4th IABS NGS Conference
December 4th, 2024
Frankfurt, Germany

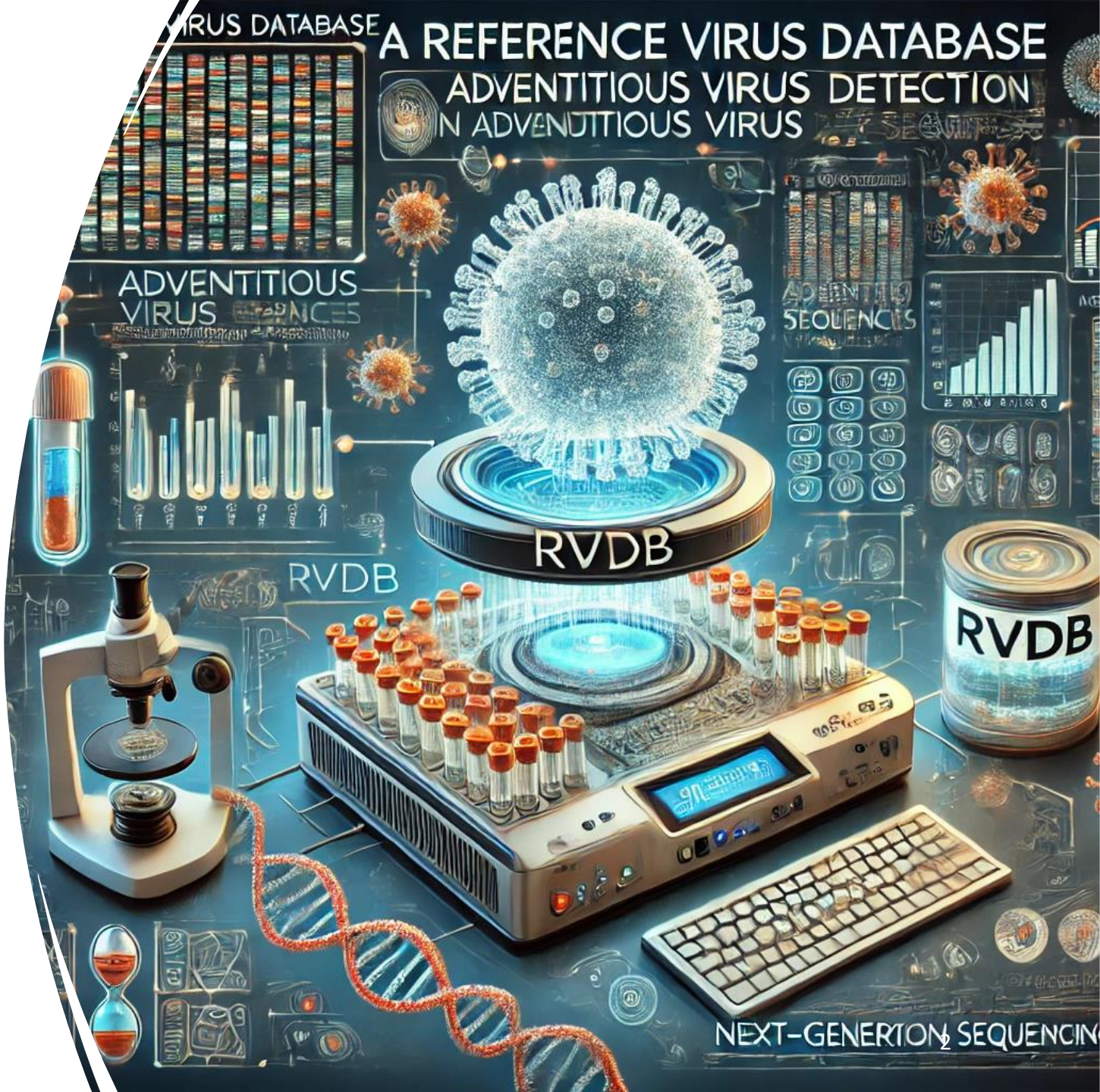
Refinement Efforts on CBER's Reference Virus Database (RVDB) to Enhance Accuracy and Specificity of Virus Detection

Pei-Ju Chin, M.S., Ph.D.
Trent J. Bosma, Ph.D.
Arifa S. Khan, Ph.D.

Division of Viral Product
Office of Vaccine Research and Review
Center for Biologics Evaluation and Research
U.S. Food and Drug Administration

Outline

- Limitations of public nucleotide databases and the need for a refined and reference viral database
- Workflow for development and updating of RVDB
- Challenges and RVDB refinements
- Machine Learning to identify non-viral and irrelevant sequences for RVDB refinement



Background: Need for a New Reference Viral DataBase (RVDB)

Some limitations in public databases

- **NCBI Viral Genomes Resource (RefSeq and Neighbors)**
  - Majority are complete viral genomes, annotated
    - *Other viruses with only partial genomes and ERV sequences are under-represented*
  - Limited in sequence diversity
    - *One record per virus species for RefSeq (some exceptions) and diversity in neighbors more, but not all inclusive*

- **NR/NT (non-redundant nucleotide collection)**
  - Includes additional viral diversity: full viral genomes and partial viral sequences
  - Also, includes abundance of cellular sequences and uncharacterized sequences
    - Large number of non-viral sequences can "bury" detection of viral signal
    - Many mis-annotated sequences (cellular sequence is annotated as viral and vice versa)

- **NCBI Protein database (non-redundant protein sequences)**
  - Only contains complete protein sequences
    - Partial virus sequences that cannot encode proteins are not represented

