Supplementary Text S1

Dataset of reference viral genomes Viral genome sequences and taxonomic information of viruses and their hosts are based on the GenomeNet/Virus-Host DB (Mihara, et al., 2016). The Virus-Host DB covers viruses with complete genomes stored in RefSeq/viral and GenBank. The sequence and taxonomic data are downloadable at http://www.genome.jp/virushostdb/. Viral genome sequences and taxonomic information used in ViPTree will be updated every few months following Virus-Host DB updates.

Calculation of genomic similarity (S_G) and proteomic tree The original Phage Proteomic Tree (Rohwer and Edwards, 2002) tested multiple approaches and parameters to calculate genomic distance for evaluation of compatibility between the Phage Proteomic Tree and taxonomical system of the International Committee on Taxonomy of Viruses. In contrast, ViPTree performs proteomic tree calculation based on tBLASTx as reported previously by other several studies (Bellas, et al., 2015; Bhunchoth, et al., 2016; Mizuno, et al., 2013). Specifically, normalized tBLASTx scores (S_G; $0 \leq S_G \leq 1$) between viral genomes are calculated as described (Bhunchoth, et al., 2016). For simple calculation of S_G of segmented viruses, sequences of each segmented viral genome were concatenated into one sequence by inserting a sequence of 100 ambiguous nucleotides (N) at each concatenation site. In addition, genome sequences longer than 100 kb are split into each 100 kb fragment just for faster tBLASTx calculation. Genomic distance is calculated as $1-S_G$. Proteomic tree generation is performed by BIONJ (Gascuel, 1997) based on the genomic distance, using the R package ‘Ape’. The root of a tree is defined by mid-point rooting.

Gene finding and gene annotation Gene finding of uploaded sequences is performed by GeneMarkS (Besemer, et al., 2001) without using its self-training method and with using an option "-p 0" to prohibit gene overlaps. Automatic functional annotations of predicted genes are performed by protein similarity searches against NCBI/nr using GHOSTX (Suzuki, et al., 2014).

References

Bellas, C.M., Anesio, A.M. and Barker, G. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Front Microbiol* 2015;6:656.

Besemer, J., Lomsadze, A. and Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 2001;29(12):2607-2618.

Bhunchoth, A.*, et al.* Two asian jumbo phages, ΦRSL2 and ΦRSF1, infect Ralstonia solanacearum and show common features of ΦKZ-related phages. *Virology* 2016;494:56-66.

