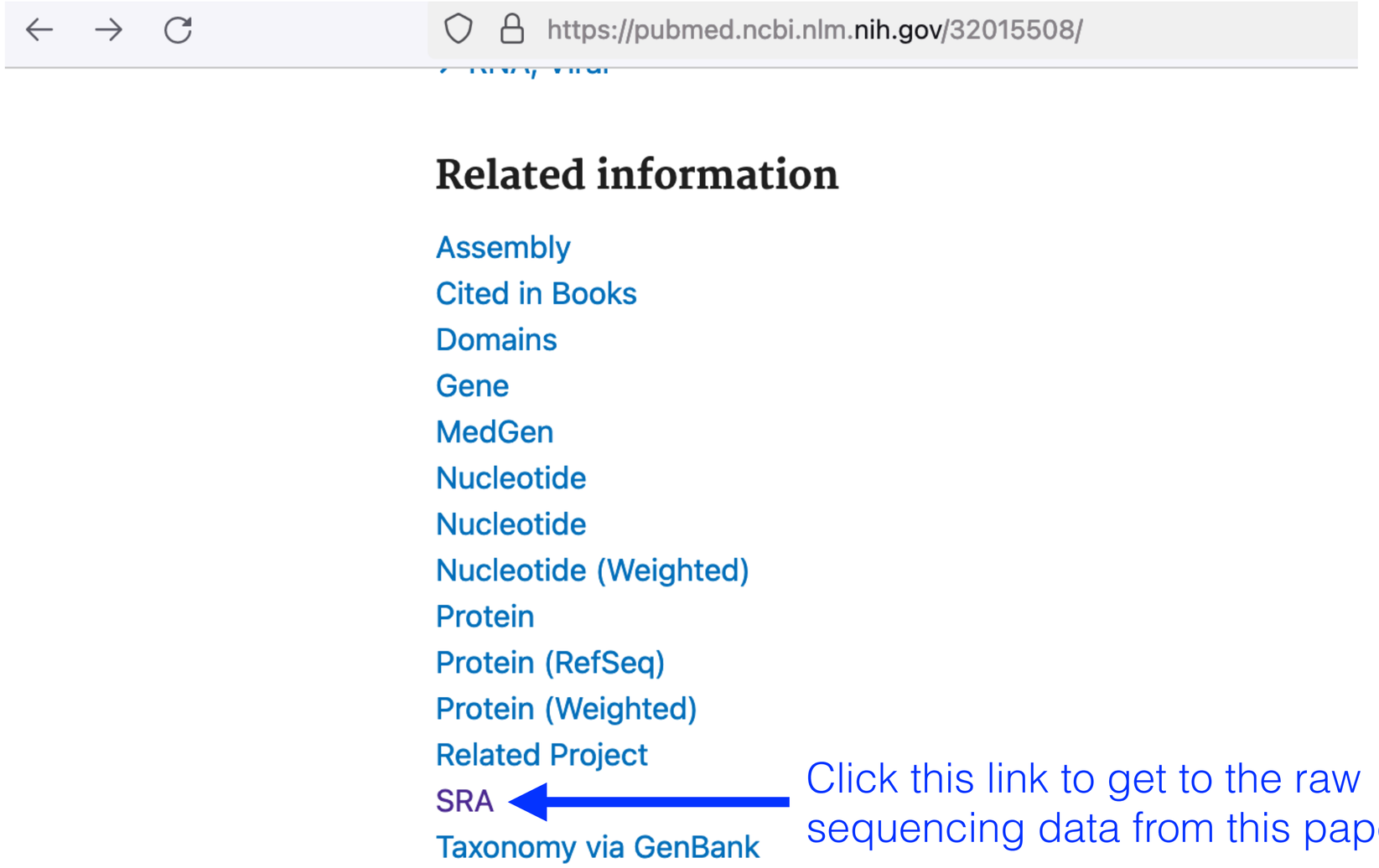


Click this link to get to the protein sequences encoded by this genome

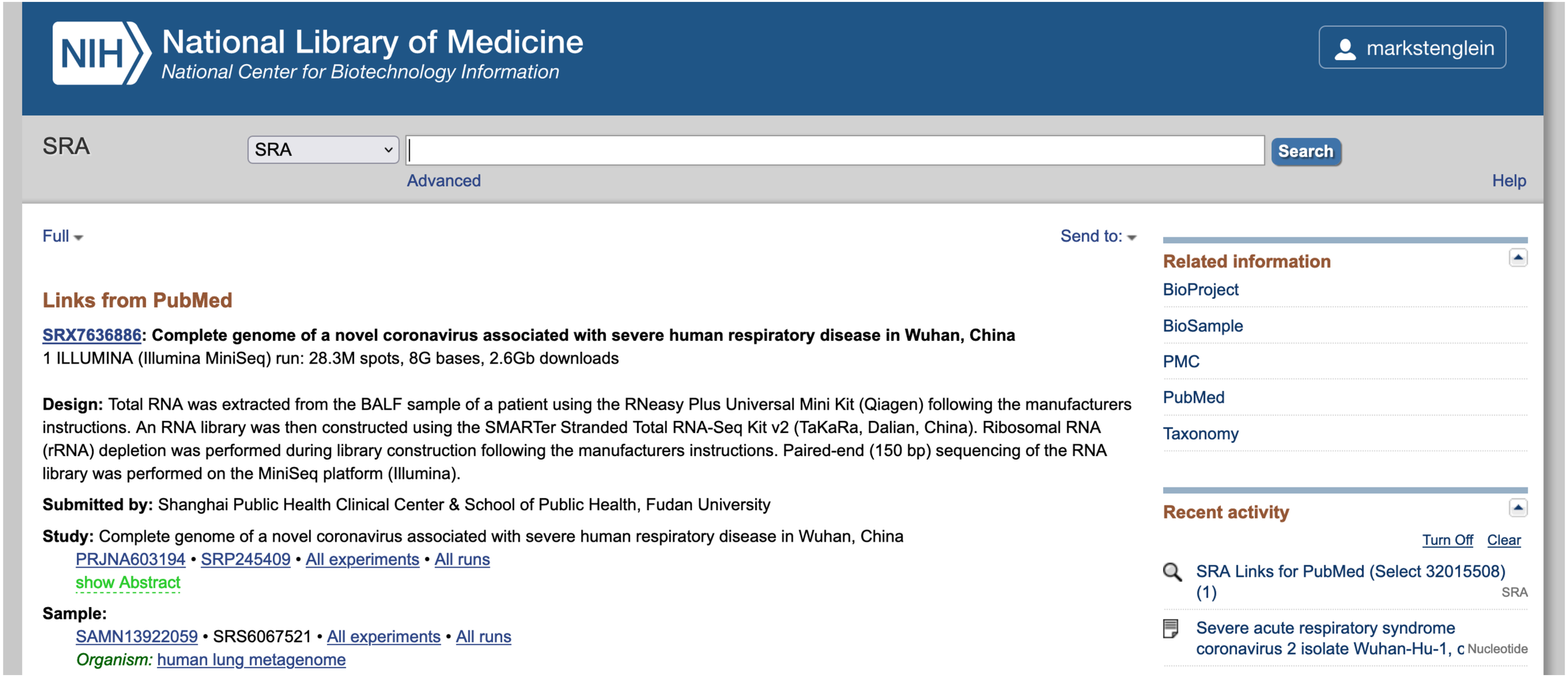
You can download the protein sequences all at once, in various formats



