NCBI Virus: Accessing sequence data and virus information

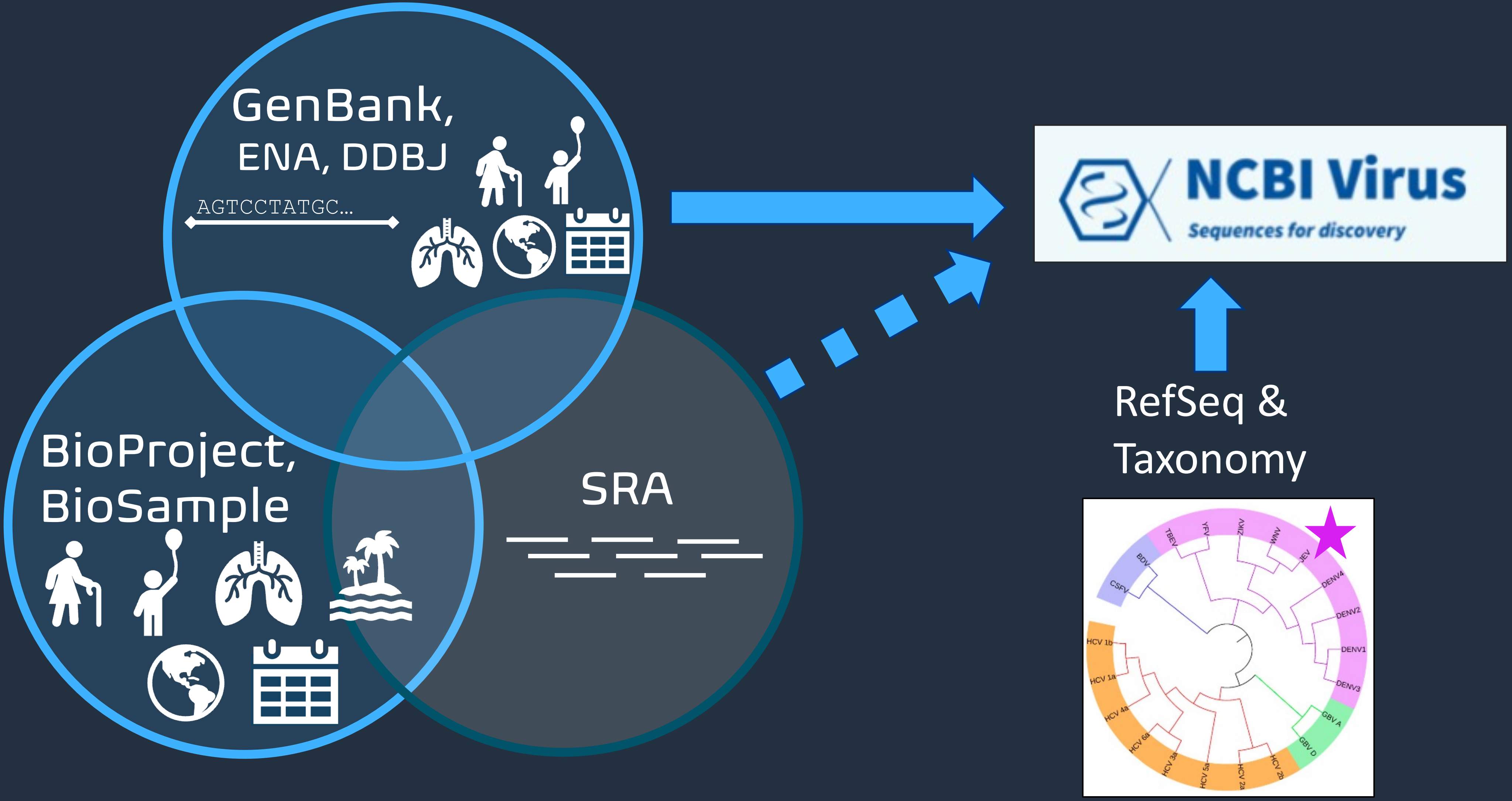
Eneida Hatcher, PhD

National Library of Medicine
National Center for Biotechnology Information

More data, more problems?

- How do you find data for viruses that have certain characteristics?
- From a geographic location, or collected recently?
- Which name do you search with?
- Part of a lineage or subtype?

NCBI Virus



GenBank, ENA, DDBJ

AGTCCTATGC...

BioProject, BioSample

SRA

NCBI Virus

Sequences for discovery

RefSeq & Taxonomy

NCBI Virus Curation

Submission to GenBank (or DDBJ or ENA)

serum

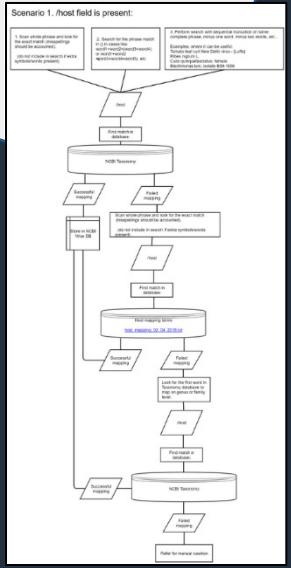
pbmc

plasma

plasma samples from hiv/hcv coinfected patients

blood

“blood”


Scenario 1. /host field is present:
1. Scan whole phrase and look for the exact match (misspellings should be accounted). (do not include in search if extra symbols/words present)
2. Search for the phrase match in () in cases like: word1+word2+(word3+word4) or Word1+word2 +word3+word4+word5), etc.
3. Perform search with sequential truncation of name: complete phrase, minus one word, minus two words, etc... Examples, where it can be useful: Tomato leaf curl New Delhi virus - [Luffa] Ribes nigrum L. Culex quinquefasciatus; female Blechmonas luni; isolate B39 1006
/host
Find match in database
NCBI Taxonomy
Successful mapping
Failed mapping
Store in NCBI Virus DB
Scan whole phrase and look for the exact match (misspellings should be accounted). (do not include in search if extra symbols/words present)
/host
Find match in database
Host mapping forms
host_mapping_02_04_2019.txt
Successful mapping
Failed mapping
Look for the first word in Taxonomy database to map on genus or family level.
/host
Find match in database
NCBI Taxonomy
Successful mapping
Failed mapping
Refer for manual curation