Reference Virus Database (RVDB) History and Current Status

- Work initiated in mid-2013

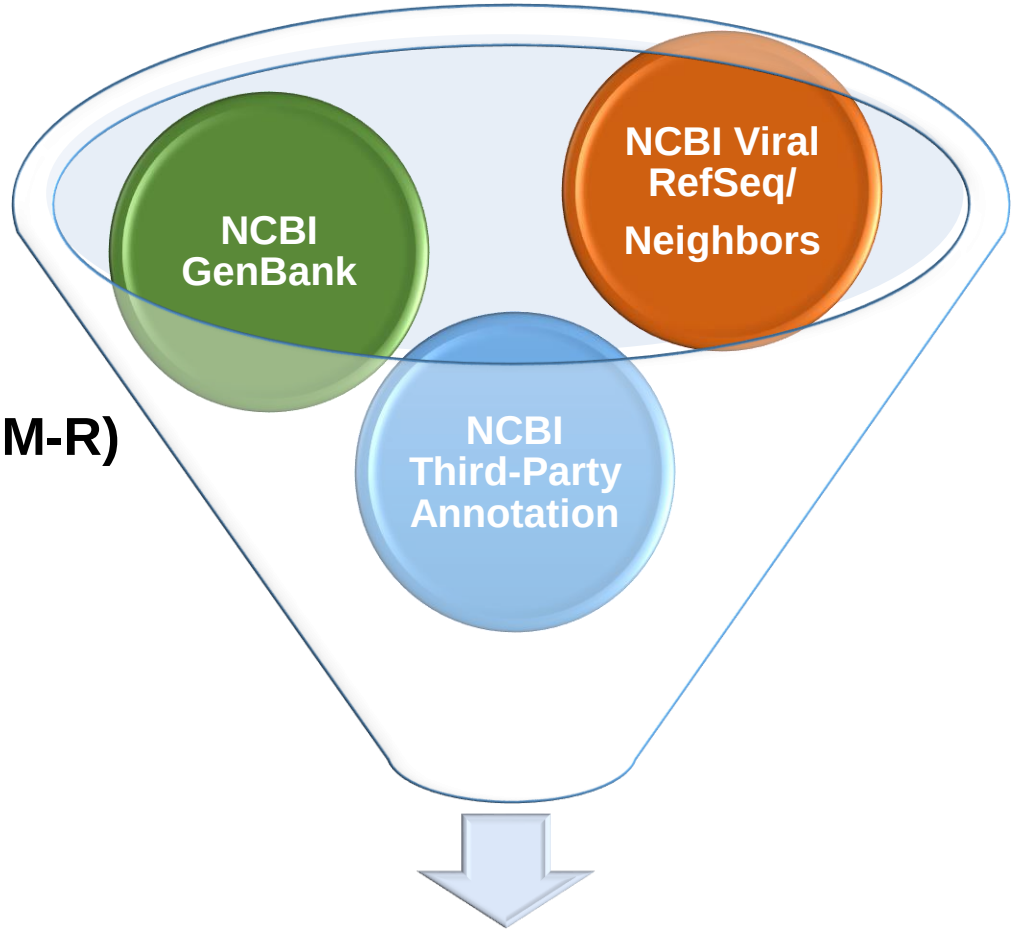
  In-house lab efforts in consultation with the AVDTWG and NCBI (Rodney Brister) resulted in development of a viral database that employed semantic data mining of various selected GenBank divisions with unique features:

  - Included all viral, viral-related, and viral-like sequences, including endogenous viruses and retroelements
  - **Had reduced non-viral/cellular content to improve virus detection specificity by Next-generation Sequencing (NGS) technologies**
  - **Initial RVDB work was published (Goodacre et. al. 2018. mSphere 3:10.1128)**

Production Workflow for Reference Virus Database (RVDB)

Semantic-Refine (SEM-R) Filter

"rRNA"
"Phage"
"Receptor"
"Cytochrome"
"non autonomous"
"Nuclear envelope"
"endogenous tripeptide"
…….≈700 Negative Keywords



Candidate Sequences

NCBI GenBank

Environmental Sampling (ENV)
HTS Sampling(HTC)
Invertebrate (INV)
Mammalian (MAM)
Plant (PLN)
Primate (PRI)
Rodent (ROD)
Viral (VRL)
Vertebrate (VRT)
Third-party Annotation (TPA)
~~Phage (PHG)~~

Crosstalk to sequencing adapters and vectors

RVDB Provides 4 Formats to Adapt Various Application Scenarios

- U-RVDB *fasta* file
  - Un-clustered, contain all viral sequences with redundancy
  - Higher computation-demanded. **Suitable for virus detection by *blastn/nhmmer***
- C-RVDB *fasta* file
  - Clustered, sequences share 98% similarity are collapse to one representative sequence for each clade
  - Lower computation-demanded. **Suitable for virus detection by *tblastx***
- SQLite DB Script
  - Create the entries (*fasta* header and the corresponding information) for advanced bioinformatic pipelines/workflows
- Proteic RVDB (provided by Institut Pasteur: https://rvdb-prot.pasteur.fr/)
  - **Hidden Markov Model (HMM)** profile of viral protein domains
  - Unknown viruses with remote homology by ***hmmsearch*** / ***hmmscan***

RVDB Version 29.0

RVDB

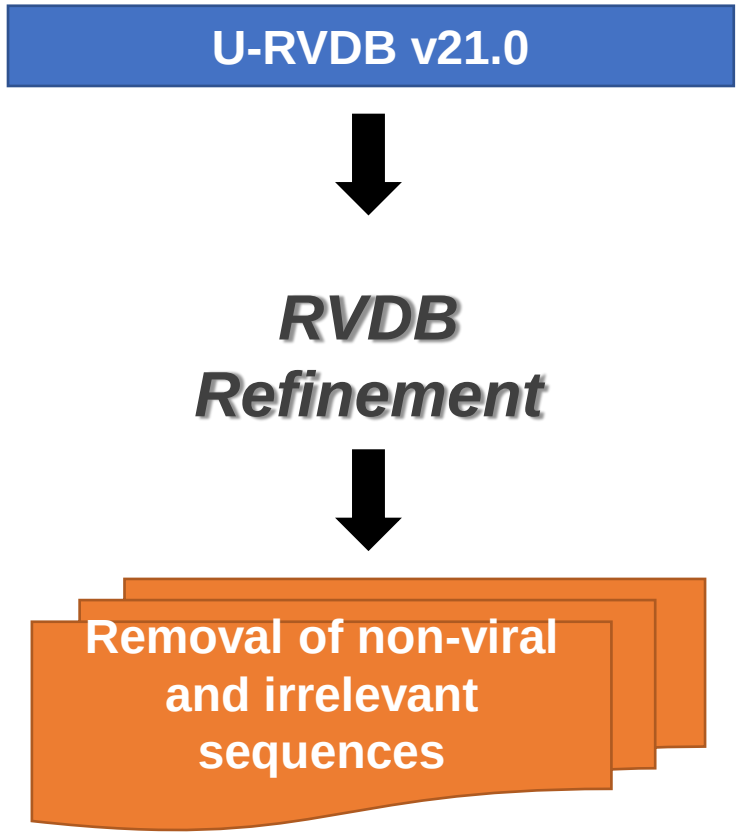
Proteic-RVDB

- Released in July 2024
- Based on GenBank May 2024, release version 260 and RefSeq May 2024, release version 224
- Number of sequences
  - Un-Clustered: 10,000,605 (*fasta and SQL*)
  - Clustered: 1,173,482 (*fasta*)
- **5,782,653sequences of SARS-CoV-2 (5,473,308 in v28.0)**
- Available at UNIV. DELAWARE RVDB site (https://rvdb.dbi.udel.edu/)
- Corresponding Proteic Databases (Proteic RVDB) were generated by Marc Eloit and Thomas Bigot, Institut Pasteur (https://rvdb-prot.pasteur.fr/)

RVDB Refinement: Viral and Non-Viral Sequences Annotation

- The collection of RVDB sequences is subset directly from NCBI GenBank, viral RefSeq and TPA **without any modifications.** Therefore, the potential issues are inherited
  - Poor quality of sequence (e.g. Poly Ns in SARS-CoV-2 sequences)
  - Sequencing vector carryover (vector/adapter/linker/primer…etc)
  - Mis-annotation of non-viral sequences as viral, and *vice versa*
  - Flanking host sequences associated with endogenous retroviruses
- Pipelines and strategies for overcoming these issues were developed