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



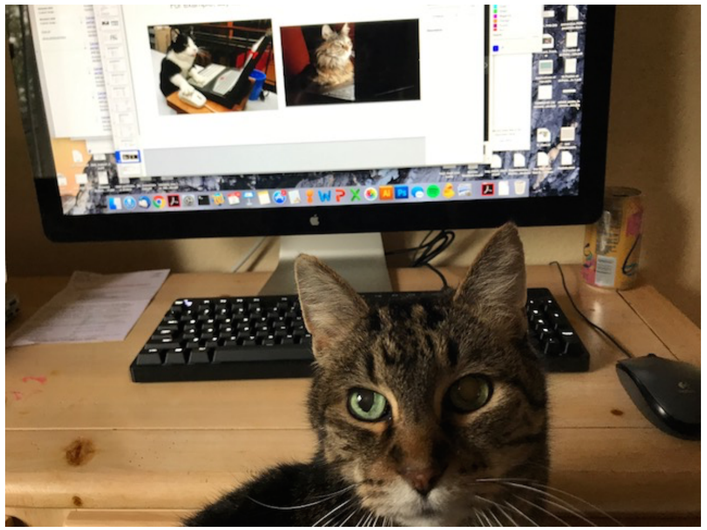
The NGS dataset that led to the initial identification of SARS-CoV-2



We will download NGS datasets from the SRA database later

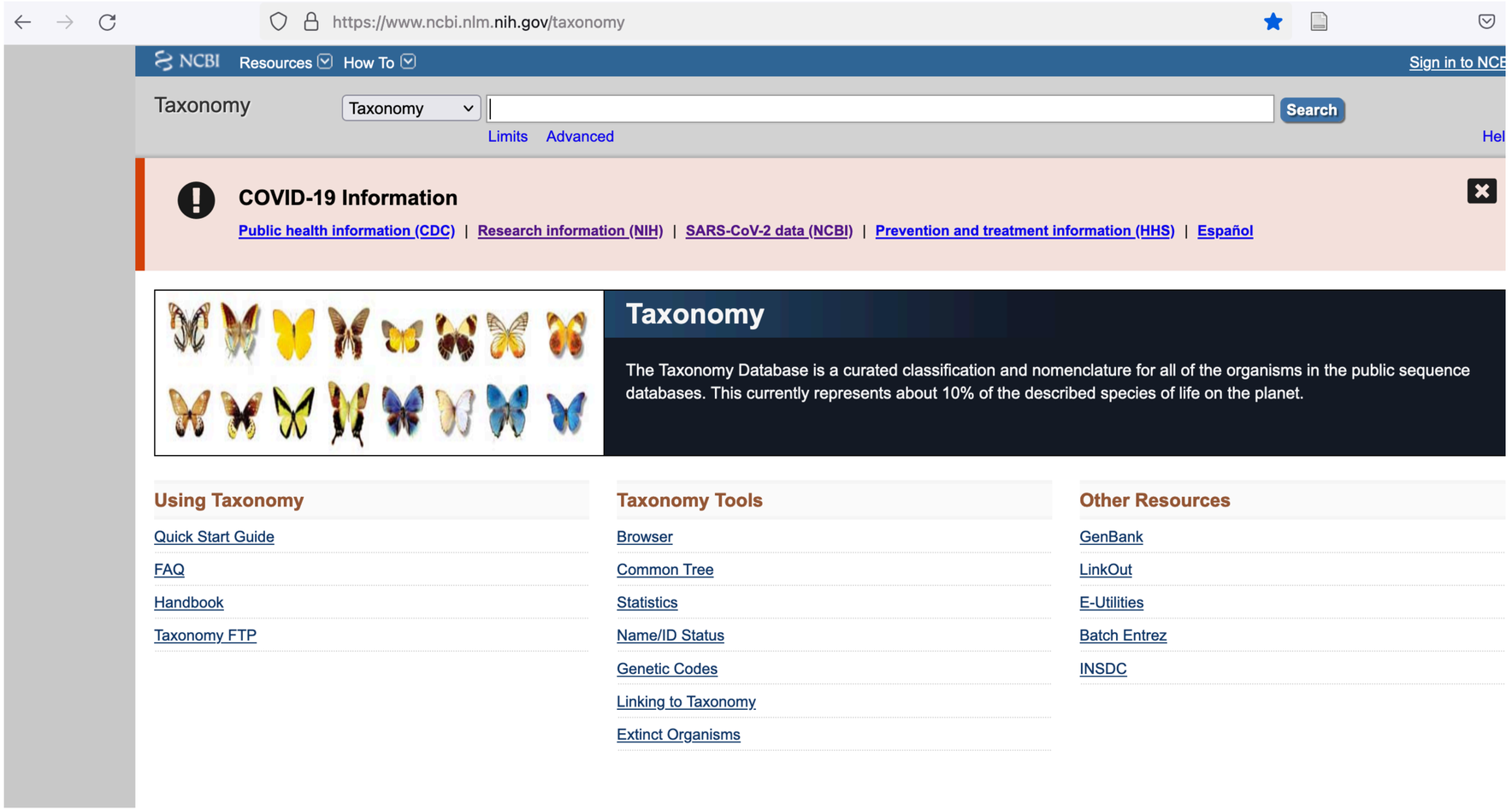
There are often many paths to the same data

