A Non-redundant, Reference Virus Database (RVDB) for Adventitious Virus Detection in Biologics Using High-Throughput Sequencing (HTS) Technologies

Pei-Ju Chin, Trent Bosma and Arifa S. Khan

Office of Vaccines Research and Review, Division of Viral Products, Laboratory of Retroviruses, CBER, U.S. Food and Drug Administration, Silver Spring, MD 20993

Background

Genomic and metagenomic analysis using high-throughput sequencing (HTS) has resulted in a great expansion of the virosphere. However, there are challenges for HTS bioinformatics for detecting distantly-related viruses due to limitations of the publicly-available databases, which are not complete with all viral sequences, and virus detection can be obscured using the NCBI nr/nt collection, which contains diverse viral sequences, due to the large amount of cellular sequences. This gap was recognized by our laboratory during HTS investigations of Sf9 insect cells, which are a new cell substrate used for several baculovirus-expressed vaccines and other products. Those studies led to the discovery of a novel rhabdovirus in Sf9 cells by extensive analysis using nr/nt, which had very limited sequence identity to any known virus (only 4% coverage in a 295-bp region with 66% nucleotide identity to a partial sequence of Taastrup virus, a new member of the family Mononegavirales) **[1]**. Therefore, we initiated efforts to create a new, non-redundant, reference viral database (RVDB) that would include all virus, virus-related, and viral-like sequences, including endogenous viruses and retroelements, and have an overall reduced cellular sequence content. This effort was in consultation with the Advanced Virus Detection Technologies Interest Group, which includes scientists from industry (vaccines, gene therapies, and biotherapeutics), regulatory and other government agencies (including NCBI), technology service providers, and academia **[2]**.

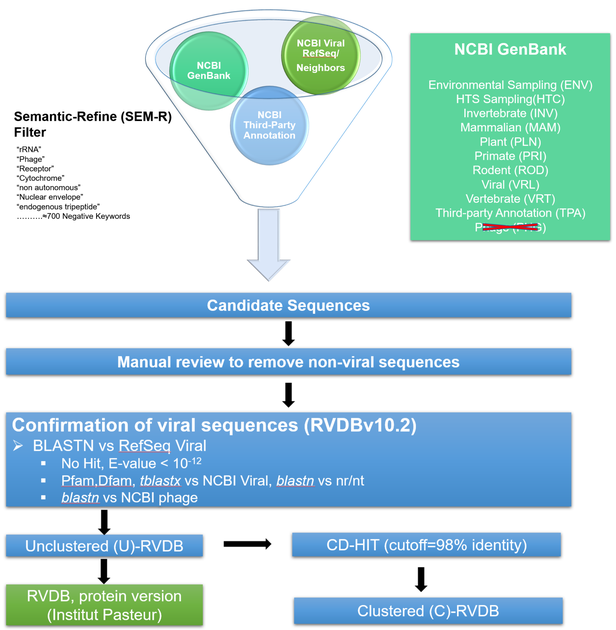
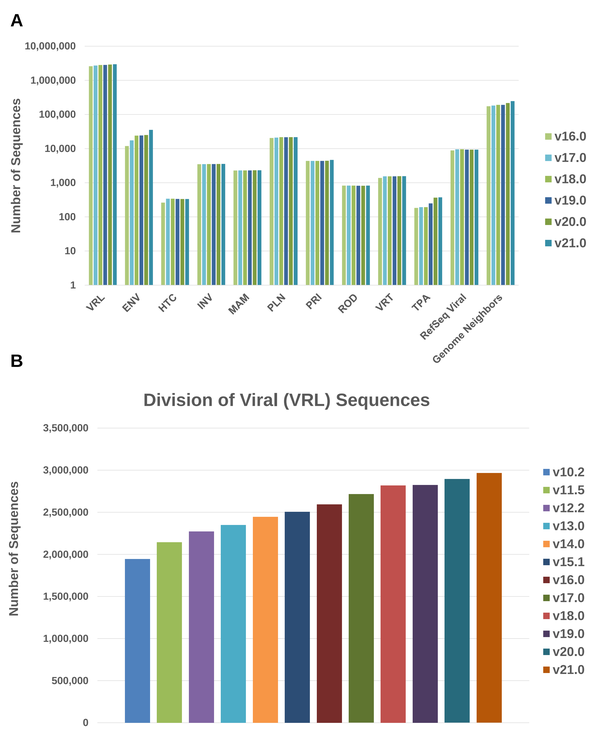


Figure 1. General workflow for development of RVDB

Materials and Methods

The overall strategy for developing RVDB included: a) development of a refined, final keyword screen for semantic selection of all virus, virus-related, and virus-like sequences from GenBank, regardless of their size;, b) verification of viral identity using various bioinformatics tools (BLAST and HMMER); c) clustering at 98% sequence identity using CD-HIT-EST to remove redundancy; and d) testing robustness of RVDB for HTS analysis using our large, in-house next-gen datasets and comparing the run time and results to those obtained using the NCBI databases (NCBI Viral RefSeq + genome neighbors and nr/nt). The overall strategy, with details of the different steps of selection and analysis, is shown in **Fig. 1**. Details of the generation and characterization of RVDBv10.2 is described by Goodacre et. al. **[3]**. The sequence collection of consecutive RVDB versions is shown in **Fig. 2**.



