Supplementary Information for

Biological Species in the Viral World

Louis-Marie Bobay and Howard Ochman

Louis-Marie Bobay
Email: ljbobay@uncg.edu

This PDF file includes:

Figs. S1 to S8
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S2

www.pnas.org/cgi/doi/10.1073/pnas.1717593115

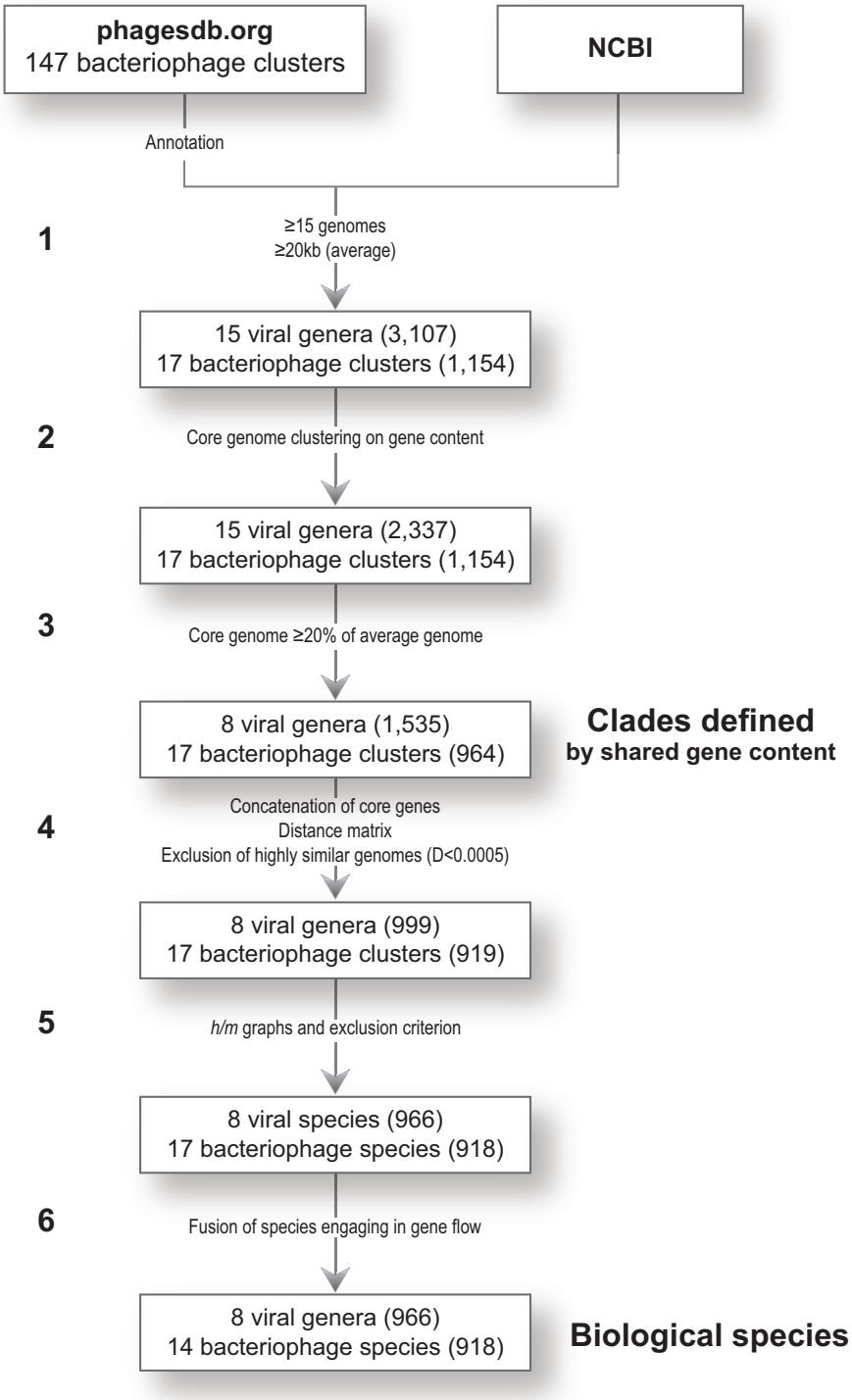

Figure S1. Flow chart depicting each analytical step in defining biological species. Steps are numbered one to six and correspond to those described in the Methods section. Numbers of viral and bacteriophage clades remaining after each filtering step are indicated within each box, and numbers in parentheses are the total number of viral and bacteriophage genomes at each step.