Host

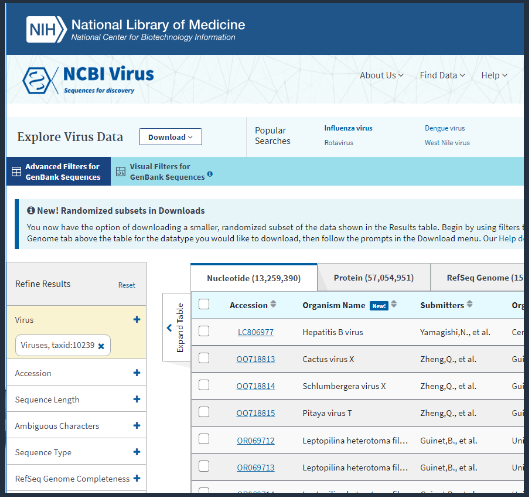
Sample location in host

Collection date

Geographic regions

Lab or vaccine strain

Environmental strain



National Library of Medicine
National Center for Biotechnology Information

NCBI Virus
Sequences for discovery

About Us Find Data Help

Explore Virus Data Download

Popular Searches: Influenza virus, Dengue virus, Rotavirus, West Nile virus

Advanced Filters for GenBank Sequences | Visual Filters for GenBank Sequences

New! Randomized subsets in Downloads

You now have the option of downloading a smaller, randomized subset of the data shown in the Results table. Begin by using filters t
Genome tab above the table for the datatype you would like to download, then follow the prompts in the Download menu. Our Help d

Refine Results Reset

- Virus +
  - Viruses, taxid:10239 ×
- Accession +
- Sequence Length +
- Ambiguous Characters +
- Sequence Type +
- RefSeq Genome Completeness +

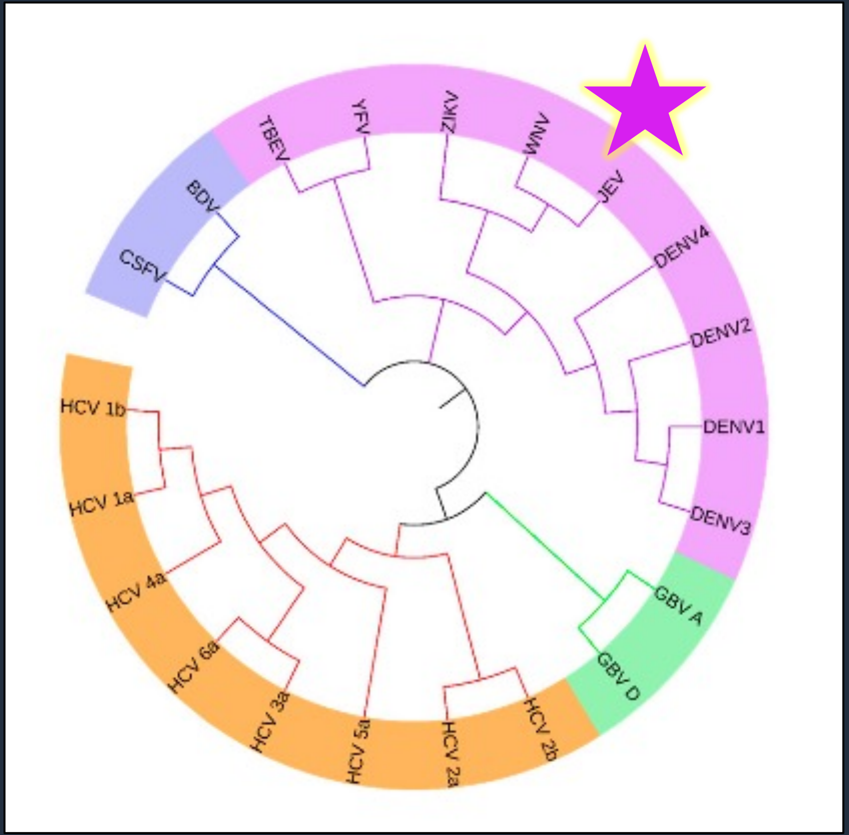
Nucleotide (13,259,390) | Protein (57,054,951) | RefSeq Genome (15

Expand Table

□	Accession	Organism Name New!	Submitters	Org
□	LC806977	Hepatitis B virus	Yamagishi,N., et al.	Cer
□	OQ718813	Cactus virus X	Zheng,Q., et al.	Gui
□	OQ718814	Schlumbergera virus X	Zheng,Q., et al.	Gui
□	OQ718815	Pitaya virus T	Zheng,Q., et al.	Gui
□	OR069712	Leptopilina heterotoma fil...	Guinet,B., et al.	Uni
□	OR069713	Leptopilina heterotoma fil...	Guinet,B., et al.	Uni

RefSeq records

- Representatives of a species or unclassified group
- Based on GenBank record
- Usually, 1 per species
- ICTV exemplars and “new” viruses
- Standard gene & protein annotation for well characterized viruses

