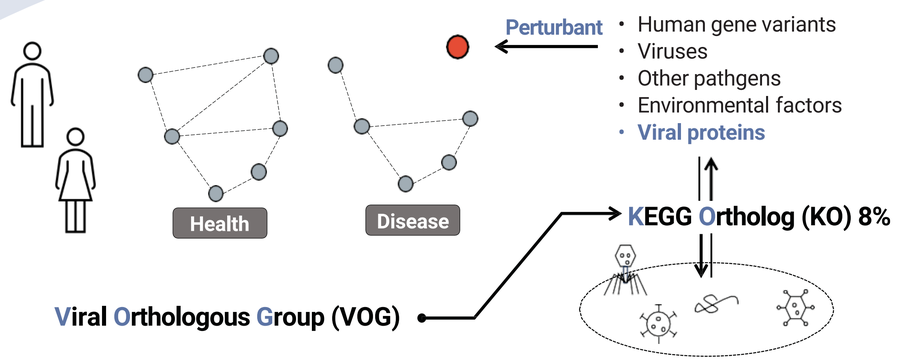
KEGG MEDICUSにおけるウイルスタンパク質のオーソログデータベース

Viral Orthologous Groups (VOG) database in KEGG MEDICUS

Lingjie Meng[1], Hiroyuki Ogata[1], Minoru Kanehisa[1]
1. Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan

Background



VOG database aims to link unkown viral proteins to known functions

Low sequence homogeneity

- High mutation rate (10^{-3} – 10^{-8} viruses; $<10^{-9}$ cellular organisms)
- Frequent recombination events
- Rapidly adapt to host environments and elicit Immune Responses.

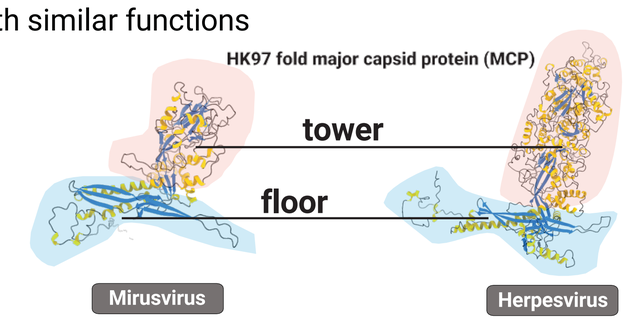
Underestimated species diversity

- 10^4 viral species have been identified.
- It is estimated that 10^7 – 10^9 virus species in the global virome.
- The isolation of virus-host system is limited by the techniques.