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



I ❤️
NCBI
Taxonomy

A database representation of the tree of life

The NCBI Taxonomy page for the domestic cat

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

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:

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:

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

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)

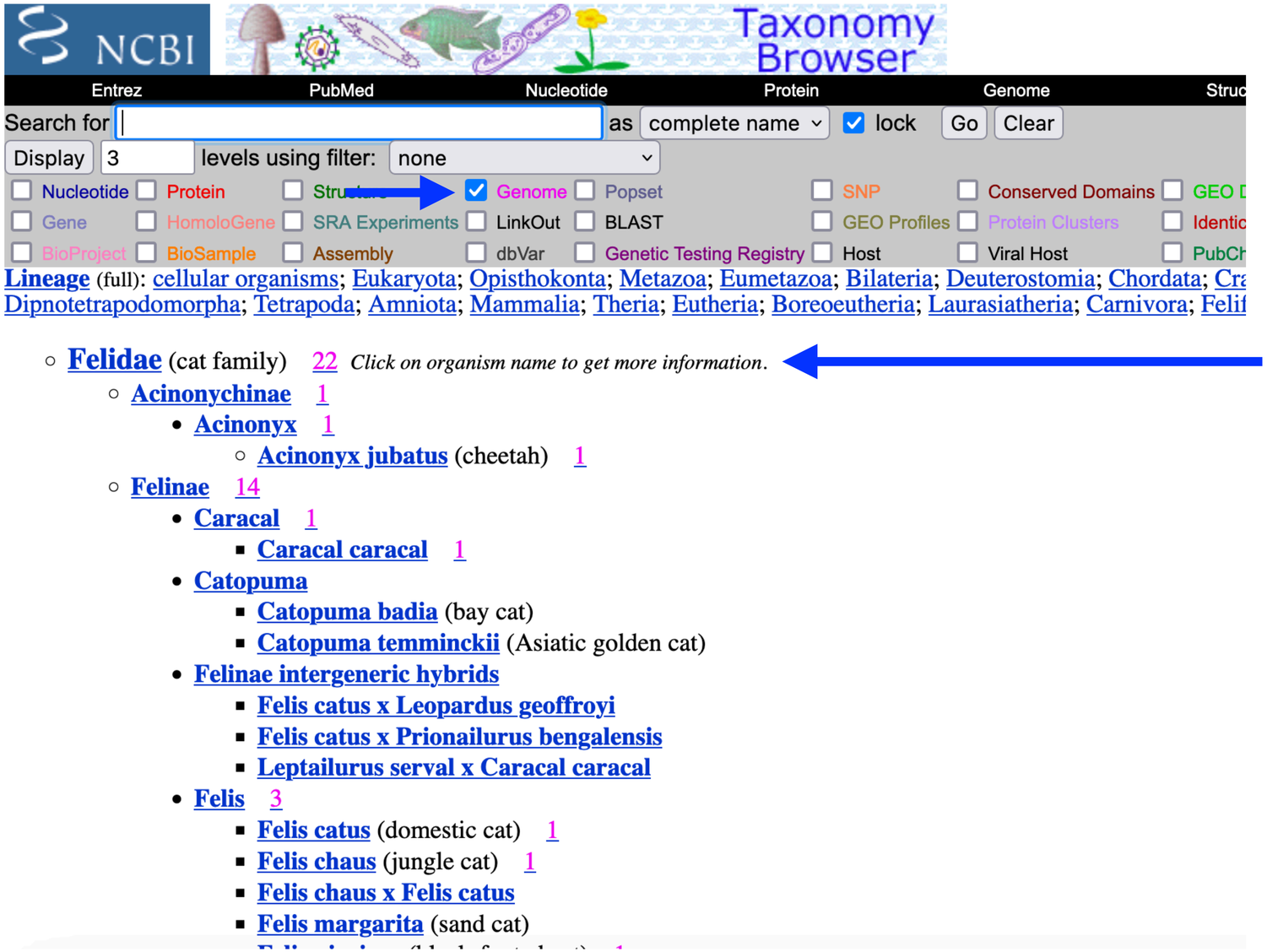
cat, **cats**

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

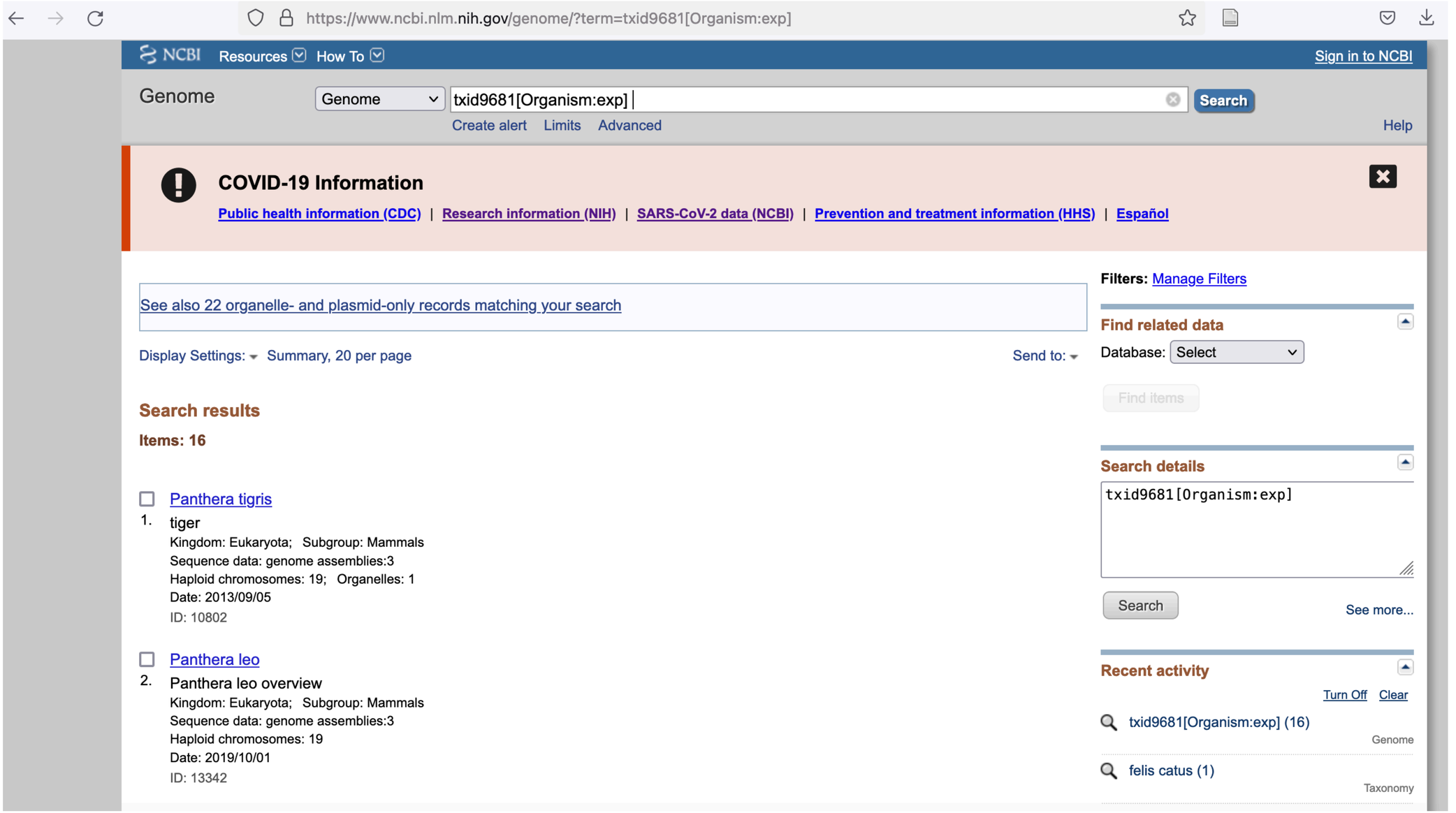
Felidae

You can go up the taxonomic tree in the Taxonomy db