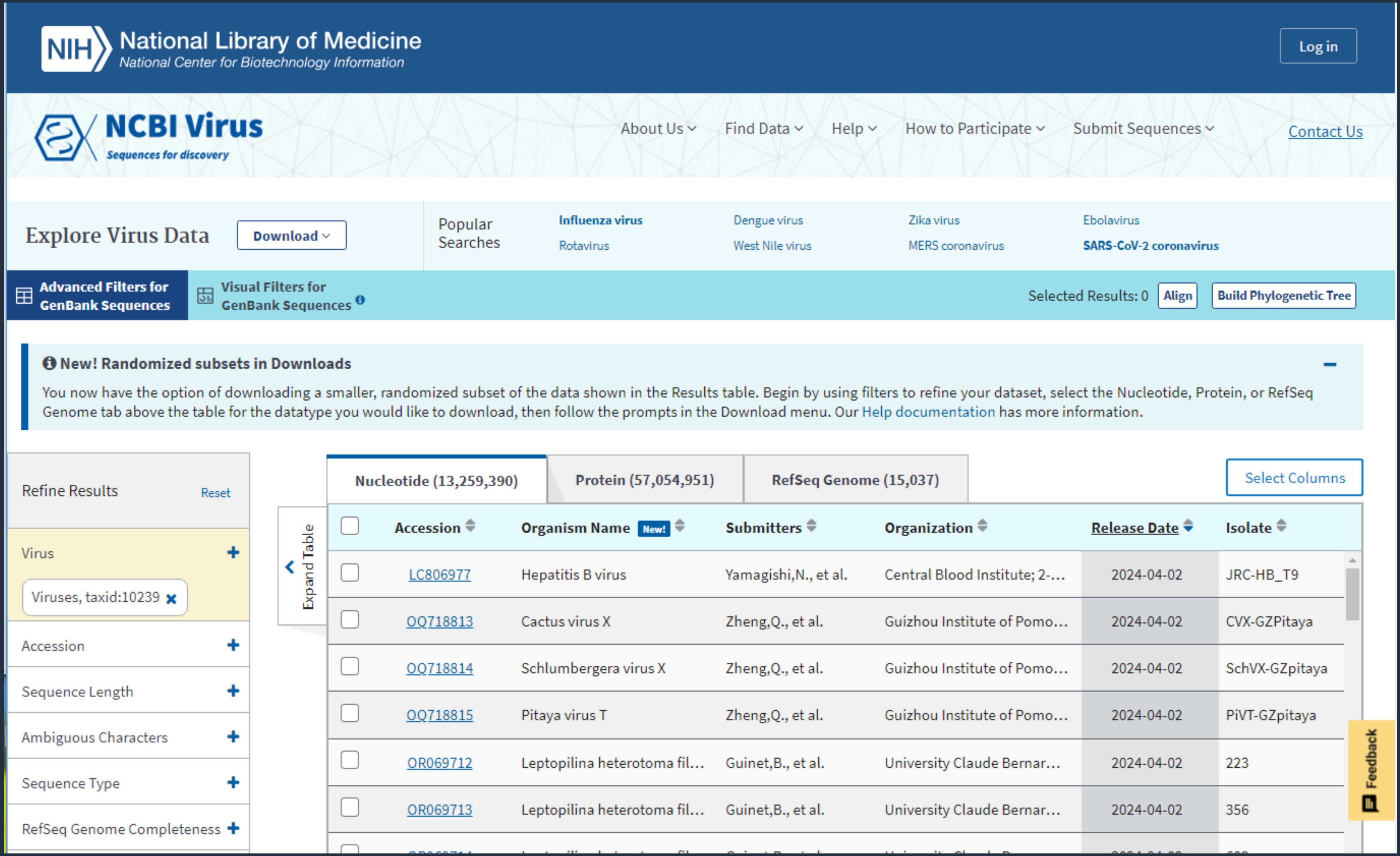


Role of RefSeqs in Public Health Response

- GenBank team + NCBI Virus team
- 24 hrs - RefSeq with improved annotation
- Standardized annotation – protein coding regions & names
- Demonstrated SARS-CoV-2 taxonomy
- RefSeq was used to create VADR models
    - Submissions are normally manually reviewed - Automated submission of sequences
- Provides widely available sequence coordinates, S:D614G

NCBI Virus Results Table

- Accession: GenBank, SRA, BioProject, BioSample
- Geographic region (North America; USA; GA)
- Submitting author, organization, location
- Several more



New! Randomized subsets in Downloads

You now have the option of downloading a smaller, randomized subset of the data shown in the Results table. Begin by using filters to refine your dataset, select the Nucleotide, Protein, or RefSeq Genome tab above the table for the datatype you would like to download, then follow the prompts in the Download menu. Our Help documentation has more information.

Nucleotide (13,259,390) | Protein (57,054,951) | RefSeq Genome (15,037)

□	Accession	Organism Name New!	Submitters	Organization	Release Date	Isolate
□	LC806977	Hepatitis B virus	Yamagishi,N., et al.	Central Blood Institute; 2-...	2024-04-02	JRC-HB_T9
□	OQ718813	Cactus virus X	Zheng,Q., et al.	Guizhou Institute of Pomo...	2024-04-02	CVX-GZPitaya
□	OQ718814	Schlumbergera virus X	Zheng,Q., et al.	Guizhou Institute of Pomo...	2024-04-02	SchVX-GZpitaya
□	OQ718815	Pitaya virus T	Zheng,Q., et al.	Guizhou Institute of Pomo...	2024-04-02	PiVT-GZpitaya
□	OR069712	Leptopilina heterotoma fil...	Guinet,B., et al.	University Claude Bernar...	2024-04-02	223
□	OR069713	Leptopilina heterotoma fil...	Guinet,B., et al.	University Claude Bernar...	2024-04-02	356

https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202%20(SARS-CoV-2),%20taxid:2697049, Just search "NCBI Virus" ☺