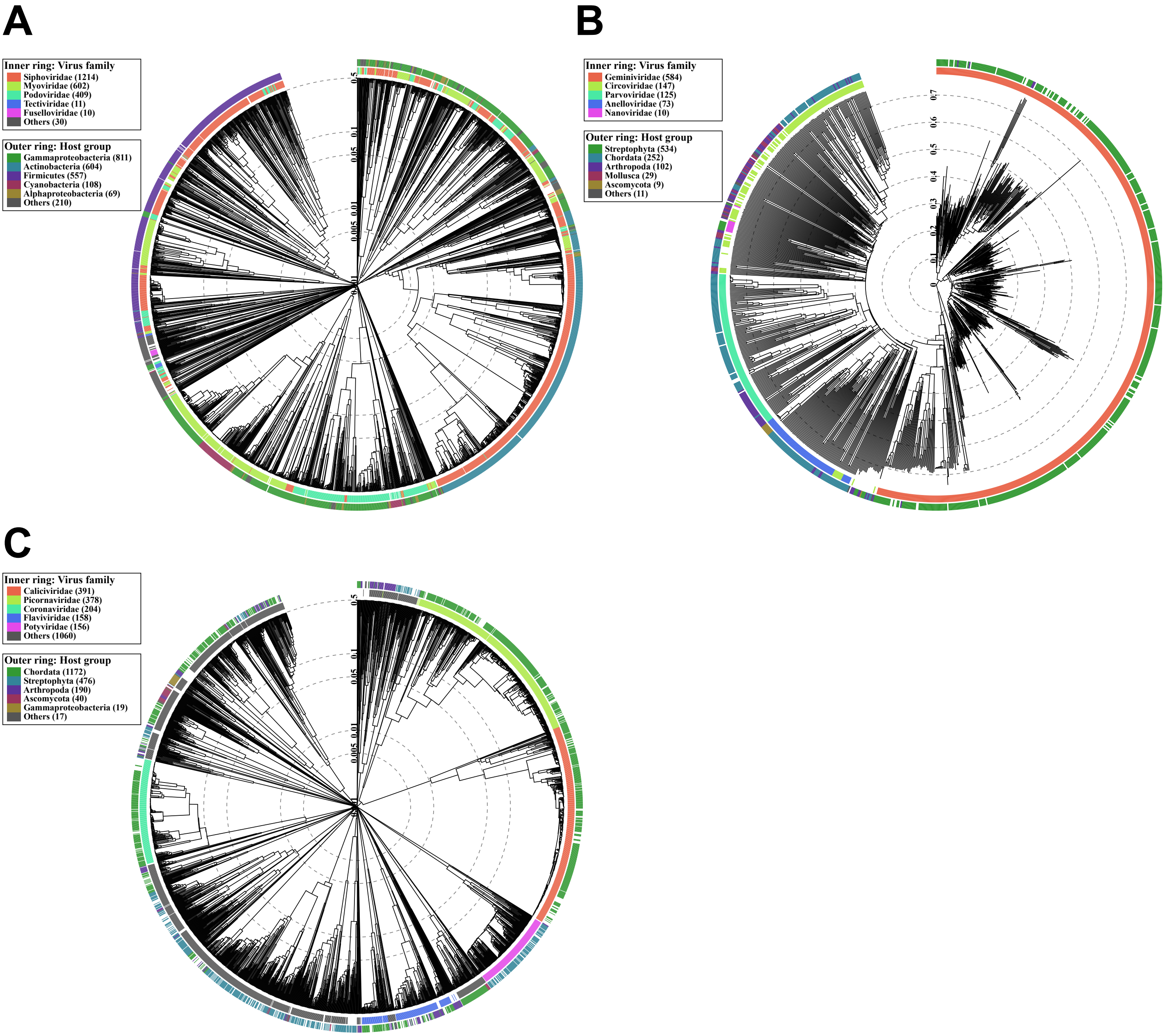
Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 1997;14(7):685-695.

Mihara, T.*, et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* 2016;8(3):66.