Challenges of clustering the viral orthologous

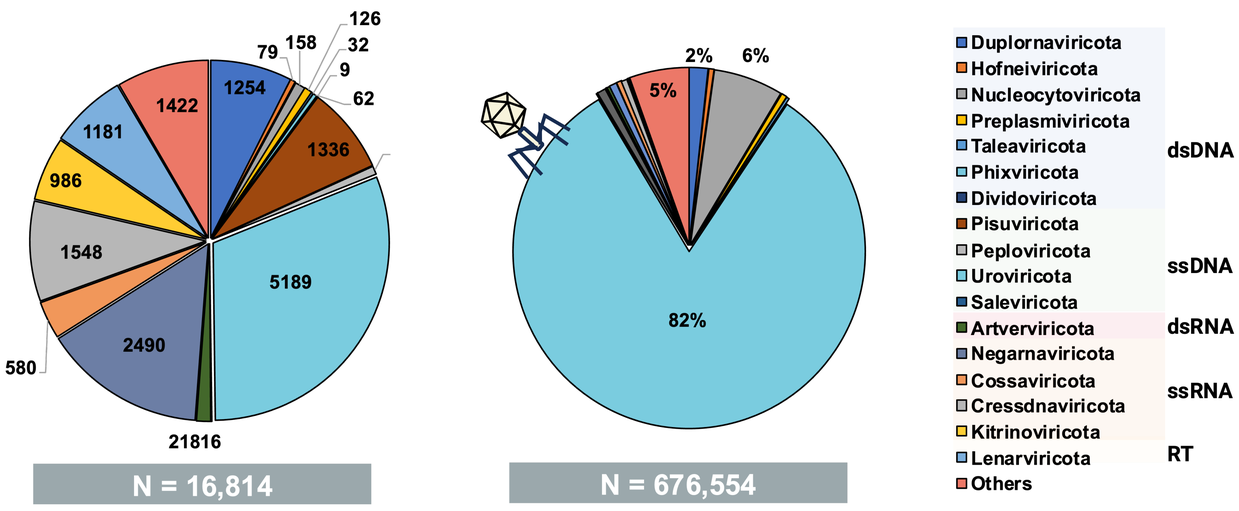
Structure-based clustering

Viral protein structure are more conserved compared to sequences Structure similarity could help identify divergent sequences with similar functions



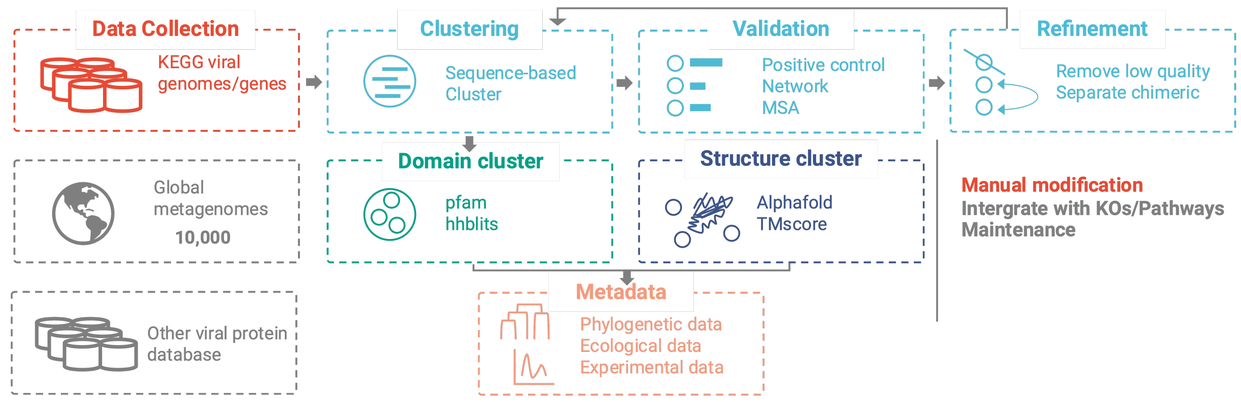
Possible solution for sequence divergency

Method and Materials



Viral genomes in the KEGG database (March, 2024) 11,420 viral species were included.

Workflow of generating the VOC database



Metadata for each viral genes

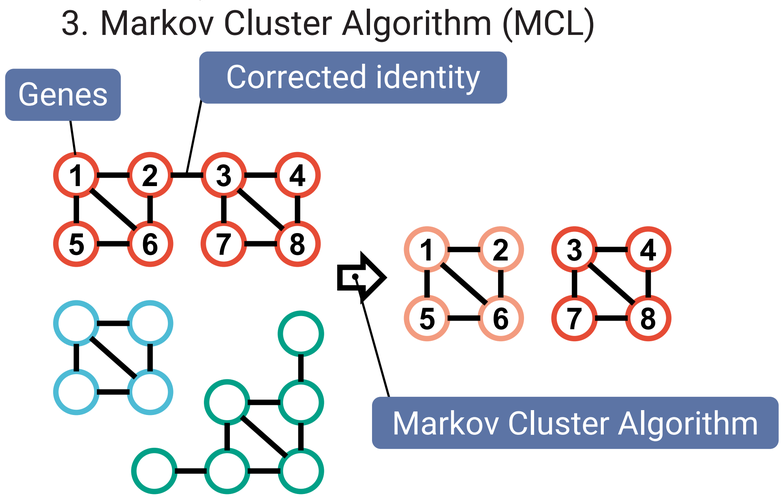
vg	RefSeq	Genome Entry	Virus Taxa ID	Phylum	Baltimore	Virus lineage	Host lineage	KO	Cluster
vg:29125313	YP_009302533	NC_031245	1792032	Uroviricota	BC1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Caudovirales; Myoviridae; Thornevirus; Bacillus virus SP15	Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus subtilis group; Bacillus paralicheniformis	K00012	cluster_5464
vg:10329251	YP_004324717	NC_015289	445685	Uroviricota	BC1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Caudovirales; Myoviridae; unclassified Myoviridae	Bacteria; Terrabacteria group; Cyanobacteria/Melainabacteria group; Cyanobacteria; Synechococcales; Prochlorococcaceae; Parasynechococcus; Parasynechococcus marenigrum	K00033	cluster_5464