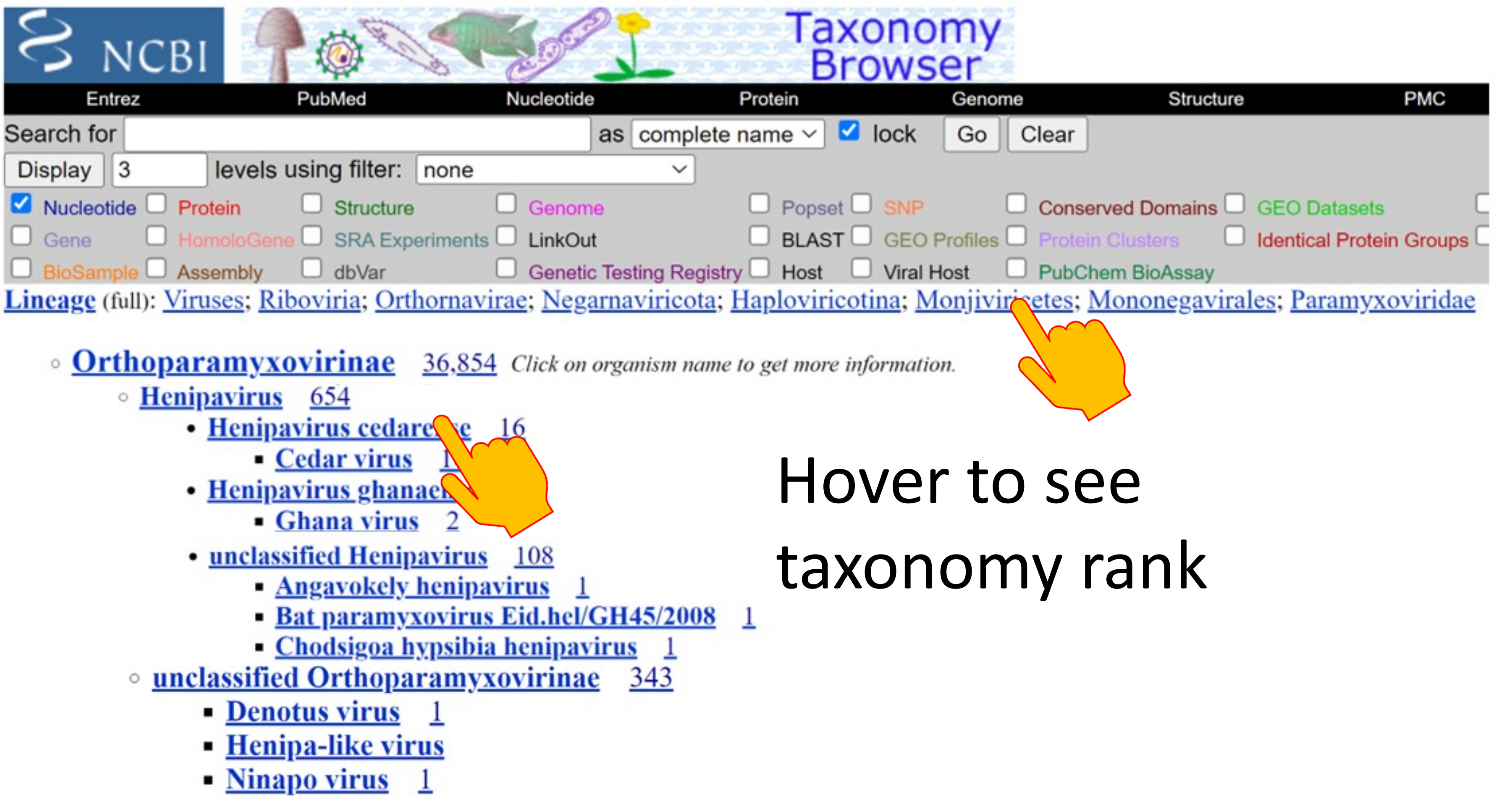
Pango Lineages in NCBI Virus

- Automated tool ☺
- Each evening: Check for updates to the tools, Run on all SARS-CoV-2 sequences
- Filter by lineage
- Include lineage in Results Table download, or define in downloads
  - Versions are also included in Results Table download:
  - 4.3.1/1.26/v0.1.12/0.3.19/0.6.2
  - pangolin 4.3.1 / pangolin-data 1.24 / constellations v0.1.12 / scorpio 0.3.19 / UShER 0.6.2

NCBI Taxonomy

- Virus species: ICTV
  - Gives us a formalized way to talk about virus groups
- Virus name, common names: Nipah virus, SARS-CoV-2
- When no official taxonomy is available yet, NCBI places sequence records into “unclassified” bins
- NCBI taxonomy also includes synonyms, virus names, common names, etc. to make data searches easier
  - Nipah-virus, 2019-nCoV, Newcastle disease virus
- Includes nodes below the species level

NCBI Taxonomy Bowser



Search for [ ] as complete name ☑ lock

Display 3 levels using filter: none

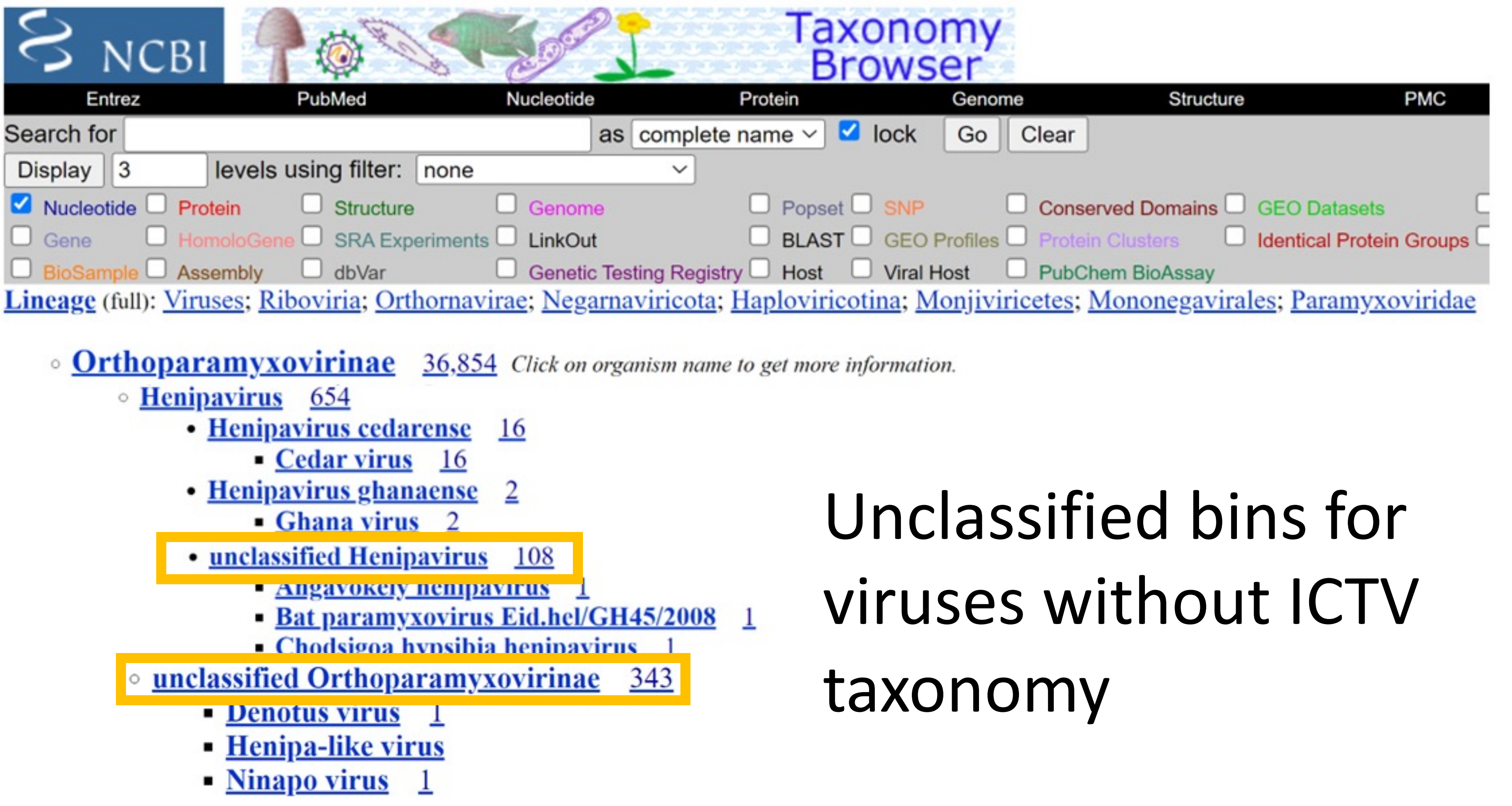
☑ Nucleotide □ Protein □ Structure □ Genome □ Popset □ SNP □ Conserved Domains □ GEO Datasets □
□ Gene □ HomoloGene □ SRA Experiments □ LinkOut □ BLAST □ GEO Profiles □ Protein Clusters □ Identical Protein Groups □
□ BioSample □ Assembly □ dbVar □ Genetic Testing Registry □ Host □ Viral Host □ PubChem BioAssay