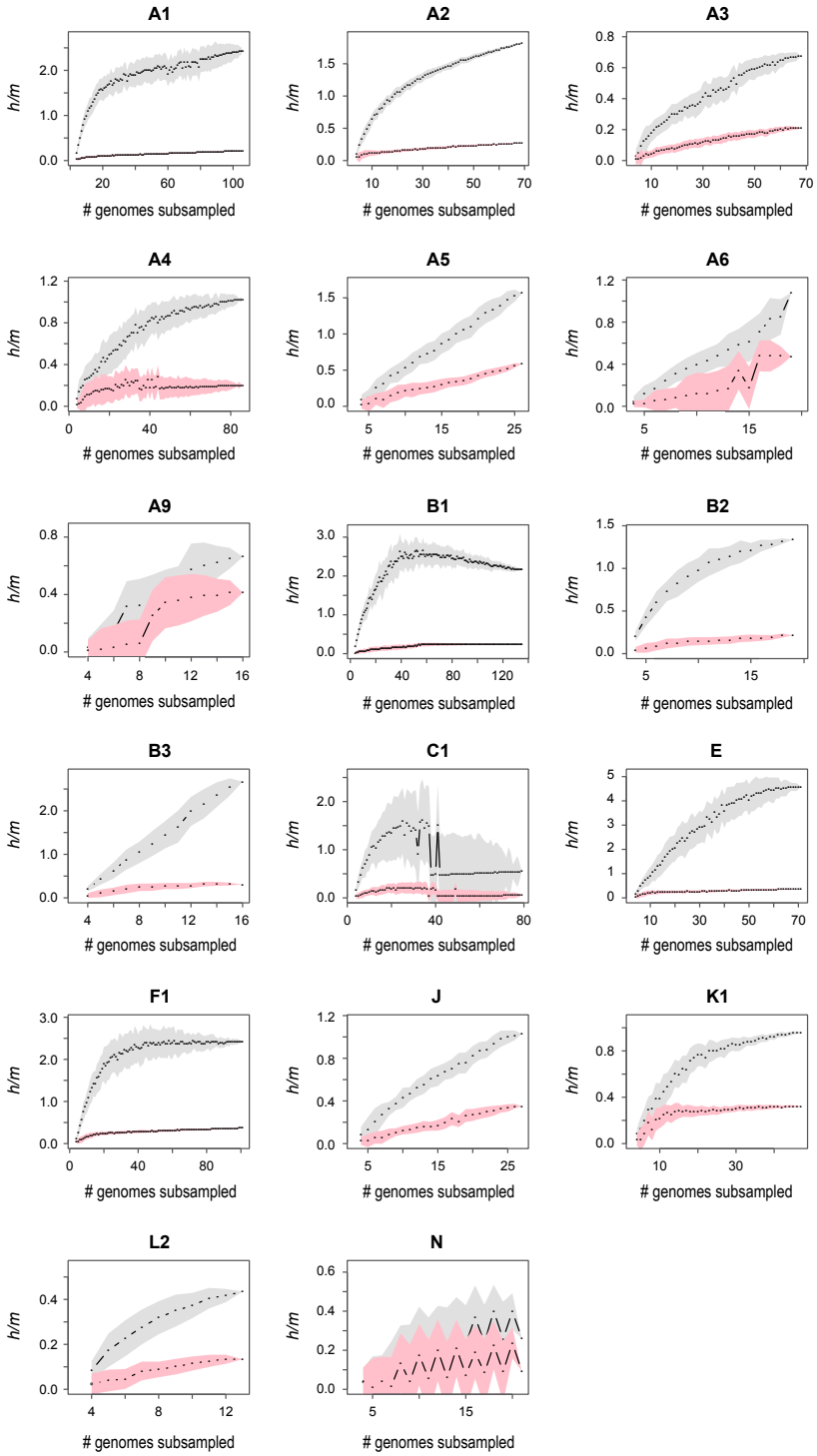

Figure S2. Patterns of gene flow in bacteriophages infecting *Mycobacterium smegmatis.* For each cluster, gene flow was estimated by the ratio of homoplasic to non-homoplasic alleles (*h/m*) with a re-sampling strategy. For each number *i* of genomes, 100 combinations of *i* genomes were randomly sampled and *h/m* was computed for each combination. Within the bivariate plots, black dots are medians and the grey-shaded region is the standard deviation for the indicated number

of subsampled combinations of strains, and red dots and pink-shaded regions denote median *h/m* values and standard deviation for simulations in which all homoplasies are introduced by