RVDB Refinement to Improve Specificity for Virus Detection



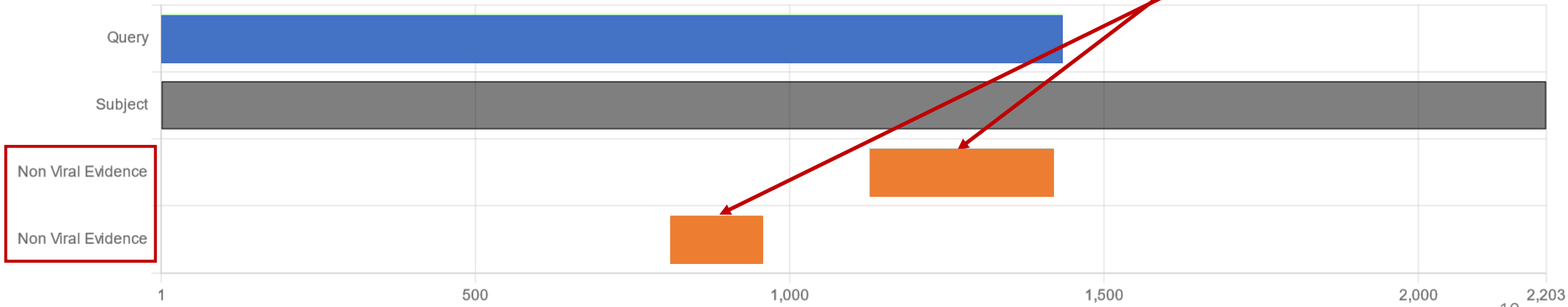
Removal Category	Sequence Group
Irrelevant Viral Sequences	**Untranslated Regions (UTRs)**
	Satellite sequences/Satellite viruses[#]
	Phages/Viroids/Virophages[#]
	No virus defined (environmental)
Mis-annotated Viral Sequences	**Stealth virus**
	Non-viral by manual curation
	"TY" and "Pol" used to name clone
	Host/Cellular IAP

[#]Satellite viruses and virophages were decided to keep according to the discussion of AVDTWG Subgroup C meeting in May 2021

RVDB Non-Viral Annotation Strategy for Improving Virus Detection Accuracy

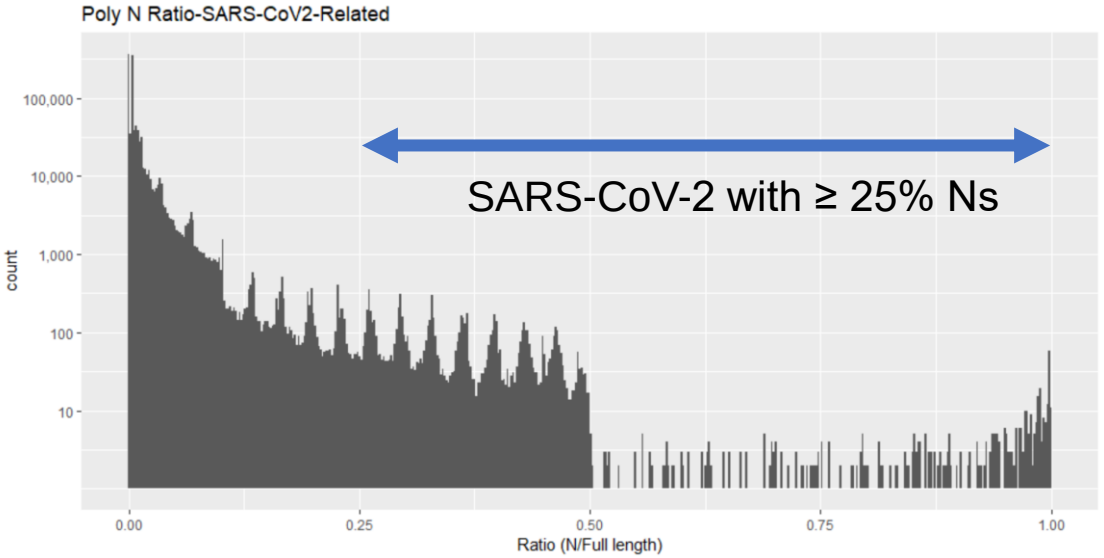
- **Automatic annotation pipeline to indicate non-viral regions in RVDB [**Trent Bosma (CBER), Madolyn MacDonald (UDEL)**]**
- RVDB is used for NCBI BLASTn against in-house, knowledge-based non-viral dataset (rRNAs, mitochondria, protozoa, bacteria, and phages)
- The accessions and coordinates of non-viral regions are provided at RVDB website
  - Unexpected hit follow-ups (resulting from the cross reactivity of non-viral segments)
  - Database masking to reduce false positive hits

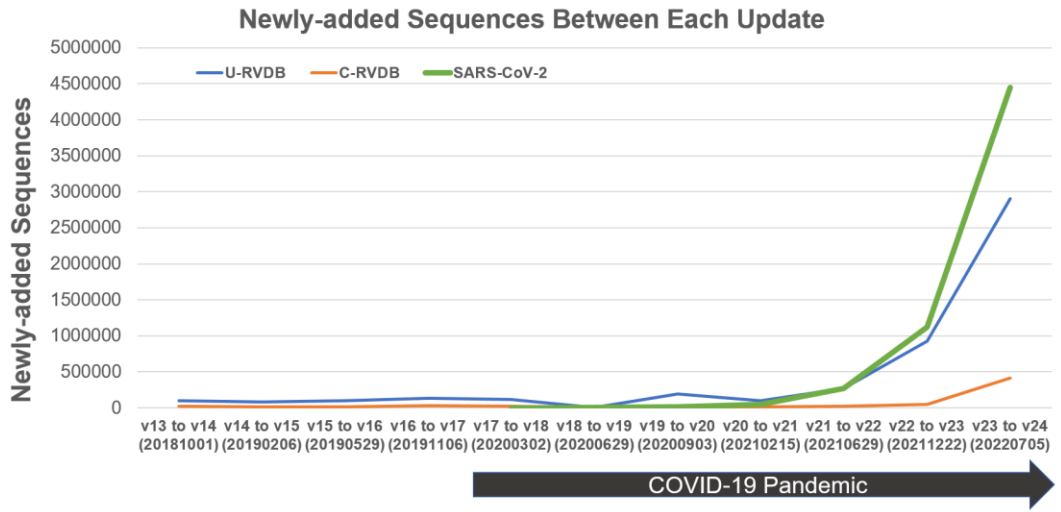
Accession	Header Info	Start	End	Nonviral Hit Category
OQ752844.1	acc|GENBANK|OQ752844.1|HIV-1 isolate BC001_proviral_073_Chimera from Canada defective provirus genomic sequence|Human immunodeficiency virus 1|VRL|20-OCT-2023	1126	1419	mitochondria
OQ752844.1	acc|GENBANK|OQ752844.1|HIV-1 isolate BC001_proviral_073_Chimera from Canada defective provirus genomic sequence|Human immunodeficiency virus 1|VRL|20-OCT-2023	810	958	mitochondria