Lineage (full): Viruses; Riboviria; Orthornavirae; Negarnaviricota; Haploviricotina; Monjiviricetes; Mononegavirales; Paramyxoviridae

- **Orthoparamyxovirinae** 36,854 *Click on organism name to get more information.*
  - **Henipavirus** 654
    - **Henipavirus cedarense** 16
      - **Cedar virus** 1
    - **Henipavirus ghanaense**
      - **Ghana virus** 2
    - **unclassified Henipavirus** 108
      - **Angavokely henipavirus** 1
      - **Bat paramyxovirus Eid.hel/GH45/2008** 1
      - **Chodsigoa hypsibia henipavirus** 1
  - **unclassified Orthoparamyxovirinae** 343
    - **Denotus virus** 1
    - **Henipa-like virus**
    - **Ninapo virus** 1

Hover to see taxonomy rank

NCBI Taxonomy Bowser



Search for [ ] as complete name ☑ lock Go Clear

Display 3 levels using filter: none

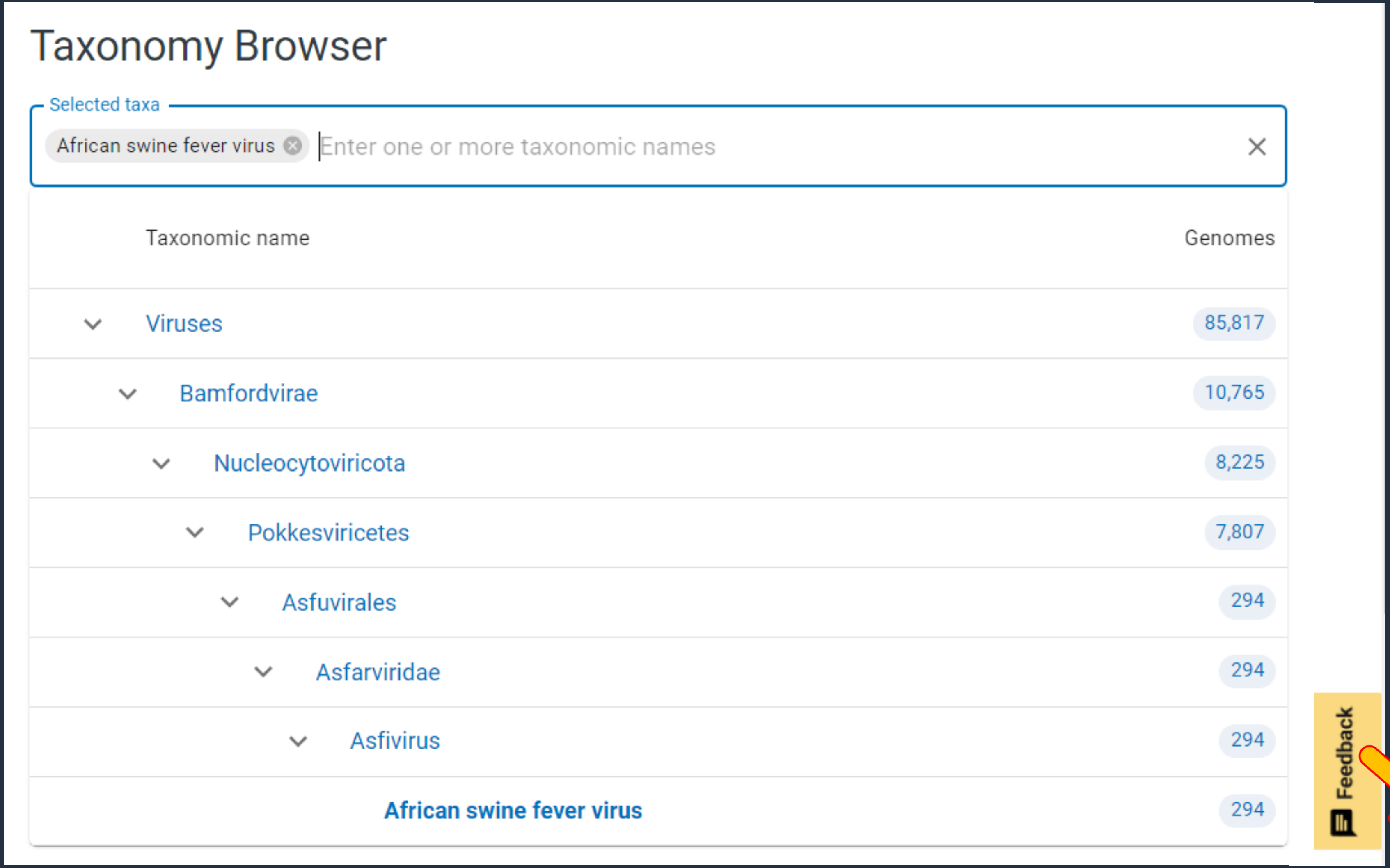
☑ Nucleotide □ Protein □ Structure □ Genome □ Popset □ SNP □ Conserved Domains □ GEO Datasets □
□ Gene □ HomoloGene □ SRA Experiments □ LinkOut □ BLAST □ GEO Profiles □ Protein Clusters □ Identical Protein Groups □
□ BioSample □ Assembly □ dbVar □ Genetic Testing Registry □ Host □ Viral Host □ PubChem BioAssay

Lineage (full): Viruses; Riboviria; Orthornavirae; Negarnaviricota; Haploviricotina; Monjiviricetes; Mononegavirales; Paramyxoviridae

- **Orthoparamyxovirinae** 36,854 *Click on organism name to get more information.*
    - **Henipavirus** 654
        - **Henipavirus cedarense** 16
            - **Cedar virus** 16
        - **Henipavirus ghanaense** 2
            - **Ghana virus** 2
        - **unclassified Henipavirus** 108
            - **Angavokely henipavirus** 1
            - **Bat paramyxovirus Eid.hel/GH45/2008** 1
            - **Chodsigoa hypsibia henipavirus** 1
    - **unclassified Orthoparamyxovirinae** 343
        - **Denotus virus** 1
        - **Henipa-like virus**
        - **Ninapo virus** 1

Unclassified bins for viruses without ICTV taxonomy

NCBI Taxonomy Bowser – New!