convergent mutations, as described in the text. Differences between the distributions of observed and simulated *h/m* values indicate the extent to which homoplasies are introduced by recombination.

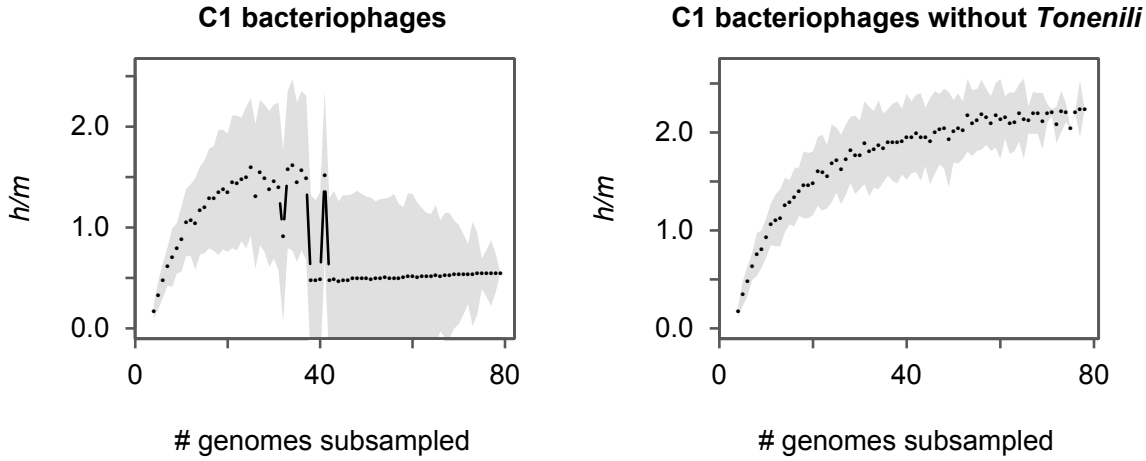

Figure S3. Redefining species membership in bacteriophage cluster C1. The discontinuity detected in the graph for the entire set of genomes classified in bacteriophage cluster C1 indicated the presence of multiple species (left). After removal of the sexually isolated genome (bacteriophage *Tonenili*), the graph was rebuilt (right). Black dots represent the median and the grey area indicates the standard deviation of *h/m* for the different combinations of genomes.