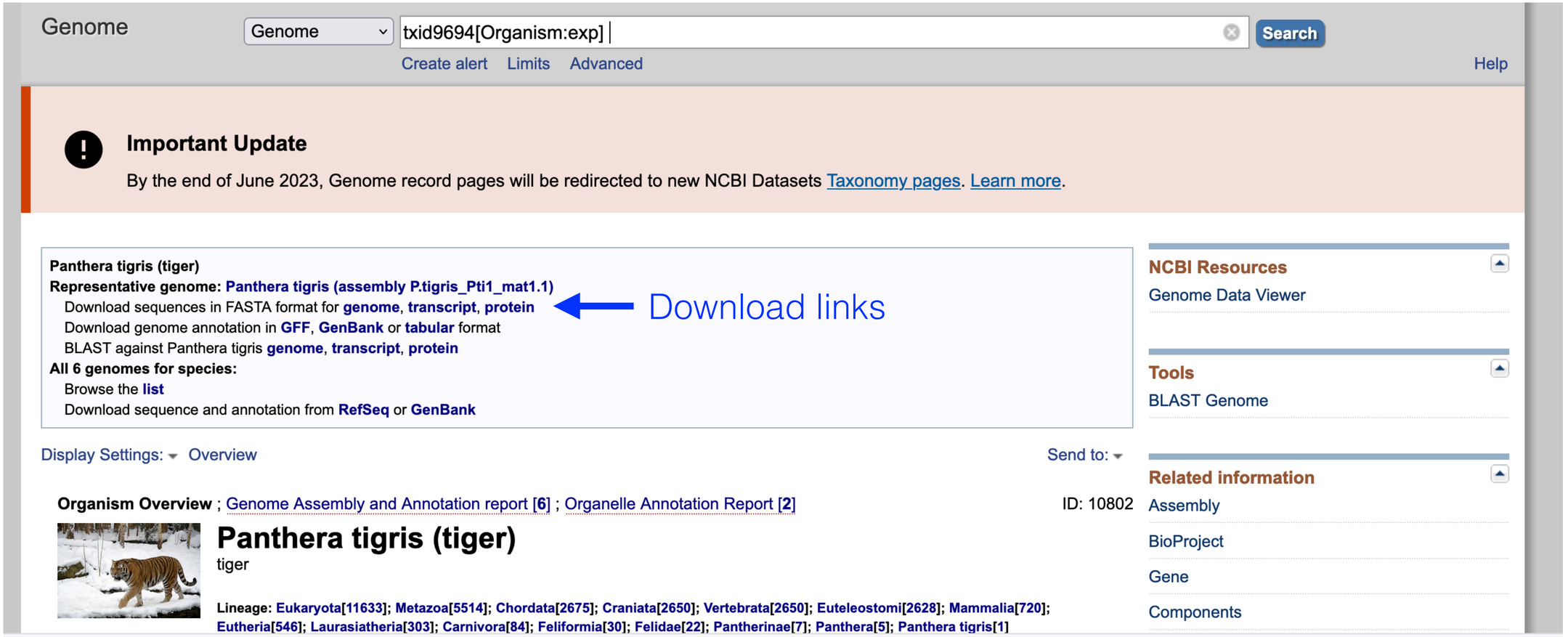


22 genomes for species in Felidae

Felidae genomes



You can download genome sequences from the genome database

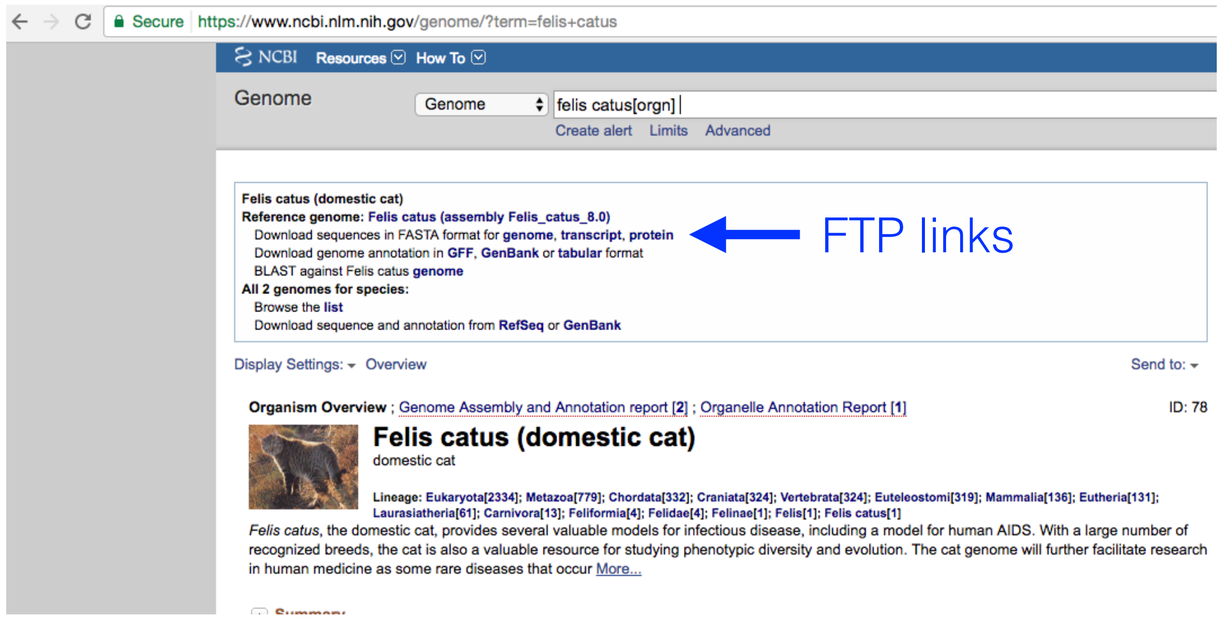


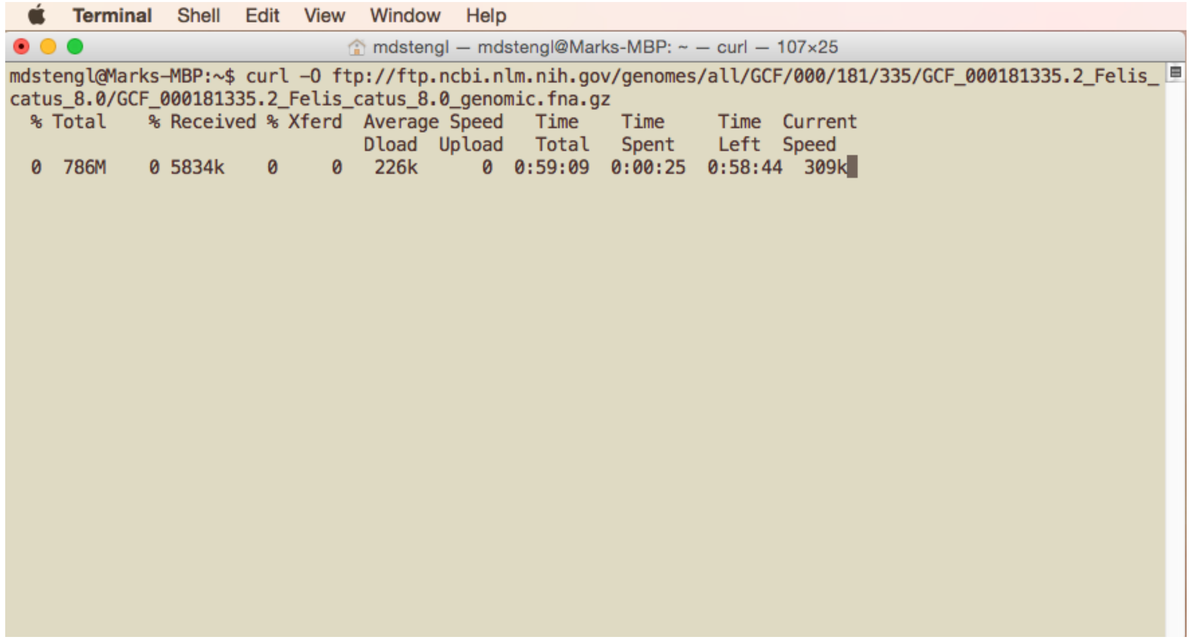
Links to genome, transcriptome, proteome, annotation

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.