Taxonomy Browser

Selected taxa

African swine fever virus ⊗ Enter one or more taxonomic names ✕

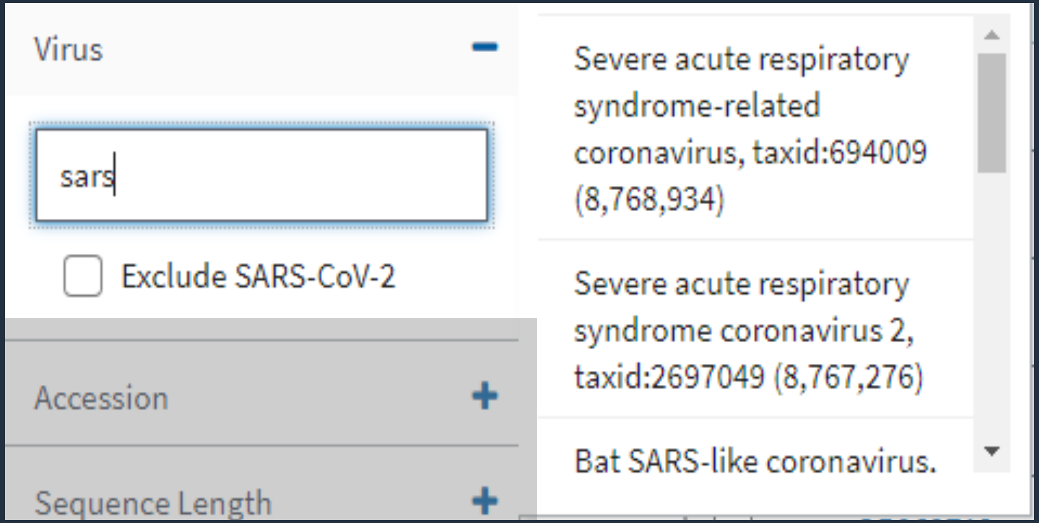
Taxonomic name	Genomes
Viruses	85,817
Bamfordvirae	10,765
Nucleocytoviricota	8,225
Pokkesviricetes	7,807
Asfuvirales	294
Asfarviridae	294
Asfivirus	294
African swine fever virus	294

Search “NCBI Datasets” → Taxonomy

Genomes ≠ Nucleotide records

Please share your feedback

NCBI Virus = NCBI Taxonomy



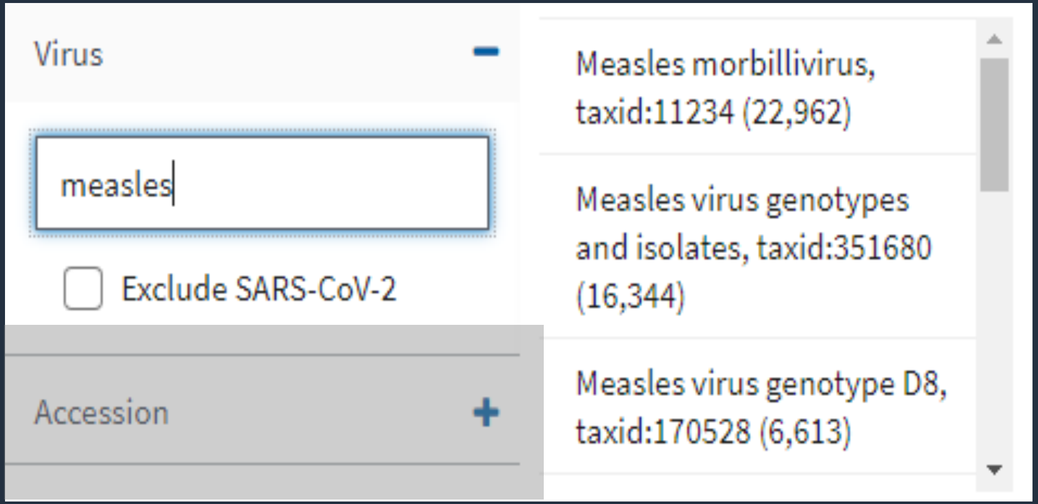
sars

□ Exclude SARS-CoV-2

Severe acute respiratory syndrome-related coronavirus, taxid:694009 (8,768,934)

Severe acute respiratory syndrome coronavirus 2, taxid:2697049 (8,767,276)

Bat SARS-like coronavirus.



measles

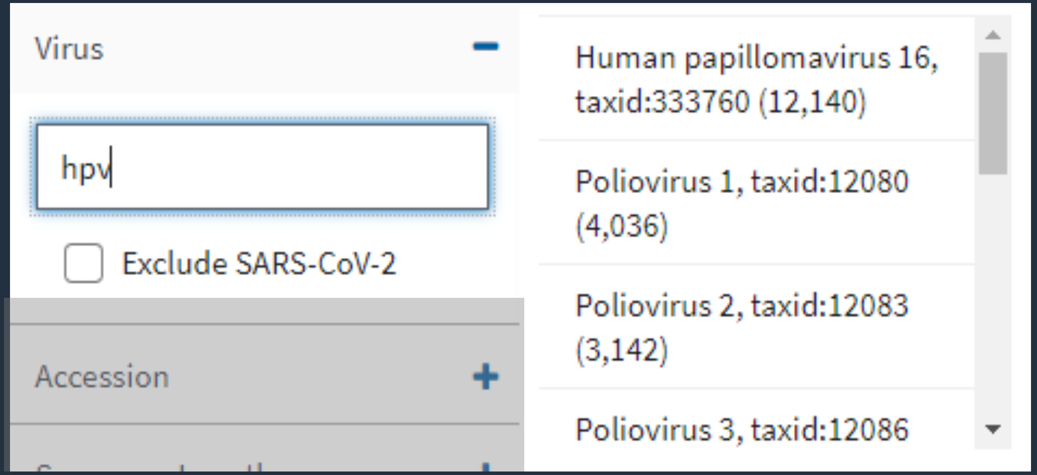
□ Exclude SARS-CoV-2

Measles morbillivirus, taxid:11234 (22,962)

Measles virus genotypes and isolates, taxid:351680 (16,344)

Measles virus genotype D8, taxid:170528 (6,613)