Results

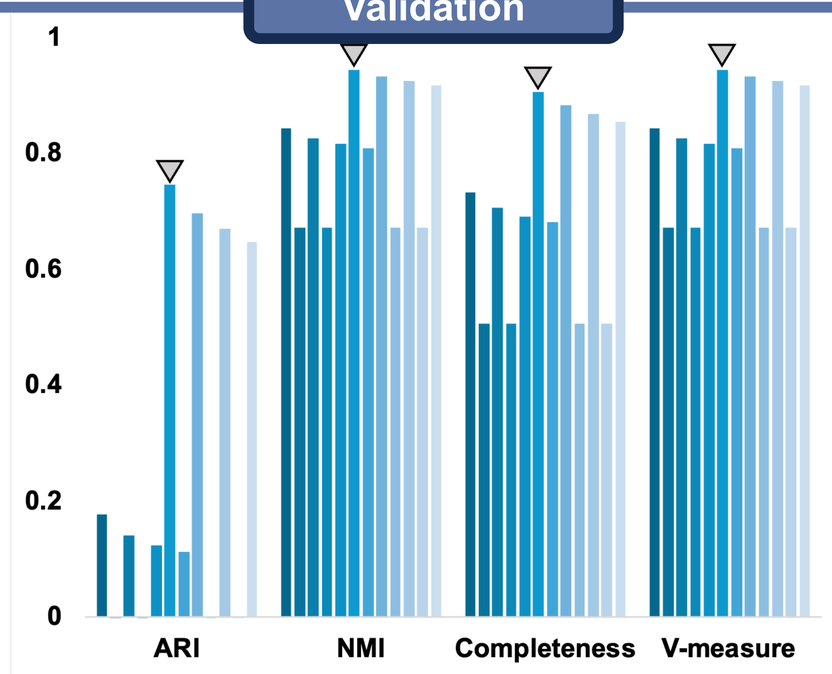
Clustering

Sequence-based clustering

1. Sequence pairwise alignment (ssearch, blastp…)
2. Identity correction
3. Markov Cluster Algorithm (MCL)



Validation

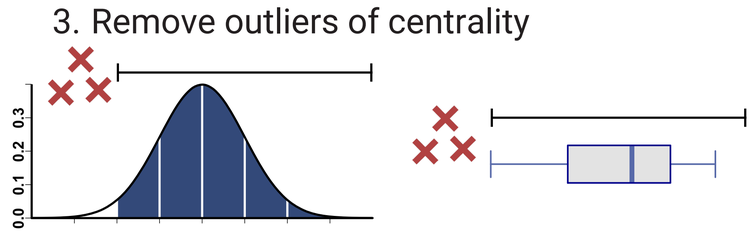


The combination of ssearch and MCL i2 (▽) outperform other clustering metrics.