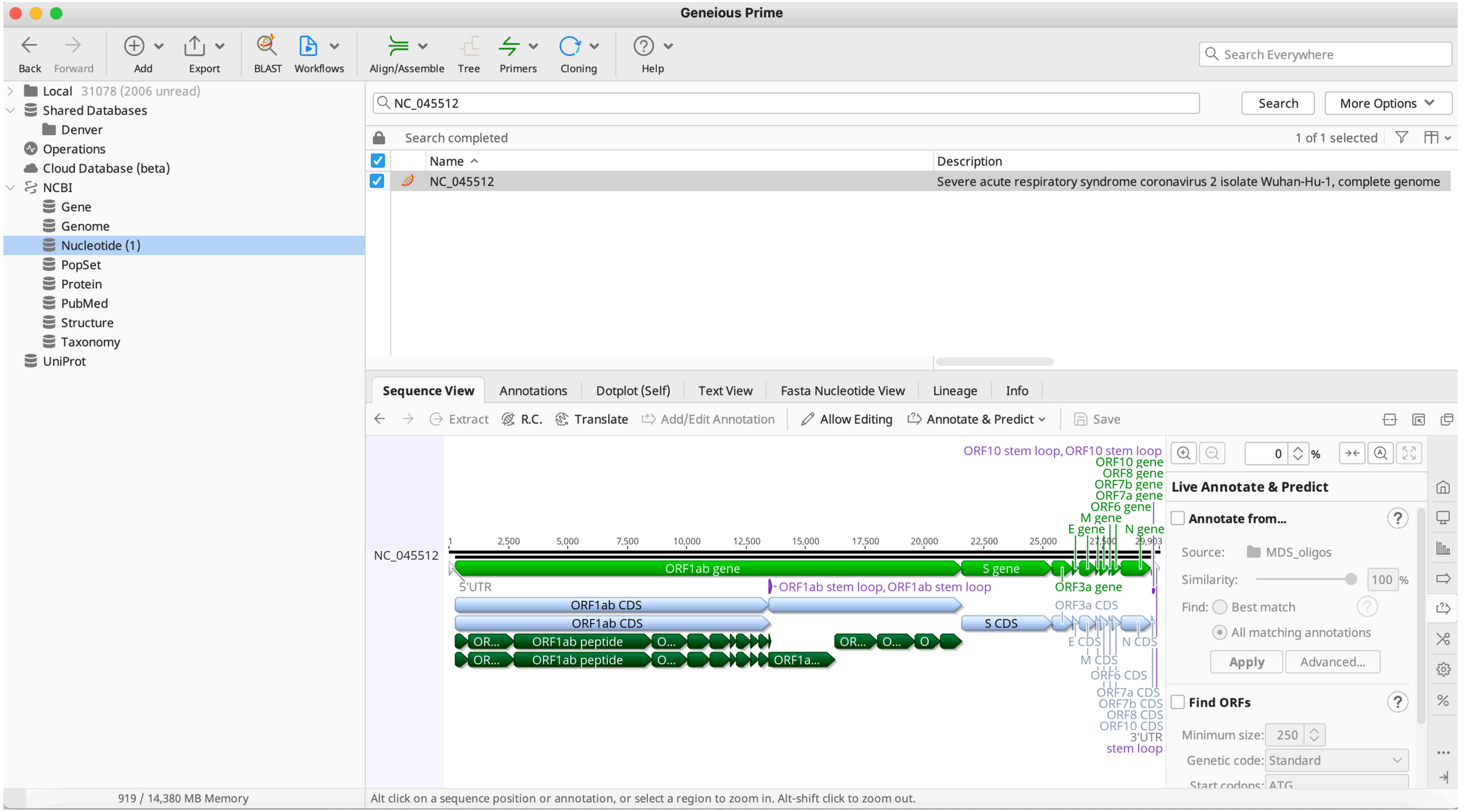




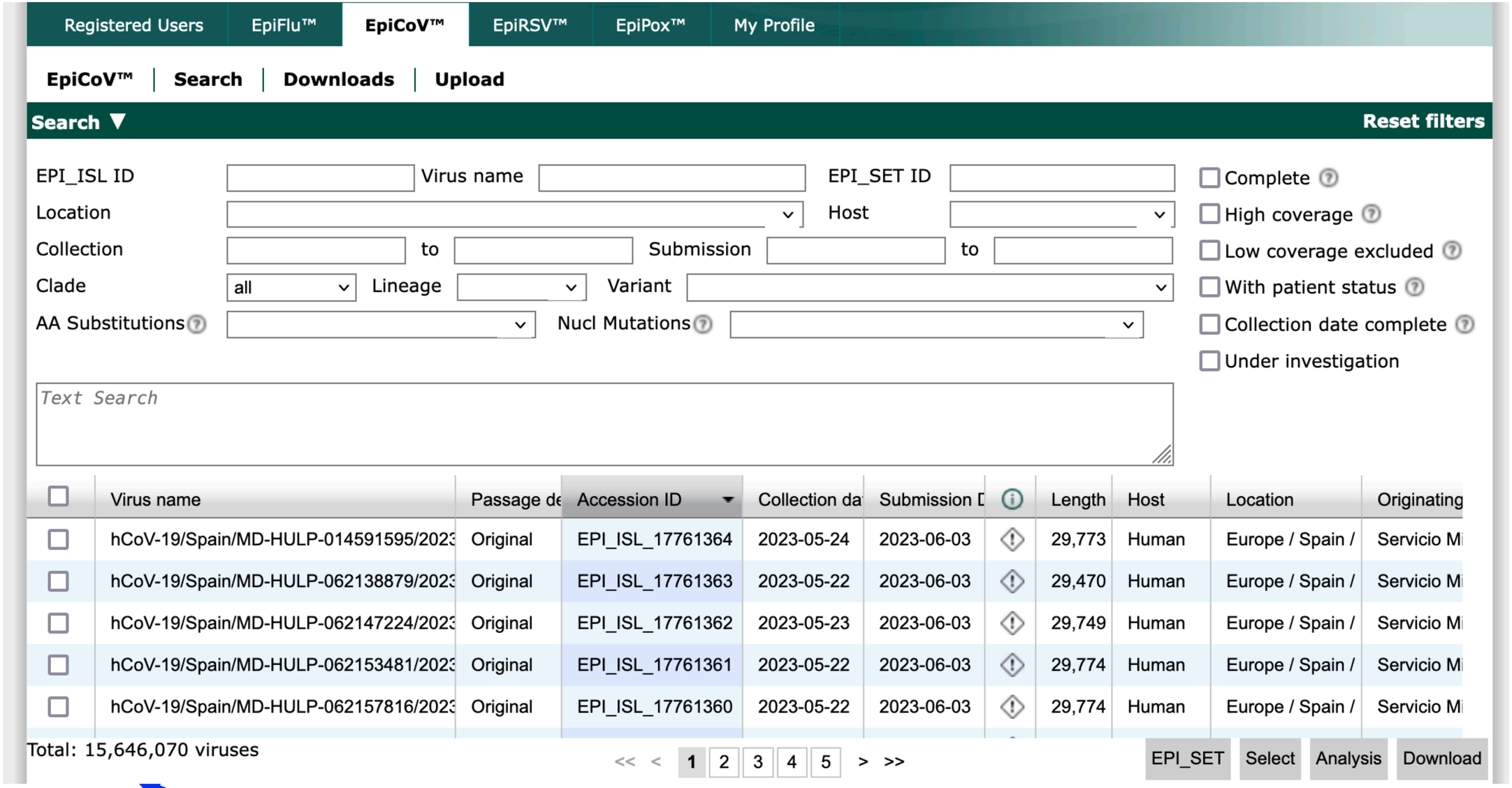
curl is a file transfer utility built into Linux, MacOS