RVDB Refinements

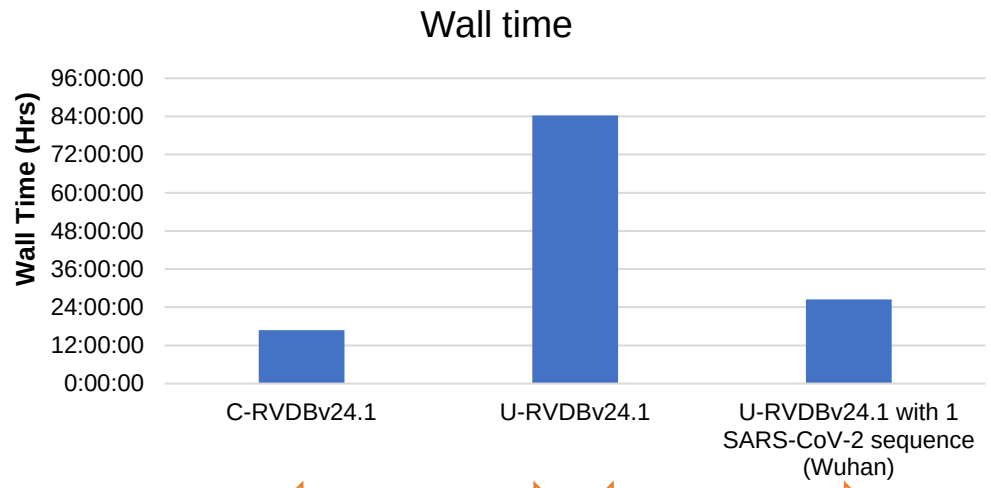
- **"Behind the scene" improvements**
  - Transition to Python 3 scripts (Jaysheel Bhavsar@UDel)
  - Automatic RVDB production pipeline (**85** Python scripts down to **2** BASH scripts)
- **Taxonomy-based phage cleanup pipeline to supplement keyword-based search**
  - Example: *Mycobacterium* phage > *Mycobacterium* virus
- **Poly N detection and filtration pipeline (for SARS-CoV-2 sequences only)**
- **New Collapsing Strategy to generate C-RVDB**
  - A burden from the flooding of SARS-CoV-2 with high redundancy and size (35Kbs)
  - A new collapsing strategy was discussed in AVDTWG subgroup C and evaluated
  - RVDB without SARS-CoV-2 redundancy lessens the computational burden for NGS bioinformatic analysis

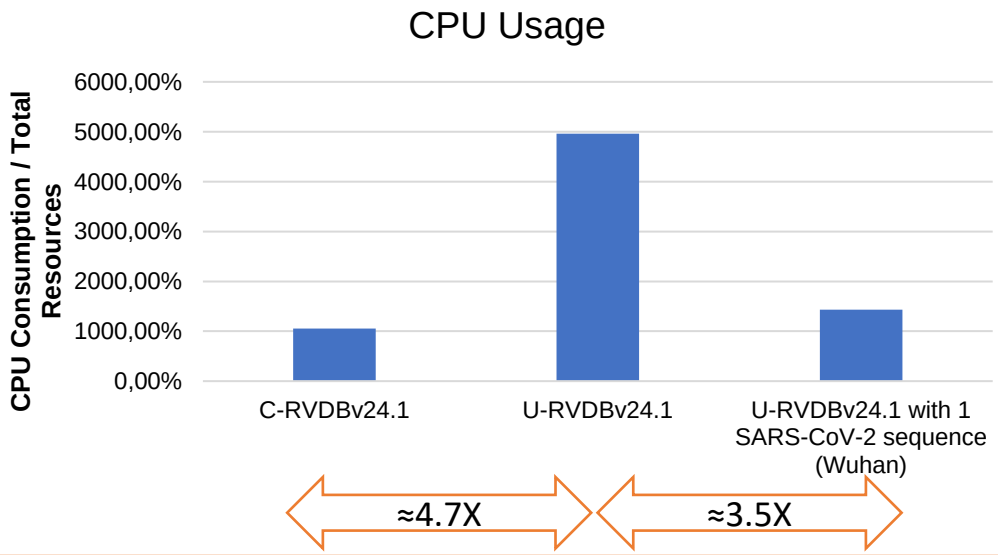




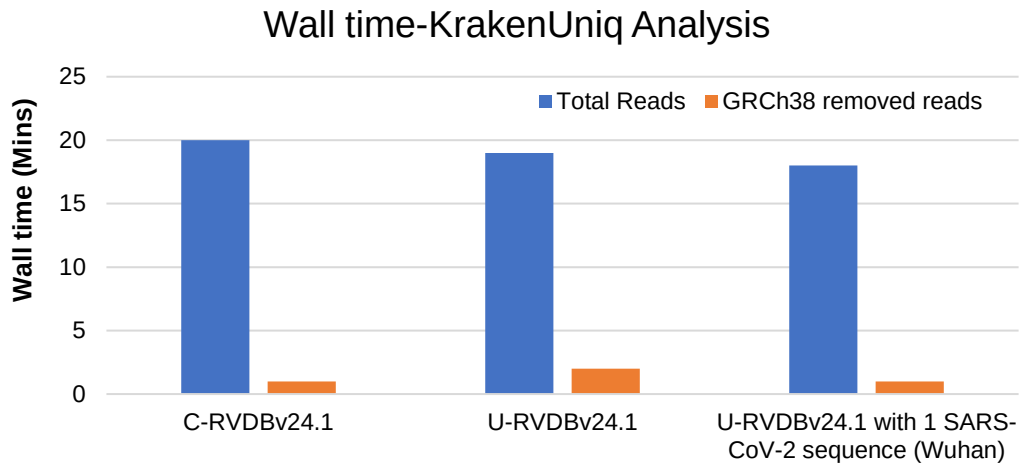
Impact of SARS-CoV-2 Redundancy in NGS Analysis Algorithms/Tools