Search with <u>any</u> level of taxonomy, common name, many abbreviations, etc



hpv

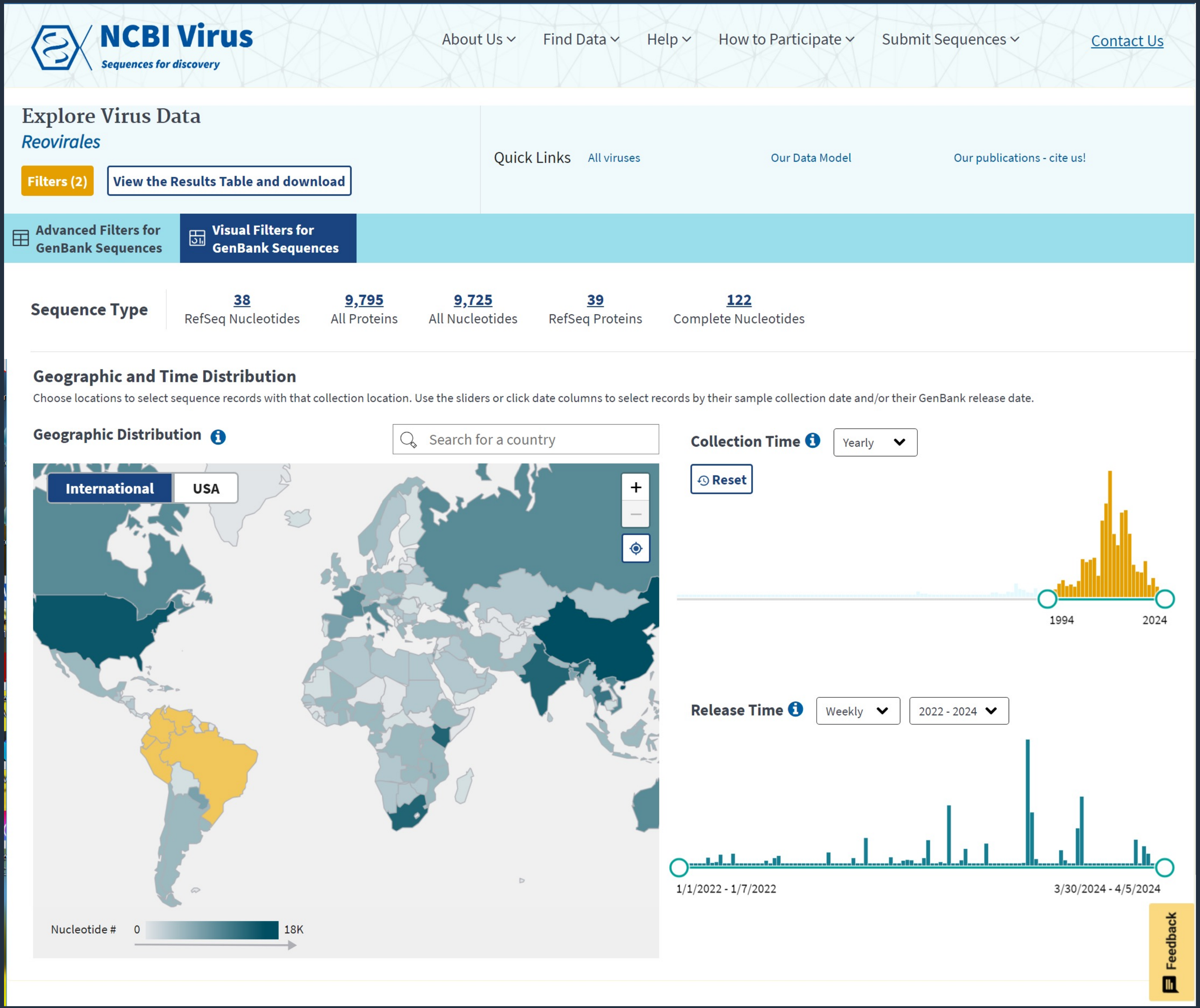
□ Exclude SARS-CoV-2

Human papillomavirus 16, taxid:333760 (12,140)

Poliovirus 1, taxid:12080 (4,036)

Poliovirus 2, taxid:12083 (3,142)

Poliovirus 3, taxid:12086



Explore Virus Data

Reovirales

Filters (2)
View the Results Table and download
Quick Links
All viruses
Our Data Model
Our publications - cite us!
Advanced Filters for GenBank Sequences
Visual Filters for GenBank Sequences

Sequence Type

38	9,795	9,725	39	122
RefSeq Nucleotides	All Proteins	All Nucleotides	RefSeq Proteins	Complete Nucleotides

Geographic and Time Distribution

Choose locations to select sequence records with that collection location. Use the sliders or click date columns to select records by their sample collection date and/or their GenBank release date.

Interactive dashboard for exploring sequence data, understanding biases, and refining selections