Refinement

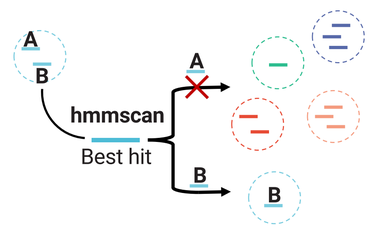
Network-based refinement

1. Sequence alignment network
2. Centrality calculation
3. Remove outliers of centrality



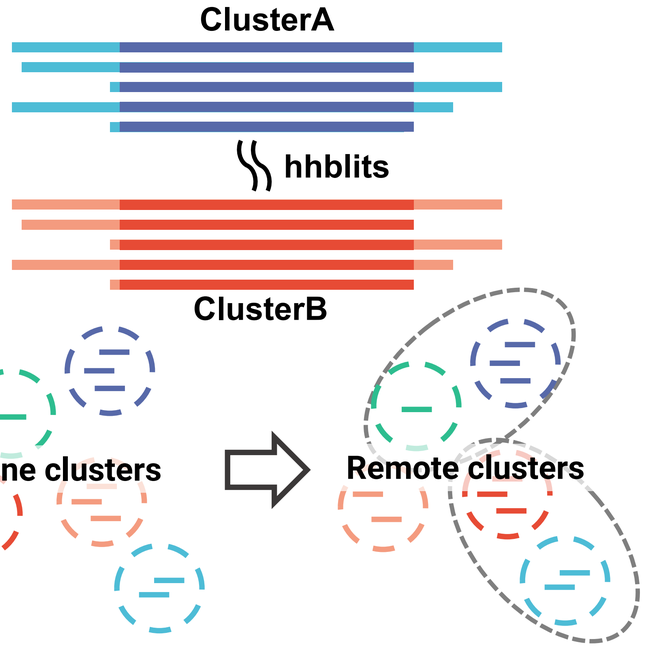
Higher than (mean - 2 std) & **Higher than (Q1 – 1.5 IQR)**

4. Remove outliers by length distribution
5. Intra-cluster domain conservation



Performed multiple iterations (**N = 8**) to remove all sequences that had the best hits to non-self clusters.

Domain cluster

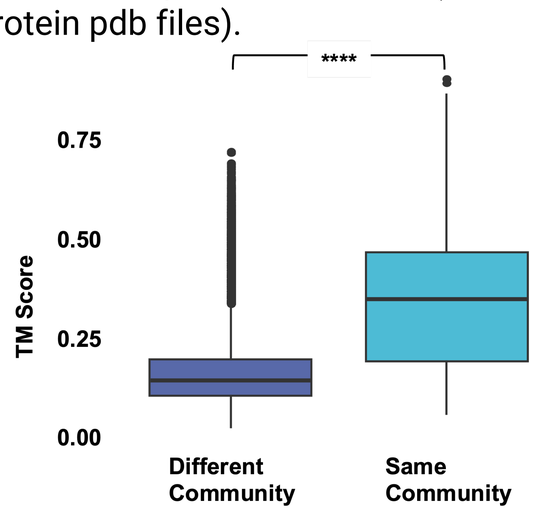


N = 53,627
33,329 > 3 genes

N = 22,925
3,042 non-singleton

Structure

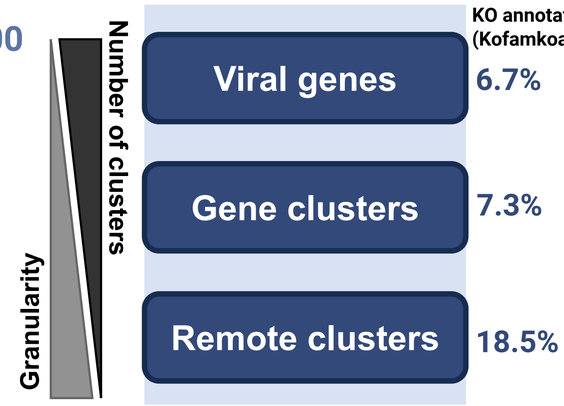
Predicted protein 3D structures for **33,329** cluster representatives using **ESMfold/Alphafold** and generated a viral protein database (over **100,000** viral protein pdb files).



Same community (remote clustering) viral genes tends to have a similar protein structure.

Summary

By far



- Generated over **30,000** viral gene clusters and **3,000** remote clusters.
- Improved the KO annotation rate to **18%**.
- An integrated database for annotations.

Future steps

Over **1,500,000** environmental genomes

500,000 high quality genomes

23,000 viral species

Environment data

Acknowledgement

- Database Integration Coordination Program (DICP): Integrated database linking human and pathogen genomes to diseases and drugs
- Computational time was provided by the SuperComputer System, Institute for Chemical Research, Kyoto University.