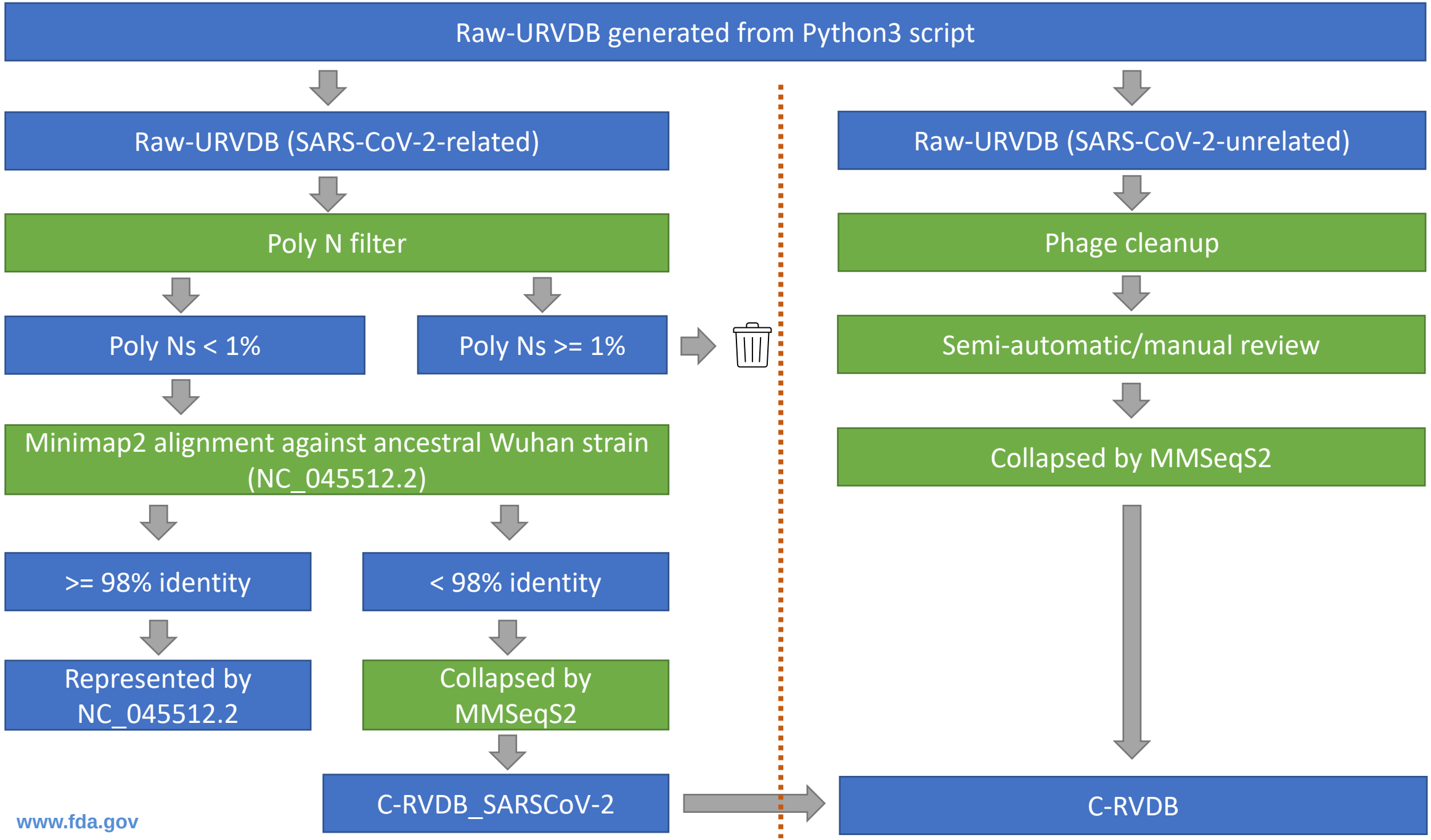
BLASTn *-num_threads 56*





KrakenUniq *--threads 100*




Raw-URVDB generated from Python3 script
Raw-URVDB (SARS-CoV-2-related)
Poly N filter
Poly Ns < 1%
Poly Ns >= 1%
Minimap2 alignment against ancestral Wuhan strain (NC_045512.2)
>= 98% identity
< 98% identity
Represented by NC_045512.2
Collapsed by MMSeqS2
C-RVDB_SARSCoV-2
Raw-URVDB (SARS-CoV-2-unrelated)
Phage cleanup
Semi-automatic/manual review
Collapsed by MMSeqS2
C-RVDB

Metadata Matrices of SARS-CoV-2 Collapsing Processes

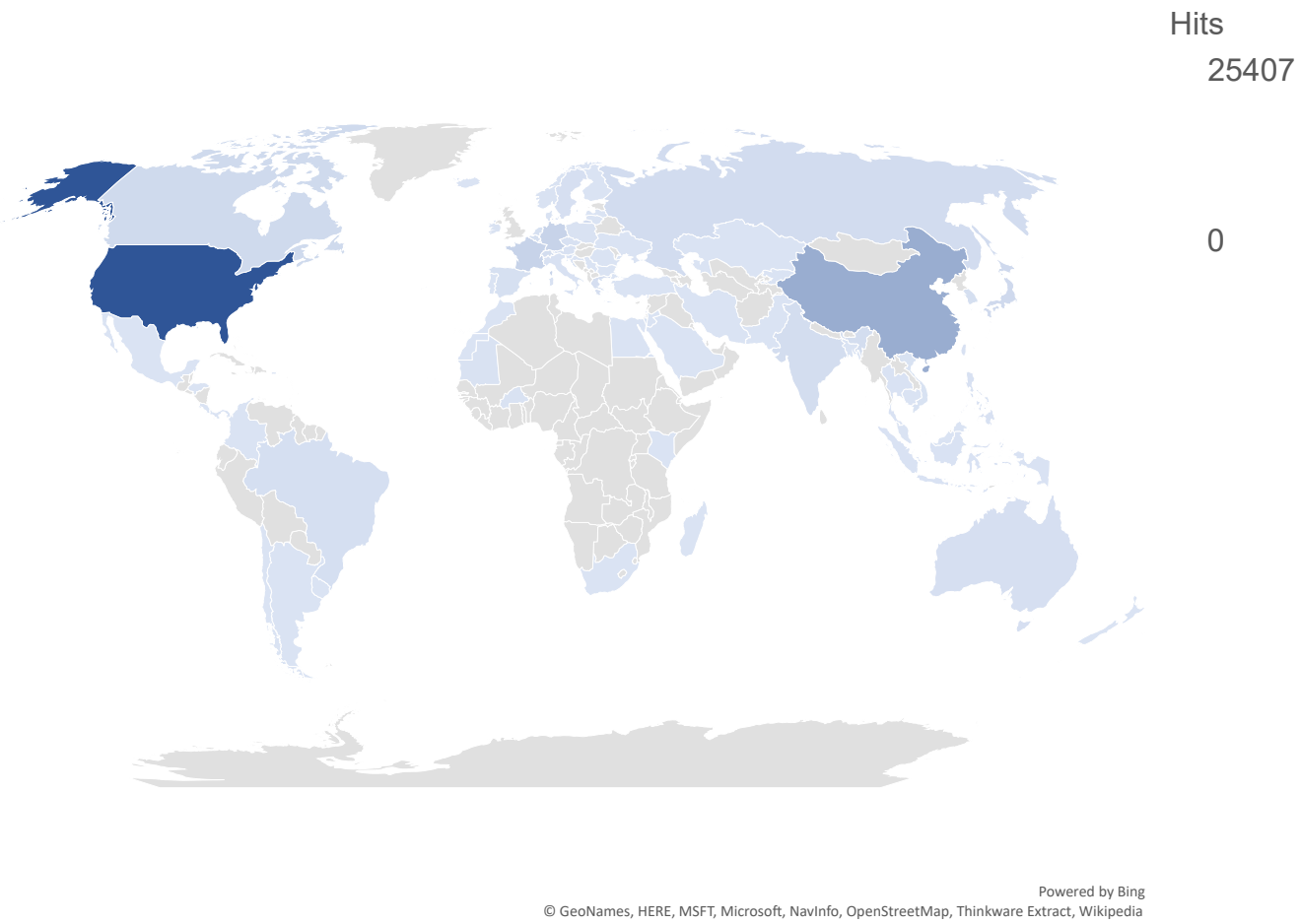
	Raw-URVDB (SARS-CoV-2-related)	Poly N Filter (<1%)		Minimap2 against Wuhan Strain (< 98% identity)		MMSeqS2 Collapsing (98%)		Collapsing Ratio
#Sequences	5,570,282		3,836,108		1,080,989		1,067	5,220X
DB Size	159 GB		109 GB		31 GB		3.1 MB	51,290X
# Unique Pango lineages	1,472		1,402		576		23	64X
# Unique Scorpio call	31		30		28		10	3.1X