Flexible download formats

FASTA: nucleotide, coding regions, or proteins

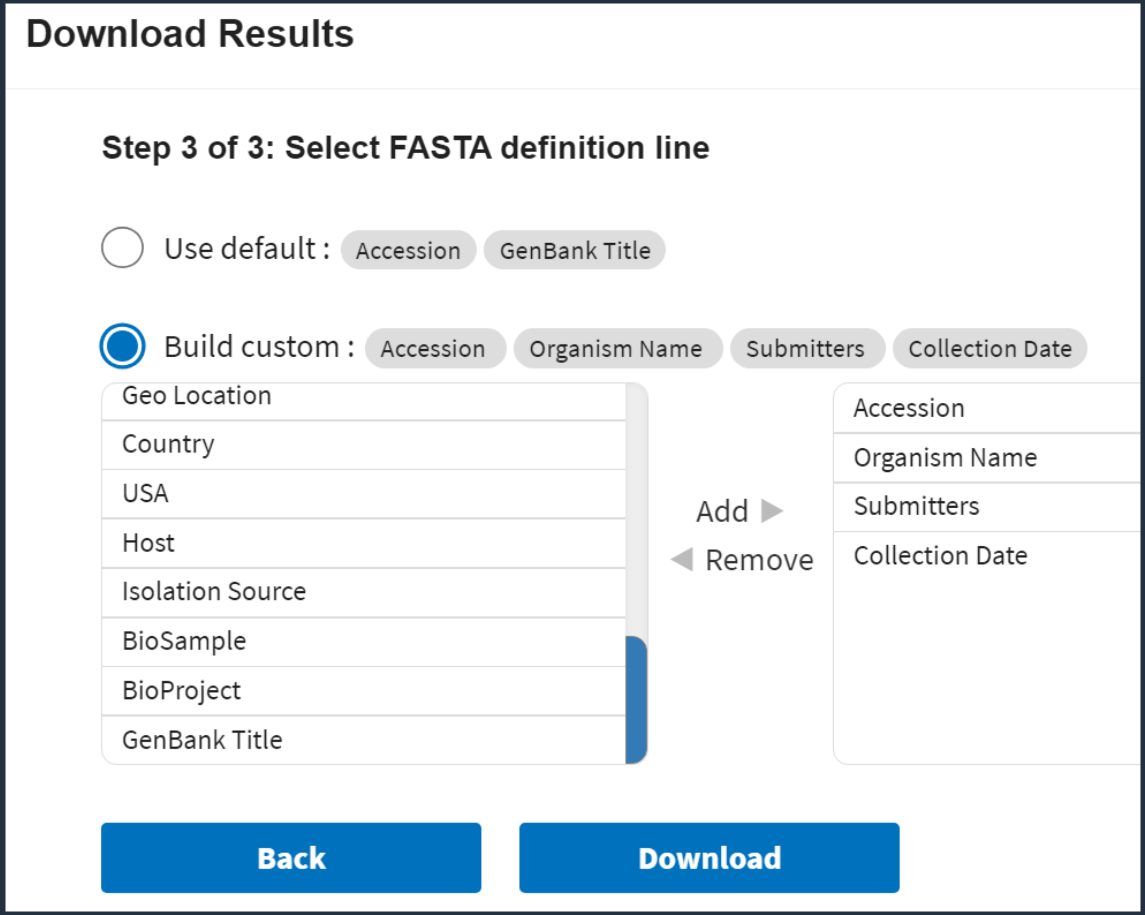
- Customize the defline/header

Accessions lists

Results table

- Choose which fields to include in the table

Randomized subsets, stratified by country, collection year, release year, or host



Download Results

Step 3 of 3: Select FASTA definition line

○ Use default : Accession GenBank Title

● Build custom : Accession Organism Name Submitters Collection Date

Geo Location
Country
USA
Host
Isolation Source
BioSample
BioProject
GenBank Title

Accession
Organism Name
Submitters
Collection Date

NCBI Datasets – Command line or API

- Programmatically access virus data
- Nucleotide & Protein sequences, annotation, and metadata report
- Same normalized metadata as NCBI Virus
- Great for large data downloads
- https://www.ncbi.nlm.nih.gov/datasets/docs/how-tos/virus/

Help us help you.

- Lineages can change over time – but data provided during submission is archival
- NCBI Virus can associate sequence records with up-to-date lineages
- & help distribute lineage references
- We need:
  - Community-accepted classification schemes
  - Automated classification tools with long-term support

Acknowledgements

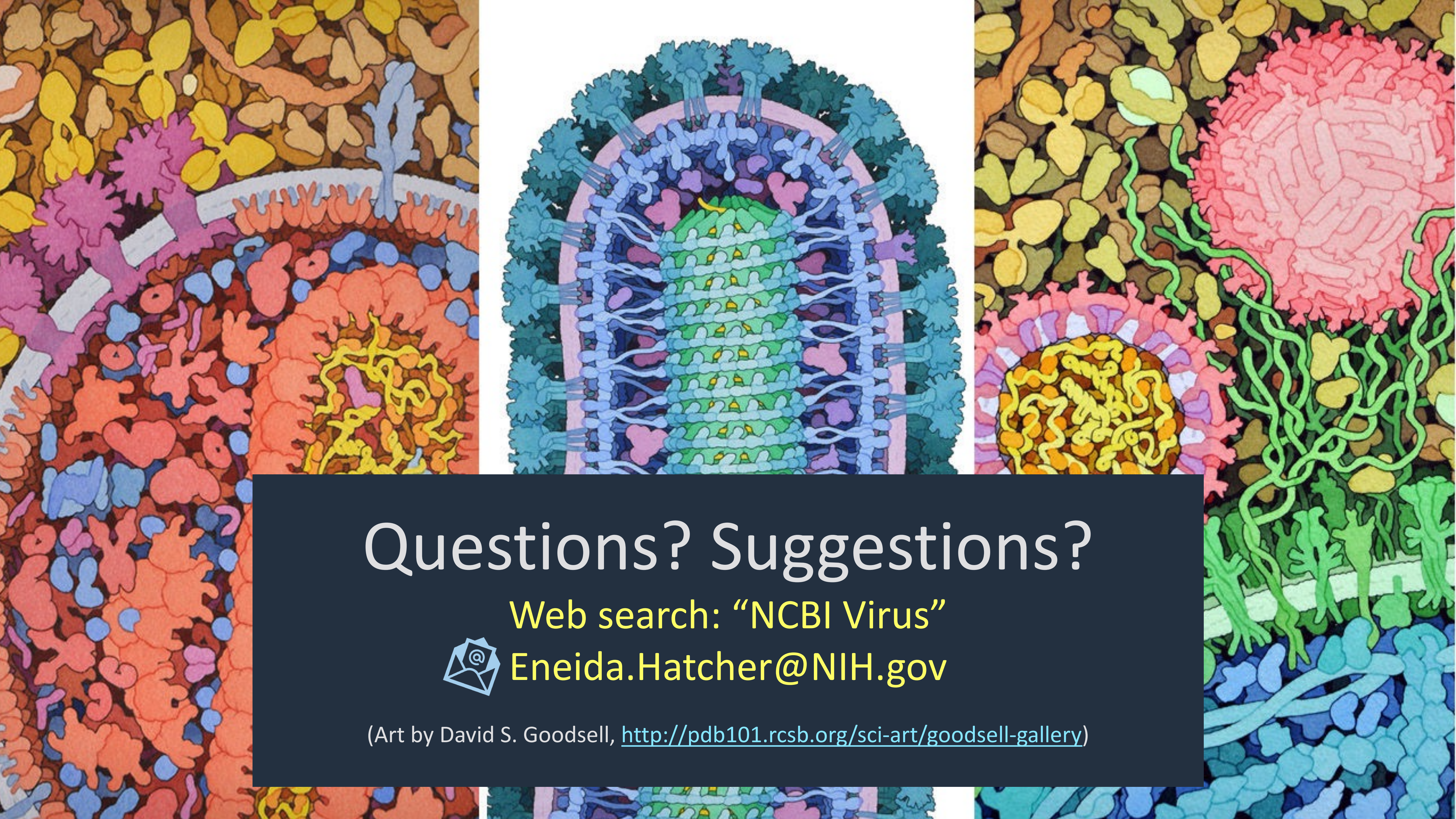
NCBI Virus team

Olga Blinkova
Brett Spurrier
David Kristensen
Yuri Ostapchuk
Sergey Resenchuk
Igor Tolstoy
Anna Glodek
Ravinder Eskandary
Ryan Connor – former
J. Rodney Brister - former

NCBI & NLM colleagues

Everyone who has contributed to the shared public data

This work was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health



Questions? Suggestions?

Web search: “NCBI Virus”

Eneida.Hatcher@NIH.gov

(Art by David S. Goodsell, http://pdb101.rcsb.org/sci-art/goodsell-gallery)