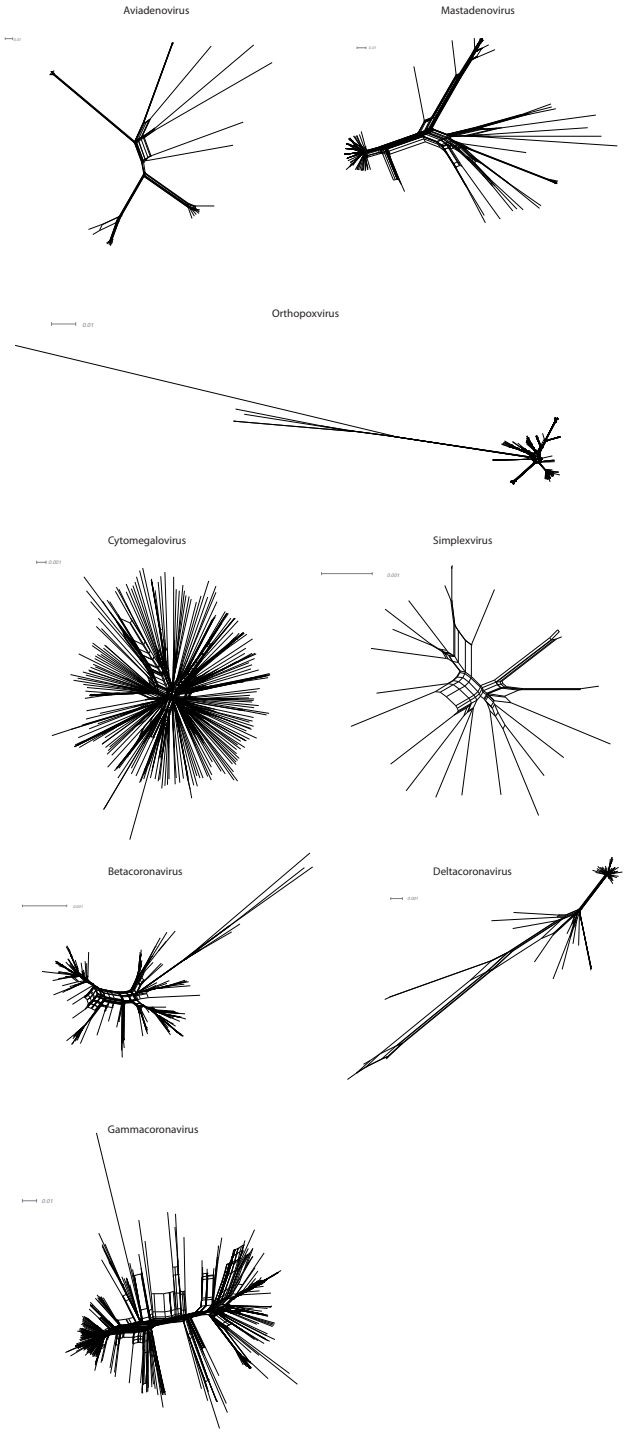

Figure S4. Phylogenetic networks of viral species. Phylogenetic networks were built using the core genome of each biological species of viruses with SplitsTree (1). The scale beside each phylogenetic network indicates substitution rate.