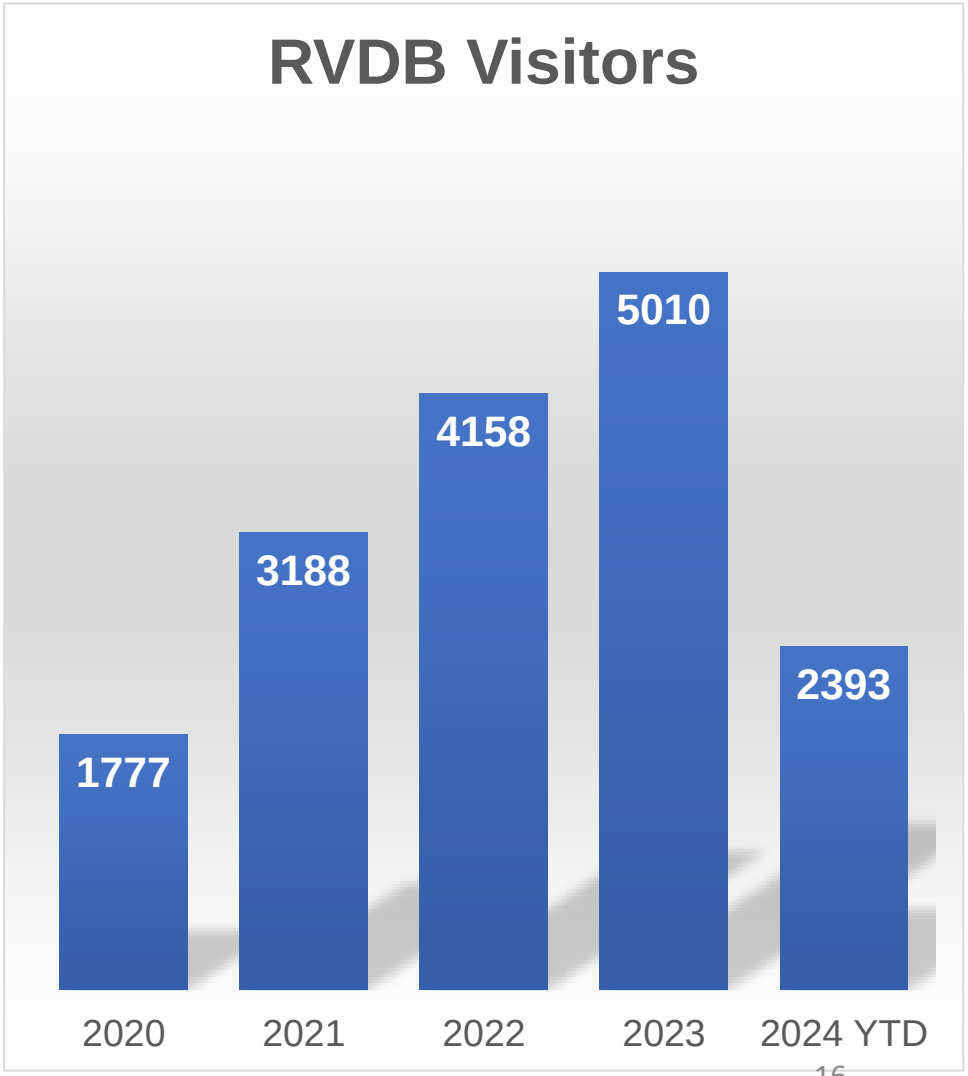
Scorpio Calls of Cascaded SARS-CoV-2 Collapsing Processes

U-RVDB (SARS-CoV-2 diversity)						U-RVDB (SARS-CoV-2 representatives)
243 A.23.1-like 53 A.23.1-like+E484K 538512 Alpha (B.1.1.7-like) 165 AV.1-like 2274 B.1.1.318-like 1090 B.1.1.7-like+E484K 863 B.1.617.1-like 14 B.1.617.3-like 7530 Beta (B.1.351-like) 108538 Delta (AY.4.2-like) 668223 Delta (AY.4-like) 1759131 Delta (B.1.617.2-like) 5290 Delta (B.1.617.2-like) +K417N 13325 Epsilon (B.1.427-like) 28457 Epsilon (B.1.429-like) 2471 Eta (B.1.525-like) 26622 Gamma (P.1-like) 33280 Iota (B.1.526-like) 1311 Lambda (C.37-like) 5444 Mu (B.1.621-like) 1213742 Omicron (BA.1-like) 676359 Omicron (BA.2-like) 95 Omicron (BA.3-like) 4818 Omicron (Unassigned) 30259 Probable Omicron (BA.1-like) 3505 Probable Omicron (BA.2-like) 17 Probable Omicron (BA.3-like) 955 Probable Omicron (Unassigned) 44 Theta (P.3-like) 1200 Zeta (P.2-like)	Poly N Filter (<1%)	172 A.23.1-like 40 A.23.1-like+E484K 438424 Alpha (B.1.1.7-like) 128 AV.1-like 1689 B.1.1.318-like 906 B.1.1.7-like+E484K 534 B.1.617.1-like 11 B.1.617.3-like 4884 Beta (B.1.351-like) 78316 Delta (AY.4.2-like) 454800 Delta (AY.4-like) 1305259 Delta (B.1.617.2-like) 3681 Delta (B.1.617.2-like) +K417N 8472 Epsilon (B.1.427-like) 19123 Epsilon (B.1.429-like) 1864 Eta (B.1.525-like) 19238 Gamma (P.1-like) 25326 Iota (B.1.526-like) 1059 Lambda (C.37-like) 4498 Mu (B.1.621-like) 643798 Omicron (BA.1-like) 498065 Omicron (BA.2-like) 54 Omicron (BA.3-like) 3038 Omicron (Unassigned) 10373 Probable Omicron (BA.1-like) 1781 Probable Omicron (BA.2-like) 464 Probable Omicron (Unassigned) 28 Theta (P.3-like) 916 Zeta (P.2-like)	Minimap2 against Wuhan Strain (< 98% identity)	2 A.23.1-like 17627 Alpha (B.1.1.7-like) 214 B.1.1.318-like 56 B.1.1.7-like+E484K 24 B.1.617.1-like 2 B.1.617.3-like 152 Beta (B.1.351-like) 1583 Delta (AY.4.2-like) 10331 Delta (AY.4-like) 206093 Delta (B.1.617.2-like) 907 Delta (B.1.617.2-like) +K417N 666 Epsilon (B.1.427-like) 2035 Epsilon (B.1.429-like) 187 Eta (B.1.525-like) 2889 Gamma (P.1-like) 775 Iota (B.1.526-like) 97 Lambda (C.37-like) 573 Mu (B.1.621-like) 320831 Omicron (BA.1-like) 497116 Omicron (BA.2-like) 36 Omicron (BA.3-like) 2731 Omicron (Unassigned) 3170 Probable Omicron (BA.1-like) 1195 Probable Omicron (BA.2-like) 153 Probable Omicron (Unassigned) 3 Theta (P.3-like) 10 Zeta (P.2-like)	MMSeqS2 Collapsing (98%)	**1 Alpha (B.1.1.7-like)** **1 Delta (AY.4.2-like)** **1 Delta (AY.4-like)** **6 Delta (B.1.617.2-like)** **1 Gamma (P.1-like)** **1 Mu (B.1.621-like)** **26 Omicron (BA.1-like)** **11 Omicron (BA.2-like)**