Total number of entries in RVDBv21.0: 3,280,994 including the sequences of NCBI RefSeq Viral (n=9,349) and NCBI Genome Neighbors (n=245,905)

Figure 2. Statistics of sequence collection in consecutive RVDB versions: The sequence number of each GenBank divisions is shown in **(A)**. The major source of RVDB collection, viral division (VRL), is shown in **(B)**.

Results and Discussion

- **RVDB GitHub and Resources**
  - https://github.com/ArifaKhanLab/RVDB, or Google "RVDB GitHub"
  - RVDB DIY Toolbox (Manual and Python code)
- **Ready-to-use RVDB is currently available** @ https://rvdb.dbi.udel.edu/ (BLAST search is available!)
- **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 unknown virus detection by *blastn*
  - C-RVDB *fasta* file
    - Clustered, sequences share 98% similarity are collapse to one representative sequence for each clade
    - Lower computation-demanded. Suitable for unknown virus detection by *tblastx*
  - SQLite DB Script
    - Create the entries (*fasta* header and the corresponding information) for advanced bioinformatic pipelines/workflows
  - Proteic RVDB (Institut Pasteur)
    - Hidden Markov Model (HMM) profile of viral protein domains
    - Unknown viruses with remote homology by *hmmsearch / hmmscan*
- **RVDB Refinement: Viral and Non-Viral Sequences Annotation**
  - The collection of RVDB is subset *bona-fide* from NCBI GenBank, viral RefSeq and TPA **without any modifications**. Therefore, the potential issues are inherited
    - Quality of sequence
    - Sequencing vector carryover (vector/adapter/linker/primer…etc)
    - Mis-annotation
    - Host carryovers in endogenous retroviruses
  - Pipelines for overcoming these issues are being developing

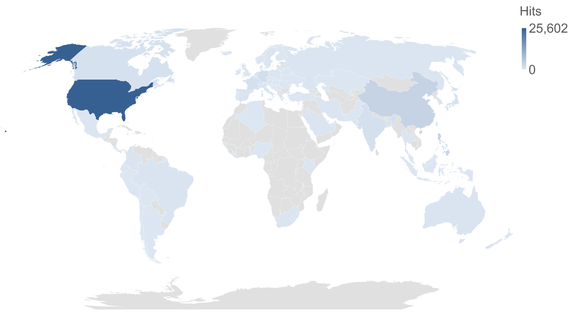


Figure 3. RVDB user demography by hits (Jan. 2020~Sept. 2020)

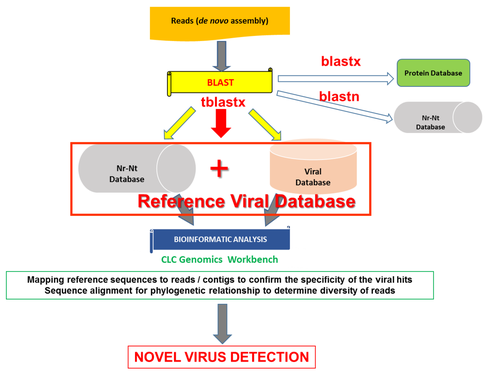


Figure 4. A Bioinformatics approach for novel virus detection

Conclusion

RVDB is expected to aid HTS investigations for known and novel viruses, thereby enhancing product safety, due to the inclusion in the database of all viral-related sequences, including sub-genomic viral fragments as well as endogenous retroviruses and retrotransposons, along with the reduction of overall cellular content. Ongoing annotation efforts will result in a high-quality database that will increase confidence in obtaining accurate results from HTS data analysis.

References

1. Ma, H., Galvin, T.A., Glasner, D.R., Shaheduzzaman, S., and Khan, A.S. 2014. Identification of a Novel Rhabdovirus in Spodoptera frugiperda Cell Lines. J Virol 88:6576-6585.
2. Khan, A.S., Vacante, D.A., Cassart, J.P., Ng, S.H.S., Lambert, C., Charlebois, R.L., and King, K.E. 2016. Advanced Virus Detection Technologies Interest Group (AVDTIG): Efforts on High Throughput Sequencing (HTS) for Virus Detection. PDA J Pharm Sci and Tech 70:591-595.
3. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A.S. 2018. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. mSphere 3:e00068-18..

SCAN ME — **RVDB**

SCAN ME — **Prot-RVDB**

SCAN ME — **RVDB GitHub**

This project was initially funded by the Medical Countermeasures Initiative. The work has continued through CBER Targeted Intramural Funding for Pandemic Influenza