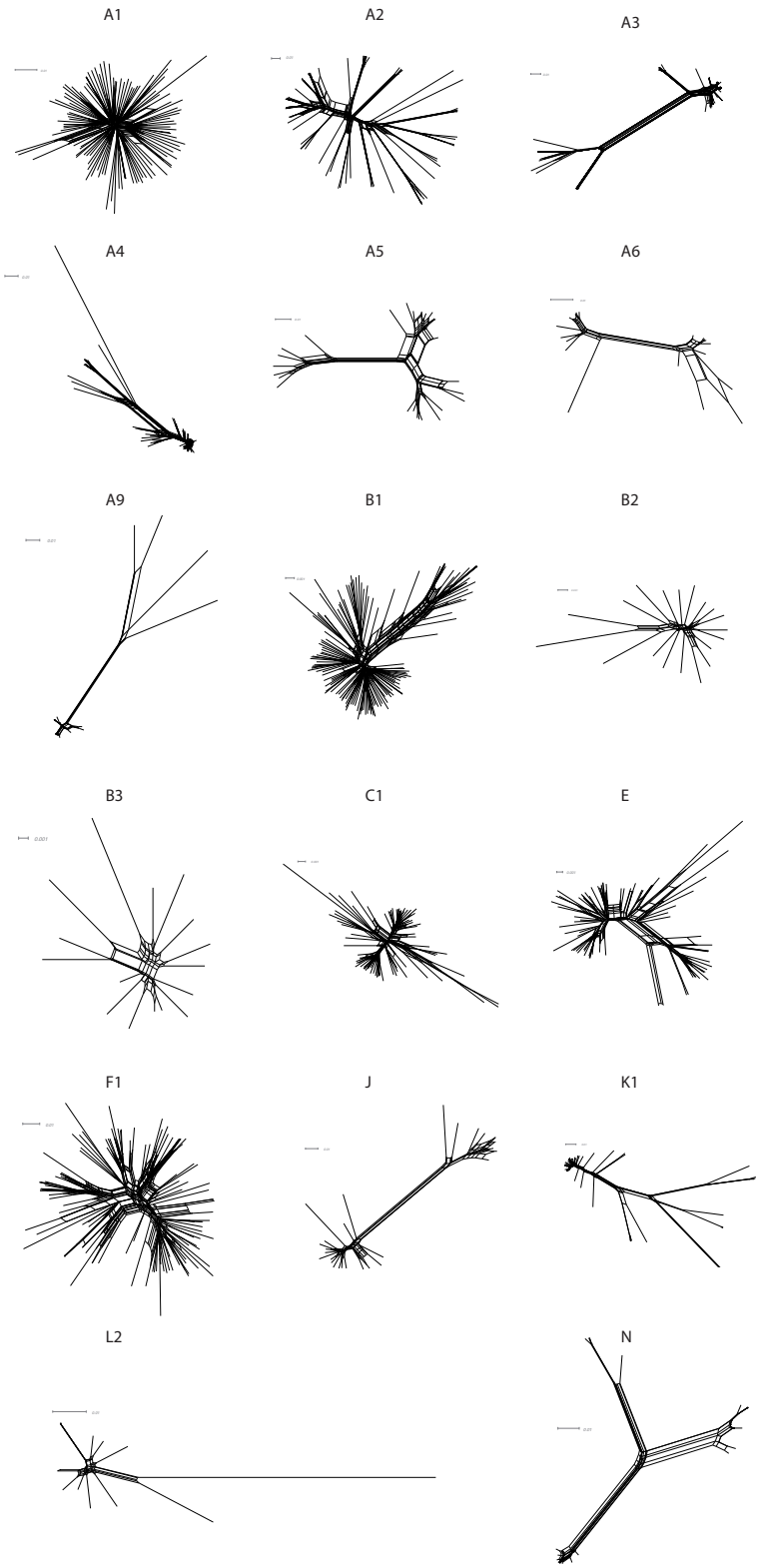

Figure S5. Phylogenetic networks of bacteriophages species. Phylogenetic networks were built using the core genome of each biological species of bacteriophages with SplitsTree (1). The scale beside each phylogenetic network indicates substitution rate.