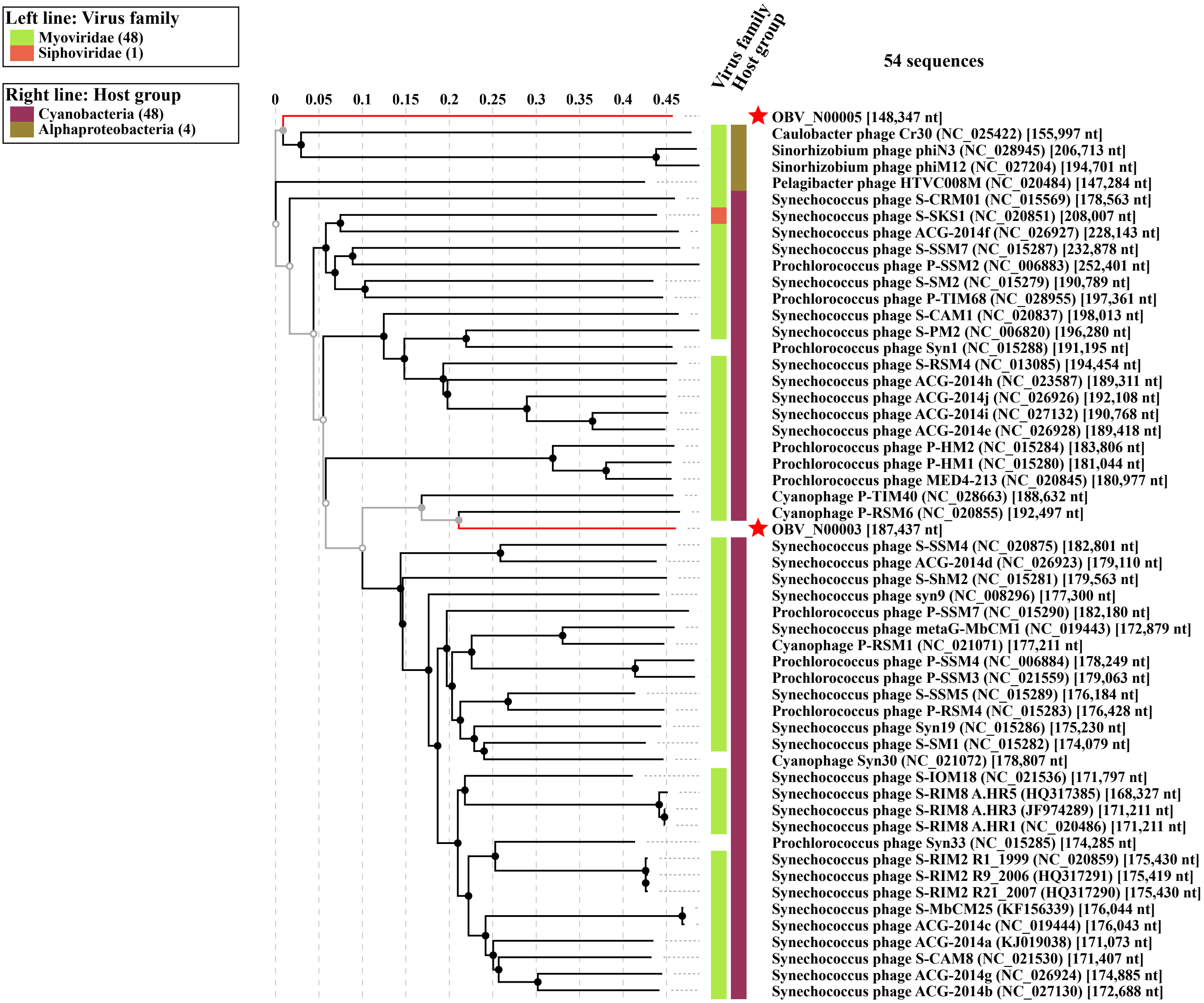
Mizuno, C.M.*, et al.* Expanding the marine virosphere using metagenomics. *PLoS Genet* 2013;9(12):e1003987.

Rohwer, F. and Edwards, R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* 2002;184(16):4529-4535.