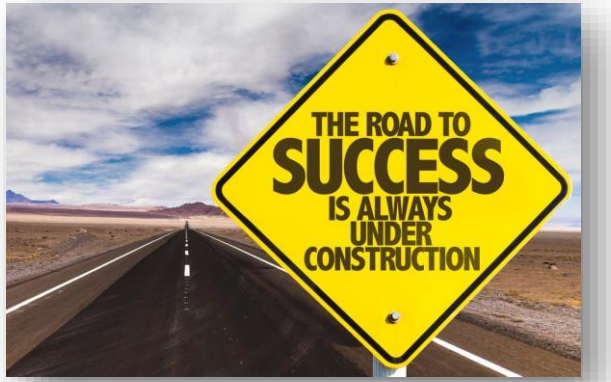
Growing RVDB Community

RVDB User Demography by Hits (YTD 2024)





Current and Future Work



- **Continuous in-house efforts for RVDB updating and refinements**
  - Quarterly update, based on NCBI's release schedule
  - Annotation of non-viral and viral but with host content
    - Expend non-viral/host dataset (rRNA mitochondria, vector, Phi X, Illumina and Nanopore adapters sequences and possible whole genome(s)) for broader detection of non-viral sequences
    - Annotation of host sequences in viruses with large genome (e.g. herpesvirus, pandoravirus)
  - Annotation of endogenous (retro)viruses
    - Sequences will be identified for distinguishing the different ERV families
    - Re-search NCBI nr/nt database to collect other ERVs from non-viral, cellular divisions
  - User experience improvement for RVDB website
    - Routine RVDB user survey to collect the feedbacks (First survey in 2020)
- **Parallelization of RVDB production pipeline**
  - Utilization of modern CPU architecture (multi-processing cores) to accelerate RVDB production pipeline
  - Identified tasks to be parallelized: Decompression of NCBI GenBank files and SEM-R pipeline

Ongoing Work: Artificial Intelligence (AI)/Machine Learning (ML) to Assist RVDB Manual Review Process

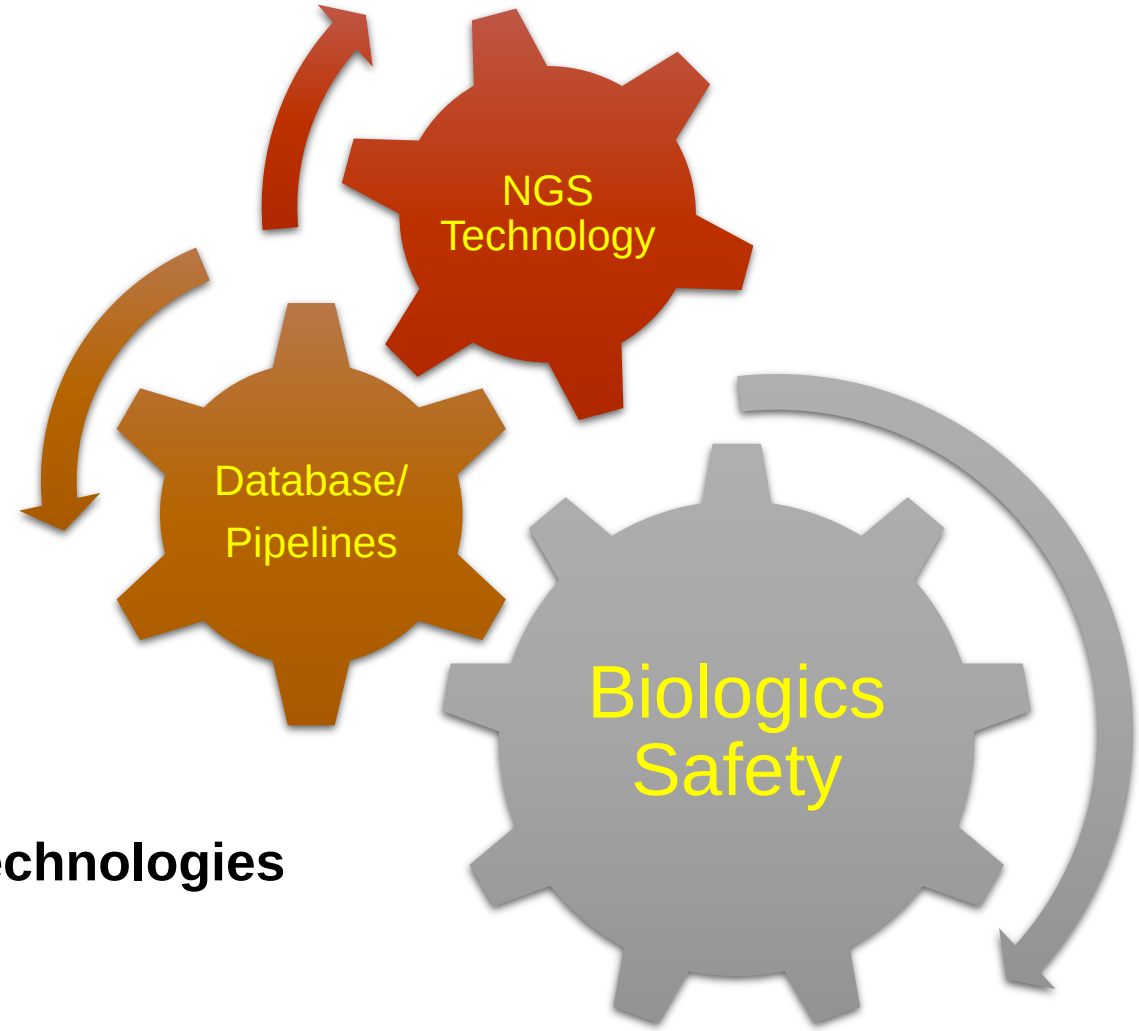
- About 100K sequences' fasta header are manually reviewed during each RVDB release update to determine if they are viral to be kept or non-viral to be removed
  - Labor intensive and time consuming
- The Large Language Models (LLMs) were finetuned by 10 millions labelled dataset with the flags of viral and non-viral/irrelevant from manual review results since 2018 (Supervised Learning)
  - **BERT has the best performance scores among 7 evaluated LLMs**
- BERT-RVDB finetuned model was implemented since v26.0
- LLMs with larger parameters (Google Gemini, Meta LLaMA, and OpenAI GPT4…etc.) will be evaluated for their potential of higher accuracy