similar utilities exist for Windows

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

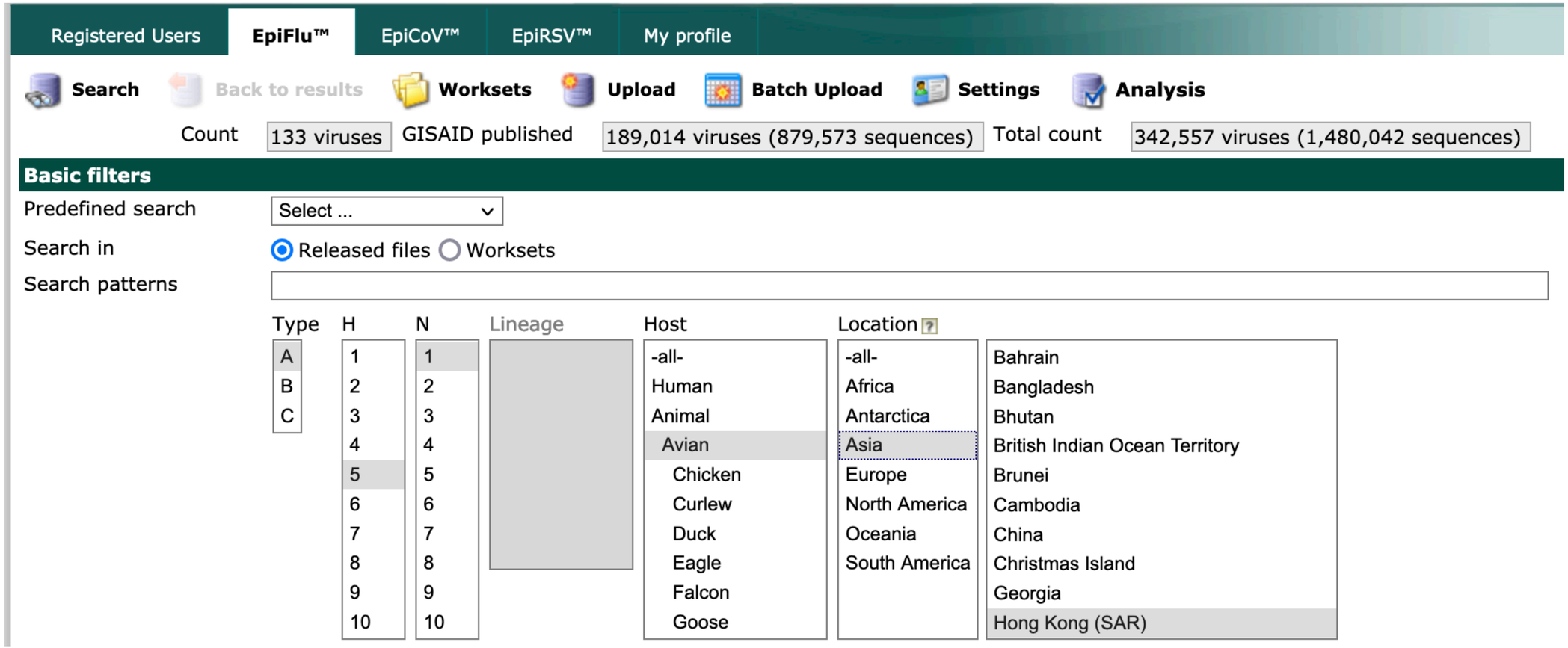


Non-NCBI databases include GISAID: a repository of virus sequences



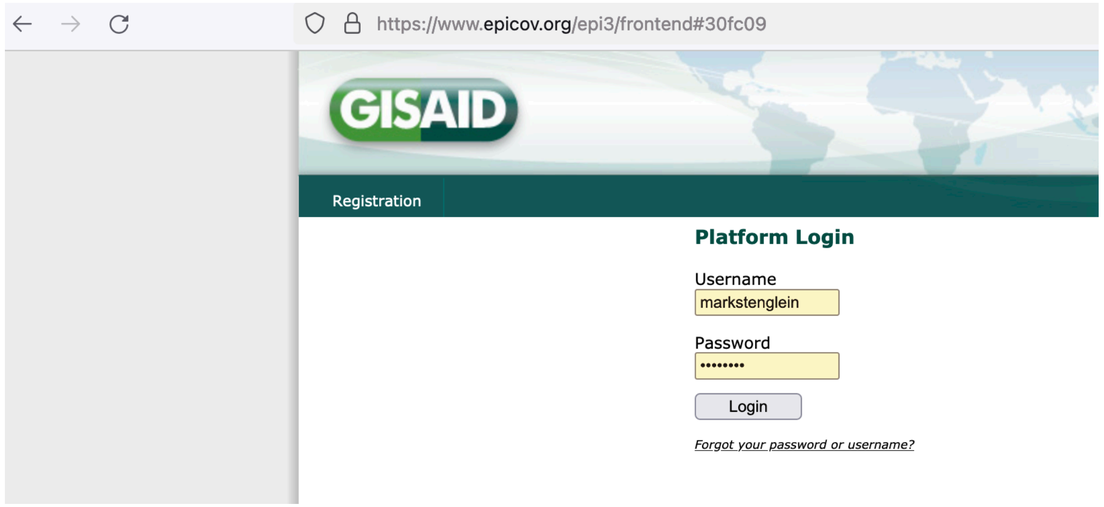
15 million SARS-CoV-2 sequences!! (~7M in Genbank)

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