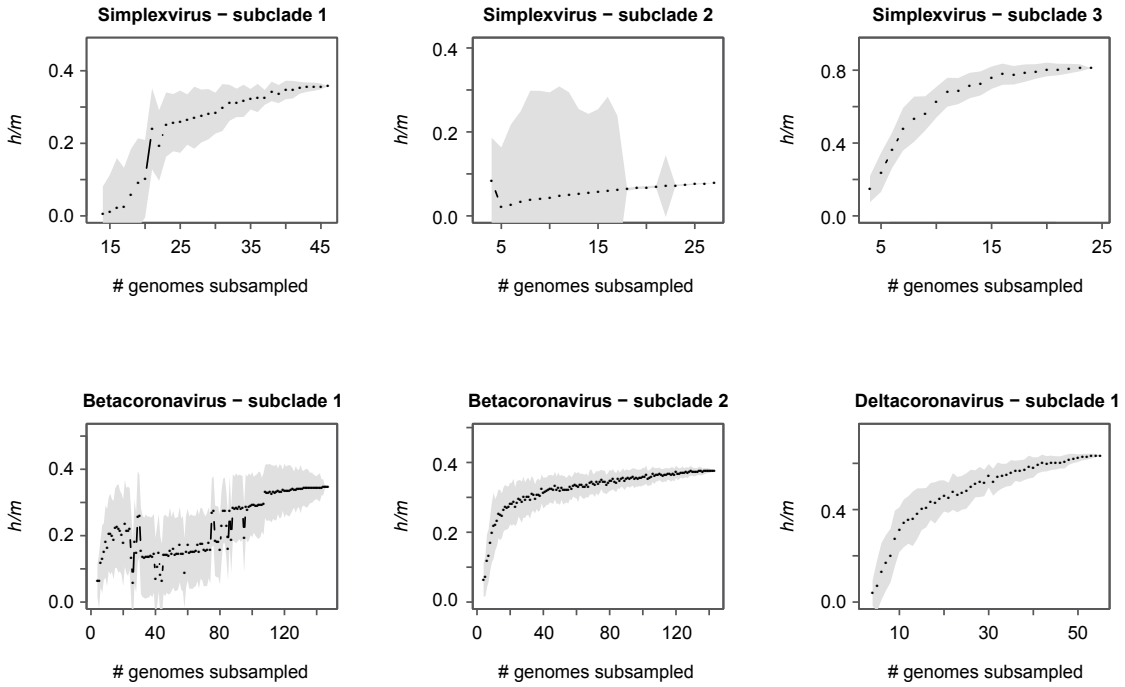

Figure S6. Patterns of gene flow in subclades of viral genera. Subclades within the three viral genera *Simplexvirus*, *Betacoronavirus* and *Deltacoronavirus* that each presented low signals of gene flow were tested independently for a signal of gene flow. Subclades were defined from phylogenetic trees that included all members of a genus, and analyses proceeding progressively examining smaller subclades, such that subclade 3 is included in subclade 2, which is included in subclade 1. Black dots represent the median and the grey area indicates the standard deviation of *h/m* of the different combinations of genomes.