BERT-RVDB differentiates mis-used Ty-named cell clone from true Ty-retrotransposons

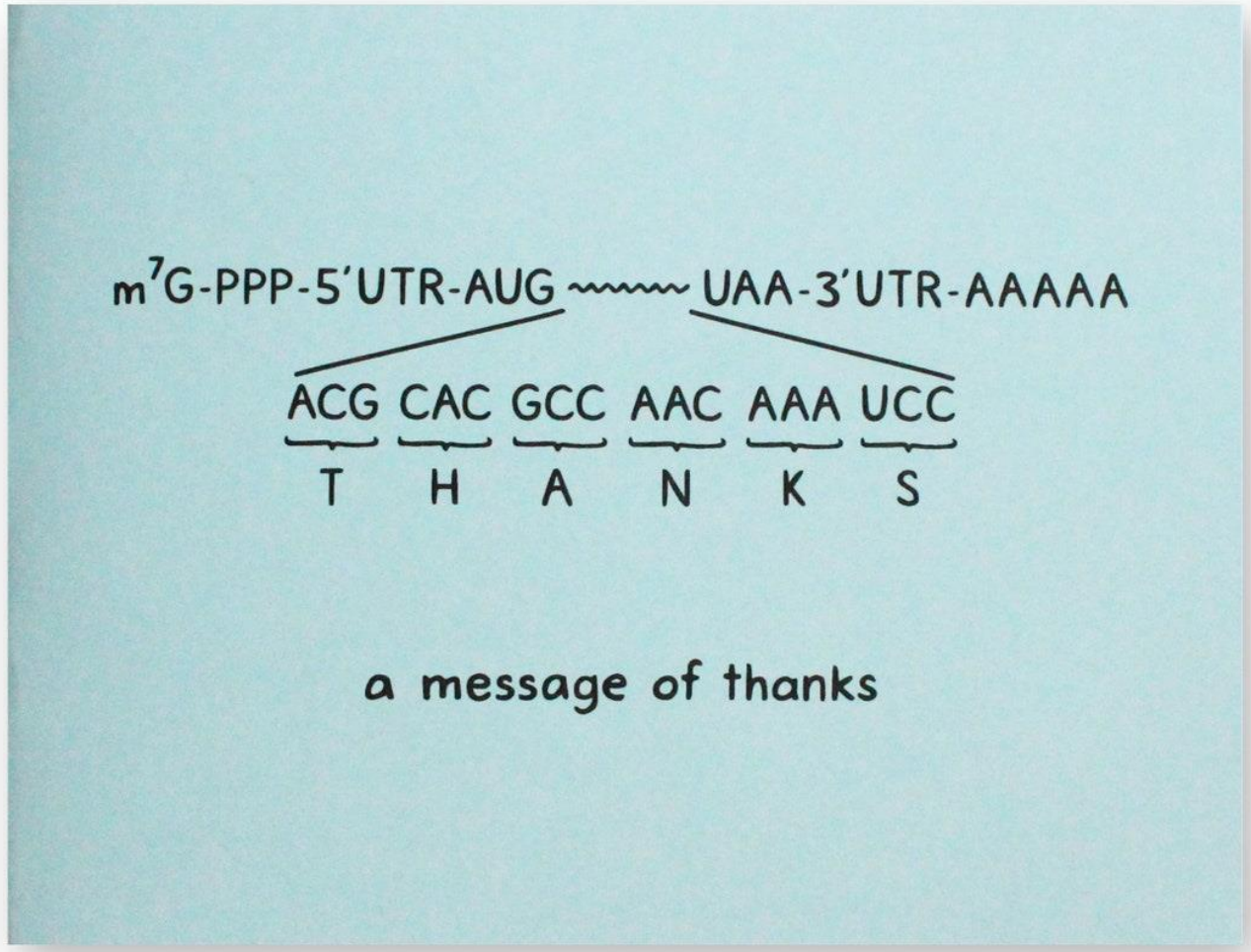
		Manual review judgement	**Inference from BERT-RVDB**
OQ613583.1	Colletotrichum truncatum strain TY1_1 glyceraldehyde-3-phosphate dehydrogenase (gapdh) gene, partial cds	host, Ty-named clone	NEGATIVE
OQ613656.1	Colletotrichum truncatum strain TY1_1 chitin synthase 1 (CHS-1) gene, partial cds	host, Ty-named clone	NEGATIVE
OR078387.1	Saccharomyces cerevisiae strain YJM193 transposon TY7, complete sequence		POSITIVE
OR079743.1	Idesia polycarpa clone IPRT1 retrotransposon Ty1-copia reverse transcriptase Ty1-like (ty1) gene, partial sequence		POSITIVE
OR079744.1	Idesia polycarpa clone IPRT2 retrotransposon Ty1-copia reverse transcriptase Ty1-like (ty1) gene, partial sequence		POSITIVE
OR079745.1	Idesia polycarpa clone IPRT3 retrotransposon Ty1-copia reverse transcriptase Ty1-like (ty1) gene, partial sequence		POSITIVE
OR079746.1	Idesia polycarpa clone IPRT4 retrotransposon Ty1-copia reverse transcriptase Ty1-like (ty1) gene, partial sequence		POSITIVE

Acknowledgements

- **Arifa Khan's Laboratory***
  - Arifa Khan
  - Trent Bosma
- **Center for Devices and Radiological Health (CDRH) /U.S. FDA**
  - Mike Mikailov
- **University of Delaware**
  - Shawn Polson
  - Jaysheel Bhavsar
  - Madolyn (Maddy) MacDonald
- **Discussion in Advanced Virus Detection Technologies Working Group (AVDTWG)**
  - Subgroup C



***Funding**:
FDA Medical Countermeasures Initiative - CBER Targeted Intramural Research - PDUFA.
NIIMBL-BMGF through Umbrella CRADA FDA No. 2018-0031-CRD


m7G-PPP-5'UTR-AUG
UAA-3'UTR-AAAAA
ACG CAC GCC AAC AAA UCC
T H A N K S
a message of thanks