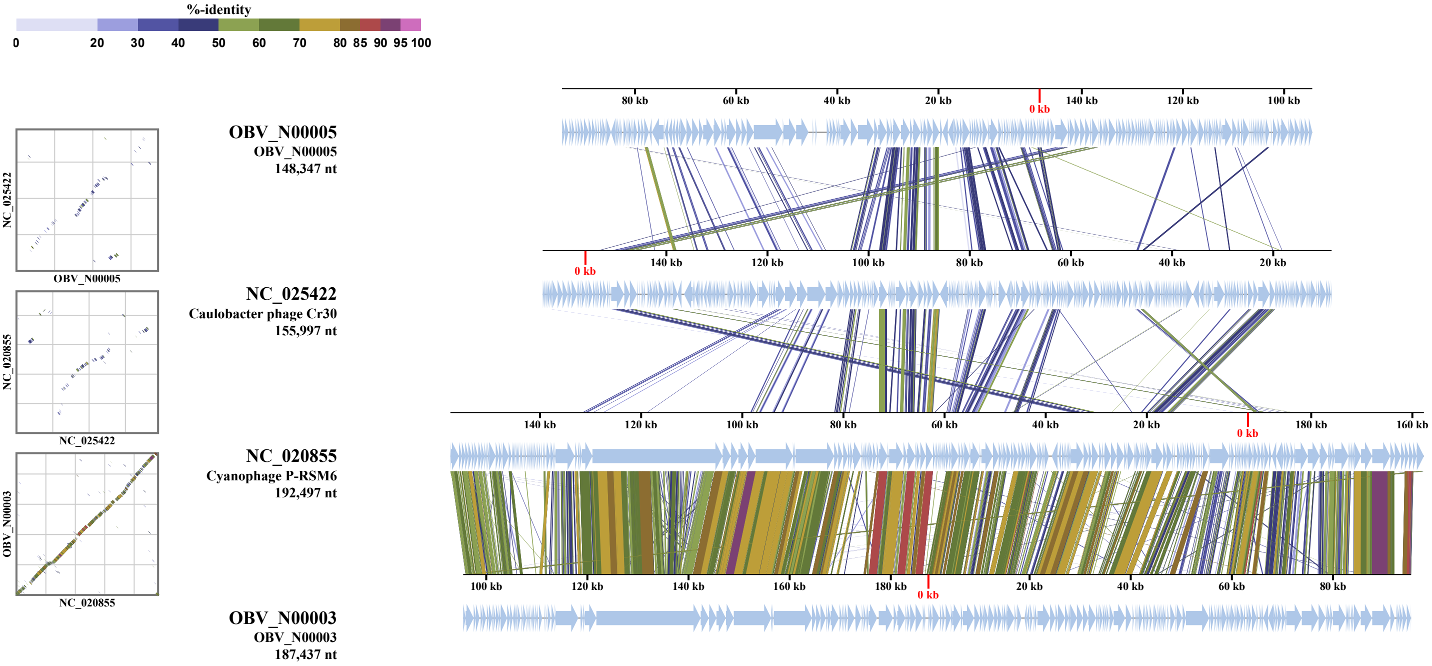
Suzuki, S.*, et al.* GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS One* 2014;9(8):e103833.



Supplementary Figure S1. Viral proteomic trees of reference genomes represented in the circular view. Color rings indicate virus families (inner rings) and host groups (at a level of phylum except for Proteobacteria; outer rings). These trees are calculated by BIONJ based on genomic distance matrixes, and mid-point rooted. Branch lengths are log-scaled (A, C) and linearly scaled (B). The sequence and taxonomic data is based on Virus-Host DB (release 78). (A) A tree of prokaryotic dsDNA viruses (2,384 genomes). (B) A tree of eukaryotic ssDNA viruses (1,013 genomes). (C) A tree of all ssRNA viruses (2,485 genomes).