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

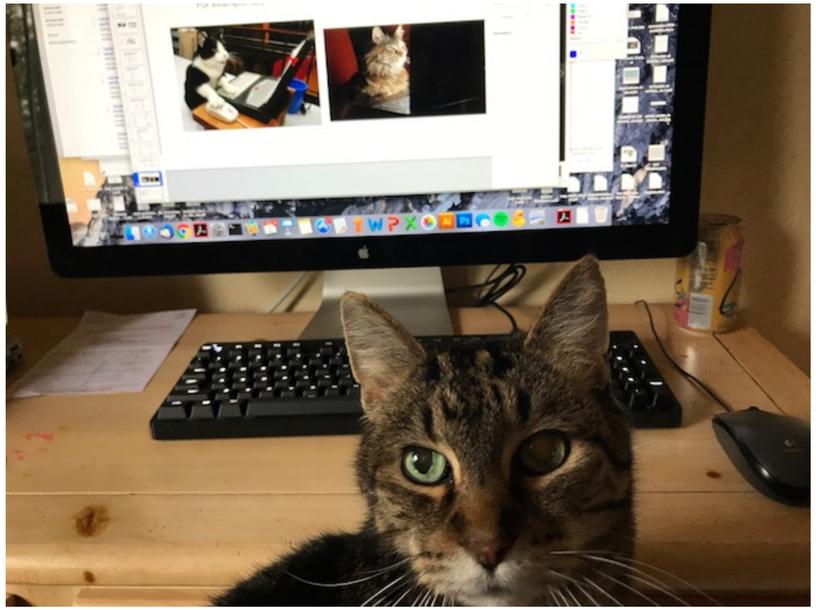
DOI 10.5524/100060

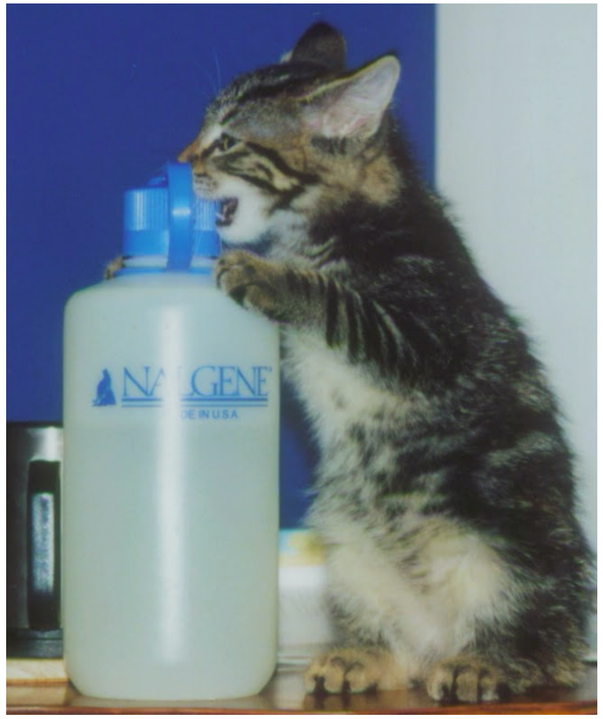
Table Settings

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).

Questions?





Kirby in 2000, wondering where his GenBank CDROMs are