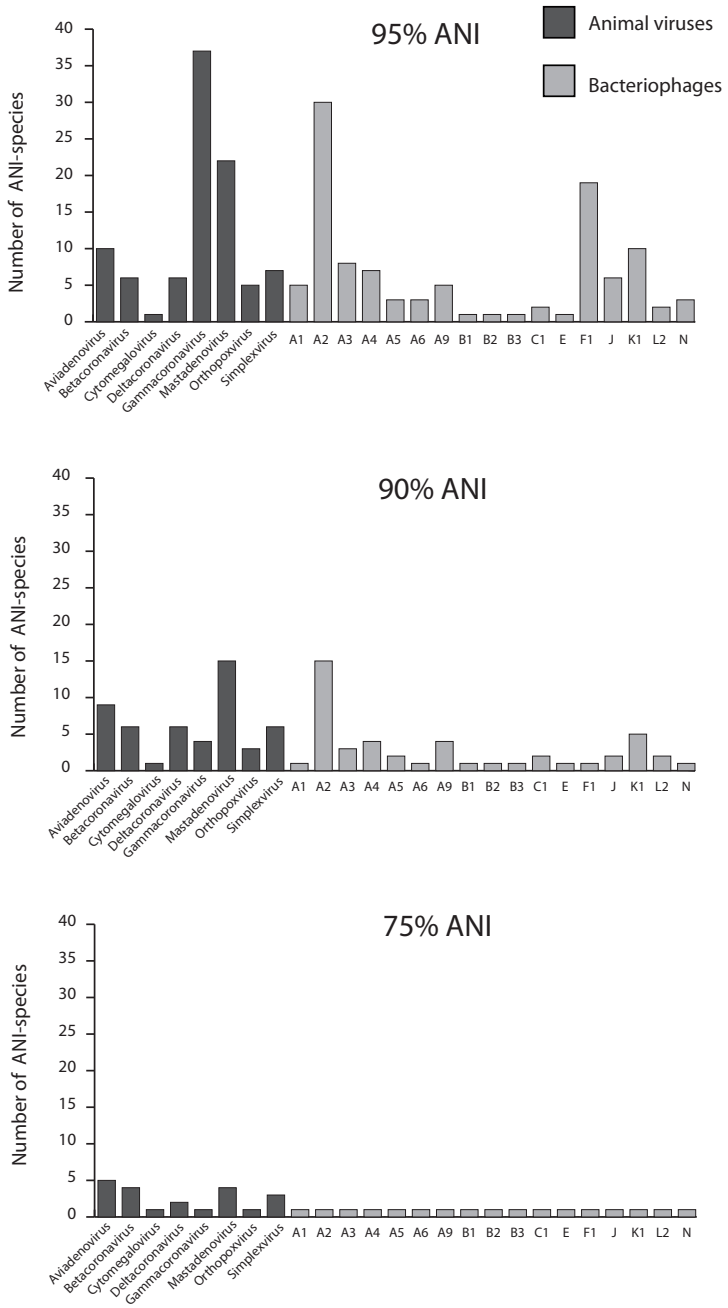

Figure S7. Number of ANI-species obtained at different sequence-identity thresholds. Average nucleotide identity (ANI) was computed along the entire core genome for members of the same biological species. Shown are numbers of groupings (ANI-species) obtained at various sequence-identity thresholds, selected as follows: 95%, threshold recommended for defining bacterial species; 90%, threshold recommended for defining bacterial genera; 75%, threshold representing the maximal divergence observed in biological species of bacteria. Note that the different sequence-identity thresholds do not partition each biological species in a uniform manner and that the maximal divergence observed between members of some viral species exceeds that observed in bacteria.

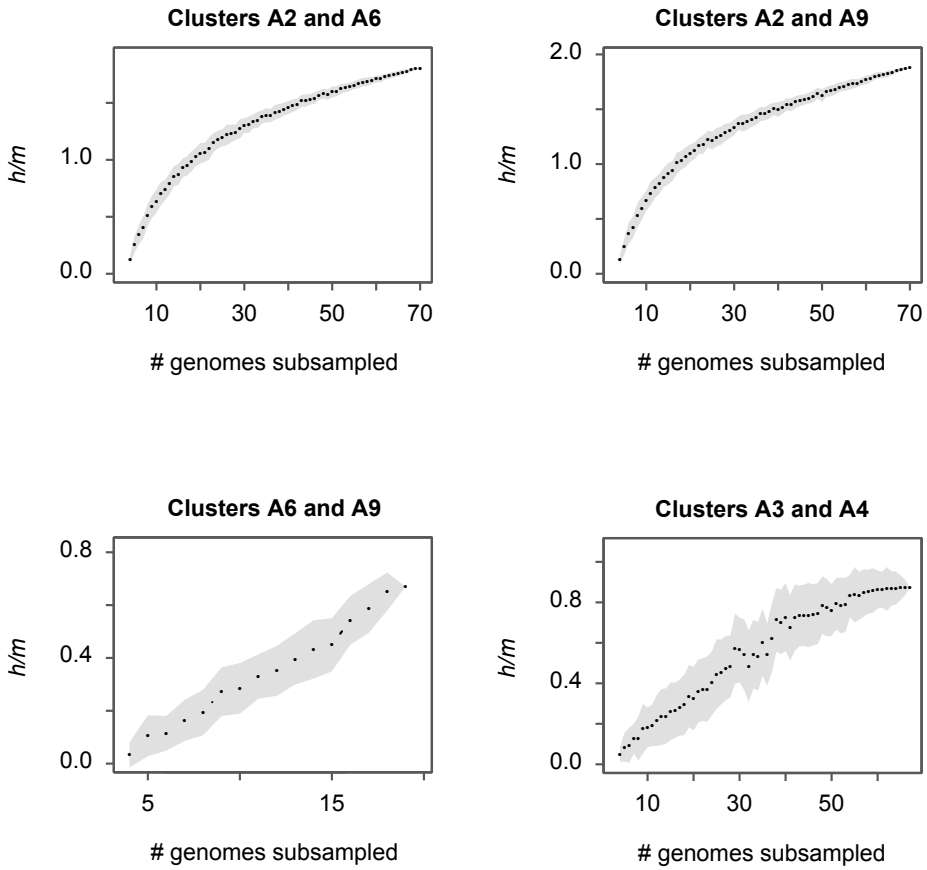

Figure S8. Patterns of gene flow between bacteriophage clusters. Patterns of gene flow were analyzed for pairs of clusters with large numbers of shared homologs ($n \geq 16$). Black dots represent the median and the grey area indicates the standard deviation of *h/m* of the different combinations of genomes.

SI References

1. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23(2):254-267.