Left line: Virus family
🟩 Myoviridae (48)
🟥 Siphoviridae (1)
Right line: Host group
🟥 Cyanobacteria (48)
🟫 Alphaproteobacteria (4)
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Virus family
Host group
54 sequences
★ OBV_N00005 [148,347 nt]
Caulobacter phage Cr30 (NC_025422) [155,997 nt]
Sinorhizobium phage phiN3 (NC_028945) [206,713 nt]
Sinorhizobium phage phiM12 (NC_027204) [194,701 nt]
Pelagibacter phage HTVC008M (NC_020484) [147,284 nt]
Synechococcus phage S-CRM01 (NC_015569) [178,563 nt]
Synechococcus phage S-SKS1 (NC_020851) [208,007 nt]
Synechococcus phage ACG-2014f (NC_026927) [228,143 nt]
Synechococcus phage S-SSM7 (NC_015287) [232,878 nt]
Prochlorococcus phage P-SSM2 (NC_006883) [252,401 nt]
Synechococcus phage S-SM2 (NC_015279) [190,789 nt]
Prochlorococcus phage P-TIM68 (NC_028955) [197,361 nt]
Synechococcus phage S-CAM1 (NC_020837) [198,013 nt]
Synechococcus phage S-PM2 (NC_006820) [196,280 nt]
Prochlorococcus phage Syn1 (NC_015288) [191,195 nt]
Synechococcus phage S-RSM4 (NC_013085) [194,454 nt]
Synechococcus phage ACG-2014h (NC_023587) [189,311 nt]
Synechococcus phage ACG-2014j (NC_026926) [192,108 nt]
Synechococcus phage ACG-2014i (NC_027132) [190,768 nt]
Synechococcus phage ACG-2014e (NC_026928) [189,418 nt]
Prochlorococcus phage P-HM2 (NC_015284) [183,806 nt]
Prochlorococcus phage P-HM1 (NC_015280) [181,044 nt]
Prochlorococcus phage MED4-213 (NC_020845) [180,977 nt]
Cyanophage P-TIM40 (NC_028663) [188,632 nt]
Cyanophage P-RSM6 (NC_020855) [192,497 nt]
★ OBV_N00003 [187,437 nt]
Synechococcus phage S-SSM4 (NC_020875) [182,801 nt]
Synechococcus phage ACG-2014d (NC_026923) [179,110 nt]
Synechococcus phage S-ShM2 (NC_015281) [179,563 nt]
Synechococcus phage syn9 (NC_008296) [177,300 nt]
Prochlorococcus phage P-SSM7 (NC_015290) [182,180 nt]
Synechococcus phage metaG-MbCM1 (NC_019443) [172,879 nt]
Cyanophage P-RSM1 (NC_021071) [177,211 nt]
Prochlorococcus phage P-SSM4 (NC_006884) [178,249 nt]
Prochlorococcus phage P-SSM3 (NC_021559) [179,063 nt]
Synechococcus phage S-SSM5 (NC_015289) [176,184 nt]
Prochlorococcus phage P-RSM4 (NC_015283) [176,428 nt]
Synechococcus phage Syn19 (NC_015286) [175,230 nt]
Synechococcus phage S-SM1 (NC_015282) [174,079 nt]
Cyanophage Syn30 (NC_021072) [178,807 nt]
Synechococcus phage S-IOM18 (NC_021536) [171,797 nt]
Synechococcus phage S-RIM8 A.HR5 (HQ317385) [168,327 nt]
Synechococcus phage S-RIM8 A.HR3 (JF974289) [171,211 nt]
Synechococcus phage S-RIM8 A.HR1 (NC_020486) [171,211 nt]
Prochlorococcus phage Syn33 (NC_015285) [174,285 nt]
Synechococcus phage S-RIM2 R1_1999 (NC_020859) [175,430 nt]
Synechococcus phage S-RIM2 R9_2006 (HQ317291) [175,419 nt]
Synechococcus phage S-RIM2 R21_2007 (HQ317290) [175,430 nt]
Synechococcus phage S-MbCM25 (KF156339) [176,044 nt]
Synechococcus phage ACG-2014c (NC_019444) [176,043 nt]
Synechococcus phage ACG-2014a (KJ019038) [171,073 nt]
Synechococcus phage S-CAM8 (NC_021530) [171,407 nt]
Synechococcus phage ACG-2014g (NC_026924) [174,885 nt]
Synechococcus phage ACG-2014b (NC_027130) [172,688 nt]

Supplementary Figure S2. A Proteomic tree of dsDNA viruses represented in the rectangular view. This tree includes 52 reference viral genomes (48 *Myoviridae*, one *Siphoviridae*, and three unclassified viruses) and two uploaded viral genomes that are highlighted by red branches and stars (OBV_N00003 and OBV_N00005; BioProject: PRJDB4437). The tree is constructed by BIONJ based on genomic distance matrixes, and mid-point rooted. Branch lengths are linearly scaled. In the ViPTree server, where inner nodes of a tree are shown as filled circles, each of them links to a genomic alignment of sequences included in its subtree.



Supplementary Figure S3. An example of the genomic alignment view of four viral genomes (two reference and two environmental viral genomes) that are included in Supplementary Figure S2. Pairwise dot plots of these genomes are also shown. Colored lines in the alignment and the dot plots indicate tBLASTx results (E-value < 1e-2). Grid lines in the dot plots indicate 40 kb intervals. Positions of each sequences are automatically adjusted (i.e., circular permuted and reverse stranded) for clear representation of colinearity between genomes.